UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Master's Degree in International Economic and Business focus on Business Organization & Strategy

An empirical analysis of the determinants influencing destination choice among Italian travelers

Advisor:

Prof. Mariateresa Ciommi

Candidate:

Laura Gazzoli

Academic year 2023 – 2024

ABSTRACT

This thesis explores the factors influencing travel destination choices among Italian tourists, integrating both descriptive and statistical analyses to provide a comprehensive understanding of tourism behavior. The research begins with a historical and conceptual overview of tourism, examining its evolution and its role as an economic driver, particularly in Italy. Using data from a survey by Bank of Italy, the second chapter offers a detailed descriptive analysis of Italian travel patterns, including demographic characteristics, preferences, and expenditure habits. In the third chapter, logistic regression is employed to help understanding how specific factors influence a destination choice. In the fourth chapter, the discriminant analysis is utilized to visualize which factors differentiate travelers that choose a destination.

TABLE OF CONTENTS

INTRODUCTION

The following thesis examines the factors that influence destination choice among Italian travelers. It explores how personal preferences and socio-economic conditions shape travel decisions, providing a detailed analysis of the Italian tourism landscape.

The first chapter presents a historical and conceptual overview of tourism, tracing its evolution from early definitions to contemporary understandings. The phenomenon of tourism, defined as the movement of people to destinations outside their usual environment for personal or professional reasons, has been a critical aspect of economic and social development. Early definitions focused on tourism as a temporary relocation for leisure or non-remunerated activities, while modern perspectives have expanded to include the psychological and motivational dimensions of travel. The chapter also explores different typologies of tourism, highlighting the distinction between domestic, inbound, and outbound tourism, as well as the economic impact of tourism on regional and global scales. Italy, in particular, has experienced significant economic benefits from tourism, which contributes to both employment and Gross Value Added (GVA).

The second chapter delves into the intricate process of selecting travel destinations, highlighting how this decision-making journey goes beyond merely picking a spot on a map. Influenced by a myriad of factors the choice of destination plays a pivotal role in shaping both individual experiences and the tourism industry's dynamics. Drawing on data from the Bank of Italy’s survey, the chapter provides an in-depth descriptive analysis of Italian travelers' preferences. It delves into demographic characteristics, travel patterns, and expenditure habits, illustrating the diverse factors that influence destination choices.

The data reveals trends in travel duration, preferred destinations, and motivations for travel, such as leisure, family visits, and business trips. This chapter underscores the importance of personal, economic, and situational factors in shaping tourism behavior and highlights the significant impact that tourism patterns have on the economy.

The third chapter employs the logistic regression model to understand the factors that influence destination choices. This method models the relationship between demographic and behavioral variables and travel decisions, providing insights into how factors such as age, gender, region of residence, and expenditure influence the likelihood of choosing specific destinations. Through logistic regression, the chapter predicts travel behavior and categorizes travelers into distinct profiles, offering valuable insights into the motivations and characteristics of Italian tourists.

The fourth chapter shifts to discriminant analysis, which help us explore how various predictors differentiate between travelers who visit certain destinations and those who do not. This analysis helps to refine the understanding of travel decision-making by highlighting the strengths and weaknesses of different predictors, and comparing their effectiveness in classifying travelers. The chapter assesses the performance of discriminant analysis in relation to logistic regression, emphasizing its utility in handling class imbalances and improving predictive accuracy.

CHAPTER 1

THE EVOLUTION AND DEFINITION OF TOURISM: HISTORICAL AND CONCETUAL OVERVIEW

Tourism is a multifaceted phenomenon that encompasses social, cultural, and economic dimensions. According to the United Nations World Tourism Organization (UNWTO) glossary, tourism is "*a social, cultural and economic phenomenon which entails the movement of people to countries or places outside their usual environment for personal or business/professional purposes*" (UNWTO, www.unwto.org). This chapter provides a historical and conceptual overview of tourism, tracing its evolution from early definition to contemporary understandings.

1.1 Early Definitions and Theoretical Foundations

The concept of tourism was discussed a long time ago, but the first definition emerged in 1941 through the work of Hunziker and Kraft, who defined it as "*the sum of the phenomena and relationships arising from the travel and stay of non-residents, insofar as they do not lead to permanent residence and are not connected with any earning activity*". In 1976, the Tourism Society proposed that "*tourism is the temporary, short term relocation of individuals to destinations beyond their usual places of residence and employment, encompassing all activities undertakes during their stay at each destination*".

In 1981, Burkart and Medlik outlined the fundamental characteristics of the tourism industry, distinguishing between conceptual and technical definitions. The conceptual definition defines the nature of tourism, as a group of activities, while the technical definition categorizes the different types of tourists and tourism activities (Buck, 1978).

1. 2 Modern Perspectives on Tourism

Kodhyat (1998) expanded on earlier definition by emphasizing the journey's purpose, highlighting the aim of seeking balance and happiness in social, cultural, natural, and knowledge-related aspects of the environment. This kind of perspective aligns with Richardson and Fluker's (2004) assertion that tourism involves temporary journeys which depart from one's original location and stop in another place, with the intention not to seek employment or stable life in the visited destinations but rather to discover the place through excursions or to fulfil various desires. Thus, tourism can be defined as the temporary vacation of tourists, individuals or groups, which travel to a different destination from their original one, with the purpose of leisure, business or other motivations all with the aim of seeking happiness and fulfilment.

From a statistical perspective, a tourist is defined as "*any person visiting a country other than that in which he has his usual place of residence, for any reason other than following an occupation remunerated from within the country visited*" (IUOTO, 1963).

It is crucial to distinguish between tourists and travelers as tourism activities entail a discretionary allocation of both time and financial resources. One fundamental concept is that tourists typically represent net consumers of economic resources within the regions

they visit, as their expenditure on various goods and services often surpasses any incidental income earned during their travels. Unlike travelers, tourists do not embark on their journeys primarily to earn remuneration from stops along the way.

1.3 Typologies and Classifications

Leiper in the Tourism system[1] stated that Tourism can be categorized into two main themes: tourism management and tourism studies, each with its focus and application. As a management discipline, tourism holds significance on various scale, such as regional, national and global. It is represented across public, private and third sectors, and is recognized as a significant economic contributor. Through tourism, we gain insights of our world, our different modes of travel and the intricate relationship between social mobility, regional cultures and the diverse place where these interactions occurs. Tourism's fundamental ideas, concepts and models originate from various fields of research, like economics, geography, history and so on. Economics pertains to resource utilization and capitalization, while geography is essential for comprehending the space, the location of the resources and movement through space. Although tourism is often discussed in the context of leisure or hospitality, it is fundamentally an economic subject. (Leiper, 2004)

[1] Leiper, Tourism, Critical concepts in the social sciences; Volume I, 2004, pg 26-27.

1.4 Economic Impact of Tourism

Tourism' s economic importance is evident in it contribution to national income, employment and development. Revenues derived from international tourism serve as a valuable source of income, especially aiding in the development efforts of all countries. Tourists expenditures not only contribute to the income of both public and private sectors but also influence wages and employment opportunities. Despite being influenced by the economic conditions of tourist-generating countries, it tends to provide more stable earnings compared to primary products. In many instances, the income generated from tourism has shown a higher growth rate than that from merchandise exports, particularly in countries with a limited industrial base. Consequently, tourism serves as a crucial income stream for numerous countries, spanning both developed and developing economies. (Padure, 2005)

The Australian Department of Tourism and Recreation (1975) identified tourism as "*an identifiable nationally important industry. The industry involves a wide cross section of component activities including the provision of transportation, accommodation, recreation, food and related services*". Wahab (1975) criticized the purely economic approach, suggesting that tourism comprises three elements: "*man, the author of the act of tourism; space, the physical element to be covered; and time, the temporal element consumed by the trip and stay*". (Leiper, 2004)

Tourism plays a pivotal role in the Italian economy, contributing significantly to its Gross Value Added (GVA) and employment landscape. In 2019, the direct contribution of tourism to Italy's GVA amounted to 6.2%, equivalent to EUR 99.9 billion. Additionally, the tourism sector directly employed 2.1 million individuals, accounting for 8.8% of total

employment. Furthermore, it provided support to over 218,000 enterprises across various segments of the economy. However, the onset of the COVID-19 pandemic resulted in a sharp decline in the direct contribution of tourism to Italy's GVA, plummeting to 4.5% in 2020. The impact of COVID-19 reflects across various facets of Italy's tourism landscape. In 2020, international arrivals witnessed a decline of 61.0%, collapsing to 25.2 million, while domestic tourism also experienced a substantial decrease of 37.1%, amounting to 34.1 million. The downturn in tourism activity led to an estimated loss of EUR 27.0 billion in expenditure from international visitors alone. Despite some signs of recovery in 2021, international arrivals remained significantly below pre-pandemic levels, registering a decline of 58.3% compared to 2019, with a total of 26.9 million tourists. Among the top source markets in 2021 were Germany (17.1%), France (14.5%), and Austria (9.3%). However, tourism expenditure from international visitors in 2021 amounted to EUR 21.2 billion, reflecting a substantial decrease of 52% compared to 2019 levels. (OECD, 2022)

In 2023 the World Travel & Tourism Council (WTTC) underscores the robust recovery of Italy's Travel & Tourism sector following the pandemic-induced downturn. As per the findings, the sector is poised to inject €194 billion into the Italian economy this year, trailing just 3% behind its pre-pandemic peak. The sector contributed 9,1% to the global GDP, increasing 23.2% from 2022. The same year there were 27 million new jobs, representing a 9.1% increase compared to 2022. These number results from and increase in domestic visitors spending, which rose by 18.1% and international visitors spending which increased by 33.1%. (WTTC, 2023)

The challenges posed by the COVID-19 pandemic underscore the resilience of Italy's tourism sector and the need for strategic measures to support its recovery and long-term sustainability. As the industry navigates through these unprecedented times, collaboration among stakeholders and innovative strategies will be imperative to revive Italy's tourism landscape and drive economic growth in the post-pandemic era. Domestic tourism constitutes a significant component of the Italian tourism sector, representing 56.4% of total tourism expenditure in 2019. Despite the challenges posed by the COVID-19 pandemic, domestic tourism has demonstrated a more robust recovery compared to international tourism. In 2021, domestic tourism experienced a notable rebound, with 37.2 million tourists recorded. Although this figure remains 31.5% below pre-pandemic levels, the resurgence of domestic tourism underscores its resilience and importance in sustaining the tourism industry during times of crisis. As travelers prioritize domestic destinations and experiences amongst ongoing uncertainties, the revitalization of domestic tourism serves as a crucial catalyst for the overall recovery of Italy's tourism sector. (OECD, 2022)

Additionally, WTTC forecasts a substantial job creation of over 65,000 positions within the sector in 2023, nearly recouping all the employment losses incurred during the COVID-19 pandemic and bringing the total workforce to almost 2.8 million. In the preceding year, the Travel & Tourism sector's contribution to Italy's GDP witnessed a remarkable surge of 33.4%, surpassing €194 billion and accounting for 10.2% of the nation's economy. This uptrend signifies an important stride towards reclaiming the pre-pandemic GDP high of €200.5 billion recorded in 2019. The resurgence of Italy's Travel & Tourism sector underscores its pivotal role in driving economic growth and

employment opportunities, signaling a promising trajectory for the country's post-pandemic recovery efforts. The sector's resilience and capacity to rebound from adversity highlight its importance as a cornerstone of Italy's economic landscape, fostering prosperity and vitality across various sectors and communities.

The Travel & Tourism sector in Italy witnessed significant growth and recovery last year, marked by notable achievements in job creation and international visitor spending. In 2022, the sector added an impressive 315,000 new jobs compared to the previous year, bringing the total employment figure to 2.7 million nationally. This translates to approximately one in every nine jobs across Italy. Moreover, the sector has successfully reclaimed 334,000 of the 477,000 jobs that were lost during the pandemic, indicating a substantial rebound in employment opportunities. The resurgence of international travel to Italy was also notable, with spending from overseas visitors experiencing a remarkable increase of 99.3%, surpassing €42 billion. Although this figure is still 11% below the levels observed in 2019, it reflects a significant step towards pre-pandemic levels of tourism activity. Julia Simpson, President & CEO of WTTC, emphasized the pivotal role of the Travel & Tourism sector in contributing to the Italian economy. She considers the sector's robust recovery as a positive development for job creation and economic prosperity throughout Italy, particularly as international visitors begin to return. Looking ahead, WTTC anticipates further growth in tourism, projecting that it will represent 12% of Italy's GDP over the next decade. This optimistic outlook underscores the enduring resilience and importance of the Travel & Tourism sector as a key driver of economic growth and prosperity in Italy.

The global tourism organization anticipates significant growth in the Travel & Tourism sector's contribution to Italy's GDP, projecting it to nearly reach €237 billion by 2033. This represents approximately 12% of the Italian economy, with over 3.3 million individuals expected to be employed within the sector nationwide. Remarkably, this means that one in seven Italians will find employment within the Travel & Tourism industry. In 2022, the European Travel & Tourism sector made a substantial contribution of €1.9 trillion to the regional economy, trailing just 7% below the peak observed in 2019. WTTC predicts that the sector's GDP contribution in the region will rocket to €2.04 trillion in 2023, coming close to reaching the apex achieved in 2019. Despite the challenges posed by the pandemic, the sector employed 34.8 million individuals across the region in 2022, reflecting an increase of 2.9 million compared to the previous year. However, this figure still falls short by 3.2 million from the peak recorded in 2019. Nevertheless, WTTC forecasts a full recovery of the jobs lost during the pandemic by the conclusion of 2024, signifying a promising outlook for employment within the sector. (WTTC, 2023)

In May 2023, there was an 8% increase in the number of Italian tourists travelling abroad compared to the previous year. During that month, approximately 4.4 million outbound travelers departed from Italy, showing a rise from the 4.1 million reported in May 2022. However, these figures remained below those recorded in May 2019, prior to the onset of the coronavirus (COVID-19) pandemic. In 2022, the primary motivation for most Italian tourists traveling abroad was for pleasure or leisure purposes. Additionally, exploring the natural landscapes of destinations and participating in cultural activities were significant factors driving outbound travel that year. Analyzing the outbound trips from Italy by

destination country, Spain emerged as the top choice for Italian travelers in 2022, followed by France, Croatia, and Greece. Moreover, findings from a May 2023 survey focusing on the travel intentions of Italians revealed that Spain and France continued to rank as the preferred European destinations for trips over the next six months (Statista, 2023[2]).

1.5 Tourism and Cultural Heritage

In 2017, the UNWTO and UNESCO joined the efforts into the Second World Conference on Tourism and Culture. The primary objectives of this conference were aimed at strengthen the collaboration among stakeholders in tourism and culture sectors to address the topics aimed at understand the cultural landscape within the realm of tourism. Specifically, the UNWTO and UNESCO outlined the conference's objectives to address governance models, sustainable development, and the protection of cultural heritage in tourism, as well as exploring the synergy between tourism and culture in urban development and creativity, along with the pivotal role of cultural tourism in sustaining tourism destinations. A central emphasis during discussions was placed on the importance of educational initiatives at cultural heritage sites. The development of international and regional tourism routes emerged as a critical imperative for the international community to prioritize. Throughout discussions, tourism routes were acknowledged for their ability to foster a sense of "*global citizenship*" and serve as a conduit for linking human civilization achievements. In line with Crompton's (1979) analysis, much of the discourse

[2] For more details visit www.statista.com

surrounding tourism motivation has revolved around the concepts of pull and push factors, which are rooted in social psychological factors rather than originating directly from tourists themselves. Push motives have been instrumental in explaining the desire to embark on a vacation, while pull motives have effectively elucidated the selection of destinations. Motivation, as conceptualized in various theories, arises when individuals seek to fulfil needs, prompting action. The disruption of their state of stability, or homeostasis, occurs when individuals become aware of deficiencies in their needs.

Tourism is widely recognized as a significant economic driver with the capacity to catalyze global economic growth. Its ability to complement other economic sectors, contribute to gross domestic product (GDP), create employment opportunities, and generate foreign exchange underscores its pivotal role in economic development (Ashley et al., 2007; Dwyer, Forsyth, & Spurr, 2004; García, 2005; Hernández & González, 2013; Rosentraub & Joo, 2009). Beyond its direct economic contributions, tourism also exerts an important impact on the economic and cultural advancement of societies, thereby enhancing the overall welfare of resident populations. However, it is evident that the nexus between tourism growth and economic development encounters significant constraints, particularly in countries with lower levels of economic development. These constraints stem from entrenched poverty, as well as deficiencies in economic, institutional, and human resources. In addition to its economic implications, tourism plays a crucial role in shaping the cultural landscape and social fabric of communities. By facilitating interactions between tourists and local residents, tourism can foster cultural exchange and mutual understanding, thereby contributing to societal progress and cohesion. The relationship between tourism and economic development is dynamic and

complex, influenced by a myriad of factors including policy frameworks, market dynamics, and socio-cultural considerations. To fully harness the potential of tourism as a catalyst for economic growth and societal advancement, certain efforts are required to address existing challenges and capitalize on emerging opportunities. This entails fostering a conducive environment for tourism development through strategic planning, investment in infrastructure and human capital, and stakeholder engagement.

