UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea Magistrale in Data Science per l'Economia e le Imprese

Implementazione di una Pipeline ETL e analisi dei dati sulla qualità dell'aria di Bologna

Implementation of an ETL Pipeline and analysis of the air quality data of Bologna

Relatore: Chiar.mo
Prof. Luca Virgili

Tesi di Laurea di:
Raffaele Di Deo

Anno Accademico 2023/2024

dedica

ELENCO DELLE FIGURE

ELENCO DELLE TABELLE

ELENCO DEI CODICI

INTRODUZIONE E SCOPO DELLA TESI

Molteplici sono le tecniche di divulgazione per informare, educare e sensibilizzare l'opinione pubblica ad una tematica così attuale e fluttuante come quella dell'inquinamento atmosferico. Negli ultimissimi anni, questo fenomeno è diventato un catalizzatore di attenzione a causa dell'incremento della sua presenza nelle città densamente popolate come quella di Bologna (Agenzia Europea dell'Ambiente, 2018). Le conseguenze che si innescano a causa dell'aria inquinata sono eterogenee ma omogeneamente destinate al pianeta terra inteso in senso stretto come ambiente e in senso lato come i viventi che lo abitano. Numerose e sempre più diffuse sono le iniziative proattive e le proposte di interventi messi in atto per arginare la minacciosa condizione odierna (Commissione Europea, 2023). Si sperimentano regolamentazioni e normative per monitorare la qualità dell'aria attraverso la distribuzione capillare di stazioni di monitoraggio sul territorio. Per procedere alla risoluzione o quanto meno alla limitazione di questo allarmante fenomeno ambientale, è bene conoscerlo dettagliatamente, scovandone le cause principali dell'emissione di sostanze inquinanti, e analizzando i livelli presenti nell'aria. Il celebre "Più persone vedono e capiscono questo enorme problema, più possibilità abbiamo di risolverlo" (Hawkins, 2018) muove l'esigenza di rendere una problematica del genere più chiara e fruibile per comprenderne a pieno il senso. A supportarci, nella trasformazione di quelli che sono dati grezzi in informazioni accurate e puntuali per le decisioni strategiche personali e collettive, ci sarà la Data Analytics. Una metodologia all'avanguardia per rappresentare intuitivamente, mediante l'utilizzo di grafici e KPIs, i dati ad addetti o meno ai lavori. La presente tesi è il risultato di un approccio graduale, suddiviso in step per la creazione di una pipeline ETL per l'analisi dei dati dell'aria di Bologna, l'area metropolitana presa come raggio d'azione. Il processo avrà come obiettivo quello di analizzare i dati disponibili dell'inquinamento dell'aria e del traffico comparando intervalli di tempo differenti in relazione a variabili esogene quale l'introduzione di regolamentazioni normative. Le analisi condotte hanno permesso di osservare la variazione dei livelli di inquinanti presenti nell'aria nei due anni presi in analisi evidenziando che le sole politiche di limitazione della circolazione non sono sufficienti per determinare un sostanziale calo delle principali sostanze inquinanti emesse nell'aria. In particolare è emerso che, nonostante le restrizioni al traffico abbiano avuto una

ripercussione sulla mole di veicoli presenti sulle strade, i livelli delle sostanze emesse principalmente dai motori a combustione sono aumentati. Inoltre, lo studio delle rilevazioni del traffico veicolare, ci ha permesso di ottenere informazioni utili riguardanti la ciclicità e le abitudini di movimento dei veicoli all'interno della città di Bologna. Questi risultati possono essere utili alla pianificazione delle strategie di gestione del traffico urbano. Per avere una panoramica completa sull'inquinamento atmosferico e sulla situazione del traffico veicolare, la tesi sarà così composta: nel capitolo 1 verrà presentata una panoramica sul fenomeno osservato con relative definizioni degli agenti inquinanti, i limiti di norma vigenti e le principali fonti di inquinamento. Nel capitolo 2 ci si addentrerà nell'ambito delle tecnologie e delle nozioni teoriche a disposizione per cercare di rappresentare al meglio il fenomeno, e quindi verrà approfondito il tema della pipeline ETL. Una volta descritte le varie fasi del processo ETL si passerà alla presentazione delle tecnologie utilizzate per massimizzare il valore dei dati precedentemente trasformati e caricati. Successivamente, quindi, si passerà alla presentazione, descrizione e approfondimento dello strumento di Power BI che permette una visualizzazione facilitata dei dati riguardanti i livelli di inquinamento atmosferico e del traffico veicolare, nell'arco degli ultimi anni. Nel capitolo 3 sarà la volta di passare alla pratica: dopo aver selezionato e presentato le fonti di dati si passerà alla creazione dell'infrastruttura e quindi alla descrizione dello sviluppo e dell'implementazione delle dashboard interattive, principale output della trattazione. Verranno presentate, in maniera statica le diverse visualizzazioni, in realtà dinamiche, che favoriscono l'engagement e l'interazione dell'utente. Parallelamente alla realizzazione delle visualizzazioni ci sarà la descrizione dei vari tooltips a supporto delle dashboard. Infine, verrà fatta una previsione della qualità dell'aria mettendo a confronto due modelli, ARIMA e ETS. Seguiranno poi le considerazioni sulla qualità dell'aria con lo scopo di analizzare criticamente la situazione attuale nell'aria di Bologna. Nell'intento di dimostrare competenze analitiche si cela anche quello di una crescente consapevolezza sul fenomeno che traini l'impegno verso un futuro che dia ampio respiro alla sostenibilità del nostro ecosistema.

Capitolo 1
PANORAMICA SULL'INQUINAMENTO ATMOSFERICO

Definito come qualsiasi alterazione dell'aria atmosferica provocata dall'introduzione di sostanze dannose in quantità e qualità tali da minacciare la salute umana e l'ambiente, l'inquinamento rappresenta una delle sfide più complesse, seconde al riscaldamento globale, che il nostro pianeta si ritrova ad affrontare[1]. Nonostante questa attuale, sia l'epoca della crescente consapevolezza della tutela dell'ambiente, l'inquinamento atmosferico continua ad emergere come la principale minaccia ambientale, alla qualità della vita e al benessere delle generazioni presente e future. La salute umana, l'ecosistema e l'economia globale sono nel mirino di questo fenomeno complesso e multiforme. Infatti, nonostante i segnali di miglioramento della qualità dell'aria, grazie alla diminuzione delle emissioni di inquinanti principali nel corso del tempo, molti esperti ritengono che l'inquinamento debba essere, ora più che mai, considerato come una delle principali preoccupazioni ambientali. L'OMS ritiene che, diminuendo il livello di un particolare tipo di inquinante (conosciuto come PM10), si potrebbe ridurre la mortalità nelle città inquinate dal 5 al 15% all'anno[2]. Infatti, secondo l'Istituto Superiore di Sanità[3], nel mondo quasi due milioni di persone ogni anno muoiono prematuramente a causa dell'inquinamento atmosferico, presente sia nell'ambiente esterno che dentro casa. In Europa, esso rappresenta il principale fattore di rischio per la salute. A darne evidenza è proprio l'European Environment Agency[4] che attesta che nel 2020 nell'Unione Europea, il 96% della popolazione urbana è stato esposto a livelli di particolato fine superiori al livello di riferimento basato sulla salute stabilito dall'Organizzazione Mondiale della Sanità (European Environment Agency, 2022). Queste linee guida, rivolte a tutti i Paesi del mondo, implicano una riduzione di oltre tre volte del livello attuale di inquinamento atmosferico finalizzata ad uniformare gli obiettivi per la qualità dell'aria in tutto

1 https://www.eea.europa.eu/it/articles/i-vantaggi-di-un2019aria-piu
2 https://www.ilfattoquotidiano.it/2022/06/16/polveri-sottili-loms-rivede-le-linee-guida-ora-rischiano-tutti-ma-non-allo-stesso modo/
3 https://www.epicentro.iss.it/ambiente/in-oms-guida06
4 https://www.eea.europa.eu/it/highlights/le-morti-premature-causate-dallinquinamento

il mondo, riducendo standard che superano di gran lunga gli attuali limiti nazionali vigenti. Il superamento dei livelli considerati sicuri in molte città ha portato l'OMS ad ampliare le sue linee guida con l'obiettivo di stabilire standard efficaci, considerando le circostanze specifiche di ciascun contesto nazionale e conseguire una significativa riduzione della mortalità e delle malattie. L'obiettivo è che diventino parte integrante delle legislazioni nazionali, offrendo una guida fondamentale per affrontare questo problema su scala globale in quanto numerosi paesi non dispongono di una regolamentazione adeguata sull'inquinamento atmosferico, rendendo difficile il controllo di questo grave rischio per la salute pubblica. L'inquinamento atmosferico, con il suo impatto deleterio sulla salute umana e sull'ambiente, richiede azioni immediate e coordinate a livello globale.

Gli agenti fisici, chimici e biologici che alterano le caratteristiche naturali dell'atmosfera terrestre costituiscono l'inquinamento atmosferico. Un inquinante atmosferico si riferisce a qualsiasi sostanza o agente che perturba l'equilibrio naturale dell'aria sia l'introduzione di nuovi composti dannosi o l'alterazione delle proporzioni di sostanze già presenti. Tali fonti di inquinamento[5] possono essere categorizzate come di origine naturale, come eruzioni vulcaniche o incendi, o di origine antropica, causate principalmente dalle attività umane come il traffico veicolare, il riscaldamento domestico e le attività industriali. L'inquinamento atmosferico è più diffuso nelle grandi città a causa delle emissioni dei veicoli e del riscaldamento degli edifici, ma è presente anche nelle aree industriali che non adottano misure per ridurre le sostanze inquinanti nell'aria. A causa della dispersione atmosferica e della diffusione degli inquinanti, l'inquinamento atmosferico può interessare anche zone lontane dalle fonti primarie di inquinamento. Tutte queste fonti diffondono numerosi agenti inquinanti ma tra le principali sostanze, introdotte nell'analisi di questa trattazione troviamo:

Monossido di carbonio: Il monossido di carbonio[6] (CO) è un gas incolore, insapore e inodore, composto da un atomo di carbonio e un atomo di ossigeno. È un sottoprodotto della combustione incompleta di materiale organico o combustibile contenente carbonio, come legna, carbone, gas naturale o benzina. A basse concentrazioni, può causare sintomi come mal di testa, vertigini, nausea e confusione. A causa della sua natura inodore, il monossido di

5 http://www.arpamoliseairquality.it/forme-di-inquinamento/

6 https://www.salute.gov.it/imgs/C_17_opuscoliPoster_283_ulteriorieallegati_ulteriorieallegato_2_alleg.pdf

carbonio è particolarmente pericoloso, poiché può accumularsi in ambienti chiusi senza che le persone se ne accorgano.

Benzene: composto organico volatile e infiammabile. Il benzene[7] è ampiamente utilizzato come solvente industriale nella produzione di plastica, resine, nylon, detergenti e altri prodotti chimici. Componente dei derivati del petrolio è presente nell'aria a causa di eventi naturali, attività umane, industriali e nei gas di scarico dei veicoli a combustione alimentati da benzina. È noto per essere altamente tossico e cancerogeno per gli esseri umani, in quanto può danneggiare il DNA.

Ossidi di azoto: Gli ossidi di azoto[8] sono composti chimici formati da azoto e ossigeno, ed include il monossido di azoto (NO) e il biossido di azoto (NO2). Sono prodotti principalmente durante i processi di combustione a elevate temperature, come quelli nei motori a combustione interna dei veicoli, nelle centrali elettriche e nelle attività industriali. Gli ossidi di azoto contribuiscono alla formazione di smog fotochimico e possono reagire con altri composti atmosferici per formare particolato fine e ozono troposferico. Può causare problemi respiratori.

Ozono: L'ozono[9] è un gas incolore e altamente reattivo, composto da tre atomi di ossigeno (O3). È presente in natura sia nell'atmosfera terrestre, dove forma lo strato di ozono, sia a livello del suolo, dove è un inquinante atmosferico. L'esposizione all'ozono può provocare irritazione delle vie respiratorie, ridotta funzione polmonare, infiammazione polmonare e peggioramento di condizioni respiratorie preesistenti come l'asma.

Le polveri sottili: Le polveri sottili[10] sono classificate in base al diametro delle particelle; le due classificazioni principali sono **PM10** (diametro inferiore a 10 μm) e **PM2.5** (diametro inferiore a 2.5 μm). Queste sono prodotte principalmente da processi di combustione, che avvengono, ad esempio, negli autoveicoli e negli impianti di riscaldamento domestico o da processi secondari come l'usura dei freni e degli pneumatici. Gli effetti di queste polveri possono essere estremamente nocivi per la salute umana, soprattutto quelli causati dalle più

7 https://www.issalute.it/index.php/la-salute-dalla-a-alla-z-menu/b/benzene

8 https://www.efficienzaenergetica.enea.it/glossario-efficienza-energetica/lettera-o/ossidi-di-azoto-nox.html

9 https://www.snpambiente.it/temi/lozono-linquinante-critico-in-estate/

10 https://osservatoriocpi.unicatt.it/ocpi-pubblicazioni-l-inquinamento-da-polveri-sottili-pm10-e-pm2-5-in-italia-e-europa

fini (PM2.5); queste, infatti, date le loro dimensioni, potrebbero raggiungere anche gli alveoli polmonari, causando gravi patologie sia all'apparato respiratorio che all'apparato cardio-circolatorio.

Data l'emergenza si è sentita la necessità di stabilire e regolamentare i limiti che regolano le quantità massime di inquinanti nell'aria sia per garantire la salute umana sia per la gestione del fenomeno. L'ARPAE[11] (Agenzia Regionale per la Prevenzione, l'Ambiente e l'Energia dell'Emilia-Romagna) si dedica intensamente al rispetto dei limiti normativi stabiliti dalla legge, mediante sistemi di monitoraggio ambientale, previsioni e valutazioni tramite modellistica matematica, stime delle emissioni in atmosfera e studi sull'impatto sulla salute.

Tabella 1-1 Limiti normativi degli agenti inquinanti in Italia

	Target protezione	**Valore limite**	**Unità di misura**	**Mediazione**	**Superamenti massimi**
CO	Salute	10	mg/m³	Giorno	Max 0 annui
Benzene	Salute	5	µg/m³	Anno	Max 5 µg/m³ media annua
NO	Salute	200	µg/m³	Ora	Max 30 µg/m³ media annua
NO2	Salute	200	µg/m³	Ora	Max 18 annui
O3	Salute	120	µg/m³	Giorno	Max 25 annui
PM10	Salute	50	µg/m³	Giorno	Max 35 annui
PM2.5	Salute	25	µg/m³	Anno	Max 25µg/m³ media annua

Fonte: personale rielaborazione dei dati estratti da "La qualità dell'aria in Emilia-Romagna. Edizione 2023"

Come abbiamo visto, le fonti di inquinamento atmosferico sono molteplici e spaziano dalle attività industriali al traffico veicolare, dagli impianti di riscaldamento alle centrali elettriche. Le emissioni inquinanti, compromettendo seriamente la salute umana e l'equilibrio degli ecosistemi, hanno un impatto particolarmente accentuato nelle aree urbane densamente popolate, come nel caso emblematico della città di Bologna nella Pianura Padana[12]. Secondo dati forniti da ARPAE, la pianura padana registra livelli preoccupanti di sostanze inquinanti come PM10, PM2.5, ossidi di azoto e biossido di zolfo. Queste particelle in sospensione, emesse principalmente da fonti antropiche come autoveicoli, industrie e attività agricole,

11 https://www.arpae.it/it/temi-ambientali/aria/report-aria/report-regionali/aria-2023.pdf

12 https://www.infodata.ilsole24ore.com/2024/02/24/la-nuvola-rossa-dellinquinamento-sulla-pianura-padana-quattro-anni-dopo/?refresh_ce=1

contribuiscono alla formazione di smog e nebbie tossiche che compromettono la qualità dell'aria e la salute dei residenti.

Caratterizzata da un intenso traffico automobilistico, da una densa concentrazione di industrie e da una conformazione geografica che non permette una ventilazione costante in quanto circondata dall'arco alpino, Bologna affronta livelli critici di inquinanti atmosferici con frequenti superamenti dei limiti di legge per PM2.5 e ossidi di azoto. Posizionata da IqAir sul podio delle 3 città più inquinate di Italia[13], registra nelle sue zone urbane e industriali concentrazioni elevate di inquinanti, con conseguenze negative sulla salute pubblica e sulla qualità della vita. Affrontare l'inquinamento atmosferico a Bologna richiede un impegno congiunto da parte delle autorità locali, delle istituzioni regionali e dei cittadini stessi, attraverso politiche e misure efficaci per ridurre le emissioni inquinanti e migliorare la qualità dell'aria. La recente introduzione della normativa "zona 30", entrata in vigore il 16 gennaio 2024[14], mira a promuovere la sicurezza stradale riducendo il limite di velocità a 30 km/h. In concomitanza del rallentamento della velocità, i sensori di proprietà dell'ARPAE hanno registrato un aumento dei livelli di inquinanti nell'aria[15]. Questo fenomeno continua ad alimentare dubbi riguardo la reale efficacia della nuova normativa stradale.