The impact of tourism is starting to be discussed in the 1990s, marked by a series of seminal studies that highlighted both the potential benefits and challenges associated with tourism development. Hazari (1993) drew attention to the inflationary effects of tourism, cautioning against the unchecked growth of the industry. Sinclair (1998) emphasized the substantial investment required for tourism development, particularly in terms of physical infrastructure and human capital. He underscored the need for destination countries to prioritize investment in skilled labor within the tourism sector to ensure sustainable growth. Dunn and Dunn (2002) brought to light the pervasive issue of crime and violence in some tourist destinations, arguing that addressing these concerns is essential for the successful implementation of tourism initiatives. They noted that improving public safety not only enhances the visitor experience but also incurs additional costs for destination management. Furthermore, Gursoy and Rutherford (2004) and Jenner and Smith (1992) highlighted the environmental impacts of tourism and advocated for the implementation of policies to promote responsible tourism development. These authors emphasized the importance of mitigating the negative effects of tourism on natural ecosystems and local communities. Collectively, the insights from these studies underscore the multifaceted nature of tourism development and the need to balance economic growth with

environmental conservation and social well-being. (Pablo Juan Cárdenas-García et al, 2015)

Sánchez-Rivero, Pulido-Fernández, and Cárdenas-García (2013) conducted a comprehensive analysis of 117 countries and arrived at a conclusion: the growth of tourism within a country does not inherently lead to economic development unless specific conducive conditions are in place to facilitate this process. They underscored the importance of recognizing that not all interventions aimed at promoting tourism growth are equally effective in fostering economic development. In essence, certain variables associated with tourism growth exhibit stronger correlations with economic development than others. Consequently, the authors advocate for directing efforts primarily towards promoting these key variables to maximize their impact on economic development. The failure to identify and prioritize factors useful to transform tourism growth into economic development carries significant opportunity costs for countries. By neglecting to focus on the essential drivers of economic development within the tourism sector, countries risk squandering valuable resources and missing out on opportunities for sustainable development and inclusive growth. In light of these findings, it becomes imperative for policymakers to adopt a different approach to tourism development, one that goes beyond the mere expansion of the tourism sector and instead prioritizes interventions that have the greatest potential to help increasing the economic development. This entails conducting thorough assessments of the underlying factors that influence the relationship between tourism growth and economic development within specific contexts. By identifying and targeting these key variables, policymakers can effectively harness the transformative power of tourism to drive economic progress and enhance the well-being

of their populations. (Does tourism growth influence economic development?, Journal of travel research,2015)

The latest report "*Capital Investment Fuels Growth in Travel & Tourism, Forecast to Reach Nearly $1 Trillion says WTTC*" from the World Travel & Tourism Council (WTTC) on the Economic Impact of Travel & Tourism in 2022 reveals an optimistic rebound in investment within the sector, overcoming the setbacks caused by the pandemic and indicating a robust return to growth. Between 2010 and 2019, investment in Travel & Tourism experienced consistent growth, achieving a compound annual growth rate (CAGR) of 4.3%. During this period, investment expanded from $754.6 billion in 2010 to $1.1 trillion in 2019, accounting for 4.5% of total economy-wide investment. However, the COVID-19 pandemic led to a sharp decline, with a 24% reduction in 2020 followed by an additional 8% decline in 2021. Nevertheless, the year 2022 marked a pivotal moment. Fueled by pent-up demand on a global scale, investment in Travel & Tourism surged to $856 billion, representing an 11.1% increase from the previous year. Although this fell short of 2019 levels by 22.5%, it still marked a remarkable 53% increase compared to the investment levels seen in 2000. This resurgence in investment signals a promising recovery for the Travel & Tourism industry, reflecting renewed confidence and momentum in the post-pandemic era.

In regions such as Asia-Pacific and Africa, the level of investment in Travel & Tourism surged by an impressive 161% in 2022 compared to the year 2000. Conversely, Europe and the Middle East experienced more subdued growth, with the pandemic reversing much of the significant progress made in these regions over the past two decades. Despite

the challenges posed by the pandemic, Travel & Tourism investment in these regions remained above the levels observed in 2000. (WTTC, 2022)

1.6 Recent trends and Developments

In 2022, the proportion of Italy's gross domestic product (GDP) attributed to the travel and tourism sector experienced a decline of 4% compared to the pre-pandemic year of 2019. During 2022, travel and tourism contributed to approximately 10.2% of the nation's GDP, both directly and indirectly. The total contribution of the travel and tourism industry to Italy's GDP exceeded 190 billion euros that year. Moreover, in 2022, the total expenditure by international tourists visiting Italy, encompassing both overnight stays and same-day visits, surpassed 44 billion euros, effectively reaching pre-pandemic spending levels. Upon analyzing the breakdown of inbound tourism expenditure in Italy by originating country, Germany emerged as the primary market, surpassing the United States, France, and the United Kingdom in terms of visitor spending. (Statista, 2024)

The United States, instead, emerged as the leader among the top ten markets in terms of absolute investment in the tourism sector in 2022, with a total investment of $213 billion, signaling a sector poised for renewed prosperity. China followed closely behind with a $146 billion investment in 2022, while Saudi Arabia secured the third position with a total investment of $42 billion during the same period. When considering the proportion of Travel & Tourism investment relative to the overall economy, island destinations claimed the top spots in 2022. The US Virgin Islands led the pack, allocating 35% of total economic investment to Travel & Tourism, closely followed by Antigua & Barbuda at

34% and Aruba at nearly 32%. This highlights the significant role played by Travel & Tourism in the economies of these island nations, underscoring their reliance on the sector for economic growth and development.

Private investment, such as establishing hotels and expanding car fleets, plays a pivotal role in increasing the capacity of the travel and tourism sector. The combination of private investments and public funding are crucial for fostering growth within the industry. The effect of these investments translates into the creation of additional jobs, the stimulation of larger economies, and the fortification of communities. As more resources are injected into the travel and tourism sector, the positive impacts reverberate throughout various facets of society, leading to enhanced prosperity and well-being. Julia Simpson, President & CEO of WTTC, emphasizes the significance of investment in travel and tourism, characterizing it as more than just a numerical endeavor—it represents the pulse of global connectivity and economic rejuvenation. Despite the setbacks endured during the pandemic, the growth witnessed in 2022 should serve to understand the sector's trajectory in the foreseeable future. Investment in travel and tourism emerges in the world's recovery and advancement. The sector's resilience and innovative capacity make it a key driver in shaping a more prosperous and interconnected global future. While remaining steadfast in their confidence, industry stakeholders recognize the importance of remaining vigilant as they navigate the evolving challenges and opportunities on the path toward a more prosperous and interconnected future. (WTTC, 2022)

Ivana Jelinic, Chair and CEO of ENIT, emphasizes the importance of annual monitoring within the tourism industry, particularly focusing on outdoor tourism. Such monitoring provides a comprehensive understanding of emerging trends, empowering operators to

proactively adjust to the evolving preferences of travelers. In Italy, renowned for its diverse landscapes offering unique opportunities, tracking the increasing popularity of various holiday types is crucial for resource optimization and the development of targeted promotional strategies. By promoting outdoor experiences, not only does it enrich the tourism offering, but it also fosters sustainable development within the industry, encouraging environmental conservation efforts and active engagement from local communities. Livio Gigliuto, Chairman of Istituto Piepoli, asserts, "*Summer holidays remain non-negotiable: despite economic uncertainties, the 'holiday movement' is experiencing notable growth compared to 2022. With the pandemic receding, two distinct trends have emerged: firstly, spending summer in Italy is the top choice among Italians, driven not solely by financial considerations. Secondly, outdoor tourism has firmly established itself as the preferred holiday option for approximately one fifth of Italians. Additionally, outdoor holidays are identified as the ones where Italians allocate the highest expenditure*." (ENIT, 2023)

The robust recovery observed in the travel and tourism sector, especially in the post-pandemic era, underscores its vital role as an economic engine and cultural bridge. However, the journey forward demands strategic planning, sustainable practices, and collaboration across sectors to ensure that tourism not only flourishes economically but also contributes positively to the social and cultural fabric of societies.

In conclusion, tourism's multifaceted nature, as explored throughout this chapter, highlights its evolution from early definitions to its current status as a dynamic global industry. The sector's significant contributions to economic growth, job creation, and cultural exchange are undeniable, but so are the challenges it faces, including environmental impact, the need for sustainable development, and the volatility introduced

by global events like the COVID-19 pandemic. Through continued innovation, investment, and a commitment to sustainable practices, tourism can remain a cornerstone of global economic and social development, fostering a more interconnected and prosperous world.

CHAPTER 2 –

UNDERSTANDING TOURISM BEHAVIOUR: A DESCRIPTIVE ANALYSIS OF TRAVEL DESTINATION CHOICE

Choosing a travel destination is far more than just picking a place on a map. It's a complex decision-making process influenced by various factors that range from personal preferences to economic considerations. Understanding this process is crucial, particularly from an economic standpoint, as it impacts not only individual travelers but also entire industries and economies.

This study aims to explore the process of how travelers reject or select certain destinations during their decision-making process. By examining various factors influencing destination choice we seek to gain an understanding of tourist behavior.

Firstly, individual preferences play a significant role in destination selection. People may be drawn to certain destinations due to cultural attractions, natural beauty, adventure opportunities, or simply because it's a place they've always dreamed of visiting. These personal inclinations can heavily influence the decision-making process and dictate where individuals ultimately choose to travel.

Moreover, practical considerations such as budget, time constraints, and accessibility also come into play. Travelers often weigh the costs associated with visiting a particular destination against their available resources. Factors like accommodation, food and activities contribute to the overall expense of a trip. Additionally, the time required to

reach the destination and the duration of the trip are important factors, especially for those with limited vacation time.

Beyond individual considerations, broader economic factors shape travel decisions on a larger scale. Destinations with strong tourism infrastructure, marketing efforts, and favorable exchange rates may attract more visitors. Conversely, destinations facing political instability, natural disasters, or negative publicity may see a decline in tourism and suffer economic setbacks.

The travel industry itself is a significant driver of economic activity, supporting millions of jobs worldwide. Destination selection impacts various sectors including transportation, hospitality, entertainment, and retail. For example, popular tourist destinations often experience increased demand for services such as hotels, restaurants, tour operators, and souvenir shops, leading to job creation and revenue generation.

2.1 The bank of Italy: survey

Since 1996, the Bank of Italy has yearly conducted a survey on international tourism based on interviews and counts of both resident and non-resident travelers crossing Italian borders. For estimating the number of international travelers, the survey data is supplemented by administrative data, where available, and, since late 2020, by mobile phone data. Moreover, the survey serves as an extensive information base for researchers and industry professionals, providing a wide range of analytical data. These datasets are made available to users both through monthly updates and at the microdata level. On April 23 of 2024, Bank of Italy released the comprehensive survey on Italian tourists

conducted in 2023, on both domestic and international, to gather crucial insights into travel patterns and expenditure. Commissioned to BVA Doxa and Scenari, the survey aimed to measure passenger traffic flows at major border points, essential for defining tourism expenditure in Italy's balance of payments. The survey is utilized by the Bank f Italy to conduct an investigation on international tourism, the one regarding 2023 was published on 18 June 2024. In this investigation emerges that 2023 saw continued grow in spending at current prices by both foreign travelers in Italy and Italian travelers abroad. The surplus in the tourism balance of payments slightly increased to 20.1 billion euros, representing 1.0% of GDP (up from 0.9% the previous year). In fact, spending by Italian travelers abroad rose by 21.3% compared to the previous year. Both business and personal travel spending increased, especially for vacations, which nearly doubled and accounted for 42.3% of total spending. Based on the same survey, this thesis delves into the intricate details extracted from the survey data conducted individually on Italian travelers, providing a comprehensive analysis of Italian tourist behaviors, preferences, and trends.

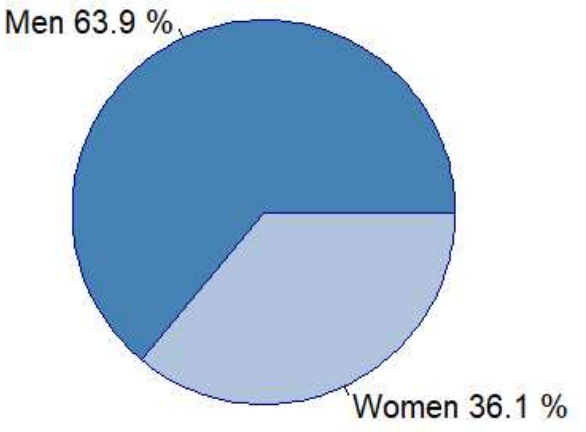
2.1.1 Demographic characteristic

The survey, commenced with an examination of the demographic characteristics of Italian travelers, aiming to capture an understanding of the diverse cohort participating in the study.

Among the 46,579 respondents a nuanced portrait of the Italian traveler emerged, revealing intriguing insights into age distribution, professional backgrounds, gender composition and regional origins. Age served as a defining parameter in delineating the traveler demographic, reflecting the varied preferences and priorities of different age

cohorts. From adventurous millennials seeking experiential journeys to seasoned retirees indulging in leisurely escapes, the spectrum of travel motivations and behaviors spanned across generations.

Figure 2.1: Gender Distribution



Source: Our elaboration on Bank of Italy's data

The survey unveiled a rich tapestry of age diversity, underscoring the universal allure of travel across the lifespan. Profession emerged as another key dimension shaping the travel landscape, offering a glimpse into the occupational profiles of Italian travelers. The intersection of work and leisure, often blurring conventional boundaries, underscored the evolving nature of contemporary travel experiences. Gender dynamics played a pivotal role in shaping the composition of the traveler cohort, revealing intriguing disparities in travel preferences and behaviors. The survey unveiled a notable gender skew, with men

comprising 64% of the respondent pool, while women accounted for the remaining 36% (*Fig. 2.1)*.

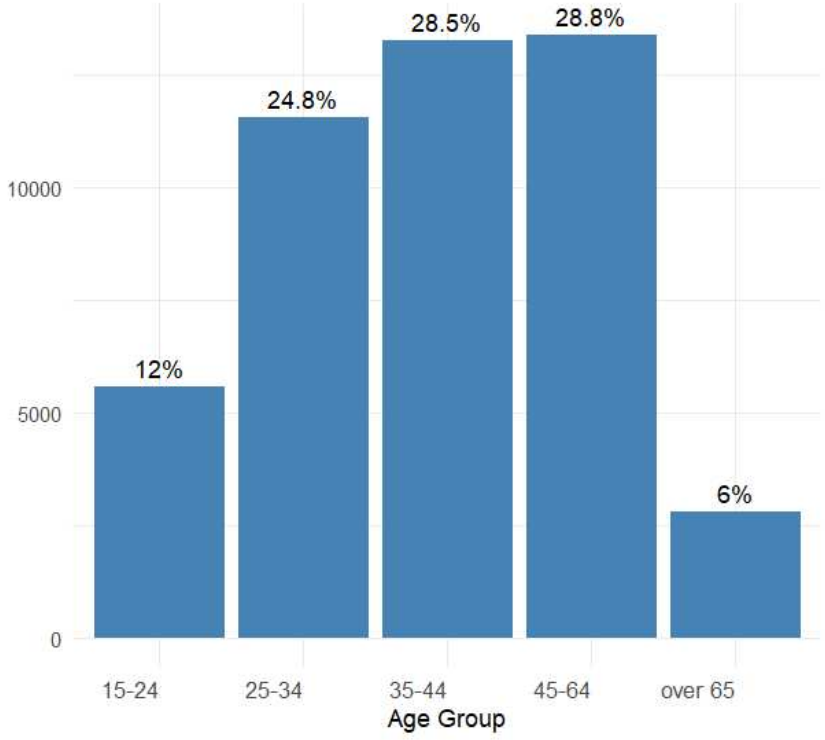
This gender asymmetry, though reflective of broader societal trends, hinted at underlying nuances in travel decision-making processes, warranting further exploration. Regional origin emerged as a critical determinant in shaping travel preferences and itineraries, reflecting the diverse cultural landscapes and geographical attractions across Italy. The survey meticulously documented the geographic distribution of respondents, offering valuable insights into regional travel patterns and preferences. In dissecting the demographic profile of Italian travelers, the survey unearthed a treasure trove of insights, illuminating the multifaceted nature of travel motivations and behaviors. By unraveling the intricacies of age, profession, gender, and regional origin, the survey provided a comprehensive foundation for subsequent analyses, paving the way for a nuanced understanding of Italian tourist behaviors and preferences and providing a more comprehensive understanding of the travel cohort involved in the study.

The survey delved deeper into the age distribution of Italian travelers, revealing intriguing patterns that illuminate on the underlying motivations driving travel behavior (*Fig 2.2)*.

The predominant age group, spanning between 45 and 64 years, emerged as the most prolific cohort, comprising a significant portion of the respondent pool. The prominence of this age group suggests a convergence of factors that make this cohort particularly predisposed to travel. As said above, with established careers and higher disposable incomes, these experienced professionals possess the financial means and leisure time to indulge in travel pursuits. Moreover, their accumulated wisdom and life experiences often manifest in an enhanced appreciation for exploration and cultural immersion, driving

them to seek enriching travel experiences. The presence of individuals aged 25 to 44, slightly younger, underscores a similar inclination towards travel, although within a different life context. Often characterized by career advancement and family responsibilities, this demographic group navigates the intricacies of work-life balance, carving out moments of respite and rejuvenation through travel. Whether it's a family vacation to bond with loved ones or a solo adventure to rediscover oneself, this age group embodies the diverse variety of motivations that propel travel decisions.

Figure 2.2: Age division among respondents

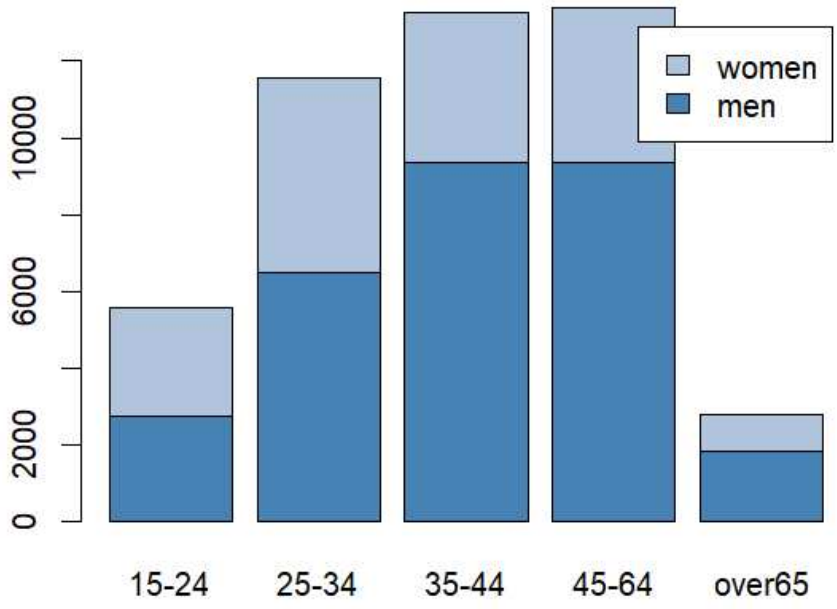


Source: Our elaboration on Bank of Italy's data

The demographic distribution observed in the survey underscores the intertwined relationship between age, life stage, and travel behavior. It reflects not only the diverse motivations driving Italian travelers but also the varying degrees of financial capability and vacation planning skills across different age cohorts. Expanding on the age distribution findings in this manner provides a nuances understanding of the motivations and behaviors driving travel decisions across different age cohorts, enriching the overall analysis of Italian tourist behaviors.

The differing age patterns in travel behavior between men and women could stem from various societal and biological factors. Among males, those within the 35-44 and 45-54 age groups, followed by the 25-34 age, are the favorite ages for travelling *(Fig. 2.3)*.

Figure 2.3: Relationship between age groups and gender



Source: Our elaboration on Bank of Italy's data

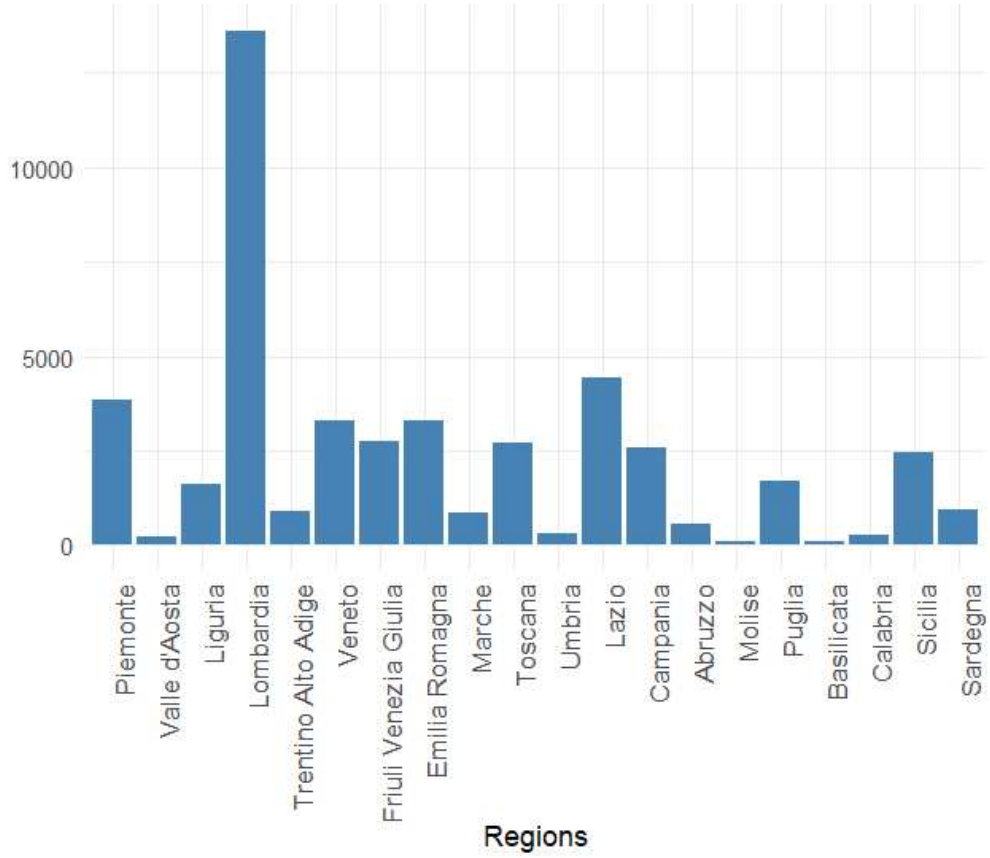
This choice may be driven by lifestyle, career and personal circumstances, like career stability combined with financial resources and less family responsibility. Traditionally, men might delay extensive travel until their mid-30s due to career advancement, financial stability, or the desire to establish themselves professionally before engaging in leisure activities. This delay could also coincide with milestones such as marriage and starting a family, prompting men to prioritize family-oriented travel during this stage of life.

They predominantly visit countries such as France, Switzerland, Germany, Spain, Slovenia, Austria, the USA, Greece, the Netherlands, Turkey, the Czech Republic, and Dubai. Conversely, women might engage in travelling before reaching their mid-30s because they still don't have any family responsibility and focus on personal growth and exploration. In fact women travel more between 25 and 34 years old. They may also prioritize travel during their late teens and twenties, seeking experiences and adventures before potentially taking on greater familial or career responsibilities in their thirties. Biologically, women may feel a sense of urgency to travel before their mid-30s due to factors such as fertility and the biological clock, which can influence decisions regarding timing for family planning and travel. Their main destinations include Spain, Switzerland, France, the United Kingdom, Germany, and the USA.

Based on these results we may suggest that both male and female travelers, in different age groups, find themselves influenced by familial obligations when choosing their destinations. It's intriguing to note that while there might be variations in the specific countries visited, the overarching theme of family-oriented travel remains consistent across genders and age brackets. This speaks to the enduring importance of family ties and responsibilities in shaping individuals' travel decisions, regardless of gender or age. Additionally, the diversity of destinations visited highlights the varied interests and

preferences among travelers, reflecting a rich tapestry of cultural experiences sought by both men and women alike.

Figure 2.4: Italian regions with most travelers



Source: Our elaboration on Bank of Italy's data

Lombardia emerged as the top region for outbound tourism *(Fig. 2.4).* Within Lombardia, Milan and Varese stand out as key contributors to this numbers, reflecting the economic expertise and cosmopolitan nature of these urban centers. Lazio, home to Italy's iconic capital Rome, accounted for 91% of the region's travelers. Piemonte follows closely

behind. In stark contrast, Basilicata emerges as a relative outlier, with the lowest participation. Regional disparities in travel behavior shed light on varying socio-economic factors and cultural preferences across Italy.

Respondents from Lombardia travelled mainly to European countries, their journeys are divided through the picturesque France, the alpine wonders of Austria, the sun-kissed shores of Spain, and the serene beauty of Switzerland. Additionally, they embraced transatlantic adventures across the Atlantic to explore the diverse landscapes and cultures of the United States. Piemonte echoed Lombardia's travel patterns, mainly visiting the same wonders.

2.2 Travel Preferences and Patterns

2.2.1 Duration of Stays

The survey revealed diverse travel preferences among Italian tourists, uncovering trip durations and patterns that underscore the nuanced complexity of vacation planning and decision-making. Among the participants 6,429 individuals opted for one-day trips, seeking brief yet immersive experiences close to their home. Other 5,752 participants, preferred week-long stays, engaging in extended getaways to unwind and explore destinations with more calm.

The medium stay among all the participants revealed to be 14,5 days. This duration encapsulates the different range of travel preferences, accommodating both short-term trips and long-term explorations. Whether it’s a city break or a cross-country journey, travelers’ preferences span between different temporal possibilities, reflecting the myriad motivation and desired that drive travel decisions. This inclination towards specific

durations may be attributed to practical considerations or cultural norms governing vacation planning. Round-numbers duration facilitate logistical arrangements and budgetary estimation. Moreover, they align with traditional concepts of time measurement and planning, resonating with cultural expectation and societal norms surrounding leisure and travel.

However, it's essential to recognize the travel preferences are subjective and multifaceted, shaped by individual circumstances, preferences and aspirations. While some may gravitate towards familiarity and predictability of round-numbered durations, others may relish the spontaneity and flexibility of unconventional itineraries.

By delving into the temporal dimensions of vacation planning, we gain valuable insights into the motivations and aspirations that drive Italian tourists' quest for exploration and discovery. Moreover, a closer examination of travel patterns in various countries reveals fascinating nuances. The survey findings offer valuable insights into the multifaceted nature of travel preferences among Italian tourists. From the duration of stays to the choice of destinations, travelers' decisions are influenced by a myriad of factors, including personal preferences, cultural influences, and practical considerations.

2.2.2 Country-Specific Preferences

In Austria, for instance, travelers tend to spend between 0 and 4 nights, reflecting a flexible approach to accommodation and itinerary planning. Conversely, in France, one-night stays are prevalent, with a maximum duration of 4 nights, indicating a preference for short, immersive experiences. Slovenia, with its proximity to neighboring destinations, sees a trend of one-day visits without overnight stays, showcasing the

convenience and accessibility of day trips. In contrast, the Czech Republic boasts a minimum stay of 4 nights, suggesting a desire among travelers to delve deeper into the country's cultural and historical treasures. Meanwhile, in Germany, stays typically range between 3 and 4 nights, creating a balance between exploration and efficiency. Switzerland presents a unique case, with travelers opting for day trips without overnight stays, maybe a decision driven by the costs of the country.

Continuing with the analysis, Turkey stands out for its preference for 4-night stays, indicative of a more difficult accessibility of the country and a desire among travelers to immerse themselves more deeply in the country's rich cultural heritage and diverse landscapes. This longer duration allows for a more comprehensive exploration of Turkey's myriad offerings.

Europe emerged as the most favored international destination among Italian tourists, Frances attracted 5,890 respondents, followed by Switzerland with 4,775 and Spain with 4,394. While outside Europe, USA emerged as the most visited country with 2,378. In contrast, countries in Africa and Asia recorded lower visitation rates, potentially due to safety concerns and accessibility challenges.

The choice of destination reflects not only preferences for cultural experiences or natural attractions but also practical considerations such as ease of travel and perceived safety.

2.2.3 Travel Motivation

Leisure emerged as the primary motive for travel among Italian tourists, with 17,681 respondents citing it as their reason for the trip. Visits to family and relatives followed closely, with 5,420 respondents, highlighting the importance of familiar bonds in shaping

travel decisions. Additionally, work-related travel accounted for 2,400 respondents, underscoring the intersection of professional commitments with leisure activities.

When delving into the motivations behind travel, it becomes evident that gender plays a significant role in shaping these preferences. For men, leisure activities often take precedence, with the desire of exploring new destinations, indulging in adventurous pursuits, and seeking relaxation away from the daily stress. Occasional travel, whether for special events, sports, or cultural experiences, provides men with an opportunity to unwind, fostering a sense of exploration and discovery.

In contrast, women's travel motivations tend to revolve around a blend of leisure and familiar responsibilities. While leisure remains a driving force, women often prioritize family visits and reunions, recognizing the importance of maintaining strong familiar bonds and nurturing relationships. Both underscore the universal allure of travel as a means of self-discovery, enrichment, and connection with the world around us.

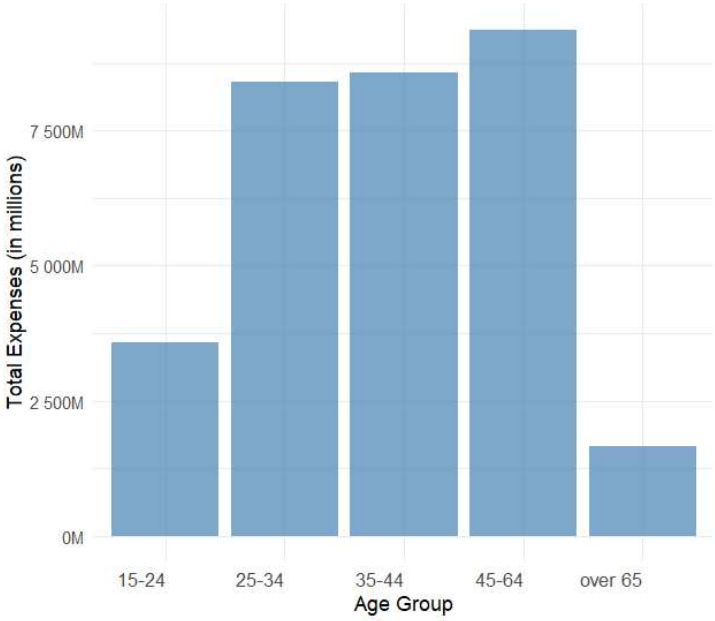
Despite these gendered differences in travel motivations, it's essential to recognize that individuals of all ages share similar motives for travel. Whether young adults embarking on solo adventures, couples seeking romantic getaways or seniors indulging in retirement travels, the desire for leisure, familiar connections, and occasional travels transcends age boundaries. Each age group brings its unique perspectives, preferences, and life stages to the travel experience, enriching the blend of human exploration and cultural exchange.

2.2.4 Expenditure Patterns

The survey also provided insights into the expenditure patterns of Italian tourists. On average, respondents reported spending 678104 euros amount per trip, with a significant

portion allocated to transportation, accommodation, dining, and entertainment. The expenditure varied depending on factors such as destination, duration of stay, and travel purpose. High-end destinations or luxury experiences tended to incur higher costs, while budget-conscious travelers sought affordable options without compromising on quality.

Figure 2.5: Expenses depending on the age group

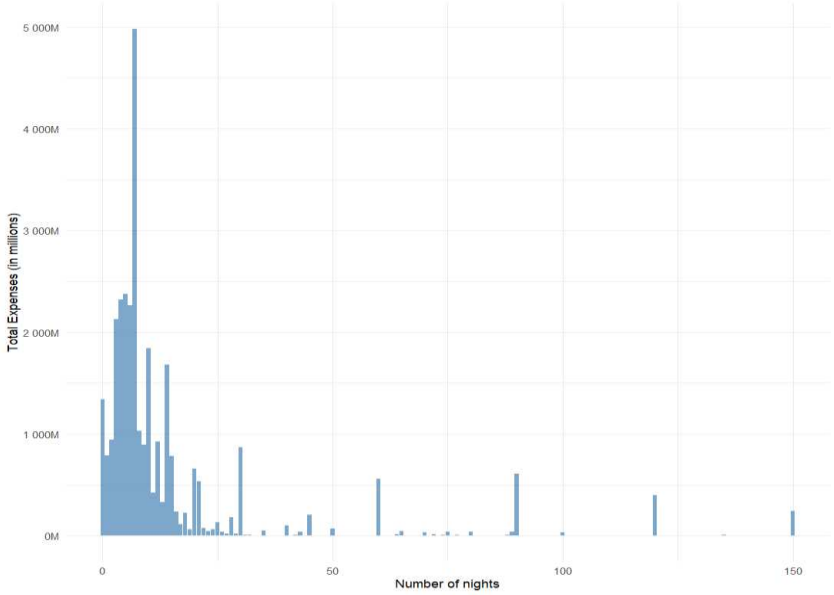


Source: Our elaboration on Bank of Italy's data

This plot visualizes the relationship between the age, plotted on the x-axis, and expenditure (in millions), plotted on the y-axis, where age is represented as a categorical variable with 5 groups: "15-24", "25-34", "35-44", "45-64", "over 64" *(Fig. 2.5).*

The x-axis represents the age groups, while the y-axis represents the total expenses in millions, ranging from 0 to 7,500 million. The highest expenses are observed in the "45-64" age group, followed by the "25-34" and "35-44" group, which show similar expenditure levels. The "15-24" group has moderate expenses, and the lower expense are in the "over 65" group.

Figure 2.6: Expenses depending on the length of the trip



Source: Our elaboration on Bank of Italy's data

Figure 2.5 provides a foundational understanding of how expenses distribute across different age categories and highlights areas for deeper analysis, especially in handling outliers and exploring non-linear relationships.

The plot illustrates the relationship between the number of night stayed, on the x-axis, and the expenses incurred on the y-axis *(Fig.2.6)*. The numbers of nights took into consideration for this plot are reduced to 150, to have a better look on the majority on the number of nights spent. The majority of data points are clustered toward the lower end of the number of nights scale, particularly between 0 to 50 nights. Expenses for these data point vary widely but are predominantly clustered at a lower end of the expenses scale.