13 https://www.ilrestodelcarlino.it/bologna/cronaca/smog-alle-stelle-correre-l9p1ilgn#?live

14 https://comunicatistampa.comune.bologna.it/2024/guida-a-bologna-citta-30-da-martedi-16-gennaio-via-a-ordinanze-e-controlli-online-la-mappa-navigabile-delle-velocita-con-i-nomi-e-i-limiti-di-tutte-le-strade-della-citta-30

15 https://gazzettadibologna.it/primo-piano/bologna-citta-30-a-gennaio-e-aumentato-lo-smog/

Capitolo 2
STRUMENTI PER LA DATA ANALYTICS

2.1 **Pipeline di ETL**

In un'era Data-Driven[16] come quella contemporanea, gli asset informativi provengono da una moltitudine di fonti diverse con un'elevata eterogeneità di formati. L'obiettivo di ogni realtà aziendale è renderli effettivamente utilizzabili e quindi disponibili per le attività di analisi ed elaborazione.

L'analisi dei dati, o Data Analytics[17], è il processo di esplorazione, interpretazione e comunicazione dei pattern significativi dei dati per trarre informazioni utili e guidare le decisioni aziendali. Coinvolge l'applicazione di metodi statistici, algoritmi di machine learning, tecniche per scoprire tendenze e correlazione tra i dati e strumenti software per la visualizzazione dei dati. L'obiettivo della Data Analytics è quello di estrarre valore dai dati grezzi, siano essi strutturati o meno, al fine di prendere decisioni informate e guidare al successo l'azienda. Questo processo si articola in diverse fasi, come la *Data Cleaning* (pulizia dei dati per rimuovere errori o valori mancanti), *Data Exploration* (l'esplorazione dei dati per identificare pattern e relazioni), *Data Modeling* (creazione di modelli per perseguire l'obiettivo dell'analisi) e la *Data Visualization* (comunicazione dei risultati attraverso report e visualizzazioni).

Nel contesto dell'analisi dei dati, le pipeline ***ETL*** (Extract, Transform, Load) svolgono un ruolo cruciale nella preparazione ed ottimizzazione dei dati per un'analisi accurata e tempestiva.

L'ETL fornisce l'infrastruttura necessaria per raccogliere, trasformare e caricare dati provenienti da diverse fonti in un formato omogeneo e utilizzabile. Questo processo di preparazione dei dati è essenziale per garantire che le analisi successive siano basate su dati

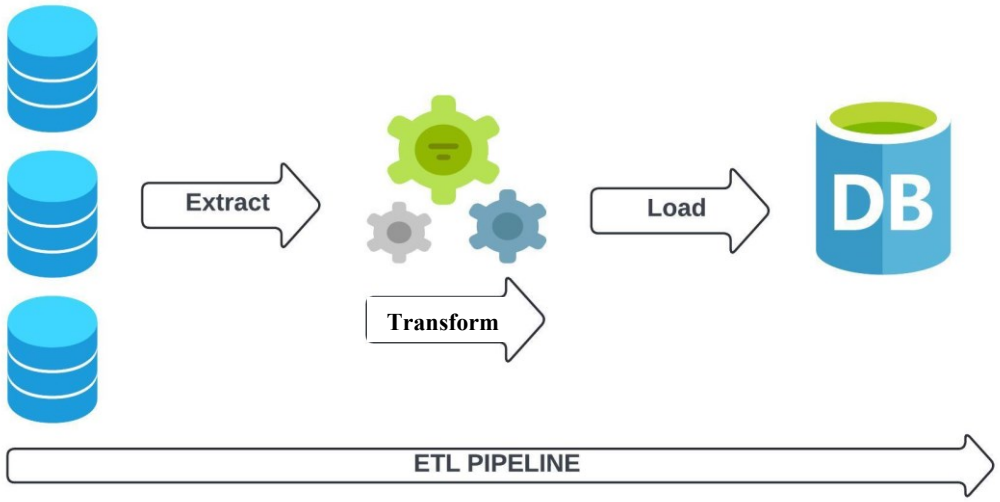
[16] Essere Data-Driven significa farsi guidare dai numeri, avere un approccio basato sui dati, per prendere decisioni informate, basate su fatti oggettivi e non su sensazioni personali.

[17] https://it.wikipedia.org/wiki/Analisi_dei_dati

accurati e coerenti. Di conseguentemente, il processo di ETL rappresenta la pietra angolare di qualsivoglia strategia di Data Analytics efficace, in quanto garante di dati affidabili e potenzialmente utili sui quali le future decisioni aziendali trovano fondamenta.

Una pipeline di ***ETL***[18] può essere comunemente definita come un processo di integrazione di dati che concretizza dati provenienti da diverse fonti verso un sistema di destinazione, sia esso un data warehouse o un database (Vassiliadis, 2002). L'obiettivo fondamentale consiste nel migliorare l'accesso e l'analisi complessiva, nonché nel potenziare la generazione di report. La preparazione dei dati permette di ottenere benefici tangibili in termini di riduzione degli errori manuali e del tempo di elaborazione. L'implementazione di questo processo garantisce la centralizzazione e standardizzazione dei dati, ottimizzando quindi la collaborazione e la gestione della loro migrazione.

Figura 2-1 *Rappresentazione grafica di una Pipeline ETL*



Fonte: Personale elaborazione di un flusso ETL

Come rappresentati graficamente nella Figura 2-1, le fasi della pipeline di *ETL*[19] sono:

[18] https://www.ibm.com/topics/etl
[19] https://www.deda.cloud/etl/

Estrazione (Extract): La fase preliminare, che dà il via al processo, è quella di estrazione.

L'estrazione dei dati tramite una pipeline di *ETL* può avvenire da una moltitudine di fonti che spaziano da database relazionali, file di testo, API web, web-scraper, sensori IoT, a molto altro. Per procedere all'estrazione è necessaria un'adeguata selezione ed individuazione delle fonti di dati in base all'obiettivo che si vuole perseguire con l'analisi. Lo scopo si concretizza nell'estrazione dei dati nel loro stato grezzo che alimentano ed innescano l'intero di processo ETL.

Il processo di estrazione può avvenire secondo tre modalità:

Estrazione completa: Un metodo che prevede l'inserimento di tutti i dati, precedentemente estratti dalle diverse fonti, e il loro trasferimento massivo all'interno della pipeline. L'estrazione completa, generalmente, viene utilizzata quando si vuole popolare il database per la prima volta oppure quando non è possibile risalire alla data di ultima modifica del dato.

Estrazione incrementale: In questa modalità, vengono caricati solo i record che sono stati aggiunti o modificati in data successiva a quella dell'ultimo caricamento. L'ETL, in questo caso, memorizza la data dell'ultima estrazione e carica solo i nuovi dati, calcolando la differenza tra il sistema di destinazione e quello di origine.

Estrazione su notifica di aggiornamento: La fonte originaria avvisa il sistema ETL quando ci sono cambiamenti nei dati, attivando il processo di ETL per estrarre unicamente i nuovi dati.

Trasformazione (Transform): Una volta estratti, i dati subiscono una serie di trasformazioni per far sì che siano adatti all'analisi e quindi alla generazione di report. La fase di trasformazione dei dati, in una pipeline ETL, comporta diverse operazioni sui dataset per garantire che i dati siano puliti, coerenti e pronti per l'analisi. La fase di *transformation* si può articolare in:

Pulizia di base: In questa fase vengono uniformati tutti i dati in un formato appropriato in termini di struttura o di formattazione del singolo dato. Questa conversione, quindi, può includere l'eliminazione di errori, individuazione e allineamento dei dati di origine nel formato

di destinazione e standardizzazione di set di caratteri quali unità di misura o formati di data e ora.

Integrazione di dati: Prevede l'unione o la combinazione di più tabelle e l'inserimento di funzioni di aggregazione come media, mediana, minimo, massimo e somma, per sintetizzare le righe dei vari gruppi definiti da una o più colonne.

Filtraggio dei dati: In questa fase è possibile scremare un sottoinsieme di dati in modo da selezionare quelli di interesse raffinando la ricerca e scartando quelli superflui.

Verifica e deduplica: Viene eseguita una verifica per individuare ed eliminare i record duplicati, verificandone quindi unicità e accuratezza dei dati. Inoltre, è la fase in cui vengono apportate correzioni agli errori derivanti dall'inserimento manuale.

Derivazione dei dati: Può essere utile o necessario dover applicare regole aziendali per calcolare nuovi valori sulla base di quelli già esistenti.

Protezione dei dati: Prevede la cifratura dei dati sensibili prima di trasferirli al database di destinazione in modo da rispettare le normative sulla privacy.

Caricamento (Load): In fondo al processo ETL, i dati prima estratti e poi trasformati vengono caricati nel sistema di destinazione, che può essere un data warehouse, un database relazionale, non relazionale o un altro tipo di archivio dati. Durante questo processo, è essenziale mantenere l'integrità e la coerenza dei dati, garantendone la qualità. Il caricamento, a seconda delle esigenze dell'applicazione, può essere effettuato in modalità batch[20] o in tempo reale.

Nelle fasi di caricamento, quindi, i dati ora strutturati vengono spostati dall'area di staging[21] e caricati nel sistema di destinazione. Si tratta di un processo ben definito che può essere tuttavia svolto secondo modalità differenti:

[20] Modalità di elaborazione secondo la quale le richieste di servizio non vengono assolte immediatamente, ma sono accodate per essere soddisfatte quando lo consentirà la disponibilità delle risorse.

[21] L'area di staging (o zona di destinazione) è un'area di archiviazione intermedia per archiviare in via temporanea i dati estratti. Le aree di staging dei dati sono spesso transitorie, vale a dire che i loro contenuti vengono cancellati una volta completata l'estrazione. (El-Sappagh, 2011)

Caricamento completo: Tutti i dati dalla fonte originaria vengono trasferiti nel sistema di destinazione una volta completata la trasformazione. Questo metodo è solitamente utilizzato quando i dati vengono caricati, per la prima volta, da un sistema sorgente al sistema destinazione. Ad esempio, quando un'azienda implementa per la prima volta un sistema di data warehousing, trasferisce tutti i dati storici in un unico processo, garantendo che il nuovo sistema parta con una base di dati completa e uniforme.

Caricamento incrementale: In questa fase, simile a quella di *Extract*, l'ETL memorizza la data dell'ultima estrazione, caricando solo i record aggiunti o modificati dopo tale data. Per piccoli volumi di dati, il caricamento può avvenire in modalità *streaming*, consentendo un aggiornamento continuo e in tempo reale. Questa procedura è particolarmente utile in settori come il commercio online in cui è cruciale il monitoraggio e l'elaborazione di flussi di dati per prendere decisioni tempestive, come l'aggiornamento delle scorte in tempo reale in base agli acquisti dei clienti. Con una grande mole di dati, per garantire l'aggiornamento periodico dei dati e ridurre l'impatto sugli asset di sistema, le modifiche possono essere accumulate e poi trasferite in batch, all'interno di un dato intervallo di tempo.

Come precedentemente accennato, la pipeline ETL standardizza e automatizza l'intero processo di raccolta dei dati grezzi da più fonti di dati e in diversi formati. Osserviamo in dettaglio quali sono le finalità e i benefici tangibili che rendono questo processo così importante. Molteplici sono i vantaggi del processo ETL come:

Integrazione dei dati: L'omogeneizzazione dei dati in un unico formato rappresenta un beneficio in termini di tempi, errori e costi associati alla gestione dei dati. Aspetto essenziale durante il popolamento di un database con dati provenienti da molteplici e differenti fonti.

Miglioramento della qualità dei dati: L'aumento dell'affidabilità e della qualità dei dati è la naturale conseguenza delle azioni che avvengono nella fase di *Trasform* quali la pulizia, la normalizzazione e standardizzazione dei dati, con annessa eliminazione di errori e duplicati.

Automazione: L'automatizzazione delle attività ripetitive di migrazione, spostamento ed elaborazione dei dati aumenta l'efficienza e il risparmio di tempo dell'intero processo. Grazie alla riduzione e limitazione dell'intervento manuale, il processo ETL minimizza il rischio di

errori umani, liberando conseguentemente risorse aziendali per attività più strategiche.

Flessibilità: Il processo ETL può essere configurato ad-hoc per soddisfare esigenze specifiche dell'azienda. Quest'ultima può permettersi l'aggiunta o la modifica delle fonti di dati e il cambiamento nei criteri di trasformazione. La possibilità di personalizzare il processo ETL, grazie alla sua flessibilità, è cruciale per rispondere rapidamente ai cambiamenti di business.

Scalabilità: Con l'aumento del volume e della complessità dei dati, l'importanza del flusso ETL diventa ancora più evidente. Le pipeline ETL possono essere facilmente scalate per garantire che l'azienda possa continuare a estrarre, trasformare e caricare dati senza ritardi o sovraccarichi dei sistemi.

Accesso tempestivo ai dati: Nelle strategie di Business Intelligence, l'accesso rapido ai dati integrati è fondamentale per supportare decisioni informate. Le pipeline ETL assicurano che i dati siano già in un formato utilizzabile, permettendo la realizzazione di report e analisi in tempi molto più rapidi.

Sicurezza dei dati: L'ETL crea un livello di astrazione tra il sistema di origine e quello di destinazione, contribuendo alla corretta governance dei dati preservandone parallelamente la qualità, la sicurezza e la privacy. Questo promuove l'accesso ai dati a tutti gli stakeholder aziendali senza comprometterne la sicurezza.

In sintesi, il processo ETL è fondamentale per centralizzare, standardizzare e migliorare la gestione dei dati aziendali, rendendo più efficiente l'accesso alle informazioni e supportando decisioni aziendali più informate e tempestive. Per questa ragione, le pipeline di ETL trovano ampio impiego in una varietà di settori (come Business Intelligence e analisi aziendale) e di processi (come l'elaborazione in tempo reale e migrazione dei dati).

Una volta implementata la pipeline ETL, l'utilizzo di strumenti e tecnologie applicative è essenziale per massimizzare il valore dei dati trasformati e caricati. Le tecnologie impiegate per la realizzazione di questo progetto, sono state varie e altamente efficaci. La loro flessibilità e adattabilità su larga scala, hanno permesso di gestire grandi volumi di dati in modo efficiente ed affidabile. Le tecnologie coinvolte sono state:

Postman[22], con la sua suite completa di strumenti per lo sviluppo delle API, ha agevolato il processo di test di quest'ultime. Questo tool permette di inviare richieste HTTP di vario tipo (GET, POST, PUT, DELETE, ecc.) in modo semplice, configurando URL e parametri, visualizzando facilmente le risposte delle API, codici di stato e tempistiche di esecuzione. Ha permesso di comprendere meglio la struttura dei dati nella fase di estrazione.

Python[23], un linguaggio di programmazione ad alto livello, ha svolto un ruolo cruciale nello sviluppo di questo progetto grazie alla sua versatilità e alle sue numerose librerie. Il suo impiego è stato fondamentale sia nella fase di estrazione che in quella di caricamento dei dati, permettendo di gestire il flusso di lavoro con efficienza e precisione.

MongoDB[24], un database non relazionale del tipo document-based, dotato di enormi capacità per la gestione dei dati. Viene generalmente utilizzato quando i dati, per volume o per struttura, non sono compatibili con i tradizionali database relazionali. I suoi punti di forza riguardano la capacità di query ad-hoc, indicizzazione, bilanciamento del carico e aggregazione dei dati senza mai compromettere le prestazioni. Si è rivelato fondamentale per immagazzinare dati contenenti diverse strutture.

[22] https://www.postman.com/product/what-is-postman/
[23] https://it.wikipedia.org/wiki/Python
[24] https://it.wikipedia.org/wiki/MongoDB

2.2 La Data Analytics tramite Power BI

Nelle culture organizzative moderne, un approccio Data Driven Decision Making[25] è fondamentale: questa corrente di pensiero utilizza i dati per guidare le decisioni future, analizzando le informazioni passate per prevedere e anticipare gli eventi prossimi. Tuttavia, senza strumenti adeguati, questi dati rischiano di rimanere inaccessibili e di difficile interpretazione in particolar modo per i non addetti ai lavori. Per far sì che i dati si trasformino in informazioni è necessario che siano accessibili e comprensibili. Gli strumenti di Data Visualization, come Power BI, rappresentano la panacea per la gestione di una grande quantità di informazioni. Quest'ultime vengono rappresentate, visualizzate e rese di facile intuizione mediante questi software. Così facendo, non si ottimizza solamente il processo di gestione ma anche, e soprattutto, l'impatto che i dati possono avere: una visualizzazione semplice e facilitata trasforma i dati a disposizione in dati utili, rendendo l'intero processo decisionale maggiormente efficace e rapido.