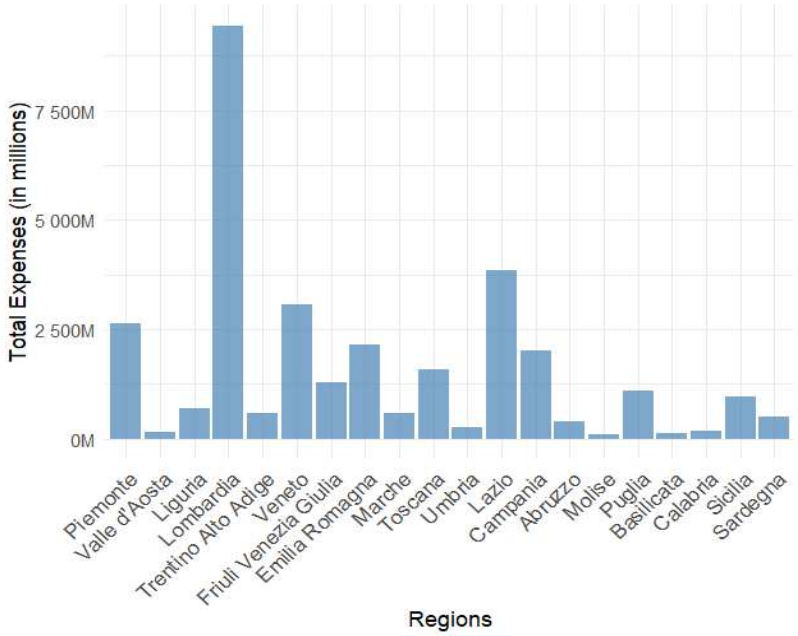
There are several outlier in the data, especially evident in the expenses variable, where some expenses are extremely high compared to the rest. Notably a few outliers also exist for higher values of nights (beyond 200 nights).

As already mentioned looking at the trend the trend line indicates on average, increases in the number of nights do not correspond with significant changes in expenses. This could suggest that for most cases, staying additional nights does not proportionally increase total expenses, which might be counterintuitive unless a fixed or discounted rate applies for longer stays.

The wide spread of expenses at lower numbers of nights highlights high variance. This variance could be attributed to different types of accommodation, varying rates, or additional spending unrelated to the number of nights (such as dining, activities, or other services).

This plot underscores the complexity of the relationship between the number of nights stayed and expenses, suggesting the influence of various factors not captured solely by the number of nights. Other factors that are indeed important to take into account like the region or residence.

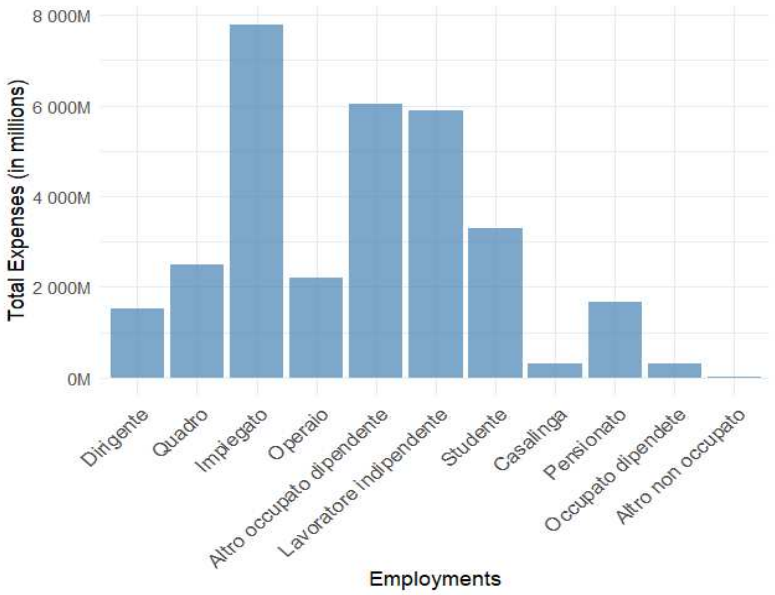
Figure 2.7: Expenses depending on the region of residence



Source: Our elaboration on Bank of Italy's data

The scatter plot visualizes the relationship between the region of residence encoded numerically and the expenses *(Fig. 2.7)*. Each numeric region code on the x-axis has a vertical spread of data points that shows the distribution of expenses for resident of each region. The spread indicated the variability in how much resident of each region spend. There are several outliers in expenses across many regions. These are data points that lie significantly above the main clusters of expenses, suggesting that there are some unusually high expenditures within each region.

Figure 2.8: Expenses depending on the employment



Source: Our elaboration on Bank of Italy's data

The plot (*Fig. 2.7)* does not exhibit a clear upward or downward trend across regions, suggesting that the region of residence might not have a strong linear relationship with the level of expenses. Expenses are scattered widely within each region, showing both low and high spenders.

The degree of spread and the presence of outliers are fairly consistent across different regions, indicating that all regions have a mix of lower and higher expense individuals.

The plot visualizes the relationship between various categories of employment, encoded numerically on the x-axis, and the expenses incurred by individuals, shown on the y-axis (*Fig. 2.8*).

There is a wide range of expense within each category of employment, as evidenced by the vertical spread of dots across all types. Notable outliers are present in several

employment categories, particularly in categories 3,6, and 9 where some individuals have expenses significantly higher than the majority within those categories.

There does not appear to be a consistent increase or decrease in expenses across the categories of employment. The trend line would likely be horizontal, suggesting no strong linear relationship between type of employment and expenses based on this plot. Despite the diversity in types of employment, the general pattern of expense distribution is consistent across different categories.

Each category shows a concentration of data points at the lower end of the expense scale, indicating that a large number of individuals in each employment category have relatively low expenses. the presence of outliers suggests that within each employment category, there may be subgroups or individuals with unique spending behaviors or circumstances, such as high earners.

External factors such as economic conditions, geopolitical events, and public health crises significantly influence travel behaviors. The survey data captured the impact of such factors on Italian tourism, with fluctuations observed in travel volumes and expenditure during periods of economic uncertainty or global crises. Moreover, changing travel restrictions and safety concerns due to health emergencies have reshaped travel preferences, leading to a shift towards domestic tourism or alternative destinations perceived as safer.

The Bank of Italy's survey provides invaluable insights into the travel behaviors and preferences of Italian tourists, offering a comprehensive understanding of demographic trends, destination choices, travel motives, and expenditure patterns. By analyzing this data, stakeholders can make informed decisions to drive sustainable growth and competitiveness in the tourism industry. Moreover, ongoing monitoring and analysis of

tourist trends are essential to adapt strategies in response to changing market dynamics and emerging opportunities. Through collaborative efforts and strategic planning, Italy can capitalize on its rich cultural heritage, natural beauty, and diverse attractions to position itself as a premier tourist destination in the global market.

CHAPTER 3

HOW DO SPECIFIC FACTORS AFFECT THE PROBABILITY OF CHOOSING AN INTERNATIONAL DESTINATION?

In this chapter, we focus on understanding how and which factors affect the probability of an Italian traveler choosing a specific destination through logistic regression, a statistical method suitable for modeling binary or multinomial outcomes. Specifically, we apply logistic regression to explore how socio-demographic and behavioral factors such as age, gender, region of residence, number of nights stayed, travel expenses, profession, and travel motive affect the likelihood of choosing popular destinations. In this analysis, we focus on three of the most frequented countries by Italian travelers: France, Switzerland, and Germany in Europe, and the most frequented one outside Europe: U.S. Each country was modeled separately, with the destination choice serving as the dependent variable (binary outcome).

3.1 Methodology and data preparation

3.1.1 Methodology

The analysis begins with the collection of relevant travel data, focusing on variables that are likely to influence a person's decision to visit a particular country. These variables

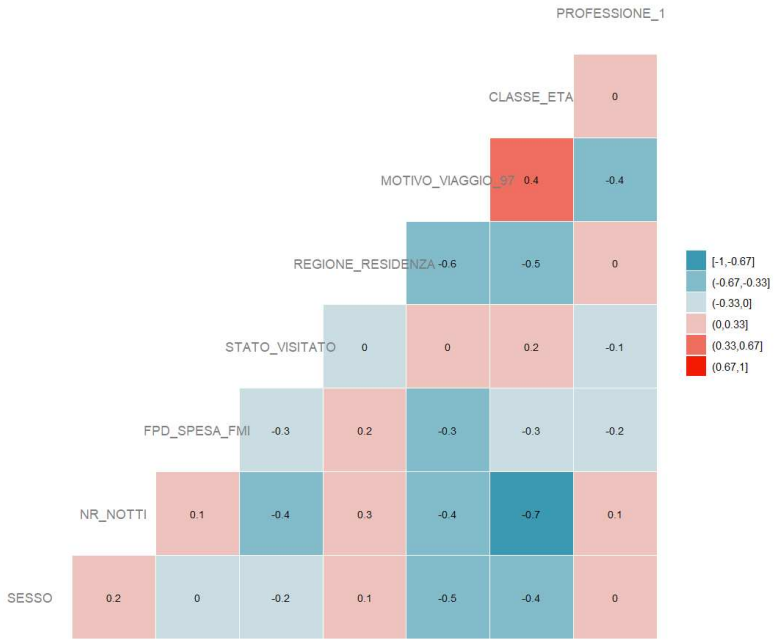
were chosen thorough a cleaning process between all the variables presented in the survey dataset, and the choice was made by focusing on the ones that could influence more a destination choice. The factors represent key demographic, socio-economic and behavioral characteristics and are age, gender, and region of residence, along with behavioral aspects such as the number of nights stayed, expenditure, profession, and travel motivation. The countries that we took into account are the most chosen by Italian travelers. To have a better look we decided to analyze three European countries and one non-European to have a wider idea.

The next step is to examine the relationships between the predictor variables through a correlation analysis. This analysis is crucial for identifying potential multicollinearity, where two or more variables are highly correlated with each other. Multicollinearity (VIF) can distort the estimates of the model coefficients, making it difficult to determine the true impact of each predictor. By calculating correlation coefficients, the analysis can reveal which pairs of variables are closely related.

3.1.2 Data Preparation

The correlation matrix reveals the linear relationships between variables such as gender, number of nights spent, daily expenditure, and the country visited. Gender shows very weak correlations with other variables, with the highest being with the region of residence and with the number of nights. There is a weak negative correlation with travel motives, suggesting a slight influence of gender on travel motivation.

Figure 3.1: Correlation Matrix



Source: Our elaboration on Bank of Italy's data

The number of nights spent shows a mild positive correlation with the region of residence and expenditure. It also has a negative correlation with age, suggesting younger travelers might stay fewer nights. There is a slight positive correlation between the country visited and age groups, indicating that the choice of destination is influenced by traveler age. Travel motivation is moderately positively correlated with age, showing a strong link between travel purpose and the traveler's age. Profession shows weak correlations with other variables, with the highest being age, suggesting that profession might be minimally influenced by age. Overall, the correlations among most variables are weak, indicating that their relationships are generally not strongly linear.

3.1.3 Logistic Regression Model

The initial modeling approach involves logistic regression, which is well-suited for predicting binary outcomes, such as whether an individual will visit a country or not (Y=1 or Y=0). The model calculates the probability of a binary outcome by transforming a linear combination of the independent variables into a probability using the logistic function. The mathematical form of the logistic regression model is:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Where:

- P(Y=1) represents the probability that the event Y (in our case, choosing a particular destination) occurs.
- $X_1, X_2, \ldots, X_n$ are the independent variables that influence the choice.
- β_0 is the intercept, it represents the log-odds of a traveler choosing the destination when all other predictors are zero. A negative intercept indicates that, in the absence of any influencing factors, the likelihood of choosing the destination is less than 0.5.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (log-odds) corresponding to the independent variables. Each coefficient reflects the change in log-odds of choosing the destination associated with a one-unit change in the corresponding independent variable. For example, a positive coefficient for age would indicate that older travelers are more likely to choose the destination, while a negative coefficient

for gender might suggest that males are less likely to choose the destination compared to females.

The coefficients β in this model are interpreted as the change in the log-odds of choosing a destination for each unit increase in the corresponding predictor variable. The odds ratios, which can be derived by exponentiating the coefficients (e^{β}), provide a more intuitive interpretation, indicating how much more likely (or less likely) a traveler is to choose a destination given a one-unit change in a predictor. Logistic regression makes several key assumptions that must be verified to ensure the validity of the model:

- Linearity of the Log-Odds, the relationship between the independent variables and the log-odds of the outcome must be linear. This does not mean that the independent variables must have a linear relationship with the dependent variable itself, but with its log-odds.
- Independence of Errors: The observations in the dataset must be independent of each other, which is particularly important in survey or cross-sectional data.
- Absence of Multicollinearity: The independent variables should not be highly correlated with each other. Multicollinearity can distort the estimates of the model coefficients. This is typically checked using variance inflation factors (VIF).

The model includes the country visited as the dependent variable and various factors as independent variables. Results are evaluated through coefficients, their statistical significance (p-values), and model fit measures like deviance and the Akaike Information Criterion (AIC). The model's performance is assessed using metrics such as accuracy,

precision, recall, and the F1 score, with the confusion matrix providing a detailed view of predictions versus actual outcomes.

Given potential data imbalance, where one outcome (e.g., non-visitors) is more common, a balanced logistic regression approach may be used. This involves adjusting the dataset to ensure equal representation of both classes during training, which helps prevent bias towards the majority class.

3.1.4 Balanced logistic regression model

The balanced logistic regression model is used to handle class imbalance, which often occurs when one class (e.g., travelers who choose a destination) is much less represented than the other (e.g., those who do not). In our case, the dataset shows that a minority of travelers select certain destinations, leading to poor performance in standard logistic regression. The balanced logistic regression model addresses this by up sampling the dataset so that both classes (choosing and not choosing a destination) are equally represented during training.

3.1.5. Random Forest

To further enhance the predictive performance, a Random Forest model is employed. Random Forest is a robust machine learning technique that builds multiple decision trees during training and aggregates their predictions to produce a final output. This method is particularly effective at capturing complex interactions between variables and is less

prone to overfitting compared to single decision trees. For a binary classification problem like destination choice, the random forest model works by training several decision trees, each considering a random sample of the *m* predictors as split candidates. The distance of each node from the predicted actual value is calculated using the Mean Square Error (MSE):

$$MSE = \frac{1}{N}\Sigma^{N}{}_{i=1}(fi - yi)^2$$

Where *yi* is the value of the data we are testing and *fi* is the value returned by the decision tree. Each tree in the random forest is built using a subset of the data (bootstrap sampling) and a random subset of the predictor variables at each split, ensuring diversity among the trees. The decision boundary created by the random forest is highly flexible, capturing complex interactions between variables like age, expenditure, and travel motivation.

In this analysis, the Random Forest model is constructed with a large number of trees (e.g., 500), each trained on a random subset of variables. The diversity among the trees helps the model generalize better to new data. The model's effectiveness is assessed through the Out-of-Bag (OOB) error rate, which offers an unbiased estimate of performance. A low OOB error rate indicates that the model is performing well and is likely to make accurate predictions on unseen data. The Random Forest model is evaluated using similar metrics as the logistic regression models, including accuracy, sensitivity, specificity, and the Kappa statistic, which measures the agreement between predicted and actual outcomes. The confusion matrix for both training and test data provides a detailed assessment of the model's performance, highlighting its strengths and areas for potential improvement.

The feature importance in random forests is typically measured by two metrics:

1. Mean Decrease in Accuracy (MDA): This indicates how much accuracy drops when a given predictor variable is removed from the model.

2. Mean Decrease in Gini (MDG): This measures the variable's contribution to reducing impurity in the decision trees, where impurity refers to how mixed the groups are in terms of classification.

3.2 Results and Interpretation

3.2.1 France

This paragraph presents a logistic regression analysis aimed at exploring the factors that influence travel decisions among the Italian population. The first country analyzed is France, the country that attracted the most Italian visitors in 2023. In the logistic regression formula France is the binary dependent variable indicating whether France was chosen as a travel destination (1 for Yes, 0 for No).

In the results of the logistic regression analysis the deviance residuals provide an indication of how well the model fits the data (Table 3.1)

The residuals in the model represent the differences between observed and predicted values, reflecting the variability in prediction errors (*Tab. 3.1).* In a well fitted model, residuals should be small and randomly distributed around zero. While the range of residuals suggests some degree of variation, extreme values, such as the maximum, indicate that certain observations are poorly predicted.

Table 3.1: France's Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-1.43	0.59	-0.44	-018	4.08

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$-7.79e^{-01}$	$3.41e^{-02}$	-22.80	<2e-16	***
Age	$1.88e^{-01}$	$4.88e^{-03}$	38.58	<2e-16	***
Gender	$-1.14e^{-01}$	$1.08e^{-02}$	-10.58	<2e-16	***
Region of Residence	$-1.28e^{-01}$	$1.03e^{-03}$	-123.43	<2e-16	***
Number of nights	$-4.46e^{-03}$	$3.01e^{-04}$	-14.83	<2e-16	***
Expenses	$-3.09e^{-07}$	$7.10e^{-09}$	-43.48	<2e-16	***
Employment	$7.81e^{-02}$	$2.44e^{-03}$	32.06	<2e-16	***
Travel Motive	$3.54e^{-02}$	$8.62e^{-04}$	41.10	<2e-16	***

Source: Our elaboration on Bank of Italy's data

Some extreme residual values suggest poor predictions for certain observations. The null deviance (353,079) and residual deviance (313,756) show that including predictors significantly improves the model's fit. The lower AIC (313,772) indicates a strong model fit while balancing complexity.

The coefficients reveal that older individuals are more likely to choose France, while males are less likely compared to females. Certain regions show lower travel likelihood to France, and longer stays slightly decrease the probability of choosing France. Spending has a negligible impact, while certain professions and travel motivations increase the likelihood of choosing France as a destination.

Table 3.2: France's Confusion matrix and Performance metrics

	Actual Negative	Actual Positive
Predictive Negative	97123	15079
Predictive Positive	70	8

Accuracy	0.8650
Precision	0.1025
Recall	0.0005
F1 Score	0.0010

Source: Our elaboration on Bank of Italy's data

Table 3.3: France's Balanced Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-2.03	-1.09	0.31	0.98	3.06

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$2.85e^{-01}$	$1.73e^{-02}$	16.485	<2e-16	***
Age	$1.39e^{-01}$	$2.57e^{-03}$	54.40	<2e-16	***
Gender	$-3.01e^{-02}$	$5.70e^{-03}$	-5.281	<2e-16	***
Region of Residence	$-7.89e^{-02}$	$4.05e^{-04}$	-194.629	<2e-16	***
Number of nights	$-2.80e^{-03}$	$1.24e^{-04}$	-22.522	<2e-16	***
Expenses	$-2.45e^{-07}$	$3.09e^{-09}$	78.988	<2e-16	***
Employment	$7.96e^{-02}$	$1.34e^{-03}$	59.177	<2e-16	***
Travel Motive	$3.29e^{-02}$	$4.31e^{-04}$	76.452	<2e-16	***

Source: Our elaboration on Bank of Italy's data

The model's accuracy of 86.51% reflects strong overall performance, but it fails to account for class imbalance. Precision for the minority class (France) is just 10.26%, indicating a high rate of false positives. The recall is extremely low at 0.05%, showing that the model rarely identifies actual cases of France selection *(Tab. 3.2)*. The F1 score of 0.0011 further highlights its poor ability to predict the minority class. While accurate for the majority class, the model struggles significantly with correctly identifying those choosing France.

In response to this, a balanced model was applied to mitigate errors stemming from class imbalance. The goal of this balanced model is to enhance the prediction of the minority class (France) and to assess how balancing the dataset impacts model performance, comparing these results with the original model. The balanced logistic regression model was built using the same predictors as the original model but applied to a dataset where the two classes (choosing France vs. not choosing France) were equally represented. This adjustment aimed to address the class imbalance present in the earlier model.

The balanced model shows less impact from extreme deviance residuals, indicating it handles outliers better.*(Fig. 3.3)* The null deviance (1,078,917 on 778,273 degrees of freedom) measures the fit of a model without predictors, while the residual deviance (964,416 on 778,266 degrees of freedom) shows improved fit with predictors included. The lower AIC (964,432) compared to the unbalanced model suggests that the balanced model provides a more accurate and efficient representation of travel decision factors, effectively managing both classes.

Table 3.4: France's Balanced Performance metrics

Accuracy	0.6292
Precision	0.2330
Recall	0.7682
F1 Score	0.3576

Source: Our elaboration on Bank of Italy's data

The balanced logistic regression model achieved an accuracy of 62.85%, lower than the 86.51% of the unbalanced model due to equal class representation *(Tab. 3.4)*. Despite the lower overall accuracy, balancing improved assessment across both classes. Precision for predicting travel to France rose to 23.25% from 10.26%, and recall increased significantly to 76.73% from 0.05%. The F1 score also improved to 0.3569 from 0.0011, reflecting better performance in predicting the minority class. The positive intercept (0.2752) indicates an increased baseline likelihood of choosing France. Overall, the balanced model enhances predictions for the minority class, demonstrating its effectiveness in providing a fair assessment of both classes.

To further enhance the analysis, a Random Forest model was applied. The Random Forest model was configured with 500 trees and 3 variables tried at each split, with feature importance enabled. The out-of-bag (OOB) error rate, which estimates how well the model generalizes to unseen data, was 1.08%. This low error rate suggests the model performs well and is highly effective at predicting outcomes for new observations.

The Random Forest model achieved an impressive accuracy of 98.86%, significantly outperforming the logistic regression model's 62.85% (*Tab. 3.5*). This high accuracy, with a narrow confidence interval (0.9880 to 0.9893), reflects the model's precision. It also surpassed the No Information Rate of 86.56%, showing its ability to perform better than

random guessing. The model's statistical significance is confirmed by an extremely low p-value (< 2.2e-16) and a Kappa value of 0.951, indicating excellent agreement between predictions and actual outcomes.

Table 3.5: France's Random Forest: Confusion Matrix and Statistics

Type of random forest	Classification
Number of trees	500
No. of variables tried at each split	3
OOB estimate of error rate	1.08%

	0	1
0	109366	241
1	34	2639

Accuracy	0.9886
95% CI	(0.988, 0.9893)
No Informaton Rate	0.8656
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.951
Mcnemar's Test P-Value	1.927e-06
Sensitivity	0.9943
Specificity	0.9521
Pos Pred Value	0.9926
Neg Pred Value	0.9630
Prevalence	0.8656
Detection Rate	0.8607
Detection Prevalence	0.8672
Balanced Accuracy	0.9732
'Positive' Class	0

Source: Our elaboration on Bank of Italy's data

The Random Forest model excelled in classification metrics: sensitivity was 99.43%, correctly identifying nearly all positive cases, while specificity was strong at 95.21%,

accurately detecting negative cases. The Positive Predictive Value (PPV) was 99.26%, ensuring most positive predictions were correct, and the Negative Predictive Value (NPV) was 96.30%, showing high accuracy in predicting negatives. The balanced accuracy of 97.32% underscores its robust performance across both classes, with a prevalence of 86.56% and a detection rate of 86.07%.

Table 3.6: France's Mean decrease accuracy and Mean decrease Gini

	0	1	MeanDecrease Accuracy	MeanDecrease Gini
Age	255.81	198.74	262.49	3723.4
Gender	222.67	115.38	189.56	1098.5
Region of Residence	330.29	737.12	579.61	40604.0
Number of nights	126.80	385.94	269.77	15064.6
Expenses	226.48	171.95	263.70	19566.3
Employment	216.94	163.98	220.09	5960.8
Travel Motive	171.13	226.72	227.98	7445.2

Source: Our elaboration on Bank of Italy's data

The feature importance analysis shows that different factors contribute variably to the predictive model(*Tab. 3.6).* Age class is a particularly significant factor, with a high Mean DecreaseAccuracy of 255.8119 and MeanDecreaseGini of 3723.412, indicating its strong impact on model accuracy and impurity reduction. Gender, though less impactful than age, is still important with MDA of 222.6878 and MDG of 1098.537. Region of residence is one of the most influential features, having the highest Gini score of 40604.044 and a MDA of 330.2965, highlighting its power in predicting travel behavior and regional imbalances. Expenditure is also significant, with MDA of 226.4769 and MDG of

19566.271, reflecting its role in reducing impurity. The number of nights stayed, while not as crucial for overall accuracy, still plays an important role. Additionally, occupation and travel motives are meaningful factors, contributing to the understanding of destination choices.

3.2.2 Switzerland

After France, we will explore the analysis of Switzerland, the second most popular destination for Italian travelers after France.

Table 3.7: Switzerland's Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max	
-1.61	-0.62	-0.34	-0.01	7.44	
	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$1.98e^{+00}$	$3.41e^{-02}$	58.038	<2e-16	***
Age	$-2.73e^{-01}$	$4.59e^{-03}$	-59.462	<2e-16	***
Gender	$4.43e^{-01}$	$9.56e^{-03}$	46.368	<2e-16	***
Region of Residence	$-1.34e^{-01}$	$1.06e^{-03}$	-126.440	<2e-16	***
Number of nights	$-1.43e^{-01}$	$1.89e^{-03}$	-75.135	<2e-16	***
Expenses	$-9.78e^{-07}$	$1.33e^{-08}$	-73.295	<2e-16	***
Employment	$-2.01e^{-02}$	$2.43e^{-03}$	-8.268	<2e-16	***
Travel Motive	$3.50e^{-02}$	$8.98e^{-04}$	38.991	<2e-16	***

Source: Our elaboration on Bank of Italy's data

The first approach employed was the logistic regression model which achieved an accuracy of 82.90%, indicating that it correctly classified the travel decisions for a

substantial majority of individuals. Precision was 61.07%, reflecting the proportion of true positives among the predicted positives, while recall stood at 21.31%, showing the model's ability to identify actual visitors. The F1 score, which balances precision and recall, was 31.59%, underscoring the challenges in achieving a balance between these two metrics.

The analysis of individual predictors shows that age has a significantly negative impact on the likelihood of visiting Switzerland, with older individuals less likely to choose it as a destination, reflecting broader trends where younger travelers are more inclined to visit (*Tab. 3.7)*. Regarding gender, males have a higher likelihood of visiting Switzerland than females, indicated by a significant positive coefficient. The region of residence is another key factor, as people from certain regions are less likely to visit Switzerland, possibly due to geographic, cultural, or economic factors. Shorter trips are more common, as indicated by the negative association between the number of nights planned and the likelihood of visiting Switzerland. Expenditure had a minimal negative effect, suggesting that higher spenders may prefer more exotic or luxurious destinations. Finally, travel motives positively influenced the choice of Switzerland, with specific reasons like tourism or business increasing the likelihood of visiting.

While the logistic regression model provides valuable insights into the relationships between predictors and the likelihood of visiting Switzerland, it faced challenges with recall, achieving only 21.31% (*Tab. 3.8)*. This low recall indicates that the model missed a significant number of actual visitors, a common issue in imbalanced datasets where the number of non-visitors far exceeds that of visitors. The F1 score of 31.59% reflects this difficulty in balancing precision and recall, highlighting the model's limitations in this context.

Table 3.8: Switzerland's Confusion Matrix and Performance Metrics

	Actual Negative	Actual Positive
Predictive Negative	88515	16474
Predictive Positive	2830	4461

Accuracy	0.8280
Precision	0.6118
Recall	0.2130
F1 Score	0.3160

Source: Our elaboration on Bank of Italy's data

Table 3.9: Switzerland's Balanced Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-2.12	-0.94	0.21	0.86	5.45

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$2.67e^{+00}$	$2.04e^{-02}$	130.811	<2e-16	***
Age	$-2.22e^{-01}$	$2.88e^{-03}$	-77.009	<2e-16	***
Gender	$3.42e^{-01}$	$6.12e^{-03}$	55.919	<2e-16	***
Region of Residence	$-1.27e^{-01}$	$6.42e^{-03}$	-197.490	<2e-16	***
Number of nights	$-7.86e^{-02}$	$7.87e^{-03}$	-99.940	<2e-16	***
Expenses	$-6.44e^{-07}$	$5.82e^{-08}$	-110.59	<2e-16	***
Employment	$1.82e^{-03}$	$1.50e^{-03}$	1.213	<2e-16	***
Travel Motive	$5.08e^{-02}$	$5.47e^{-04}$	93.151	<2e-16	***

Source: Our elaboration on Bank of Italy's data

To address the class imbalance issue, a balanced logistic regression approach was employed. This method ensured that both classes—visitors and non-visitors—were weighted equally, enhancing the model's ability to correctly identify the minority class (visitors).

In the balanced model, the coefficients analysis mainly confirmed the results of the unbalanced model as we can see in *Tab. 3.9.*

Table 3.10: Switzerland's Balanced Confusion matrix and Performance Metrics

	Actual Negative	Actual Positive
Predictive Negative	65834	2329
Predictive Positive	25511	18606

Accuracy	0.7520
Precision	0.4217
Recall	0.8887
F1 Score	0.5720

Source: Our elaboration on Bank of Italy's data

The balanced logistic regression model demonstrated a substantial improvement in recall, reaching 88.58% (*Tab. 3.10)*. This indicates that the model was significantly more effective in identifying actual visitors compared to the unbalanced model. However, this improvement in recall came with a trade-off in precision, which decreased to 41.71%. Consequently, the model's overall accuracy also dropped slightly to 74.94%, reflecting the typical trade-off between accuracy and recall when dealing with imbalanced datasets. Despite the reduction in precision and accuracy, the F1 score increased to 56.72%, showing a better balance between precision and recall.

The final model employed was a Random Forest, which demonstrated exceptional performance across all metrics.

Table 3.11: Switzerland's Random Forest: Confusion Matrix and Statistics

Type of random forest	Classification
Number of trees	500
No. of variables tried at each split	3
OOB estimate of error rate	0.21%

	0	1
0	91251	151
1	94	20784

Accuracy	0.9978
95% CI	(0.9975, 0.9981)
No Informaton Rate	0.8135
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.9928
Mcnemar's Test P-Value	0.0003466
Sensitivity	0.9990
Specificity	0.9928
Pos Pred Value	0.9983
Neg Pred Value	0.9955
Prevalence	0.8135
Detection Rate	0.8127
Detection Prevalence	0.8141
Balanced Accuracy	0.9959
'Positive' Class	0

Source: Our elaboration on Bank of Italy's data

The model achieved an impressive 99.75% accuracy, demonstrating its high effectiveness in classifying both visitors and non-visitors (*Tab. 3.11)*. With a sensitivity of 99.91%, it excels at identifying actual visitors, and a specificity of 99.05% ensures accurate exclusion of non-visitors. The balanced accuracy of 99.48% and a Kappa statistic of 0.9918 indicate near-perfect agreement between predictions and actual outcomes. The confusion matrix shows minimal misclassifications, with a near-zero class error rate for both categories. This performance highlights the Random Forest model as a top choice

for accurately predicting visits to Switzerland. (Table 12) Overall, the Random Forest model significantly outperformed both logistic regression models, demonstrating its robustness and precision in predicting travel decisions.

Table 3.12: Switzerland's Mean decrease Accuracy and Mean decrease Gini

	0	1	MeanDecrease Accuracy	MeanDecrease Gini
Age	123.49	178.66	160.99	2285.320
Gender	55.75	136.84	81.22	1141.06
Region of Residence	381.89	488.83	579.66	56740.86
Number of nights	220.82	330.83	353.02	26107.47
Expenses	80.02	93.99	102.26	27963.20
Employment	90.57	173.52	120.18	4376.33
Travel Motive	106.78	149.33	128.18	13729.52

Source: Our elaboration on Bank of Italy's data

The analysis reveals the varying importance of predictors in the model, particularly their impact on accuracy and impurity reduction (*Tab. 3.12)*. Age is highly significant, with a high Mean Decrease Gini (MDG) score of 2,285,320, indicating its strong influence on travel preferences. Gender has a more modest impact, reflected by a lower MDG score of 1,141.061. Region of residence is the most influential variable, with high MDG and Mean Decrease Accuracy (MDA) scores, highlighting the importance of geographic factors. The number of nights stayed and expenditure also play key roles, differentiating types of travelers. Travel motives, with a notable positive coefficient (0.0873), significantly influence the likelihood of choosing Switzerland.

3.2.3 Germany

The last European country analyzed is Germany, as for the other countries, a logistic regression model was employed to predict the likelihood of travelers visiting Germany based on the same factors of the above analyzed destinations. The model was configured with a binomial family, appropriate for binary outcomes.

The model's deviance residuals show a generally good fit with some outliers. The negative intercept suggests low baseline log-odds of visiting Germany (*Tab. 3.13*) Older individuals and females are significantly less likely to visit, while certain regions show a higher likelihood of visiting.

Table 3.13: Germany's Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-1.15	-0.32	-0.26	-0.19	3.34

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$-3.96e^{+00}$	$5.52e^{-02}$	-71.743	<2e-16	***
Age	$-1.81e^{-01}$	$7.72e^{-03}$	-23.331	<2e-16	***
Gender	$-6.45e^{-01}$	$2.01e^{-02}$	-32.079	<2e-16	***
Region of Residence	$6.59e^{-02}$	$8.59e^{-04}$	76.783	<2e-16	***
Number of nights	$-8.78e^{-04}$	$2.47e^{-04}$	-3.556	<2e-16	***
Expenses	$-1.01e^{-07}$	$6.56e^{-09}$	-15.475	<2e-16	***
Employment	$-1.45e^{-02}$	$4.23e^{-03}$	-3.427	<2e-16	***
Travel Motive	$8.72e^{-02}$	$1.39e^{-03}$	62.414	<2e-16	***

Source: Our elaboration on Bank of Italy's data

The number of nights spent and spending have minimal negative effects on visit likelihood. Certain professions slightly reduce, while specific travel motives strongly

increase the likelihood of visiting. All predictors are statistically significant, improving the model's fit, as evidenced by a reduction in deviance and a balanced AIC of 145,778.