Per passare dai dati grezzi a quelli rappresentati graficamente, si è ricorso quindi alla Data Visualization[26] intesa come l'esplorazione visuale e interattiva di dati con le più svariate dimensioni e origini. La visualizzazione dei dati permette di trasformare numeri e statistiche in grafici e immagini facili da interpretare, favorendo la comprensione delle tendenze e dei modelli. Power BI[27] è una piattaforma di Business Intelligence e Analytics composta da servizi software (A. Ferrari, 2016), app e connettori che interagiscono per trasformare origini dati non correlate in informazioni dettagliate interattive, coerenti e visivamente accattivanti. GARTNER Inc.[28] , nota società americana che si occupa di consulenza strategica ed analisi di mercato nel campo

[25] https://www.tableau.com/it-it/learn/articles/data-driven-decision-making
[26] https://www.agendadigitale.eu/cittadinanza-digitale/data-management/data-visualization-cose-perche-funziona-come-farla-in-modo-etico/
[27] https://learn.microsoft.com/it-it/power-bi/fundamentals/power-bi-overview
[28] https://it.wikipedia.org/wiki/Gartner#:~:text=Gartner%20Inc.,negli%20Stati%20Uniti%20d'America

della tecnologia, ha eletto[29] per il sedicesimo anno consecutivo, Power BI come leader nel Gartner Magic Quadrant 2023 per le piattaforme di analisi e Business Intelligence (Figura 2-2):

Figura 2-2 *Magic Quadrant per Analytics and Business Intelligence*

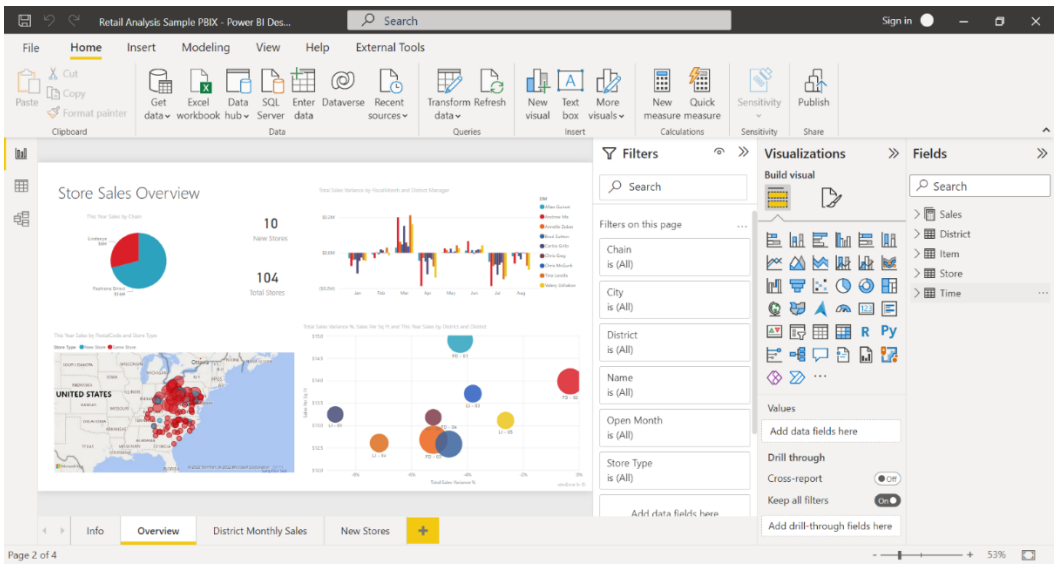
Fonte: Gartner

Lo strumento ideato e realizzato da Microsoft ha il potere di creare connessioni di dati e grafici interattivi da visualizzare attraverso pratiche dashboard personalizzate, consentendo agli utenti un'esplorazione dinamica. L'utilizzo di mappe, grafici a barre, linee e istogrammi fa di Power BI il detentore del beneficio più ambito: l'immediatezza della comprensione. Solo acquisendo rapidamente e dettagliatamente informazioni cruciali è possibile passare alla fase di azione. Infatti, questo strumento spopola tra le aziende che hanno bisogno di decodificare i loro complessi dati grezzi. Inoltre, Power BI favorisce la democratizzazione dei dati, rendendoli

[29] https://Power_BI.microsoft.com/it-it/blog/microsoft-named-a-leader-in-the-2023-gartner-magic-quadrant-for-analytics-and-bi-platforms/

accessibili a tutti gli stakeholder aziendali, indipendentemente dal loro livello di competenza tecnica. Questa accessibilità promuove una maggiore collaborazione tra i vari reparti e una più ampia condivisione delle informazioni a tutti i livelli aziendali.

Figura 2-3 *Esempio di dashboard in Power BI Desktop*



Fonte: Microsoft learn

Power BI offre numerosi strumenti e vantaggi[30] nell'analisi dei dati:

Misure[31]*:* sono formule in linguaggio DAX (Data Analisys Expression) utilizzate per eseguire calcoli sui dati a disposizione. Vengono utilizzate per eseguire calcoli aggregati come somma, media, conteggi, massimi, minimi e numerose altre operazioni. A differenza delle colonne presenti in una tabella, le misure calcolano in maniera dinamica i risultati in base al contesto della visualizzazione.

KPI[32]*:* i Key Performance Indicator sono strumenti fondamentali utilizzati per misurare e valutare l'entità di un determinato fenomeno. Per essere efficaci devono essere tempestivi, fornire informazioni aggiornate, semplici da calcolare e da comprendere per tutte le varie figure aziendali. Per avere un impatto significativo, e quindi trasmettere informazioni di valore, devono essere strettamente connesse al fenomeno che si vuole analizzare.

[30] https://www.analyticsinsight.net/data-analysis/how-power-bi-enables-real-time-data-visualization
[31] https://www.visualitics.it/misure-rapide-power-bi/
[32] https://www.digitaldictionary.it/blog/kpi-cosa-sono-e-come-si-misurano

Visualizzazione chiara e intuitiva: la rappresentazione dei dati sull'inquinamento, come nel caso della presente trattazione, in grafici, mappe o altri formati visivi rendono l'intuizione immediata e di più semplice interpretazione. In questo modo è possibile ottenere una panoramica generale, chiara e intuitiva delle tendenze e dei modelli.

Identificazione rapida dei pattern: i suoi strumenti avanzati per l'analitica, consentono l'individuazione e quindi l'aumento di consapevolezza di pattern e anomalie relative all'inquinamento. Per mezzo della sua abilità di connessione a diverse fonti di dati e di eseguire analisi complesse, Power BI permette di scoprire insight significativi che potrebbero altrimenti passare come superflui.

Monitoraggio in tempo reale: Utilizzando Power BI è possibile avere un riscontro real-time sia sui livelli di inquinamento che delle variazioni nella qualità dell'aria. Questa funzionalità è particolarmente utile per gli enti preposti alla gestione dell'ambiente e della salute pubblica, consentendo loro di prendere tempestivamente misure correttive quando necessario.

Creazione di report interattivi: Grazie alla possibilità di interagire con i dati dei report, filtrarli, eseguire drill-down e ottenere specifiche dettagliate, Power BI permette una navigazione user-friendly e dinamica, ingaggiando l'utente e facendolo interagire. Quest'ultimo vantaggio, insieme alla visione panoramica ma sintetica delle informazioni, alla flessibilità e alla customizzazione delle sezioni, ricopre un ruolo primario e attivo nella fase di comprensione dell'analisi.

Tuttavia, l'utilizzo della Data Analytics tramite Power BI, non solo facilita la comprensione dei dati per un vasto pubblico, ma aumenta anche la consapevolezza e la sensibilizzazione riguardo all'inquinamento ambientale. Superando le barriere linguistiche e tecniche, rende la comunicazione più efficace anche per persone con diversi livelli di alfabetizzazione. La facilità di esplorazione e di interazione rende i dati più accessibili e ne aumenta anche l'impatto emotivo, fondamentale per generare consapevolezza e contribuire a sensibilizzare il pubblico sulle cause dell'inquinamento e sulla necessità di adottare comportamenti più sostenibili e responsabili. Infine, viene scelto questo tipo di narrazione, denominata anche Data Storytelling, per ridurre i divari informativi ed evidenziare il link tra azioni umane e conseguenze ambientali. In questa trattazione, Power BI promuove l'importanza dei dati come risorsa potente e influente del cambiamento futuro.

Capitolo 3
ANALISI DEI DATI SULL'INQUINAMENTO DELL'ARIA

3.1 **Definizione delle sorgenti dei dati**

Come spiegato nel capitolo precedente, il processo di ETL inizia con l'estrazione dei dati dalla fonte. Per l'analisi oggetto di questa tesi, la fonte primaria è il sito web *Open Data Bologna*[33], un portale creato e gestito direttamente dal Comune di Bologna che mira a rendere i dati sempre più trasparenti e accessibili alla cittadinanza per aumentarne il coinvolgimento. L'analisi si concentra su due fenomeni di rilievo per la città di Bologna: l'inquinamento atmosferico e il traffico veicolare. Al fine di esaminare in modo approfondito entrambi questi aspetti, sono stati impiegati cinque dataset distinti.

Il primo dataset (Figura 3-1), inerente all'inquinamento atmosferico, è quello relativo alla qualità dell'aria dell'anno in corso, denominato *'Centraline qualità dell'aria (misurazioni giornaliere)'*[34], rinominato nel corso di questa analisi come *'Centraline_QA_Current'*. Questo dataset contiene le rilevazioni dei principali inquinanti atmosferici relative alle tre centraline di rilevamento dislocate nel Comune di Bologna: Giardini Margherita, Via Chiarini, e Porta San Felice.

La tabella comprende l'ID della rilevazione, la data e l'orario di rilevamento, il nome della stazione di rilevamento, il valore rilevato espresso in µg/m³ (mg/m³ per l'inquinante CO) e infine il nome specifico dell'inquinante atmosferico monitorato. Da notare che il dataset viene aggiornato quotidianamente con le rilevazioni del giorno precedente.

Figura 3-1 *Dataset 'Centraline qualità dell'aria (misurazioni giornaliere)'*

_id	reftime	stazione	value	agente_atm
311400	28 maggio 2024 00:00	PORTA SAN FELICE, BOLOGNA PIAZ...	13	PM10
313742	28 maggio 2024 00:00	PORTA SAN FELICE, BOLOGNA PIAZ...	0,2	CO (Monossido di carbonio)
311132	28 maggio 2024 00:00	GIARDINI MARGHERITA, BOLOGNA ...	35	O3 (Ozono)
311399	28 maggio 2024 00:00	GIARDINI MARGHERITA, BOLOGNA ...	9	PM10

Fonte: Open Data Bologna

[33] https://opendata.comune.bologna.it/pages/home/
[34] https://opendata.comune.bologna.it/explore/dataset/centraline-qualita-aria/information/?disjunctive.agente_atm

Il secondo dataset (Figura 3-2), relativo all'inquinamento atmosferico, riguarda la qualità dell'aria dal 2017 al 2023 ed è denominato *'Centraline qualità dell'aria (storico dal 2017)'*[35], successivamente rinominato *'Centraline_QA_Storico'*. Questo dataset presenta una struttura leggermente diversa rispetto al primo. Include il nome della stazione di rilevamento, il nome dell'agente inquinante rilevato, la data di inizio e fine del periodo di rilevamento, il valore rilevato e l'unità di misura per ciascun agente inquinante.
Analogamente al primo dataset, anche questi dati vengono aggiornati annualmente, con i dati relativi all'anno precedente resi disponibili nel mese di gennaio.

Figura 3-2 *Dataset 'Centraline qualità dell'aria (storico dal 2017)'*

COD_STAZ	AGENTE	DATA_INIZIO	DATA_FINE	VALORE	UM
VIA CHIARINI	NO2 (BIOSSIDO DI AZOTO)	31 dicembre 2023 23:01	1 gennaio 2024 00:00	26	ug/m3
PORTA SAN FELICE	NOX (OSSIDI DI AZOTO)	31 dicembre 2023 23:01	1 gennaio 2024 00:00	69	ug/m3
GIARDINI MARGHERITA	NO2 (BIOSSIDO DI AZOTO)	31 dicembre 2023 23:01	1 gennaio 2024 00:00	21	ug/m3
PORTA SAN FELICE	NO2 (BIOSSIDO DI AZOTO)	31 dicembre 2023 23:01	1 gennaio 2024 00:00	29	ug/m3

Fonte: Open Data Bologna

Terzo ed ultimo dataset per il fenomeno dell'inquinamento riguarda il dataset denominato *'Anagrafica_Centraline'*[36] che contiene le coordinate geografiche delle varie centraline di rilevamento, informazioni relative al Comune, Provincia e Regione, agenti inquinanti rilevati da ciascuna centralina e relativa unità di misura.
Parallelamente ai dataset relativi all'inquinamento atmosferico, nel contesto di questa ricerca sono stati inclusi due dataset dedicati al fenomeno del traffico veicolare. Il primo dataset fornisce dati riguardanti il traffico veicolare rilevato nell'anno in corso, denominato *'Rilevazione flusso veicoli tramite spire – anno 2024'*[37], in seguito rinominato *'Rilevazioni_SPIRE_2024'*. Il dataset contiene le rilevazioni per fascia oraria del traffico veicolare, data e giorno della settimana della rilevazione, codice identificativo dei singoli rilevatori di traffico, nome e codice identificativo della via in cui è installato il rilevatore, le relative coordinate geografiche ed altre informazioni che non sono state prese in considerazione durante questa analisi. I dati relativi al traffico veicolari vengono aggiornati con una frequenza mensile con le rilevazioni del mese precedente.

[35] https://opendata.comune.bologna.it/explore/dataset/dati-centraline-bologna-storico/information/?sort=data_inizio&disjunctive.agente
[36] https://dati.arpae.it/dataset/qualita-dell-aria-rete-di-monitoraggio/resource/21a9464d-c91a-4f17-b5c7-f3ee7560ff7e
[37] https://opendata.comune.bologna.it/explore/dataset/rilevazione-flusso-veicoli-tramite-spire-anno-2024/information/?disjunctive.codice_spira&disjunctive.tipologia&disjunctive.nome_via&disjunctive.stato

Figura 3-3 *Dataset 'Rilevazione flusso veicoli tramite spire – anno 2024'*

ID_univoco_stazione_spira	data	giorno settimana	codice spira	00:00-01:00	01:00-02:00	02:00-03:00	03:00-04:00	04:00-05:00
14	30 aprile 2024	Martedì	0.127 1.3 2 1	9	7	3	1	3
15	30 aprile 2024	Martedì	0.127 1.4 2 1	43	16	4	10	11
18	30 aprile 2024	Martedì	0.127 1.5 8 1	21	19	23	16	15
2	30 aprile 2024	Martedì	0.127 1.12 8 1	28	12	11	6	5

Fonte: Open Data Bologna

Infine, è stato incluso il dataset *'Rilevazione flusso veicoli tramite spire – anno 2023'*[38] *(Figura 3-3)*, rinominato *'Rilevazioni_SPIRE_2023'*. Questo dataset è identico per struttura e contenuto al precedente, ma contiene le informazioni relative esclusivamente all'anno 2023. È importante notare che questo dataset non viene aggiornato poiché include solo le informazioni riguardanti il 2023.

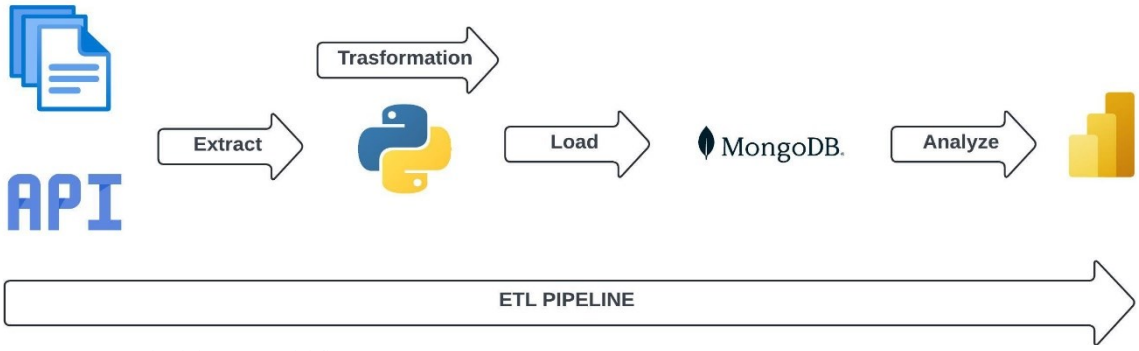
I dataset relativi all'inquinamento atmosferico e al traffico veicolare forniscono una solida base per condurre un'analisi completa e approfondita degli impatti ambientali dell'attività umana sulla città di Bologna. L'utilizzo di questi dataset, ciascuno con le proprie caratteristiche e periodi di rilevazione, permette di esaminare aspetti chiave e di ottenere una visione generale del contesto urbano, costruendo una solida base per le analisi e le conclusioni che seguiranno.

[38] https://opendata.comune.bologna.it/explore/dataset/rilevazione-flusso-veicoli-tramite-spire-anno-2023/information/?disjunctive.codice_spira&disjunctive.tipologia&disjunctive.nome_via&disjunctive.stato&sort=-chiave

3.2 **Creazione dell'infrastruttura dei dati**

Dopo aver identificato e descritto le fonti primarie dei dati utilizzati in questa tesi, si passa alla definizione e alla realizzazione della pipeline ETL. La costruzione di questa pipeline è stata progettata per assicurare che ogni fase del processo venga eseguita con la massima precisione ed efficienza.

Figura 3-4 *Struttura della Pipeline ETL implementata*



Fonte : Personale elaborazione del flusso ETL

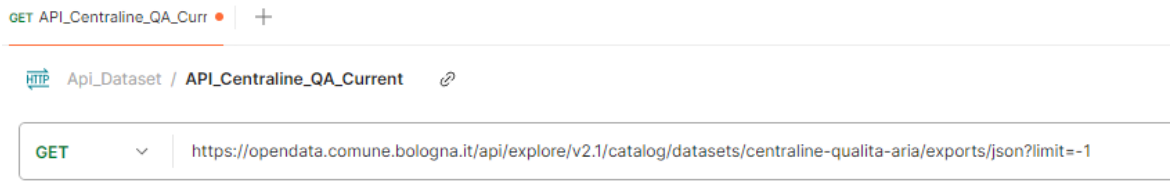
Come si può osservare nella Figura 3-4, inizialmente i dati vengono estratti dalle diverse fonti tramite chiamate *API*, testate preliminarmente con il tool *Postman* ed eseguite da uno script in *Python*. Successivamente, i dati grezzi vengono sottoposti a un processo di trasformazione che include la rimozione di variabili inutili per l'analisi, l'allineamento della formattazione dei dati tra i vari dataset e la derivazione di variabili da quelle già esistenti, garantendo così l'accuratezza e la coerenza dei dati.

Infine, i dati trasformati vengono inseriti nel database *MongoDB*. Questo passaggio finale assicura che i dati siano pronti per essere trasferiti, tramite protocollo *ODBC*[39], su Power BI, facilitando così l'analisi e la generazione di insight utili per la ricerca. Di seguito verranno descritti in dettaglio la metodologia adottata per l'estrazione, le tecniche impiegate per la trasformazione dei dati grezzi, il processo di inserimento dei dati nel database, e la progettazione e creazione delle dashboard per rappresentare il fenomeno.

[39] https://www.geekandjob.com/wiki/odbc

Le *API* permettono di accedere ai dati in tempo reale e di automatizzare il processo di raccolta. Le chiamate API utilizzate sono state sviluppate e messe a disposizione dal sito OpenData del Comune di Bologna, che offre un'interfaccia programmabile per accedere ai dataset pubblici. Per garantire l'efficacia delle chiamate API, sono stati utilizzati strumenti come Postman per testare e validare le singole richieste, prima della loro effettiva implementazione (Figura 3-5).

Figura 3-5 *URL chiamata API su Postman per il dataset 'Centraline_QA_Current'*

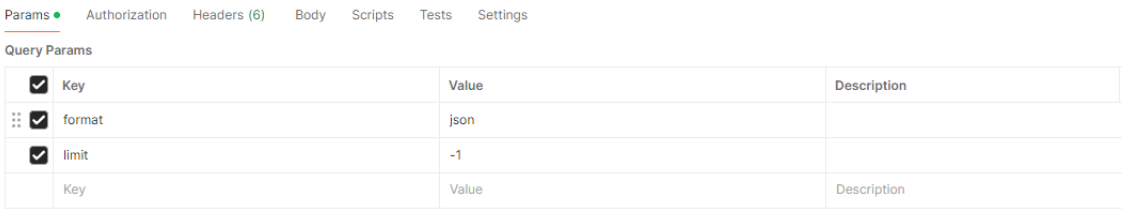


Dopo aver selezionato il tipo di richiesta HTTP *'GET'*, utilizzato per recuperare le informazioni da una risorsa, il passo successivo è stato l'analisi del '*Base URL'* (URL di base) il quale è standard per tutti i dataset consultabili sul sito Open Data del Comune di Bologna: https://opendata.comune.bologna.it/api/explore/v2.1/catalog/datasets. Ogni *'End-Point'*[40] fornisce l'accesso agli specifici dataset, rispettivamente:

- *'/centraline-qualita-aria/export/json' per* il dataset *'Centraline_QA_Current'*
- *'/dati-centraline-bologna-storico/export/json'* per il dataset *'Centraline_QA_Storico'*
- *'/rilevazione-flusso-veicoli-tramite-spire-anno-2024/export/json'* per il dataset *'Rilevazioni_SPIRE_2024'*
- *'/rilevazione-flusso-veicoli-tramite-spire-anno-2023/export/json'* per il dataset *'Rilevazioni_SPIRE_2023'*

[40] https://www.microsoft.com/it-it/security/business/security-101/what-is-an-endpoint#:~:text=Un%20endpoint%20API%20%C3%A8%20l'URL%20di%20un%20server%20o%20servizio.

Figura 3-6 *Query Params su Postman per il dataset 'Centraline_QA_Current'*

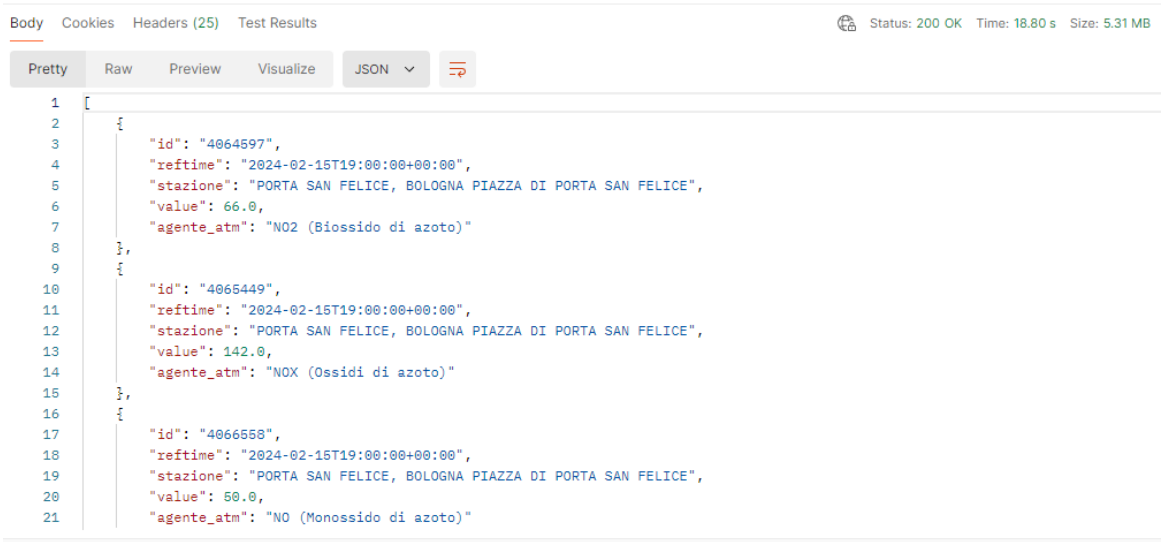


A completare l'URL troviamo i parametri di query della chiamata API che consentono di applicare dei filtri sui dati. In questo caso, sono stati utilizzati i seguenti parametri:

- *'/json?format= json'*, che imposta il corpo della risposta in formato JSON
- *'Limit=-1'*, che indica il numero massimo di record da esportare

Un aspetto importante da considerare è che ci sono dei limiti sulle richieste per i dati, imposti per evitare sovraccarichi del server. Questi limiti possono influenzare la frequenza delle richieste e la quantità di dati che possono essere scaricati in una singola chiamata. Tuttavia, non ci sono restrizioni se si desidera scaricare l'intero dataset, infatti in questo caso il parametro è stato impostato sul valore *'-1'* che è il valore di default per indicare il download della totalità dei record.

Figura 3-7 *Corpo della risposta API su Postman per il dataset 'Centraline_QA_Current'*



Le risposte delle API sono formattate in *JSON* (JavaScript Object Notation), un formato di dati leggero e facilmente leggibile sia per gli esseri umani sia per le macchine. Questo facilita la manipolazione dei dati nei processi di trasformazione successivi, permettendo di estrarre, trasformare e caricare i dati con maggiore efficienza. Inoltre, *Postman* fornisce informazioni

relativi ai codici di stato HTTP relativa alla richiesta, tempo di esecuzione e le dimensioni della risposta.

L'uso di strumenti come *Postman* ha permesso di verificare che le chiamate *API* fossero corrette e che i dati ottenuti fossero quelli attesi. Una volta validata la chiamata, lo script in Python è stato configurato per eseguire queste richieste in modo automatizzato, garantendo l'aggiornamento continuo dei dati.

Nella fase di estrazione, sono stati programmati attraverso l'ambiente di sviluppo integrato (IDE) *Pycharm* degli script *Python*, uno per ogni dataset, per effettuare chiamate API periodiche, consentendo di recuperare i dati dai vari endpoint del sito Open Data del Comune di Bologna. Questo approccio permette di accedere ai dati in tempo reale e di assicurare che le informazioni più aggiornate siano sempre disponibili per l'analisi.

Dopo la fase di estrazione, i dati grezzi vengono spostati nell'area di staging e sottoposti a un processo di trasformazione. È importante notare che per garantire una gestione ottimale dei dati, è stato creato uno script Python ad hoc per ciascun dataset. Questo approccio personalizzato consente di adattare il processo ETL alle specifiche esigenze e alla struttura unica di ciascun insieme di dati.

Di seguito si riporta, a titolo di esempio, il codice *Python* dell'intero processo ETL utilizzato per il l'estrazione, la trasformazione e il caricamento del dataset *'Centraline_QA_Current'*.

Codice 3-1 *Script ETL in Python per il dataset 'Centraline_QA_Current'*

```
import time
import requests
import tqdm
from pymongo import MongoClient
from tqdm import tqdm
import pandas as pd

# Configura la connessione a MongoDB
mongo_uri = "mongodb://localhost:27017/GetMyAQ"
client = MongoClient(mongo_uri)
database = client.get_database()

def mid(s, i, f):
    return s[i:f]


   def ultimo_record_centraline_current():
    ultimo_record =
   database.Centraline_QA_current.find_one(sort=[("record_id", -1)])
    print('Data ultimo aggiornamento Centraline_QA_current: ',
```

```
ultimo_record["Data_rilevazione"])
  #restituisce il record_id più recente
   return ultimo_record["record_id"]


def delta_aggiornamento_centraline_current(api):
   #Fase di estrazione
    last_record = ultimo_record_centraline_current()
    response = requests.get(api_url_Centraline_Current)
    if response.status_code == 200:
        data = response.json()

        data_filtered =[record
                        for record in data
                        if record["id"] >= last_record]
        print("dati correttamente ricevuti dall'API")

         total_iterations = len(data)

        # Fase di trasformazione
        for record in tqdm(data_filtered, desc="Aggiornamento dati"):
            id_record = record.get("id")
            nuovi_dati = {
                "Data_rilevazione": record.get("reftime"),
                "stazione": record.get("stazione"),
                "Giorno_rilevazione":       mid(record.get("reftime"), 0, 10),
                "Orario_Rilevazione": mid(record.get("reftime"), 11, 19),
                "valore_rilevato": record.get("value"),
                "agente_atm": record.get("agente_atm")
            }

            #Fase di caricamento
                    database.Centraline_QA_current.update_one(
                {"record_id": id_record},
                {"$set": nuovi_dati},
                upsert=True
            )

        print("Aggiornamento completato dataset delle Centraline [Current].")
        time.sleep(2)
    else:
        print("Errore nella chiamata API: ", response.status_code, " per il dataset
delle Centraline [current]")

    return

api_url_Centraline_QA_Current =
```

```
("https://opendata.comune.bologna.it/api/explore/v2.1/catalog" "/datasets/centraline-qualita-aria/exports/json?Format=json&limit=-1")

aggiornamento_centraline_current(api_url_Centraline_QA_Current)
```

L'estrazione dei dati viene eseguita attraverso una richiesta GET all'API URL specificato nella variabile **api_url_Centraline_QA_Current**. Innanzitutto, viene effettuato un controllo sullo status della risposta ricevuta dalla chiamata all'*API*: se la risposta ha uno *'status code'* pari a 200 (che indica una risposta di successo), i dati vengono salvati, in formato JSON, nella variabile **'data'**. Per evitare che ad ogni aggiornamento l'intero dataset debba essere processato, la funzione **'ultimo_record_centraline_current()'** interroga il database di destinazione estraendo il **'record_id'** dell'osservazione più recente. L'id dell'ultimo record viene salvato nella variabile **'last_record',** il quale servirà da filtro per il ciclo for che memorizzerà nella variabile **'data_filtered'** solamente i record con id maggiore all'ultimo presente nel database. Verrà stampato un messaggio di avvenuta estrazione dei dati al termine del processo.
Da questo momento si entra nella fase di trasformazione dei dati. Vengono iterati i record dei dati grezzi e per ciascun record vengono estratti i valori rilevanti come:

- data di rilevazione
- Giorno e Orario di rilevazione, derivate dalla data di rilevazione tramite la funzione **'mid()'**
- stazione
- valore rilevato
- Agente inquinante rilevato

Questi valori vengono poi strutturati in un dizionario. Di seguito le trasformazioni apportate agli altri dataset:

- Dataset *'Centraline_QA_Storico'* sono stati modificati i nomi delle variabili e uniformati i nomi degli agenti inquinanti nella variabile **"agente_atm"** tramite la funzione:

Codice 3-2 *Funzione Python per trasformazione dei valori del dataset*

```
def formatting(chiave):
lower_case = {'O3 (OZONO)': 'O3 (Ozono)', 'C6H6 (BENZENE)': 'C6H6 (Benzene)',
  'CO (MONOSSIDO DI CARBONIO)': 'CO (Monossido di carbonio)',
```

```
'NO (MONOSSIDO DI AZOTO)': 'NO (Monossido di azoto)',
'NO2 (BIOSSIDO DI AZOTO)': 'NO2 (Biossido di azoto)',
'NOX (OSSIDI DI AZOTO)': 'NOX (Ossidi di azoto)',
'PM 10': 'PM10',
'PM 2.5': 'PM2.5'}
```

- Dataset sono *'Rilevazioni_SPIRE_2024'* stati modificati i nomi dei campi e derivate delle variabili
- Dataset *'Rilevazioni_SPIRE_2023'* sono stati modificati i nomi dei campi e derivate delle variabili

Un aspetto significativo è la progettazione della struttura dei dati all'interno del database. Questo include la definizione di collezioni e documenti che meglio si adattano ai requisiti del progetto. Ad esempio, nel contesto dell'analisi delle centraline di qualità dell'aria, potrebbe essere utile organizzare i dati in modo che ciascun documento rappresenti una singola rilevazione di qualità dell'aria, con campi che includono la data di rilevamento, il valore misurato, l'agente inquinante rilevato e tutte le altre variabili (Figura 3-8).
Dopo aver implementato con successo la connessione al database MongoDB e aver definito le procedure per l'aggiornamento, questi dati strutturati vengono quindi caricati nel database MongoDB utilizzando il metodo **'update_one()'** della libreria *pymongo* di *Python*, il quale controlla, per ogni 'record_id' se esiste gia all'interno del database oppure necessita la creazione di un nuovo documento.

Figura 3-8 *Esempio dei documenti contenuti nella collection 'Centraline_QA_Current' in MongoDB Compass*


```
_id: ObjectId('6640c992a936856cbe305f97')
record_id : "4064597"
Data_rilevazione : "2024-02-15T19:00:00+00:00"
Giorno_rilevazione : "2024-02-15"
Orario_Rilevazione : "19:00:00"
agente_atm : "NO2 (Biossido di azoto)"
stazione : "PORTA SAN FELICE, BOLOGNA PIAZZA DI PORTA SAN FELICE"
valore_rilevato : 66

_id: ObjectId('6640c992a936856cbe305f99')
record_id : "4065449"
Data_rilevazione : "2024-02-15T19:00:00+00:00"
Giorno_rilevazione : "2024-02-15"
Orario_Rilevazione : "19:00:00"
agente_atm : "NOX (Ossidi di azoto)"
stazione : "PORTA SAN FELICE, BOLOGNA PIAZZA DI PORTA SAN FELICE"
valore_rilevato : 142

_id: ObjectId('6640c992a936856cbe305f9b')
record_id : "4066558"
Data_rilevazione : "2024-02-15T19:00:00+00:00"
Giorno_rilevazione : "2024-02-15"
Orario_Rilevazione : "19:00:00"
agente_atm : "NO (Monossido di azoto)"
stazione : "PORTA SAN FELICE, BOLOGNA PIAZZA DI PORTA SAN FELICE"
valore_rilevato : 50
```


È stata aggiunta una barra di progresso per seguire in tempo reale il caricamento dei dati, che fornisce informazioni aggiuntive come tempo impiegato, tempo residuo e iterazioni al secondo. Infine, viene stampato un messaggio di fine processo di ETL.

Figura 3-9 *Output dello script Python al termine del processo di ETL*

```
Dati ricevuti dall'API
Aggiornamento dati: 100%|██████████| 32688/32688 [15:20<00:00, 35.51it/s]
Aggiornamento completato dataset delle Centraline [Current].
```

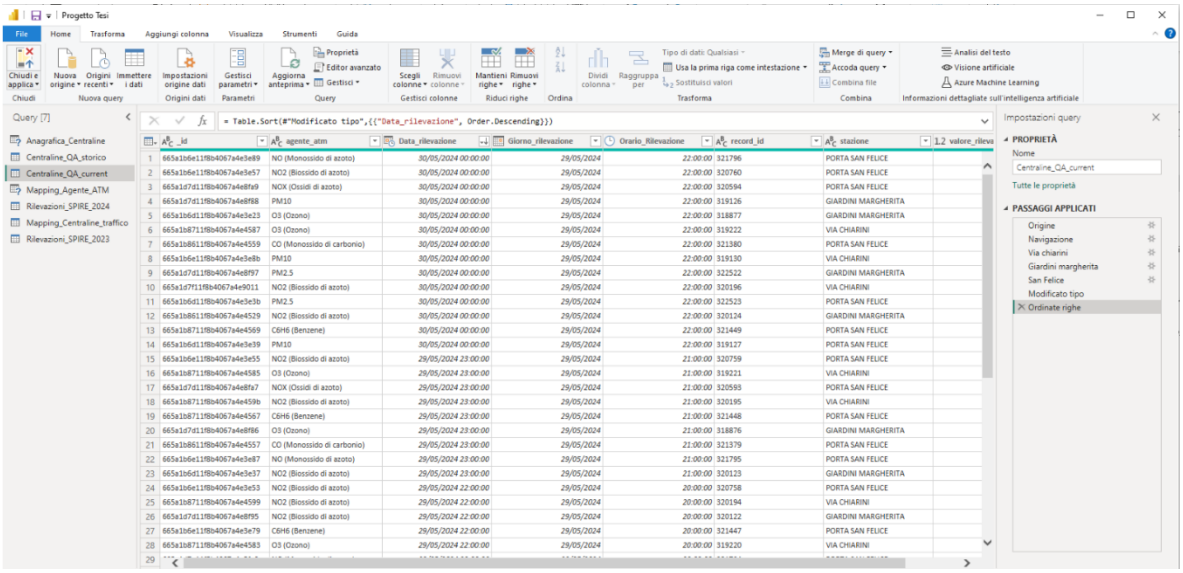
Questo processo garantisce che i dati siano aggiornati e pronti per essere analizzati.