Table 3.14: Germany's Confusion Matrix and Performance Metrics

	Actual Negative	Actual Positive
Predictive Negative	107574	4706

Accuracy	0.9580
Precision	NaN
Recall	0
F1 Score	NaN

Source: Our elaboration on Bank of Italy's data

Table 3.15: Germany's Balanced Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-2.55	-1.09	0.19	1.04	2.22

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$-4.49e^{-01}$	$1.62e^{-02}$	-27.79	<2e-16	***
Age	$-2.01e^{-01}$	$2.42e^{-03}$	-83.10	<2e-16	***
Gender	$-6.51e^{-01}$	$5.71e^{-02}$	-113.99	<2e-16	***
Region of Residence	$6.01e^{-01}$	$2.80e^{-03}$	214.66	<2e-16	***
Number of nights	$1.91e^{-03}$	$7.89e^{-04}$	24.16	<2e-16	***
Expenses	$-9.46e^{-07}$	$1.89e^{-09}$	-49.91	<2e-16	***
Employment	$-3.32e^{-02}$	$1.28e^{-03}$	-25.95	<2e-16	***
Travel Motive	$7.82e^{-02}$	$4.11e^{-04}$	190.15	<2e-16	***

Source: Our elaboration on Bank of Italy's data

The model's evaluation for visits to Germany shows it correctly identified 107,574 cases as non-visitors (TN) but missed 4,706 actual visitors (FN) (*Tab. 3.14)*. It failed to identify any true positives or false positives, leading to NaN values for precision and F1 score. With a recall of 0%, the model did not detect any visits to Germany. Despite high overall accuracy of 95.80%, the lack of true positives indicates significant issues with the data or model setup.

The balanced logistic regression model was developed by adjusting the dataset to correct the initial class imbalance, where there were significantly more individuals who did not visit Germany compared to those who did.

Table 3.16: Germany's Balanced Confusion matrix and Performance metrics

	Actual Negative	Actual Positive
Predictive Negative	68275	1362
Predictive Positive	39229	3344

Accuracy	0.6378
Precision	0.0784
Recall	0.7105
F1 Score	0.1412

Source: Our elaboration on Bank of Italy's data

The model's performance is highlighted by a notable reduction in deviance from the null model, with the residual deviance decreasing from 1,193,037 to 1,082,281, indicating a substantial improvement (*Tab. 3.15)*. The Akaike Information Criterion (AIC) of 1,082,297 suggests a good fit, though there is room for refinement. The coefficients from the logistic regression model reveal how various factors influence the likelihood of

traveling to Germany. The negative intercept of -0.4387 indicates the baseline log-odds of visiting Germany when all predictors are at their reference levels. Regarding the coefficients' results, the balanced model confirms the results of the unbalanced one.

The balanced logistic regression model shows a 63.82% accuracy in predicting travel to Germany (*Tab. 3.16)*. However, its precision is low at 7.87%, indicating many false positives. The model's recall is 71.29%, reflecting its effectiveness in identifying actual visits, but this comes at the expense of precision. With an F1 score of 0.1417, it struggles to balance precision and recall. The confusion matrix reveals 3,355 true visits and 68,305 true non-visits, but also 39,269 false positives and 1,351 missed visits. While the model is good at detecting actual visitors, it frequently misclassifies non-visitors as visitors, impacting precision. A Random Forest model was then used to improve predictions.

Our Random Forest model comprised 500 decision trees (*Tab. 3.17)*. Each tree was built by randomly selecting three variables at each split, which ensured diversity among the trees and helped prevent overfitting to the training data. The model's effectiveness is highlighted by its Out-of-Bag (OOB) error rate, which was exceptionally low at 0.45%. This rate, calculated using data not included in the training of individual trees, offers an unbiased estimate of the model's performance. The low OOB error signifies that the Random Forest model performed exceptionally well on the training data, making accurate predictions for most cases.

The confusion matrix shows that the model effectively differentiates between visitors and non-visitors to Germany. It correctly identified 429,813 non-visitors and 17,295 visitors in the training data, though it had 538 false positives and 1,471 false negatives, with a higher error rate for visitors (7.84%) compared to non-visitors (0.12%). In the test data,

it maintained strong performance with 107,382 true negatives and 4,397 true positives, alongside 138 false positives and 363 false negatives.

Table 3.17: Germany's Random Forest: Confusion Matrix and Performance Metrics

Type of random forest	Classification
Number of trees	500
No. of variables tried at each split	3
OOB estimate of error rate	0.42%

	0	1
0	107441	372
1	133	4334

Accuracy	0.9955
95% CI	(0.9951, 0.9959)
No Informaton Rate	0.9581
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.9426
Mcnemar's Test P-Value	< 2.2e-16
Sensitivity	0.9988
Specificity	0.9210
Pos Pred Value	0.9965
Neg Pred Value	0.9702
Prevalence	0.9581
Detection Rate	0.9569
Detection Prevalence	0.9602
Balanced Accuracy	0.9599
'Positive' Class	0

Source: Our elaboration on Bank of Italy's data

The model achieved an overall accuracy of 99.55%, with a Kappa statistic of 0.9438 indicating near-perfect agreement. Its sensitivity of 99.87% shows excellent

identification of non-visitors, while specificity at 92.37% reveals some occasional misclassification of visitors. The Positive Predictive Value (99.66%) and Negative Predictive Value (96.96%) highlight the model's effectiveness in both classes, and a balanced accuracy of 96.12% underscores its robust performance.

Table 3.18: Germany's Mean Decrease Accuracy and Mean Decrease Gini

	0	1	MeanDecrease Accuracy	MeanDecrease Gini
Age	226.09	227.42	257.59	1903.25
Gender	103.50	83.77	98.23	609.65
Region of Residence	99.13	256.85	138.94	3762.16
Number of nights	115.09	424.02	173.24	5919.31
Expenses	216.79	388.07	310.37	12812.97
Employment	16.39	312.77	258.19	3223.03
Travel Motive	123.73	342.18	183.09	13729.52

Source: Our elaboration on Bank of Italy's data

In summary, the Random Forest model excels in accuracy, sensitivity, and specificity but struggles with predicting the minority class of travelers to Germany. Age is a key feature across all countries, with high importance in both MDA and MDG (*Tab. 3.18)*. Region of residence and the number of nights spent also significantly enhance the model's performance. Gender has less impact compared to other features, as seen in France. Expenditure plays a crucial role in distinguishing between classes, while profession and travel motive have moderate importance, with travel motive being particularly vital for Switzerland.

3.2.4 U.S

After analyzing the most visited European countries, we delve into the analysis of the most visited country outside Europe, U.S.. The process began with a logistic regression model, which estimated the probability of an individual visiting the U.S. based on several predictor variables.

Table 3.19: U.S.' Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-2.21	-0.22	-0.19	-0.17	3.22

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	-2.49^{e+00}	$6.67e^{-02}$	-37.293	<2e-16	***
Age	$-1.21e^{-01}$	$1.03e^{-02}$	-11.830	<2e-16	***
Gender	$-3.65e^{-01}$	$2.34e^{-02}$	-15.635	<2e-16	***
Region of Residence	$4.18e^{-03}$	$1.15e^{-03}$	3.623	<2e-16	***
Number of nights	$1.08e^{-02}$	$1.78e^{-04}$	60.503	<2e-16	***
Expenses	$1.87e^{-07}$	2.77e-09	67.385	<2e-16	***
Employment	$-1.44e^{-01}$	$5.62e^{-03}$	-25.573	<2e-16	***
Travel Motive	$-1.02e^{-02}$	$1.57e^{-03}$	-6.493	<2e-16	***

Source: Our elaboration on Bank of Italy's data

The logistic regression model reveals key factors influencing the likelihood of visiting the U.S. Age has a negative coefficient (-0.1215), showing that older individuals are less likely to visit (*Tab. 3.19)*. Gender also has a negative impact (-0.3651), with females being less likely to visit compared to males. Region of residence has a modest positive effect (0.0042), while a longer stay increases the likelihood (0.0108). Although higher spending

is linked to visiting, the effect is small (1.869e-07). Certain professions (-0.1436) and specific travel motives (-0.0102) are associated with a lower probability of visiting.

Table 3.20: U.S.' Performance Metrics

Accuracy	0.9744
Precision	0.5285
Recall	0.0385
F1 Score	0.0718

Source: Our elaboration on Bank of Italy's data

Despite the model's high overall accuracy of 97.45%, this figure is somewhat misleading due to severe class imbalance (*Tab. 3.20)*. The model's precision was low at 52.86%, indicating many predicted visits were actually non-visits. The recall rate was only 3.85%, revealing difficulty in identifying actual visitors. The F1 score, which balances precision and recall, was just 7.18%, underscoring the model's challenges with the minority class of visitors.

To address the imbalance issue, a balanced logistic regression model was developed. This model adjusted the training process to give equal weight to both visitors and non-visitors, improving its ability to identify true visits.

The coefficients from the balanced logistic regression model confirm the results from the unbalanced regression, some coefficients like gender reduced the effect size and others like the number of nights stayed and the expenditure increased the effect (*Tab. 3.21)*. Overall, the balanced logistic regression model showed improved performance in

identifying visitors to the U.S., reflecting the importance of addressing class imbalance in predictive modeling.

Table 3.21: U.S.' Balanced Deviance Residuals and Coefficients

Min	1Q	Median	3Q	Max
-5.88	-0.89	-0.22	1.04	2.06

	Estimate	Std.Error	z.value	Pr (>|z|)	
(intercept)	$4.86e^{-01}$	$1.63e^{-02}$	29.889	<2e-16	***
Age	$-1.32e^{-01}$	$2.50e^{-03}$	-52.959	<2e-16	***
Gender	$-3.14e^{-01}$	$5.70e^{-03}$	-55.065	<2e-16	***
Region of Residence	$5.76e^{-03}$	$2.98e^{-04}$	19.339	<2e-16	***
Number of nights	$1.59e^{-02}$	$9.17e^{-05}$	173.522	<2e-16	***
Expenses	$5.24e^{-07}$	$2.13e^{-09}$	245.706	<2e-16	***
Employment	$-1.43e^{-01}$	$1.29e^{-03}$	-110.912	<2e-16	***
Travel Motive	$-7.02e^{-04}$	$3.93e^{-04}$	-1.784	<2e-16	***

Source: Our elaboration on Bank of Italy's data

Table 3.22: U.S.' Balanced Performance Metrics

Accuracy	0.8281
Precision	0.0901
Recall	0.6267
F1 Score	0.1576

Source: Our elaboration on Bank of Italy's data

The balanced logistic regression model showed notable improvements in recall, which increased to 62.67%, demonstrating a greater effectiveness in identifying actual visitors to the U.S. compared to the original model (*Tab. 3.22)*. However, this enhancement in recall came with a trade-off, as precision decreased to 9.01%, indicating a higher rate of

false positives. Consequently, while the F1 score improved to 15.76%, reflecting a more balanced performance between precision and recall, the overall accuracy of the model fell to 82.82%.

To further refine predictive performance, a Random Forest model was employed. The Random Forest model significantly outperformed the logistic regression models across several metrics.

Table 3.23: U.S.' Random Forest: Confusion Matrix and Statistics

Type of random forest	Classification
Number of trees	500
No. of variables tried at each split	3
OOB estimate of error rate	0.21%

	0	1
0	109366	241
1	34	2639

Accuracy	0.9976
95% CI	(0.9972, 0.9978)
No Informaton Rate	0.9743
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.9492
Mcnemar's Test P-Value	< 2.2e-16
Sensitivity	0.9997
Specificity	0.9163
Pos Pred Value	0.9978
Neg Pred Value	0.9873
Prevalence	0.9743
Detection Rate	0.9740
Detection Prevalence	0.9762
Balanced Accuracy	0.9580
'Positive' Class	0

Source: Our elaboration on Bank of Italy's data

The Random Forest model achieved an impressive accuracy of 99.76%, showcasing its exceptional ability to accurately classify both visitors and non-visitors (*Tab. 3.23)*. It demonstrated near-perfect sensitivity at 99.97%, highlighting its effectiveness in correctly identifying non-visitors and reliably ruling out individuals unlikely to visit the U.S. Although its specificity was slightly lower at 91.63%, the model still performed strongly in identifying actual visitors.

The Kappa statistic of 0.9492 indicates near-perfect agreement between predicted and actual outcomes, underscoring the model's robustness. The confusion matrix revealed minimal misclassifications, with an Out-of-Bag (OOB) error rate of just 0.21%. The positive predictive value of 99.78% signifies that nearly all predicted visits to the U.S. were accurate, while the negative predictive value of 98.73% demonstrates strong performance in predicting non-visitors. The Balanced Accuracy of 95.8% further emphasizes the model's capability to effectively handle both classes, making it particularly suitable for predicting visits to the U.S., especially in scenarios involving imbalanced datasets.

Table 3.24: U.S.' Mean Decrease Accuracy and Mean Decrease Gini

	0	1	MeanDecrease Accuracy	MeanDecrease Gini
Age	126.99	165.16	151.88	1261.34
Gender	122.54	99.50	118.85	528.27
Region of Residence	140.16	164.65	151.41	2585.75
Number of nights	136.89	449.39	212.20	4287.35
Expenses	170.77	358.94	237.72	8544.37
Employment	189.05	242.36	237.33	1480.39
Travel Motive	174.67	236.68	216.15	1320.50

Source: Our elaboration on Bank of Italy's data

The results for the random forest analysis illustrate the varying impact of different features on the model's performance (*Tab. 3.24)*. Age shows notable values in both MenDecreaseAccuracy and MeanDecreaseGini, highlighting it significant role in improving the model's accuracy and its ability to differentiate between classes. Region of residence also demonstrates considerable importance, especially in MeanDecreaseGini, which suggests it is a key factor in the model's classification capability. The number of nights and the expenditure stand out for its high scores, both enhancing the model's predictive performance. The profession and the travel motive show moderate but notable contributions, both playing a significant role in the model.

3.3 Comparison between countries

The comparative analysis of travel patterns between the U.S., France, Switzerland, and Germany reveals important insights into the factors influencing the likelihood of visiting each destination, with key variations in visitor profiles across these countries.

For the U.S., the results show that age has a significant negative impact, with older individuals less likely to visit, aligning with trends of reduced travel among older populations. Gender also plays a crucial role, as females are less likely to travel to the U.S. than males, indicating a higher propensity for male travelers. The region of residence has a modest but positive effect, suggesting that geographic proximity or regional travel trends influence U.S. travel. Travelers planning longer stays are more likely to visit, as shown by the positive coefficient for the number of nights, which could be due to the U.S.'s appeal as a destination for extended vacations. Interestingly, higher spending is linked to an increased likelihood of visiting, likely due to the greater costs of travel and

accommodation in the U.S. Despite this, the diversity of traveler profiles makes predicting exact visitor types more complex, with motivations ranging from leisure tourism to business, education, and family visits.

In Germany, gender again plays a significant role, with males more likely to visit than females. The slight positive correlation with duration of stay suggests that travelers planning longer visits are more inclined to choose Germany, likely due to its wide range of attractions. The negative effect of expenditure, however, is notable, indicating that higher spending slightly reduces the likelihood of visiting, which contrasts with trends seen in the U.S. In Germany, profession and travel motives also show moderate importance in influencing travel decisions. Germany's strong balance between precision and recall in predictive models highlights its reliability in identifying both who will visit and who will not, making it one of the most predictable destinations among the countries analyzed.

For France, cultural experiences, luxury travel, and historical tourism stand out as the main drivers of visits. The country attracts a specific group of well-defined travelers, contributing to the model's high precision in identifying likely visitors. However, the lower recall suggests that France may miss capturing broader traveler segments. This specificity contrasts with the U.S., where more varied motivations lead to greater visitor unpredictability.

In Switzerland, the results show strong recall, particularly identifying potential travelers interested in nature, adventure tourism, and global business. Switzerland, like the U.S., attracts a diverse visitor base, but the predictive models excel at identifying a broad range of potential visitors. The longer stays, similar to Germany, indicate that extended visits to

enjoy Switzerland's outdoor and adventure tourism offerings are a strong determinant of travel decisions.

In summary, the U.S. stands out for its wide appeal but faces challenges in predicting the exact type of visitor due to the diversity of its tourist base. Germany presents a balanced and predictable profile, with strong indicators from gender and length of stay. France excels in drawing specific traveler groups, especially those interested in its rich cultural heritage, while Switzerland demonstrates strength in identifying a broad spectrum of visitors, particularly those focused on nature and adventure tourism. Each country's unique appeal is reflected in how different features like age, gender, spending, and duration of stay influence the likelihood of a visit.

CHAPTER 4

WHICH FACTORS DIFFERENTIATE TRAVELERS WHO CHOOSE A SPECIFIC DESTINATION?

After analyzing which factor influence most a destination choice, we wanted to investigate what factors best differentiate travelers who choose a specific destination and we have chosen discriminant analysis as the primary tool to investigate this. Discriminant analysis, specifically Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), is well-suited for examining how demographic and behavioral factors differentiate travelers according to their destination choices. By utilizing the same variable chosen for the logistic regression model, we aim to determine which of these factors most effectively classify travelers based on their destination.

4.1 Discriminant analysis

4.1.1 Linear Discriminant Analysis

Linear Discriminant Analysis is a popular classification technique used when the goal is to separate two or more classes (e.g., visitors and non-visitors) by finding a linear combination of the predictor variables that maximizes class separability. It assumes that

the classes share a common covariance matrix and that the relationships between the predictors and the response variable are linear. μ_k

$$\delta_k(X) = X^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k{}^T \Sigma^{-1} \mu_k + log(\pi_k)$$

where:

- X is the vector of predictor variables (e.g., age, gender, expenditure, etc.)
- μ_k is the mean vector of group k
- Σ is the common covariance matrix across groups
- π_k is the prior probability of group k (e.g., proportion of travelers to a destination)
- $\delta_k(X)$ is the discriminant score for group k

The group to which a traveler is classified is determined by the group with the highest discriminant score. This is the group k for which $\delta_k(X)$ is maximized. In this case, LDA is applied to model the likelihood that a person will visit a country based on factors such as age, gender, profession, and travel motivations. The method calculates a linear boundary between visitors and non-visitors by maximizing the variance between classes while minimizing the variance within each class. The effectiveness of LDA is evaluated by metrics such as accuracy, precision, recall, and the confusion matrix.