L'ultimo componente della pipeline *ETL* è rappresentato dal software di reportistica *Power BI*. L'integrazione di *Power BI* nel processo *ETL* offre numerosi vantaggi. Prima di tutto,

consente di connettersi direttamente al database *MongoDB* tramite il protocollo *ODBC*, garantendo che i dati più recenti siano sempre disponibili per l'analisi. Inoltre, *Power BI* supporta una vasta gamma di funzionalità di trasformazione e modellazione dei dati che permettono di preparare i dati per l'analisi in modo intuitivo ed efficiente.

Il primo passo su *Power BI* è quello di recuperare i dati dalla fonte. Dopo aver creato un nuovo progetto, tramite il pulsante 'Recupera dati' nella barra degli strumenti, è possibile selezionare una delle numerosissime connessioni supportate da questo strumento. Come precedentemente accennato, bisogna selezionare il tipo di connessione *ODBC* e selezionare la connessione precedentemente configurata, in questo caso la connessione '*BI Connector'*.
In questo momento Power BI cercherà di stabilire una connessione al database. Al termine della connessione sarà possibile selezionare le *collections* da importare da MongoDB. Una volta selezionata la collection desiderata, si aprirà automaticamente lo strumento 'Trasforma dati', utile per applicare le ultime modifiche o formattazioni ai dati prima di essere importati ed analizzati (Figura 3-10).

Figura 3-10 *Strumento 'Trasforma dati' su Power BI*



Ad esempio, per il dataset *'Centraline_QA_Current'* sono state formattate tutte le colonne in base al tipo di dato che contengono, come le colonne 'Data_rilevazione', 'Giorno_rilevazione' e 'Orario_Rilevazione' modificate in formato 'Data/Ora' oppure la colonna 'valore_rilevato' modificato in numero decimale. Inoltre, per i nomi delle stazioni, contenuti nella colonna 'stazione', si è deciso di rimuovere il nome della via lasciando solamente il nome della zona

in cui la stazione è situata. Questo tipo di modifica può essere agevolmente fatta tramite codice DAX (A. Ferrari, 2015), riportato di seguito:

Codice 3-3 *Codice DAX trasformazioni valori su Power BI*

```
= Table.ReplaceValue(Centraline_QA_current_Table,"VIA CHIARINI, BOLOGNA VIA CHIARINI","VIA CHIARINI",Replacer.ReplaceText,{"stazione"})
```

Anche per tutte le altre tabelle sono state effettuate piccole trasformazioni, aggiustamenti nel tipo di dato e modifiche all'intestazione delle colonne.

Una volta conclusa la fase di trasformazione, Power BI inizierà ad importare i dati dal database, applicando automaticamente le trasformazioni su tutti i nuovi dati. È un processo che può durare da pochi minuti ad alcune ore, in base alla mole di dati da importare, dalle trasformazioni da applicare e dalla capacità di trasferimento dell'hardware sulla quale risiede l'infrastruttura.

Dopo il completamento del processo di caricamento dei dati, l'attenzione si sposta sulla creazione di dashboard interattive e informative. Queste dashboard rappresentano l'ultima fase del processo ETL, dove i dati trasformati e caricati vengono utilizzati per generare visualizzazioni significative che facilitano l'analisi e l'interpretazione.

3.3 **Implementazione di dashboard interattive**

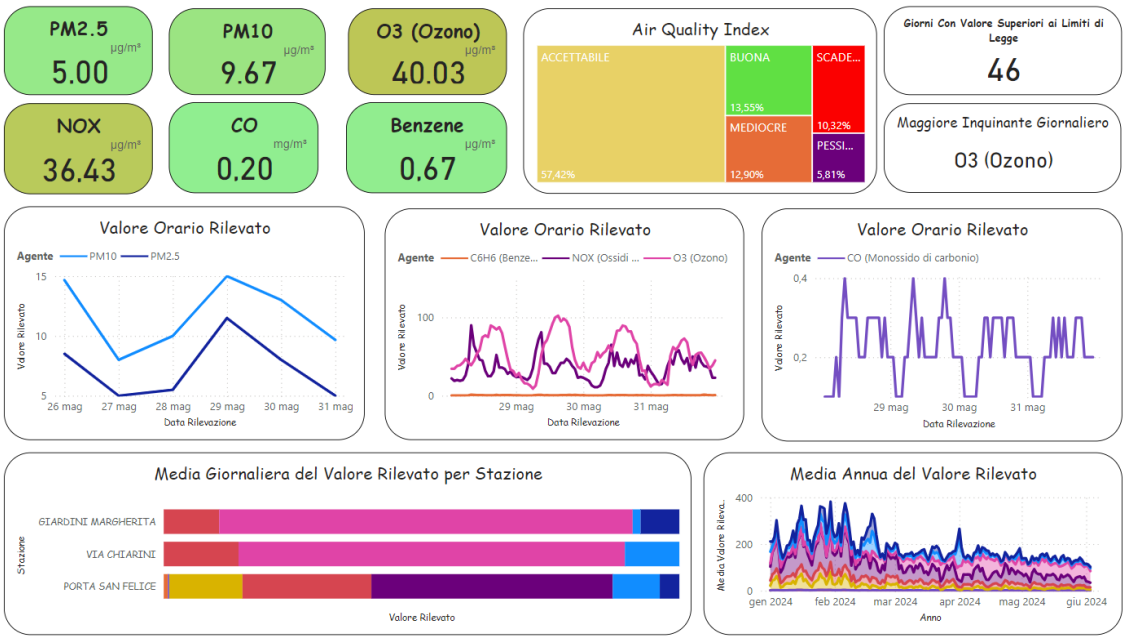
Power BI permette di riassumere grandi quantità di dati in dashboard contenenti rappresentazioni visive, come grafici, mappe e tabelle, che forniscono una panoramica chiara e immediata delle informazioni. Questo software permette di esplorare i dati attraverso visualizzazioni dinamiche che possono essere modellate sulla base delle specifiche esigenze di analisi.

Nel seguente paragrafo verranno illustrati i passaggi e le metodologie applicate per la progettazione e lo sviluppo dei cruscotti. Saranno descritti i principali KPI, gli schemi relazionali alla base di ogni dashboard e come queste visualizzazioni possano contribuire ad una migliore comprensione dei fenomeni presi in analisi in questa trattazione.

3.3.1 *Dashboard Inquinamento*

Il fenomeno preso in analisi per la creazione della prima dashboard (Figura 3-11) è quello dell'inquinamento atmosferico.

Figura 3-11 *Dashboard Inquinamento Atmosferico*



La seguente dashboard mira a fornire una panoramica dettagliata sulla qualità dell'aria della città di Bologna, rappresentando i dati sui livelli degli inquinanti atmosferici rilevati giornalmente dalle diverse centraline dislocate in diverse aree della città. Verranno sfruttati i dati raccolti nel dataset *'Centraline_QA_Current'*. Questa dashboard permetterà di monitorare in tempo reale (nei limiti delle tempistiche di aggiornamento dei dati) le variazioni dei principali inquinanti, facilitandone l'individuazione di eventuali picchi ed anomalie.
La fase di progettazione inizia dalla definizione del modello relazionale[41], al fine di una corretta propagazione dei filtri attraverso tutte le visualizzazioni interessate. La Figura 3-12 evidenzia:

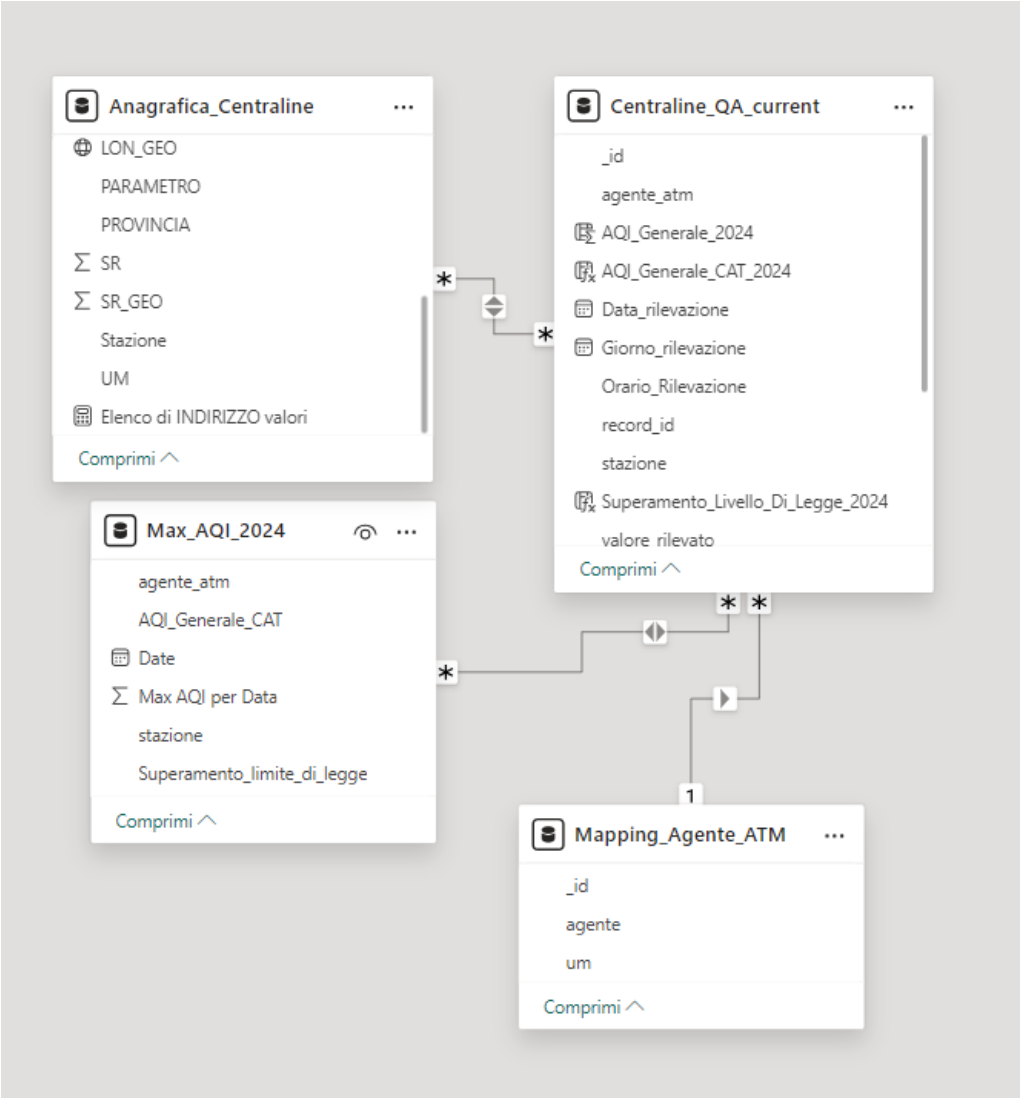
- Una relazione molti-a-molti tra la tabella '*Anagrafica_Centraline*' e la tabella *'Centraline_QA_Current'*.

- Una relazione molti-a-molti tra la tabella 'Max_AQI_2024*'*, tabella di supporto appositamente creata per lo sviluppo di alcuni KPI che verranno introdotti successivamente, e la tabella *'Centraline_QA_Current'*.

- Una relazione uno-a-molti tra la tabella 'Mapping_Agente_ATM', tabella che contiene il mapping tra l'agente inquinante e la relativa unità di misura, e la tabella *'Centraline_QA_Current'*.

[41] https://learn.microsoft.com/it-it/power-bi/transform-model/desktop-relationships-understand

Figura 3-12 *Modello Relazionale della dashboard 'Inquinamento'*

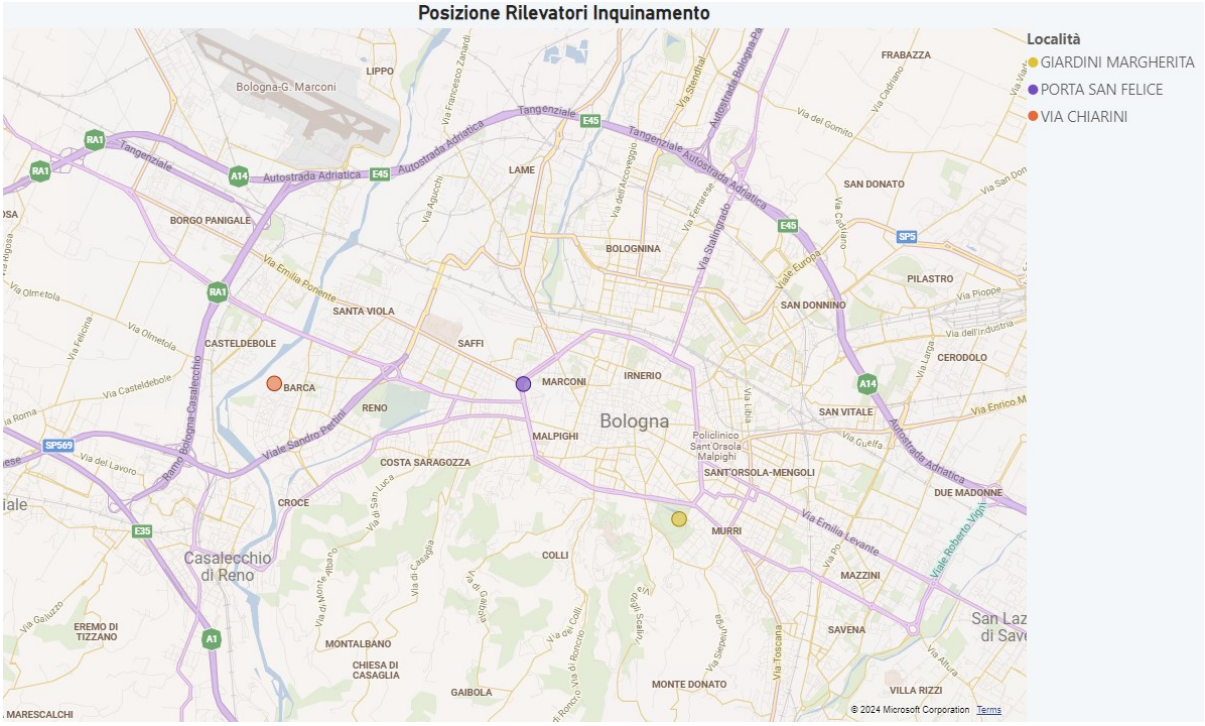


Il primo aspetto da considerare è la dislocazione delle centraline all'interno della città di Bologna. Come precedentemente accennato durante l'introduzione dei dataset utilizzati, le centraline di rilevamento sono tre e sono dislocati in tre punti della città con caratteristiche altamente diverse l'una dall'altra (Figura 3-13).

La prima è posizionata nel cuore dei Giardini Margherita, uno dei parchi più grandi nella città di Bologna, posto inaccessibile ai veicoli a combustione e frequentato principalmente da famiglie, ragazzi e sportivi. La centralina di rilevamento è dunque circondata da un'ampia zona di verde.

La seconda, invece, è posizionata nei pressi della Porta San Felice, uno degli snodi più trafficati di tutta la città, in tutte le fasce orarie. La centralina è dunque circondata da strade altamente trafficate e costellate di semafori che rallentano il traffico.
L'ultima centralina presa in considerazione è quella di Via Chiarini, in zona Barca. Essa è posizionata in una zona relativamente periferica della città di Bologna ma ai confini della zona industriale della città e circondate da alcune delle arterie principali di Bologna.
Dunque, i valori rilevati dalle tre centraline, se combinati insieme, possono dare una buona approssimazione della qualità dell'aria sull'intero contesto urbano della città di Bologna.

Figura 3-13 *Posizione rilevatori inquinamento nella città di Bologna*



La presenza capillare di queste centraline ha permesso di analizzare la situazione attuale del fenomeno dell'inquinamento atmosferico, oggetto della successiva dashboard. Questa visualizzazione ha lo scopo di rappresentare lo scenario corrente dei livelli di inquinamento dell'aria. Per poter fare ciò, sono stati scelti e implementati dei KPIs che ci aiutano a comprendere al meglio l'andamento del fenomeno.
Gli oggetti visivi rappresentati all'interno della dashboard sono:

- KPIs degli agenti inquinanti e i loro valori rilevati dalle centraline. In particolare, abbiamo PM 2.5, PM10, O3, NOX, CO e Benzene
- Mappa ad albero per i valori di Air Quality Index

- KPI numero di giorni con valore superiori ai limiti di legge
- KPI maggiore inquinante giornaliero rilevato
- Grafici a linee sull'andamento orario dell'inquinamento rilevato, per ogni inquinante, in un intervallo di tempo settimanale
- Grafici a barre in pila relativa alla media giornaliera del valore dell'agente inquinante rilevato per stazione
- Grafico ad area in pila che rappresenta la media annua del valore rilevato dell'agente inquinante dall'inizio del 2024

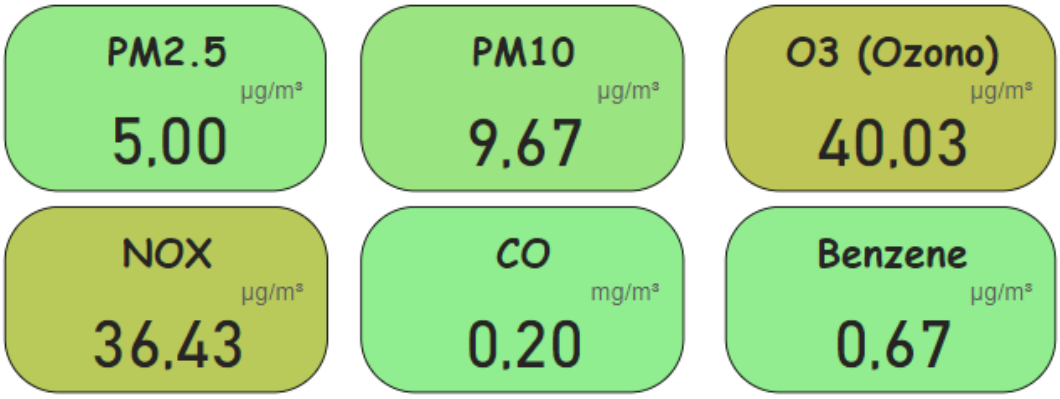
Scendendo più nel dettaglio, all'interno della dashboard (Figura 3-14) destinata alla situazione attuale dell'inquinamento, il primo elemento visivo che troviamo è quello del gruppo dei KPIs relativi alla media tra tutte le centraline, per ogni agente inquinante, aggiornato dell'ultima rilevazione disponibile. Questi KPIs sono stati sviluppati grazie alla creazione di una nuova misura all'interno della tabella *'Centraline_QA_Current',* tramite il codice DAX:

Codice 3-4 *Codice DAX per la creazione della misura 'valore medio rilevato'*

```
Media di valore_rilevato = AVERAGE('Centraline_QA_Current'[valore_rilevato])
```

I KPIs possono assumere, grazie alla possibilità di inserire funzione di sfondo, varie gradazioni di colore, dal verde chiaro al viola scuro, per indicare l'intensità della presenza dell'agente inquinante nell'aria. La scala dei valori va da 0 fino al massimo valore mai registrato per quell'agente inquinante, permettendo così una rappresentazione visiva chiara della sua concentrazione. L'utilizzo dei colori, nella Data Visualization, migliora e facilita la comprensione. Grazie alla loro attrattività visiva, i colori aiutano a soffermarsi su elementi chiave del fenomeno. Questo gruppo di KPIs ci fornisce una panoramica sulla situazione, aggiornata all'ultimo giorno disponibile, delle rilevazioni per ogni agente inquinante.

Figura 3-14 KPIs agenti inquinanti



Il secondo elemento visivo (Figura 3-15) è la mappa ad albero per i valori dell'AirQualityIndex[42], un indice che rappresenta sinteticamente lo stato complessivo dell'inquinamento atmosferico. La costruzione viene normata dall'ARPAE ed avviene tramite il calcolo di un sottoindice, su una scala adimensionale, numerico per ogni inquinante, per poi essere raggruppato in un indice sintetico categorico unico.

L'indice su scala adimensionale viene calcolato dividendo il valore rilevato per il singolo agente inquinante considerato, per il limite di legge previsto dalla legislazione (Tabella 1-1) e moltiplicando il risultato per 100. Di seguito, a titolo di esempio, una parte di codice DAX utilizzato per la costruzione di una colonna contenente l'indice adimensionale per alcuni degli agenti inquinanti presi in esame:

Codice 3-5 *Codice DAX per la creazione dell'indice AQI*

```
   AQI_Generale_2024 =
IF(
    Centraline_QA_Current[agente_atm] = "PM2.5",
    ROUNDUP((Centraline_QA_Current[valore_rilevato] / 25) * 100, 0),
    IF(
        Centraline_QA_Current[agente_atm] = "PM10",
        ROUNDUP((Centraline_QA_Current[valore_rilevato] / 50) * 100, 0),
        IF(
            Centraline_QA_Current[agente_atm] = "CO (Monossido di carbonio)",
            ROUNDUP((Centraline_QA_Current[valore_rilevato] / 10) * 100, 0)
```

[42] https://www.arpae.it/it/temi-ambientali/aria/scopri-di-piu/inquinanti-e-iqa/indice-della-qualita-dell-aria-iqa

Il passo successivo è quello di raggruppare gli indici adimensionali in classi. L'ARPAE , in linea con l'approccio adottato dalla maggior parte degli indici a livello internazionale, ha scelto di definire il valore dell'indice categorico come il valore dell'indice adimensionale peggiore. Come è possibile vedere dalla Tabella 3-1, le classi scelte sono cinque e con ampiezza uniforme e pari a 50:

Tabella 3-1 *Classi indice Air Quality Index*

Valore indice adimensionale	Cromatismi	Qualità dell'aria
< 50	verde	Buona
50 - 99	Giallo	Accettabile
100 – 149	Arancione	Mediocre
150 – 199	Rosso	Scadente
> 200	Viola	Pessima

Fonte: Personale Rielaborazione delle classi AQI presenti sul sito dell'ARPAE

L'indice categorico è stato calcolato all'interno di una nuova colonna nel relativo dataset tramite il seguente codice DAX:

Codice 3-6 *Codice DAX per trasformazione dell'indice AQI in valori categorici*

```
   AQI_Generale_CAT_2024 =
IF(
    Centraline_QA_Current[AQI_Generale_2024] <= 50,
    "BUONA",
    IF(
        AND(Centraline_QA_Current[AQI_Generale_2024] > 50,
Centraline_QA_Current[AQI_Generale_2024] <= 99),
        "ACCETTABILE",
        IF(
            AND(Centraline_QA_Current[AQI_Generale_2024] > 99,
Centraline_QA_Current[AQI_Generale_2024] <= 149),
            "MEDIOCRE",
            IF(
                AND(Centraline_QA_Current[AQI_Generale_2024] > 149,
Centraline_QA_Current[AQI_Generale_2024] <= 199),
                "SCADENTE",
                "PESSIMA"
```

```
                )
            )
        )
)
```

Infine, dato che l'indicatore generale per l'intera giornata prende in considerazione solamente il peggior valore registrato, si è deciso di costruire una tabella ausiliaria. Quest'ultima, derivata dalla tabella *'Centraline_QA_Current'*, contenente solamente l'agente inquinante, la stazione di rilevamento, la data di rilevamento, l'indice adimensionale e l'indice AQI categorico per la peggior rilevazione relativa ad ogni giorno, è stata denominata *'Max_AQI_2024'*. Questo oggetto visivo ha lo scopo di mostrare qual è la composizione delle classi di AQI dall'inizio dell'anno fino all'ultima rilevazione disponibile. Una rappresentazione di tipo mappa ad albero è visivamente attraente, efficiente dal punto di vista dell'ottimizzazione dello spazio e da quello esplicativo delle dimensioni di ogni singola categoria.

Figura 3-15 *Mappa ad albero per AirQualityIndex*

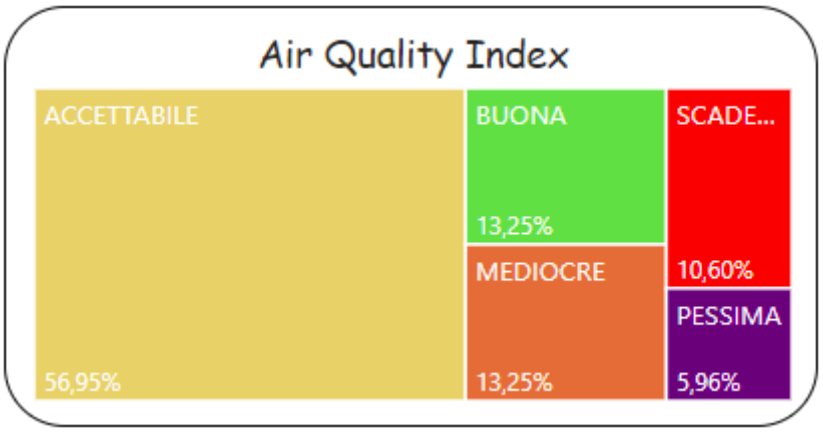
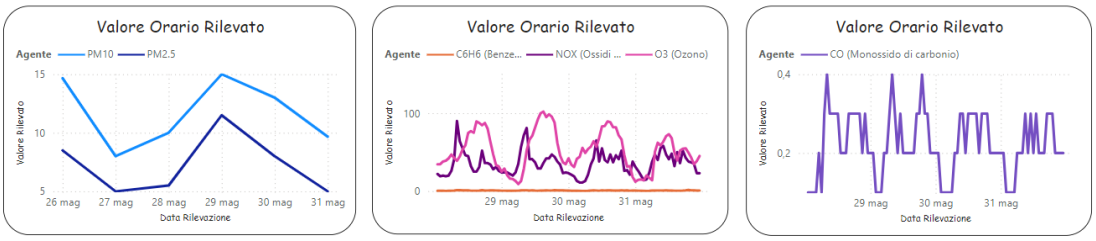


Figura 3-16 *KPIs relati al maggiore inquinate e al conteggio dei giorni con livelli superiori ai limiti di legge*

Dalla costruzione dell'AQI, sono derivati ulteriori due KPIs che forniscono informazioni sul conteggio del numero di giorni in cui si è superato il limite di legge per almeno un inquinante da inizio anno all'ultima rilevazione disponibile e qual è il maggiore inquinante presente nell'aria, sulla base del valore dell'indice adimensionale e non sul valore puntuale della rilevazione, nell'ultima giornata disponibile (Figura 3-16).

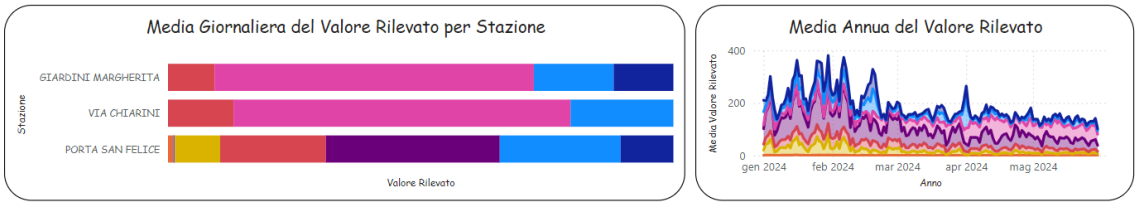
A completare la dashboard sull'inquinamento troviamo tre grafici a linee, i quali rappresentano i valori orari, mediati tra tutte le centraline, delle rilevazioni per ogni agente inquinante all'interno di un arco temporale di una settimana. Lo scopo di questi grafici è quello di mostrare qual è la situazione e l'andamento degli agenti inquinanti negli ultimi giorni (figura 3-17).

Figura 3-17 *Valore orario delle rilevazioni per tutti gli agenti inquinanti*



Infine, sulla parte bassa del cruscotto troviamo rispettivamente un grafico a barre in pila, che fornisce informazioni visive sulla composizione degli agenti inquinanti rilevati per ogni singola stazione di rilevamento per l'ultimo giorno disponibile, e un grafico ad aria in pila il quale ci mostra l'andamento medio giornaliero degli inquinanti da inizio anno 2024 fino al 31 maggio 2024, data di ultima rilevazione disponibile (Figura 3-18).

Figura 3-18 *Grafico a barre in pila e grafico ad aria in pila*



Con gli ultimi due grafici, l'intento di questa dashboard è quello di rappresentare, quanto più fedelmente possibile, la situazione oraria e giornaliera dei livelli di inquinamento rilevati ma anche fornire una panoramica a livello annuo e con indicatori sintetici globali che descrivono visivamente il fenomeno.

3.3.2 *Dashboard Traffico veicolare*

Il fenomeno preso in analisi per la creazione della seconda dashboard (Figura 3-19) è quello del traffico veicolare nella città di Bologna.

Figura 3-19 *Dashboard Traffico veicolare*

La seguente Dashboard mira a fornire una panoramica sulle condizioni di viabilità lungo le strade che interessano il tratto urbano. Verranno sfruttati i dati contenuti nel dataset *'Rilevazioni_SPIRE_2024'*, raccolti grazie alle centinaia di sensori sparsi lungo tutte le strade urbane di Bologna. Per esigenze di analisi, i dati relativi al numero di veicoli sono stati aggregati a livello di *'nome_via'* e ponderate per il numero di sensori presenti sulla stessa via. Questa operazione è stata fatta al fine di mitigare il problema della molteplice rilevazione di uno stesso veicolo che percorre l'intera via. I dati aggregati sono stati inseriti all'interno della tabella, appositamente creata, denominata *'Media_Giornaliera_Agg'*. Questa dashboard sintetizza graficamente informazioni utili relativi alla quantità di veicoli transitati in una strada, alle fasce orarie più trafficate e alla creazione di una visione generale dell'andamento del traffico da inizio anno.Anche la progettazione di questa dashboard inizia con la definizione del modello relazionale. La Figura 3-20 mostra:

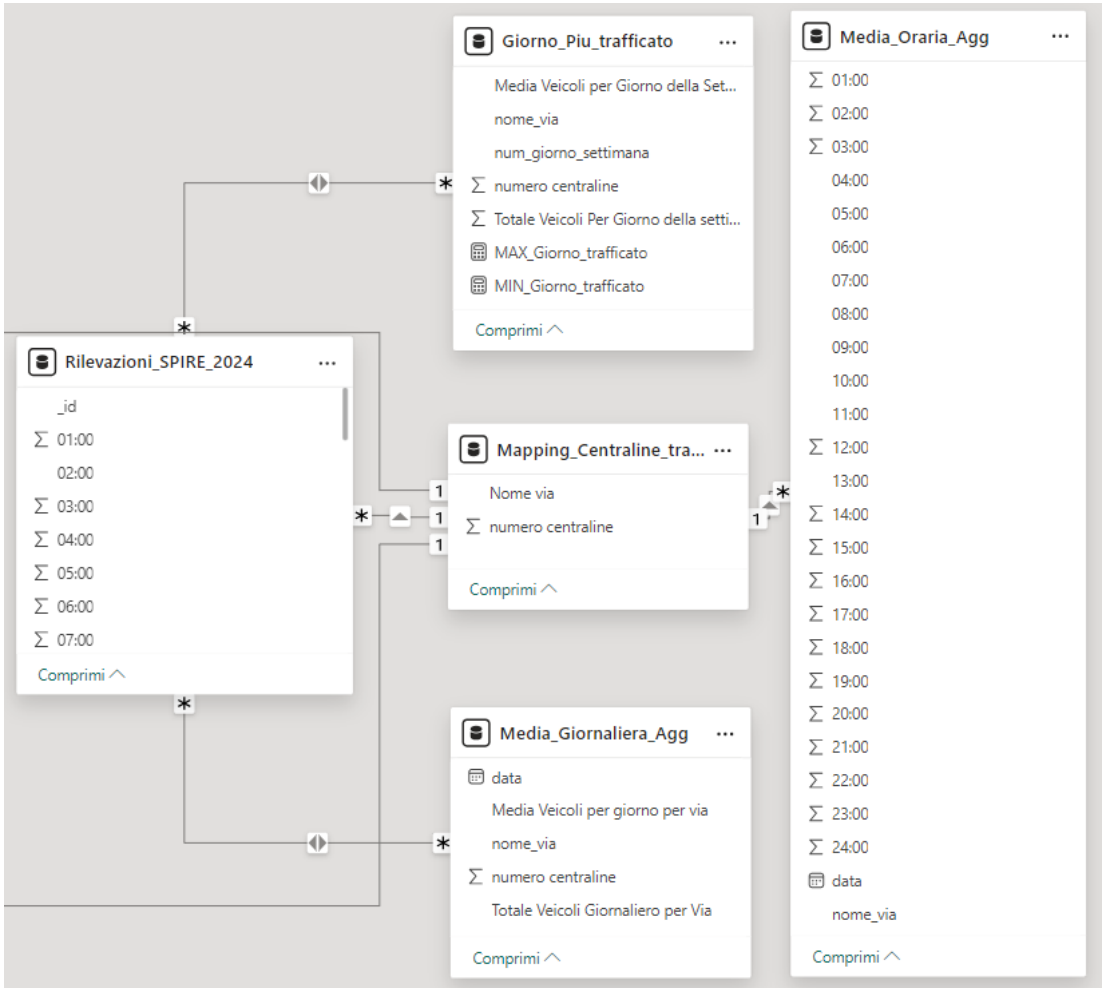
- Una relazione molti-a-uno tra la tabella *'Rilevazioni_SPIRE_2024'* e la tabella *'Mapping_Centraline_traffico',* tabella di supporto appositamente creata per porter mediare il valore delle rilevazioni per il numero di sensori presenti in ogni via.