4.1.2 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis is an extension of LDA, but it is more flexible as it does not assume that the covariance matrices of the classes are identical. Instead, QDA allows

for different covariance structures between the visitor and non-visitor groups. This makes QDA more appropriate for datasets where class distributions have different variances and covariances.

The discriminant function for QDA is:

$$\delta_k(X) = -\frac{1}{2}(x-\mu_k)^T {\Sigma_k}^{-1}(x-\mu_k) - \frac{1}{2} log \mid \Sigma_k \mid + log(\pi_k)$$

Where $\delta_k(X)$ is the discriminant score for group k, X is the vector of predictor variables (e.g., age, gender, expenditure, etc.), μ_k is the mean vector of group k, Σ_k is the covariance matrix specific for the k group and π_k is the prior probability of group k (e.g., proportion of travelers to a destination).

In QDA, the quadratic term $(X-\mu_k)^T {\Sigma^{-1}}_k (X-\mu_k)$ models the differences in spread (variance) between groups. This allows QDA to capture more complex decision boundaries, making it better suited for data where variance between groups differs significantly, for example the variance in spending between budget travelers and luxury travelers. (G. James,

4.2 Results and Interpretations

4.2.1 France

As we already said above, to delve more into the analysis we decided to perform a discriminant analysis, both linear and quadratic, which should help us to understand better the data. The first country analyzed is France. Analyzing travel predictions to France

using Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) reveals distinct insights into how these models differentiate between visitors and non-visitors. Both methods apply various predictors, such as age, gender, residence region, number of nights stayed, expenditure, profession, and travel motive, to classify individuals as either having visited France (coded as 1) or not (coded as 0).

The Linear Discriminant Analysis aims to distinguish between those who have visited France and those who haven't, based on the above mentioned predictors.

Table 4.1: France's LDA Probabilities of groups and Group means

0	1
0.8662	0.1337

	Age	Gender	Region of Residence	Number of nights	Expenditure	Employment	Travel Motives
0	4.83	1.32	21.33	10.83	945144.9	4.83	10.41
1	4.36	1.25	16.04	4.07	438239.7	5.05	12.9

Source: Our elaboration on Bank of Italy's data

Table 4.2: France's LDA Coefficients

	LD1
Age	$1.88e\text{-}^{01}$
Gender	$-1.19e^{-01}$
Region of Residence	$-9.42e^{-02}$
Number of nights	$-3.13e^{-03}$
Expenditure	$-5.71e^{-08}$
Employment	$1.32e^{-01}$
Travel Motives	$5.06e^{-02}$

Source: Our elaboration on Bank of Italy's data

The model shows a notable class imbalance, with only about 13.37% of the data representing visitors to France, while the remaining 86.63% are non-visitors (*tab. 4.1)*. This imbalance shapes the model's behavior, making it more inclined to predict non-visitors due to their overwhelming majority.

In terms of group means, LDA identifies that visitors to France are typically slightly older, spend less money, and stay for fewer nights compared to non-visitors. Additionally, differences in gender and region of residence also emerge, suggesting these factors influence the likelihood of visiting France.

Being the majority non-visitors, the coefficients of the discriminant function highlight that age and profession have positive coefficients, meaning that older individuals and those in certain professions are more likely to be classified as non-visitors (*Tab. 4.2)*. Conversely, gender and region of residence have negative coefficients, indicating these factors are more associated with visiting France.

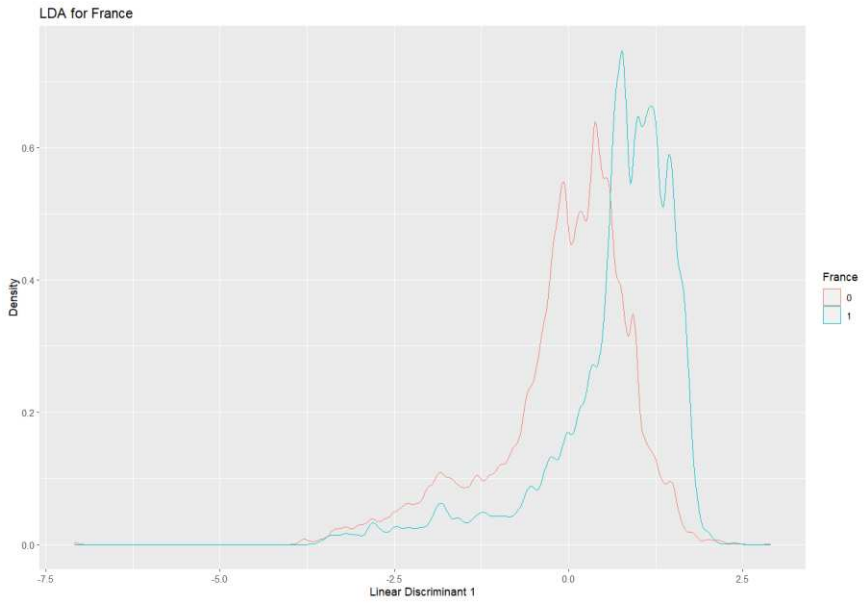
Table 4.3: France's LDA Confusion Matrix

	Actual Negative	Actual Positive
Predictive Negative	486302	75067
Predictive Positive	28	0

Source: Our elaboration on Bank of Italy's data

The confusion matrix for LDA reveals its limitations, especially in dealing with the class imbalance (*Tab. 4.3)*. While the model successfully identifies many non-visitors, it performs poorly in correctly classifying visitors. In fact the high rate of misclassification for the minority class (visitors) reflects the struggle. (Table 28)

Figure 4.1: LDA for France



Source: Our elaboration on Bank of Italy's data

The *Figure 4.1* shows the LDA classification for France, comparing visitors (blue line) and non-visitors (red line).

The Quadratic Discriminant Analysis extends the capabilities of LDA by allowing each class to have its own covariance matrix, offering a more flexible approach to modeling

the relationship between predictors and the outcome. This flexibility helps QDA handle complex relationships and better manage class imbalances.

Table 4.4: France's QDA Probabilities of groups

0	1
0.8662	0.1337

Source: Our elaboration on Bank of Italy's data

Like LDA, QDA deals with a significant class imbalance where only 13.37% of the data represents visitors (*Tab. 4.4)*. However, QDA's ability to model different covariance structures for each class improves its performance. The group means for QDA align with those identified by LDA, showing that visitors are generally older, spend less, and stay fewer nights compared to non-visitors. Nevertheless, QDA's confusion matrix demonstrates notable improvements in classifying visitors (*Tab. 4.5)*. It correctly identifies a considerable number of visitors, outperforming LDA in handling the minority class.

Table 4.5: France's QDA Confusion Matrix

	Actual Negative	Actual Positive
Predictive Negative	354576	24107
Predictive Positive	131754	50960

Source: Our elaboration on Bank of Italy's data

Despite some misclassifications where non-visitors are incorrectly labeled as visitors, QDA's enhanced modeling capabilities lead to a more accurate prediction of who is likely to visit France. Thus, while both models provide insights into the factors influencing travel to France, QDA stands out as the more robust option for handling complex data structures and improving classification accuracy.

4.2.2 Switzerland

Moving on with the analysis of Switzerland, LDA is used to classify individuals as either having visited Switzerland (coded as 1) or not (coded as 0).

The model reveals a notable class imbalance, with approximately 18.6% of the dataset representing visitors to Switzerland, while 81.4% are non-visitors (*Tab. 4.6)*. This disparity affects the model's performance, as it tends to predict the majority class (non-visitors) more frequently.

Table 4.6: Switzerland's LDA Probabilities of means and Group means

0	1
0.8139	0.1860

	Age	Gender	Region of Residence	Number of nights	Expenditure	Emplo yment.	Travel Motives
0	4.07	1.30	21.65	11.97	1019129	4.91	10.12
1	4.01	1.33	16.38	0.98	257311	4.61	13.46

Source: Our elaboration on Bank of Italy's data

Table 4.7: Switzerland's LDA Coefficients

	LD1
Age	$-3.03e^{-01}$
Gender	$5.17e^{-01}$
Region of Residence	$-8.11e^{-02}$
Number of nights	$-1.15e^{-02}$
Expenditure	$-9.79e^{-08}$
Employment	8.86^{-03}
Travel Motives	$7.90e^{-02}$

Source: Our elaboration on Bank of Italy's data

Regarding group means, LDA shows that visitors to Switzerland are generally younger, stay fewer nights, and spend less compared to non-visitors. Additionally, differences in gender and region of residence are observed, suggesting these factors play a role in predicting travel to Switzerland.

The coefficients of the discriminant function indicate that age has a negative coefficient, implying that younger individuals are more likely to visit Switzerland (*Tab. 4.7)*. Gender has a positive coefficient, highlighting its significant role in the prediction. Other factors such as region of residence and number of nights stayed also influence the likelihood of visiting.

Table 4.8: Switzerland's LDA Confusion Matrix

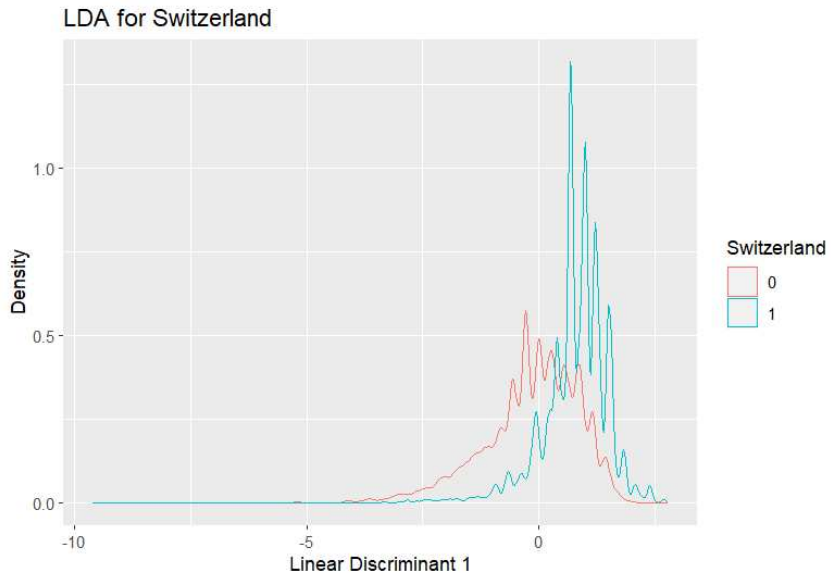
	Actual Negative	Actual Positive
Predictive Negative	456210	102313
Predictive Positive	718	2156

Source: Our elaboration on Bank of Italy's data

The confusion matrix for LDA reflects the model's struggle with classifying visitors accurately (*Tab. 4.8)*. The model correctly classifies 2,156 visitors but incorrectly labels 102,313 non-visitors as visitors. This performance highlights LDA's limitations in dealing with the class imbalance and its assumption of equal covariance matrices across groups.

The *Fig. 4.1* visualizes a linear discriminant function's ability to separate visitors (blue) from non-visitor (red). QDA improves upon LDA by allowing each class to have its own covariance matrix, offering greater flexibility in modeling the relationships between predictors and the outcome.

Figure 4.2: LDA for Switzerland



Source: Our elaboration on Bank of Italy's data

Like LDA, QDA faces a significant class imbalance, with 18.6% of the dataset representing visitors. The group means for QDA are consistent with those observed in LDA, indicating that visitors to Switzerland are generally younger, stay fewer nights, and spend less. Nonetheless, QDA shows a marked improvement in identifying visitors.

It correctly classifies 96,754 out of 253,672 actual visitors, demonstrating its superior performance compared to LDA (*Tab. 4.9)*. Despite some misclassification where non-visitors are incorrectly identified as visitors, QDA's ability to manage the class imbalance more effectively results in a better overall prediction of who is likely to visit Switzerland.

Table 4.9: Switzerland's QDA Confusion Matrix

	Actual Negative	Actual Positive
Predictive Negative	203256	7715
Predictive Positive	253672	96754

Source: Our elaboration on Bank of Italy's data

4.2.3 Germany

The third European country took into analysis is Germany where LDA, as for the already analyzed countries, categorizes individuals based on predictors such as age, gender, region of residence, number of nights stayed, expenditure, profession, and travel motive.

The model's prior probabilities reveal a striking class imbalance, with only 4.2% of the dataset consisting of visitors to Germany and 95.8% as non-visitors (*Tab. 4.10)*.

Looking at the group means, LDA identifies some clear differences between visitors and non-visitors.

Table 4.10: Germany's LDA Probabilities of means and Group means

0	1
0.96	0.42

	Age	Gender	Region of Residence	Number of nights	Expenditure	Employment	Travel Motives
0	4.07	1.31	20.52	9.86	883931.4	4.87	10.63
1	3.95	1.18	24.05	11.45	72722.8	4.62	13.29

Source: Our elaboration on Bank of Italy's data

Visitors tend to be slightly younger, stay longer, spend less, and show distinct gender and regional patterns. These variations in characteristics help the model form predictions.

Table 4.11: Germany's LDA Coefficients

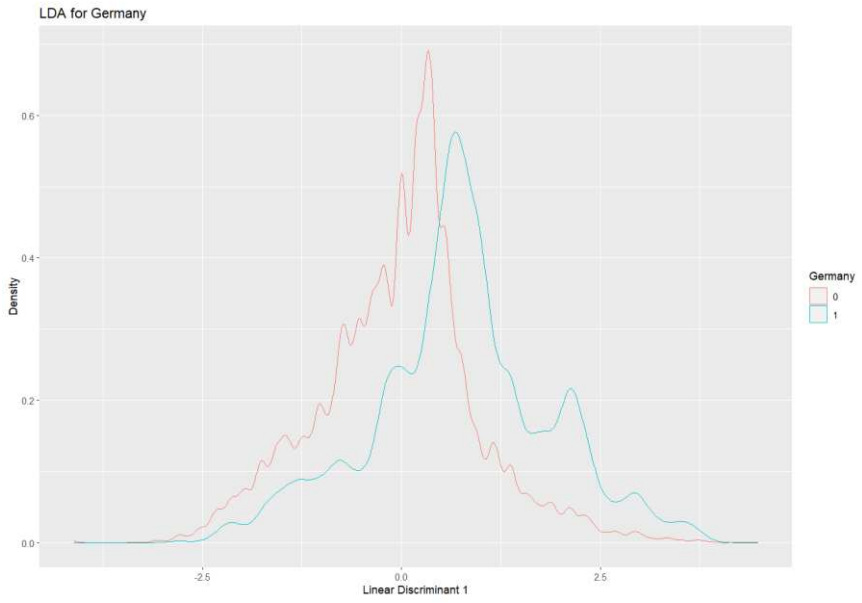
	LD1
Age	$-2.15e^{-01}$
Gender	$-7.16e^{-01}$
Region of Residence	$9.73e^{-02}$
Number of nights	$3.68e^{-04}$
Expenditure	$-4.51e^{-08}$
Employment	$-1.89e^{-02}$
Travel Motives	$1.08e^{-01}$

Source: Our elaboration on Bank of Italy's data

The LDA coefficients reveal that younger individuals and certain gender profiles are more likely to visit Germany, with age and gender significantly impacting the likelihood (*Tab. 4.11)*. Age has a negative coefficient, suggesting younger people are more inclined to visit. Gender also influences the likelihood, as shown by its large negative coefficient. However, LDA's performance is hampered by class imbalance and the assumption of equal covariance, which affects its ability to identify the minority class (visitors). Here we can see the difference between visitors and non-visitors (*Fig. 4.3)*.

QDA's results are similar to LDA's, with visitors being younger, staying longer, and spending less. The confusion matrix for QDA shows improved performance over LDA, with better identification of visitors while still occasionally misclassifying non-visitors.

Figure 4.3: LDA for Germany



Source: Our elaboration on Bank of Italy's data

4.2.4 USA

Then we moved to U.S. where the prior probabilities of LDA show a significant imbalance, with 97.45% of the dataset representing non-visitors and only 2.55% representing visitors.

Table 4.12: U.S.' LDA Probabilities of groups and Group means

0	1
0.97	0.03

	Age	Gender	Region of Residence	Number of nights	Expenditure	Employment	Travel Motives
0	4.07	1.31	20.61	9.32	797050.6	4.87	10.77
1	3.72	1.28	22.77	33.01	4185141.8	4.52	9.60

Source: Our elaboration on Bank of Italy's data

The group means provide insights into the characteristics of visitors versus non-visitors (*Tab. 4.12*). Visitors to the USA tend to be younger, stay fewer nights, and spend more compared to non-visitors. For instance, the average age class for non-visitors is higher than for visitors, and the average expenditure for visitors is significantly greater.

The coefficients of the linear discriminant function indicate that the number of nights stayed has a positive coefficient, suggesting that more nights are associated with a higher likelihood of visiting (*Tab. 4.13)*.

Table 4.13: U.S.' LDA Coefficients

	LD1
Age	$-3.12e^{-02}$
Gender	$-1.93e^{-01}$
Region of Residence	$-1.84e^{-03}$
Number of nights	$1.54e^{-02}$
Expenditure	$3.74e^{-07}$
Employment	$-6.54e^{-02}$
Travel Motives	$4.78e^{-03}$

Source: Our elaboration on Bank of Italy's data

Table 4.14: U.S.' LDA Confusion Matrix

	Actual Negative	Actual Positive
Predictive Negative	540916	12672
Predictive Positive	6156	1653

Source: Our elaboration on Bank of Italy's data

Conversely, variables such as age class and gender have negative coefficients, implying that older age and certain gender categories are less likely to visit the USA.

The confusion matrix for LDA shows that while the model accurately predicts a large number of non-visitors (540,916 correct predictions), it struggles with identifying visitors, correctly classifying only 1,653 out of 14,325 actual visitors (*Tab. 4.14)*.

QDA, despite his flexibility, shows similar group means to LDA, with visitors being younger, staying fewer nights, and spending more.

Table 4.15: U.S.' QDA Confusion Matrix

	Actual Negative	Actual Positive
Predictive Negative	528336	11493
Predictive Positive	18736	2832

Source: Our elaboration on Bank of Italy's data

QDA's confusion matrix demonstrates improved performance over LDA in some areas (*Tab. 4.15)*. It correctly identifies 2,832 visitors out of 15,305, indicating a better capability to capture the nuances of the visitor class. Overall, the analysis underscores the challenge of predicting the minority class in a heavily imbalanced dataset. While LDA provides a more interpretable model, QDA offers greater flexibility but with higher misclassification rates for non-visitors.

4.3 Comparison between European and non-European countries

After this analysis we can understand the demographic factors, spending patterns, and travel behaviors that influenced Italian traveler's destination choice. Age emerged as a key demographic variable, with younger individuals generally being more likely to visit countries like Switzerland, Germany, and the USA, while older individuals were more inclined toward France. Gender differences also played a role, with variations in likelihood of visiting based on gender, particularly in Germany and the USA, where

specific gender profiles showed higher tendencies to visit. For Germany the model indicated that the likelihood of visiting Germany was higher for male travelers compared to females. For the USA, the gender variable also had a negative coefficient, suggesting that males were less likely to visit the USA compared to females.

In terms of spending patterns, travelers visiting the USA and Germany tended to spend more on average, while visitors to France and Switzerland generally spent less. Interestingly, despite higher expenditure, visitors to the USA stayed fewer nights, suggesting that trips to the USA might be shorter but more expensive, possibly due to factors such as higher travel costs or more expensive activities. In contrast, travelers to France, Switzerland, and Germany showed patterns of staying longer but spending less, indicating a different kind of travel experience.

These factors collectively shaped the distinct travel patterns for each destination, giving us a broader idea on the factors influencing a destination choice.

CONCLUSION

This thesis set out to explore the factors influencing Italian tourists' destination choices, combining descriptive data analysis with robust statistical methods to offer a multi-faceted view of the tourism landscape. Through a detailed examination of the historical and conceptual framework of tourism, it became clear that this phenomenon is not only an economic driver but also a complex interplay of social and demographic factors.

The descriptive analysis in the second chapter highlighted key travel trends among Italian tourists, such as their preference for nearby destinations, with a particular inclination toward northern countries. Popular destinations included France, Switzerland and Germany, with leisure and family visits being the predominant motivations behind travel. Outside of Europe, the United States was the most frequently visited non-European destination, though travel to regions such as Asia and Africa remained less common due to accessibility challenges and safety concerns. Furthermore, the study underscored the impact of specific demographic variables—such as age, gender, and profession—on travel choices, showing that these factors play a substantial role in determining both the destination and the type of travel experience sought, in fact the study showed that older tourists, particularly those above 65, tended to prefer shorter trips.

The third chapter employed logistic regression model statistical models to gain deeper insights into the factors influencing travel decisions. It is important to note that the analysis was conducted on selected countries, focusing primarily on destinations frequently visited by Italian tourists. Logistic regression provided valuable predictions about travel behavior, identifying age, gender, and income as critical variables. For example, the probability of visiting France increases with age, while males were found to

be less likely than females to choose France as a destination. Similarly, younger individuals were more likely to visit the USA, but gender differences showed that females were less likely to visit the USA than males. Spending patterns revealed that travelers to the USA and Germany tended to spend more but stayed fewer nights, indicating a preference for shorter, more expensive trips. In contrast, trips to France and Switzerland involved longer stays but lower average expenditures. The analysis also highlighted the role of professional status, with professionals and managers generally demonstrating a higher propensity for long-haul travel compared to retirees or manual workers, who preferred destinations within Europe.

The fourth chapter focused on how discriminant analysis farther clarified the distinctions between travelers visiting specific destinations and those who did not. This analysis reinforced and complemented the findings from logistic regression, improving the model's ability to handle class imbalances and enhance predictive accuracy.

Overall, this thesis offers a comprehensive examination of the complex interplay between demographic, economic, and motivational factors in destination selection. The findings underscore the importance of considering not only individual preferences but also broader economic conditions when analyzing travel behavior.

Potential focuses for future research could involve expanding the scope of the analysis to include a broader range of countries, particularly those in less common travel regions such as Asia and Africa. Additionally, integrating emerging trends such as the impact of climate change on travel preferences could provide a deeper understanding of evolving tourist behavior, furthermore the role of technological advancements, particularly in relation to online booking platforms and social media influence, presents a possible area for further investigation. Finally, a longitudinal study observing changes in travel

behavior over time could offer valuable insights into the long-term effects of global events on tourism.

REFERENCES

Banca d'Italia. (2024). *Indagine sul turismo internazionale*.

Banca d'Italia. *Turismo internazionale.* www.bancaditalia.it/statistiche/tematiche/rapporti-estero/turismo-internazionale/index.html

Buck, R. (1978). Toward a synthesis in tourism theory. *Annals of Tourism Research*, 1, 110-111.

Cárdenas-García, P. J., Sánchez-Rivero, M., & Pulido-Fernández, J. I. (2015). Does tourism growth influence economic development? *Journal of Travel Research*, 54(2), 206-221.

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, 19(1), 1-14.

Crompton, J. L. (1979). Motivations for pleasure vacation. *Annals of Tourism Research*, 6(4), 408–424.

ENIT (2023). *Tourism: nine out of ten Italians are already making plans for next season*.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons, 7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R* (2nd ed.). Springer., 131-135, 156-163,319-321, 327-330

Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). Descriptive statistics. *International Journal of Academic Medicine*, 4(1), 60-63.

Leiper, N. (2004). *Tourism, critical concepts in the social sciences: Volume I*. Routledge, 26-27.

Makhlouf, H. H. (2012). The multi-dimensional impact of international tourism. *International Business & Economics Research Journal*, 11(2), 265-274.

OECD. (2022). *Tourism trends and policies*. *OECD iLibrary*. https://www.oecd-ilibrary.org/sites/698aaf87-en/index.html?itemId=/content/component/698aaf87-en

Padure, G., & Turtureanu, I. A. (2005). Economic impact of tourism. *Economics*, 1, 129-138.

Richardson, J., & Fluker, M. (2004). *Understanding and managing tourism*. Pearson Education.

Runchi, Z., Liguo, X., & Qin, W. (2023). An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting. Journal of Credit Risk.

Salas-Eljatiba, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A., & Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*.

Statista. (2024). *GDP share generated by travel and tourism in Italy 2019-2023*.

Statista. (2024). *Monthly number of Italian outbound tourists 2018-2024*.

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*.

WTTC. (2022). *Capital investment fuels growth in travel & tourism, forecast to reach nearly $1 trillion says WTTC*.

WTTC. (2023). *Travel & tourism sector shows strong recovery in Italy says WTTC*.

WTTC. (2024). Economic impact research.

APPENDIX

This appendix provides the R code used for the statistical analyses conducted in this thesis, including the descriptive analysis, logistic regression models, and discriminant analysis. The code demonstrates the use of various R packages for data manipulation, visualization, and modeling.

```
#Descriptive Analysis:
#Count
table(data$SESSO)
#Gender pie chart
labels <- c("Men", "Women")
sizes <- c(29761, 16818)
percentages <- round(sizes / sum(sizes) * 100, 1)
pie(sizes, labels = paste(labels, percentages, "%"),
    col = c("steelblue", "lightsteelblue"), main = "Gender
Distribution", border="darkblue")
#Gender and Age
table(data$SESSO, data$CLASSE_ETA)
barplot(table(data$SESSO, data$CLASSE_ETA), names.arg = c("15-24", "25-
34", "35-44", "45-64", "over65"), legend = c("men", "women"), col =
c("steelblue", "lightsteelblue"))
#Bar chart for regions
regioni_labels <- c("Piemonte", "Valle d'Aosta", "Liguria", "Lombardia",
"Trentino Alto Adige", "Veneto", "Friuli Venezia Giulia", "Emilia
Romagna", "Marche", "Toscana", "Umbria", "Lazio", "Campania", "Abruzzo",
"Molise", "Puglia", "Basilicata", "Calabria", "Sicilia", "Sardegna")
#Assign labels
#Plot
```

```
ggplot(principali2023, aes(x = as.factor(REGIONE_RESIDENZA))) +
  geom_bar(fill = "steelblue") +
  scale_x_discrete(labels = c("Piemonte", "Valle d'Aosta", "Liguria",
"Lombardia", "Trentino Alto Adige", "Veneto", "Friuli Venezia Giulia",
"Emilia Romagna", "Marche", "Toscana", "Umbria", "Lazio", "Campania",
"Abruzzo", "Molise", "Puglia", "Basilicata", "Calabria", "Sicilia",
"Sardegna")) +
  labs(x = "Regions", y = NULL) +   # 'NULL' rimuove la label sull'asse
y
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 10))
*
*
*
#Logistic Regression model:
library(dplyr)
library(corrplot)
library(ggplot2)
#Correlation
cor_matrix <- cor(expanded_data[c("SESSO", "NR_NOTTI", "FPD_SPESA_FMI",
"STATO_VISITATO"    ,    "REGIONE_RESIDENZA"    ,"MOTIVO_VIAGGIO_97",
"CLASSE_ETA",  "PROFESSIONE_10"  )],  use  =  "complete.obs")  #  use
complete.obs to handle missing values
print(cor_matrix)
ggcorr(cor_matrix,
       nbreaks = 6,
       label = TRUE,
       label_size = 3,
       color = "grey50")
```

```
str(expanded_data)
#Binary variable
expanded_data$France <- ifelse(expanded_data$STATO_VISITATO == 29, 1,
0)
table(expanded_data$France)
print(expanded_data$France)
# Convert the response variable to a factor
expanded_data$France <- as.factor(expanded_data$France)
set.seed(123)  # For reproducibility
# Training set
trainIndex <- sample(seq_len(nrow(expanded_data)), size = 0.8 *
nrow(expanded_data))
trainData <- expanded_data[trainIndex, ]
testData <- expanded_data[-trainIndex, ]
# Fit the logistic regression model for France
model <- glm(France ~ CLASSE_ETA + SESSO + REGIONE_RESIDENZA + NR_NOTTI
+ FPD_SPESA_FMI + PROFESSIONE_10 +MOTIVO_VIAGGIO_97, data = trainData,
family = binomial)
summary(model)
vif(model) #multicllinearity

# Predict probabilities
probabilities <- predict(model, testData, type = "response") #to predict
value in input data
# Convert probabilities to binary predictions
predictions <- ifelse(probabilities > 0.5, 1, 0)
predictions <- as.factor(predictions)
# Actual values
```

```
actuals <- as.factor(testData$France)
# Create the confusion matrix
confMatrix <- table(Predicted = predictions, Actual = actuals)
# Print the confusion matrix
print(confMatrix)

# True Positives, True Negatives, False Positives, False Negatives
TP <- confMatrix["1", "1"]
TN <- confMatrix["0", "0"]
FP <- confMatrix["1", "0"]
FN <- confMatrix["0", "1"]

# Calculate accuracy, precision, recall, and F1 score
accuracy <- (TP + TN) / sum(confMatrix)
precision <- TP / (TP + FP)
recall <- TP / (TP + FN)
f1_score <- 2 * (precision * recall) / (precision + recall)
print(paste("Accuracy:", accuracy))
print(paste("Precision:", precision))
print(paste("Recall:", recall))
print(paste("F1 Score:", f1_score))

#Balaced logistic regression model
trainData$France <- as.factor(trainData$France)
# Use upSample to balance the training data
trainData_balanced <- upSample(x = trainData[, -ncol(trainData)], y =
trainData$France)
table(trainData_balanced$France)
```

```
names(trainData_balanced)[names(trainData_balanced) == "Class"] <-
"France"
# Fit the balanced logistic regression model
model_balanced <- glm(France ~ CLASSE_ETA + SESSO + REGIONE_RESIDENZA +
NR_NOTTI + FPD_SPESA_FMI + PROFESSIONE_10 + MOTIVO_VIAGGIO_97, data =
trainData_balanced, family = binomial)
# Summarize the model
summary(model_balanced)

# Predict probabilities
probabilities_balanced <- predict(model_balanced, testData, type =
"response")
# Convert probabilities to binary predictions
predictions_balanced <- ifelse(probabilities_balanced > 0.5, 1, 0)
predictions_balanced <- as.factor(predictions_balanced)
# Actual values
actuals <- as.factor(testData$France)

# Create the confusion matrix
confMatrix_balanced <- table(Predicted = predictions_balanced, Actual =
actuals)
# Print the confusion matrix
print(confMatrix_balanced)
# Handle cases with missing classes
if (!("1" %in% rownames(confMatrix_balanced))) {
  confMatrix_balanced <- rbind(confMatrix_balanced, "1" = c(0, 0))}
if (!("0" %in% rownames(confMatrix_balanced))) {
  confMatrix_balanced <- rbind("0" = c(0, 0), confMatrix_balanced)}
if (!("1" %in% colnames(confMatrix_balanced))) {
```

```
  confMatrix_balanced <- cbind(confMatrix_balanced, "1" = c(0, 0))}
if (!("0" %in% colnames(confMatrix_balanced))) {
  confMatrix_balanced <- cbind("0" = c(0, 0), confMatrix_balanced)}

# Calculate performance metrics
TP <- confMatrix_balanced["1", "1"]
TN <- confMatrix_balanced["0", "0"]
FP <- confMatrix_balanced["1", "0"]
FN <- confMatrix_balanced["0", "1"]

accuracy <- (TP + TN) / sum(confMatrix_balanced)
precision <- ifelse((TP + FP) == 0, 0, TP / (TP + FP))
recall <- ifelse((TP + FN) == 0, 0, TP / (TP + FN))
f1_score <- ifelse((precision + recall) == 0, 0, 2 * (precision * recall)
/ (precision + recall))

print(paste("Accuracy:", accuracy))
print(paste("Precision:", precision))
print(paste("Recall:", recall))
print(paste("F1 Score:", f1_score))
# Train the Random Forest model

set.seed(123)  # For reproducibility
rf_model <- randomForest(France ~ CLASSE_ETA + SESSO + REGIONE_RESIDENZA
+ NR_NOTTI + FPD_SPESA_FMI + PROFESSIONE_10 + MOTIVO_VIAGGIO_97,
                         data = trainData,
                         ntree = 500,
                         mtry = 3,
                         importance = TRUE)
```

```
# Print the model summary
print(rf_model)
# Extract feature importance
importance_scores <- importance(rf_model)
# View importance scores
print(importance_scores)
# Predict on the test data
rf_predictions <- predict(rf_model, testData)
# Create the confusion matrix
confMatrix_rf <- confusionMatrix(rf_predictions, testData$France)
# Print the confusion matrix
print(confMatrix_rf)
# Plot feature importance
varImpPlot(rf_model)
*
*
*
#Discriminant Analysis:
library(MASS)
library(klaR)
library(ggplot2)
expanded_data$France <- ifelse(expanded_data$STATO_VISITATO == 29, 1,
0)
# Perform LDA for france
lda_fr <- lda(France ~ CLASSE_ETA + SESSO + REGIONE_RESIDENZA +
                NR_NOTTI   +   FPD_SPESA_FMI   +   PROFESSIONE_10   +
MOTIVO_VIAGGIO_97, data = expanded_data)
print(lda_fr)
# Predict using the LDA model
```

```
lda_fr_predictions <- predict(lda_fr, expanded_data)
# Add predictions to the original dataset
expanded_data$LDA_Pred_fr <- lda_fr_predictions$class
# Confusion matrix to evaluate the LDA model
lda_fr_conf_matrix <- table(Predicted = lda_fr_predictions$class, Actual
= expanded_data$France)
print(lda_fr_conf_matrix)

# Perform QDA
qda_fr <- qda(France ~ CLASSE_ETA + SESSO + REGIONE_RESIDENZA +
                NR_NOTTI   +   FPD_SPESA_FMI   +   PROFESSIONE_10   +
MOTIVO_VIAGGIO_97, data = expanded_data)
print(qda_fr)
# Predict using the QDA model
qda_fr_predictions <- predict(qda_fr, expanded_data)
# Add predictions to the original dataset
expanded_data$QDA_Pred_fr <- qda_fr_predictions$class
# Confusion matrix to evaluate the QDA model
qda_fr_conf_matrix <- table(Predicted = qda_fr_predictions$class, Actual
= expanded_data$France)
print(qda_fr_conf_matrix)
# Add the LDA results to the dataset
expanded_data$LDA2 <- lda_fr_predictions$x[,1]
# Visualize LDA results for France
ggplot(expanded_data, aes(x = LDA2, color = France)) +
  geom_density() +
  labs(title = "LDA for France",
       x = "Linear Discriminant 1",
      y = "Density")
```