- Una relazione molti-a-molti tra la tabella *'Rilevazioni_SPIRE_2024'* e la tabella *'Media_Giornaliera_Agg'*, tabella appositamente creata per ospitare i dati aggregati, come spiegato sopra.

- Una relazione molti-a-uno tra la tabella *'Rilevazioni_SPIRE_2024'* e la tabella *'Giorno_piu_trafficato'*, tabella di supporto creata per esigenze di sviluppo di grafici che verranno introdotti successivamente.

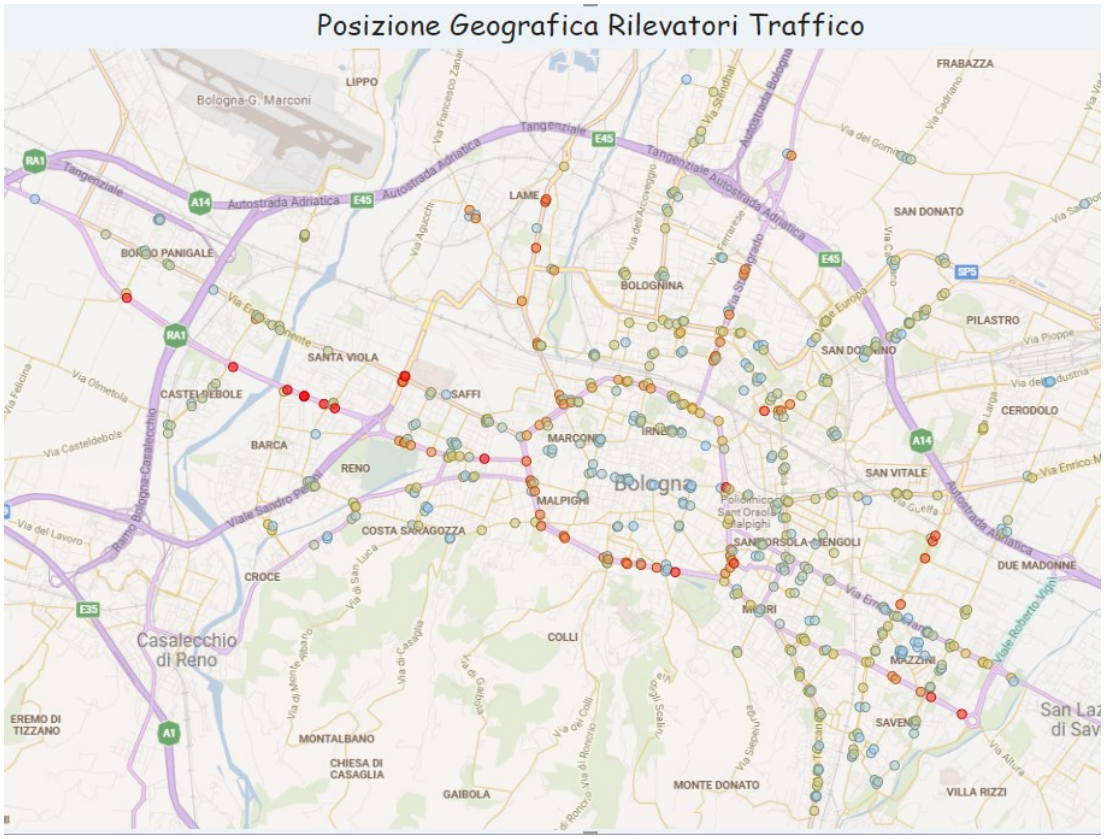
- Una relazione uno-a-molti tra la tabella *'Mapping_Centraline_traffico'* e la tabella *'Media_Oraria_Agg'*, tabella di supporto creata per esigenze di sviluppo di grafici che verranno introdotti successivamente.

Figura 3-20 *Modello Relazione della dashboard 'Traffico veicolare'*



Come prima visualizzazione, analogamente alla precedente dashboard, si rappresenta visivamente la dislocazione dei rilevatori del traffico su una mappa (Figura 3-21). È immediato notare come i rilevatori siano distribuiti sull'intero territorio urbano. La differente colorazione delle bolle, ciascuna rappresentante un rilevatore, indica il livello di affluenza lungo quella determinata via. Un colore tendente al celeste indica una bassa presenza giornaliera di veicoli mentre un colore tendente al rosso una maggiore presenza giornaliera di veicoli. Grazie a questi rilevatori è possibile individuare quali sono le strade più densamente trafficate, in quali orari e giorni della settimana.

Figura 3-21 *Posizione rilevatori traffico nella città di Bologna*



Lo scopo di questa dashboard è quello di analizzare le informazioni relative alla situazione del traffico veicolare nell'ultimo periodo, in modo tale che possano essere sfruttate per conoscere le strade più trafficate, e dunque al bisogno, evitarle per risparmiare tempo ma soprattutto emettere meno agenti inquinanti nell'atmosfera a causa delle congestioni stradali. Nello specifico in questa dashboard sono stati sviluppati ed implementati:

- Grafico a linee sull'andamento giornaliero del flusso dei veicoli
- KPI che rappresenta il numero di veicoli giornalieri
- Istogramma che mostra le vie più trafficate
- Oggetto visivo in Python per rappresentare le fasce orarie più trafficate
- Filtri da poter applicare per navigare tra i dati a disposizione

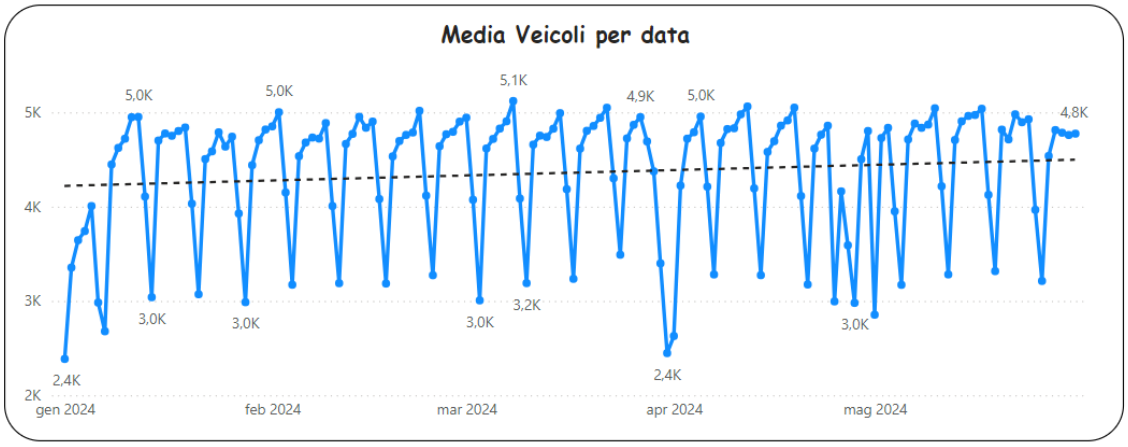
Come primo oggetto visivo troviamo il grafico a linee (Figura 3-22) che rappresenta l'andamento medio del flusso dei veicoli, con aggregazione giornaliera, in un dato periodo di tempo selezionabile dall'apposito filtro 'Data'. I dati rappresentati in questo grafico sono

contenuti all'interno della tabella *'Media_Giornaliera_Agg'*, la quale è stata derivata dalla tabella *'Rilevazioni_SPIRE_2024'* e contiene i dati aggregati tramite codice DAX per via, data di rilevazione e la media dei veicoli transitati ponderata per il numero di rilevatori lungo la stessa strada.

Il grafico in questione è utile per poter individuare eventuali picchi e anomalie riguardanti il traffico veicolare, sia dal punto di vista globale che per la singola strada. Inoltre, è presente una linea tratteggiata che indica il trend di periodo.

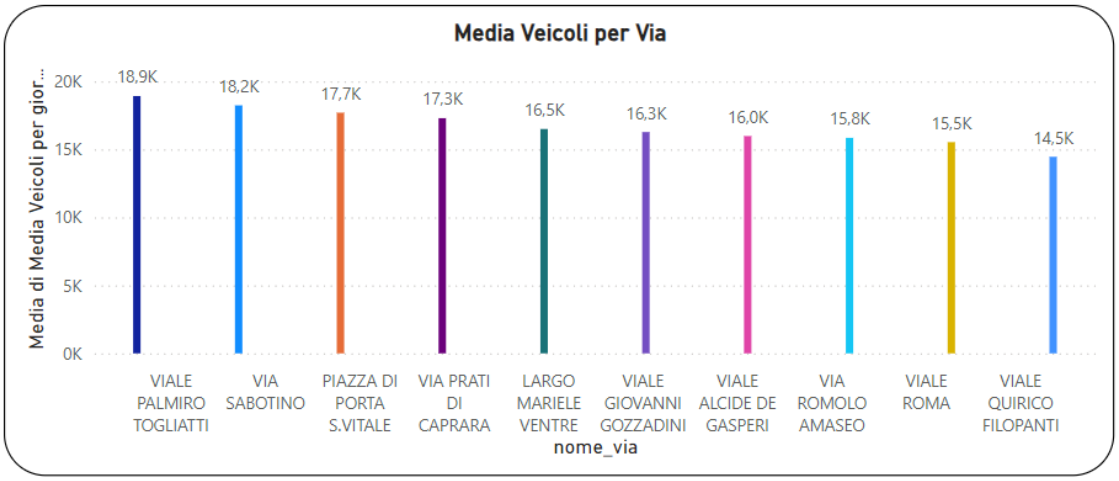
Per rendere più chiara e fruibile l'interpretazione del grafico precedente, viene in ausilio il KPI della media dei veicoli giornalieri, che in assenza di selezioni di filtri, sarà quella che prende in considerazione la totalità dei dati a disposizioni per tutte le vie. Il KPI si aggiornerà automaticamente dopo la selezione de filtri 'Data' o 'Nome Via'.

Figura 3-22 *Grafico a linee media dei veicoli giornalieri*



Il secondo oggetto visivo è l'istogramma che rappresenta le dieci vie, in ordine decrescente, con la maggior mole di traffico giornaliero (Figura 3-23). In questo caso, la scelta dei colori adottata da Power BI ha lo scopo di differenziare le strade prese in esame.

Figura 3-23 *Istogramma dieci vie più trafficate*



A concludere la rappresentazione grafica della situazione del traffico veicolare, troviamo un oggetto visivo sviluppato tramite il seguente codice Python (Figura 3-24):

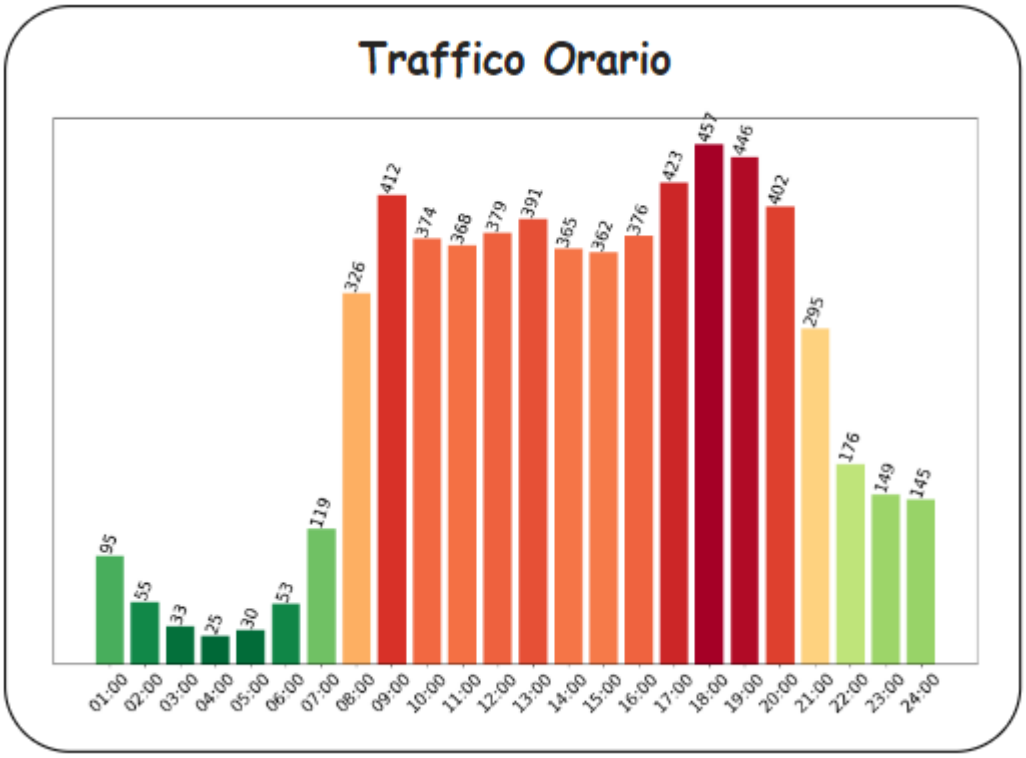
Codice 3-7 *Codice Python per la creazione dell'oggetto visivo 'Traffico Orario'*

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Calcolare i valori medi e arrotondarli
mean_values = round(dataset.mean()).astype(int)

# Colormap per il valori

norm = plt.Normalize(vmin=mean_values.min(), vmax=mean_values.max())
cmap = plt.cm.get_cmap('RdYlGn_r')
colors = cmap(norm(mean_values.values))

# Creazione del grafico a barre
plt.figure(figsize=(15, 10))
bars = plt.bar(mean_values.index, mean_values.values, color=colors)
plt.gca().axes.get_yaxis().set_visible(False)

# Etichette sui valori delle barre
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval, round(yval, 2),
va='bottom', ha='center', fontsize=19, rotation=70)
```

```
# Impostazioni grafico
plt.xticks(fontsize=19, rotation=45)
plt.tight_layout()
plt.show()
```

Quest'ultimo grafico a barre illustra e sintetizza la distribuzione del traffico durante l'intera giornata che si vuole filtrare, evidenziando le fasce orarie in cui il traffico raggiunge dei picchi di affluenza. Inoltre, grazie alla notevole flessibilità e adattabilità del linguaggio Python, è possibile colorare le barre del grafico, con sfumature dal verde al rosso che stanno ad indicare l'intensità del traffico veicolare.

Figura 3-24 *Oggetto visivo 'Traffico Orario' in Python*

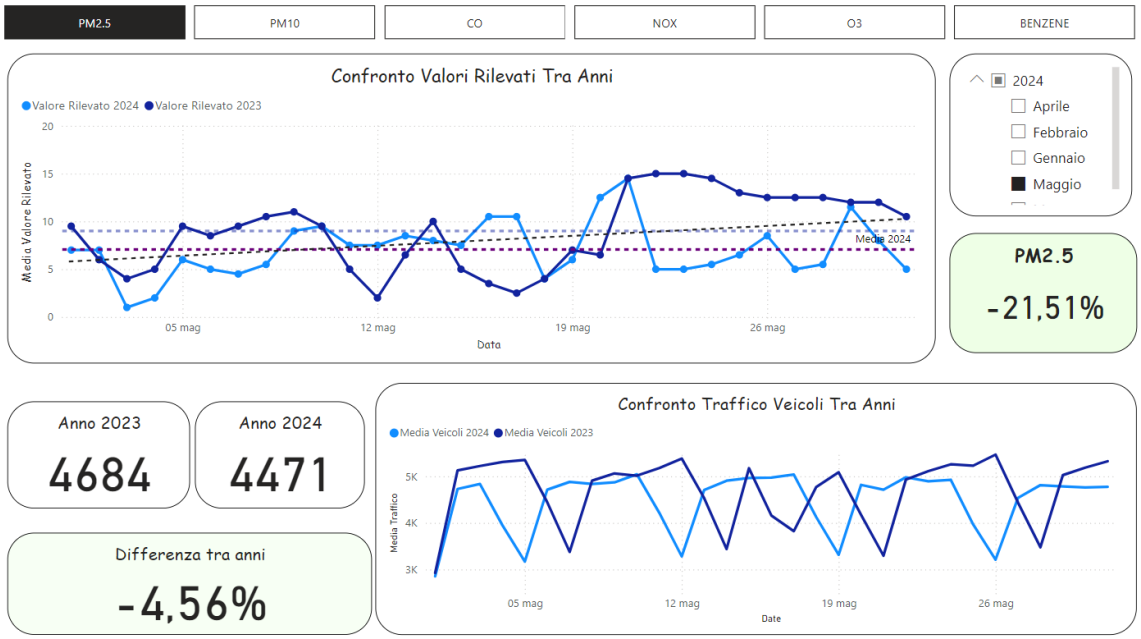


La visualizzazione sul traffico orario conclude la dimostrazione della panoramica sulla situazione attuale del traffico nelle strade di Bologna, variabile che successivamente verrà messa a paragone con gli anni precedenti in modo da poterne comprendere il comportamento.

3.3.3 *Dashboard Confronto*

Dopo aver analizzato, attraverso due precedenti dashboard, la situazione dell'inquinamento atmosferico e le condizioni di viabilità all'interno della città di Bologna, nel corso del 2024, è utile mettere a confronto i valori rilevati con quelli dell'anno precedente al fine di scovare anomalie tra i dati registrati nei due differenti anni (Figura 3-25).

Figura 3-25 *Dashboard Confronto tra anni*



Per la realizzazione di questa dashboard sono state create due ulteriori tabelle, la tabella '*Rilevazioni_inquinamento_full'* e *'Rilevazioni_SPIRE_full'*, rispettivamente il risultato dell'unione delle coppie di tabelle *'Centraline_QA_Current'* e *'Centraline_QA_Storico',* e le tabelle *'Rilevazioni_SPIRE_2024'* e *'Rilevazioni_SPIRE_2023'.* Si riporta, a titolo di esempio, il codice DAX utilizzato per la creazione della tabella *'Rilevazioni_inquinamento_full':*

Codice 3-8 *Codice DAX per unione datasets sull'inquinamento*

```
Rilevazioni_Inquinamento_full = UNION(
  SELECTCOLUMNS(
      Centraline_QA_current,
      "data", [Data_rilevazione],
      "stazione", Centraline_QA_current[stazione],
      "agente_atm", [agente_atm],
      "valore_rilevato", [valore_rilevato]
  ),
```

```
    SELECTCOLUMNS(Centraline_QA_storico,
        "data", [Data rilevazione],
        "stazione", Centraline_QA_storico[stazione],
        "agente_atm", [agente_atm],
        "valore_rilevato", [valore_rilevato]
    )
)
```

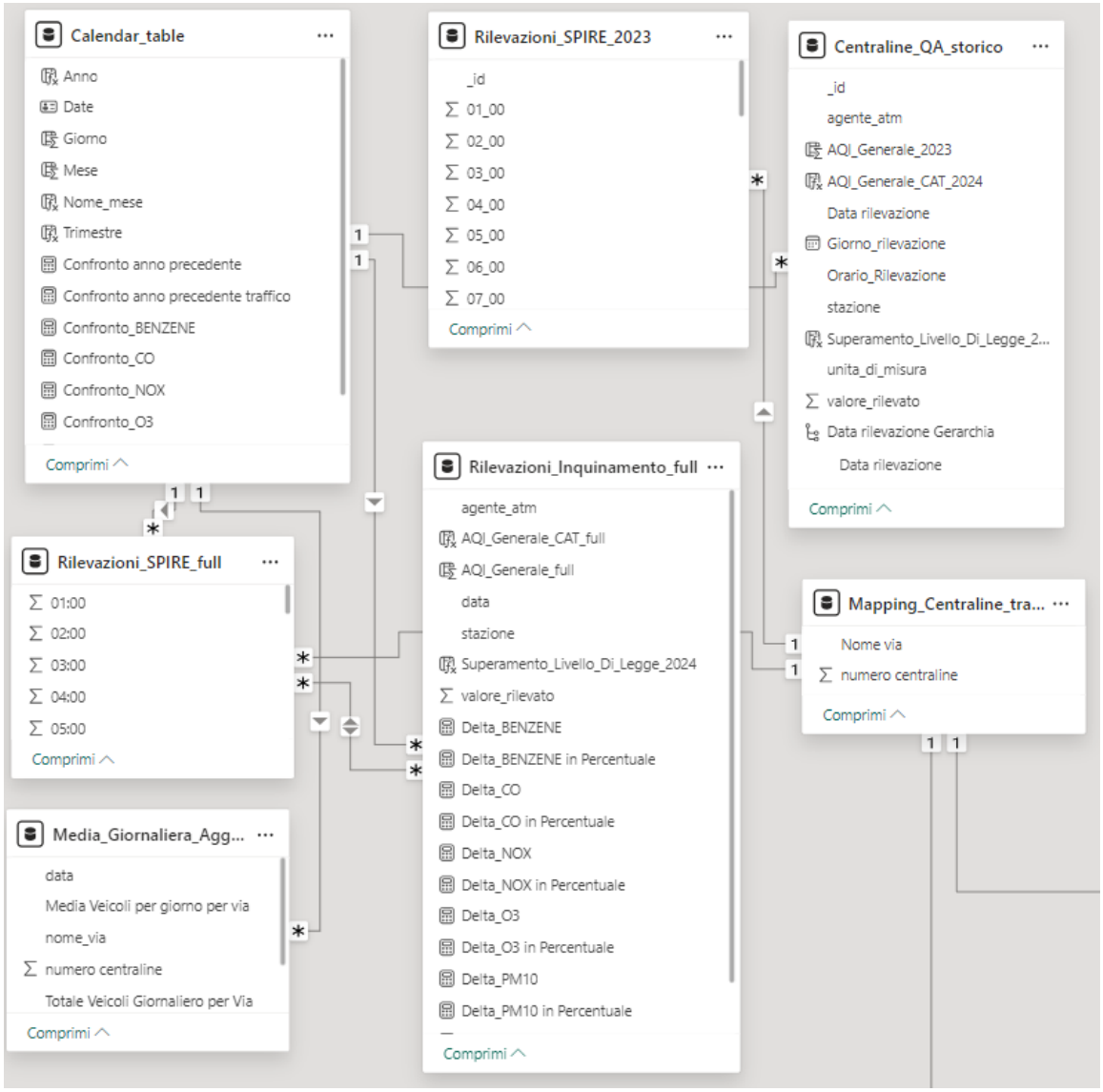
Esattamente come per le altre dashboard, in Figura 3-26 è riportato lo scherma relazionale della dashboard dedicata al confronto tra anni. Tuttavia, lo schema relazionale è ben più complesso rispetto a quelli visti finora. Oltre le tabelle già menzionate negli altri schemi relazionali, troviamo la tabella *'Calendar_table'* collegata con la relazione una-a-molti con tutte le altre tabelle. La *'Calendar_table'* è una particolare tabella che contiene una serie di date utilizzate per organizzare e gestire i dati temporali all'interno del modello di dati. Questa tabella, essenziale per molte operazioni analitiche, copre l'intervallo temporale rilevante per l'analisi, includendo ogni giorno tra la data iniziale e quella finale del periodo di interesse. È stata creata tramite codice DAX per poter essere sempre allineata alle date delle ultime rilevazioni con il seguente codice:

Codice 3-9 *Codice DAX per la creazione della 'Calendar Table'*

```
   Calendar_table                                                    =
CALENDAR("01/01/2017",MAX(Centraline_QA_current[Data_rilevazione]))
```

Alla *calendar table* sono state poi aggiunte colonne per l'anno, il mese, il giorno, il trimestre, il giorno della settimana e nomi specifici dei mesi e dei giorni della settimana. Queste colonne facilitano calcoli complessi, come aggregazioni per periodi specifici ed è essenziale per i confronti temporali, permettendo analisi dettagliate come quelle anno su anno o mese su mese.

Figura 3-26 *Scherma relazionale dashboard confronto tra anni*



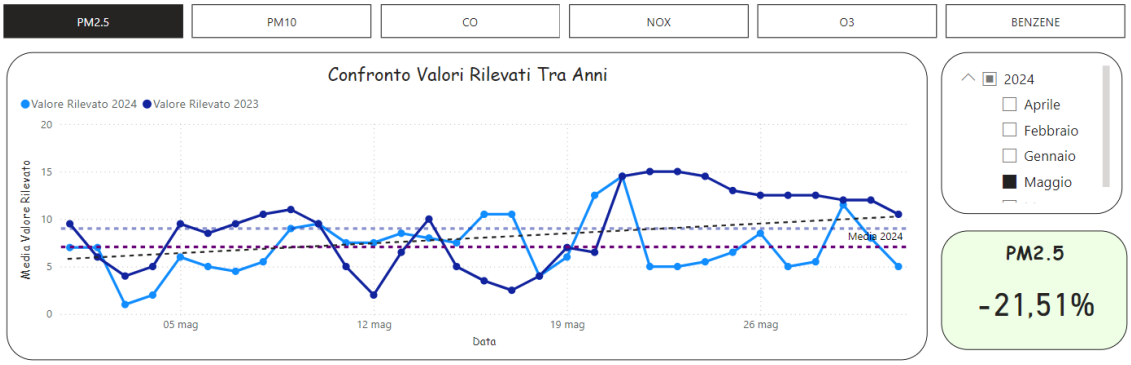
La calendar table migliora significativamente le prestazioni del modello di dati, ottimizzando le query e semplificando i calcoli per la creazione dei KPI e grafici successivi.

La Dashboard in questione è suddivisa in sei segnalibri, uno per ogni agente inquinate, al fine di alleggerire la visualizzazione evitando di aggiungere ulteriori filtri da dover selezionare. Inoltre, sono presenti:

- Un grafico a linee per visualizzare l'andamento dei valori rilevati dell'inquinante tra i due anni
- KPI che mostrano, sia numericamente che in termini percentuali, la variazione dei valori, sia per l'inquinamento che per il traffico veicolare
- Un grafico a linee per visualizzare l'andamento dei valori rilevati del traffico tra i due anni

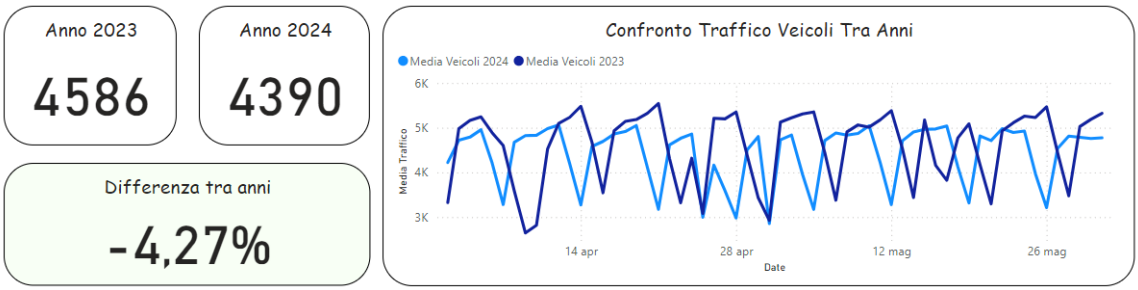
Sulla parte alta della figura 3-27 sono presenti i sei segnalibri relativi agli agenti inquinanti. Selezionando uno di questi si verrà reindirizzati alla dashboard con i grafici e KPI correlati all'agente inquinate. Subito dopo, è presente il grafico a linee che rappresenta il confronto dell'andamento del valore medio rilevato, su base giornaliera, tra i due anni. È possibile modificare il periodo visualizzato tramite il filtro che si trova nelle vicinanze del grafico. Dinamicamente, il KPI si aggiornerà con variazione percentuale dei valori medi tra i due anni. Le tre linee tratteggiate all'interno del grafico indicano rispettivamente la media del 2024, la media dei valori nel 2023 e una linea che mostra la tendenza dei valori nel 2024.

Figura 3-27 *Grafico a linee confronto tra anni valori rilevati agenti inquinanti*



A completare la dashboard troviamo (Figura 3-28), analogamente per i livelli di inquinamento, il grafico a linee che mette a confronto i dati relativi al traffico veicolare nei due anni presi in considerazione. Per rendere più chiara la comprensione sono stati aggiungi tre KPI's che mostrano rispettivamente la media del traffico nel 2023, la media del traffico nel 2024 e la variazione percentuale tra le due medie.

Figura 3-28 *Grafico a linee sul traffico veicolare per confronto tra i due anni*



La dashboard offre una visione strutturata, integrata e dettagliata delle variazioni avvenute nell'arco dei due anni. Di seguito, nella tabella, vengono riportate tutte le variazioni degli agenti inquinanti e del traffico tra il 2023 ed il 2024:

Tabella 3-2 *Variazioni percentuali traffico e inquinamento tra i due anni*

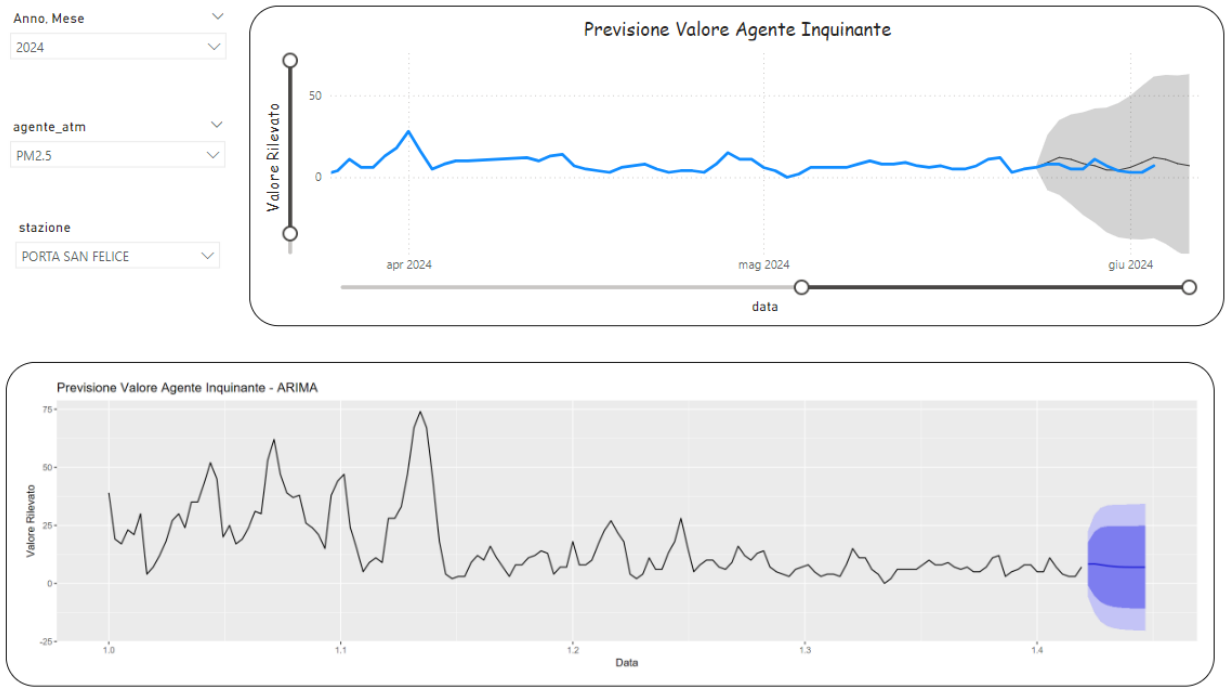
Mese	**Traffico**	**PM2,5**	**PM10**	**CO**	**NOX**	**O3**	**Benzene**
Gennaio	-2,86%	+36,12%	+33,63%	-10,60%	1,93%	+1,70%	+14,05%
Febbraio	-4,74%	+3,45%	-1,08%	-13,14%	-7%	-8,76%	-11,64%
Marzo	-7,47%	+3,25%	+11,30%	+70,08%	-8,41%	-19,85%	+9,95%
Aprile	-4,25%	+16,54%	+54,40%	-29,66%	-26,39%	+0,56%	-29,56%
Maggio	-4,56%	-21,51%	-5,08%	-53,79%	-47,98%	+3,35%	-35,10%

Questo strumento di analisi avanzata facilita la comprensione delle dinamiche legate all'inquinamento e alla viabilità urbana, consentendo di identificarne eventuali anomalie e trend emergenti.

3.3.4 *Dashboard Forecast*

Le previsioni dei valori futuri svolgono un ruolo fondamentale nell'analisi dei dati ambientali, consentendo di anticipare i trend dell'inquinamento. Nell'ultima dashboard (Figura3-29), grazie alle potenzialità di Power BI, sono stati sviluppati due grafici di previsione del valore futuro dell'inquinamento utilizzando due modelli diversi. Le definizioni, le caratteristiche e i risultati saranno trattati nella sezione 3.4.

Figura 3-29 *Dashboard previsioni inquinamento atmosferico*



Il primo grafico è un oggetto visivo realizzato tramite script R. Lo script, per esigenze di sviluppo, contiene i comandi per la realizzazione di modello *auto ARIMA*, il quale selezionerà automaticamente i parametri *P,D,Q* necessari per la creazione del modello (Figura 3-30):

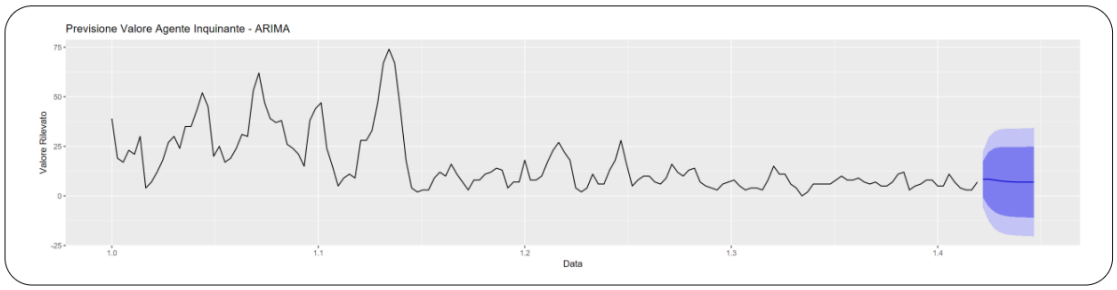
Codice 3-10 *Codice in R per la creazione del modello di previsione ARIMA*

```
library (tidyverse)
library(forecast)
library(zoo)

dataset$valore_rilevato <- na.approx(dataset$valore_rilevato)
dataset$valore_rilevato <- ts(dataset$valore_rilevato, frequency= 365)
# modello autoarima
modello <- auto.arima(dataset$valore_rilevato)
```

```
# Previsioni per i prossimi 10 passi
forecast <- forecast(modello, h = 10)
autoplot(forecast) +
  labs(
    title = "Previsione Valore Agente Inquinante - ARIMA",
    x = "Data",
    y = "Valore Rilevato"
  )
```

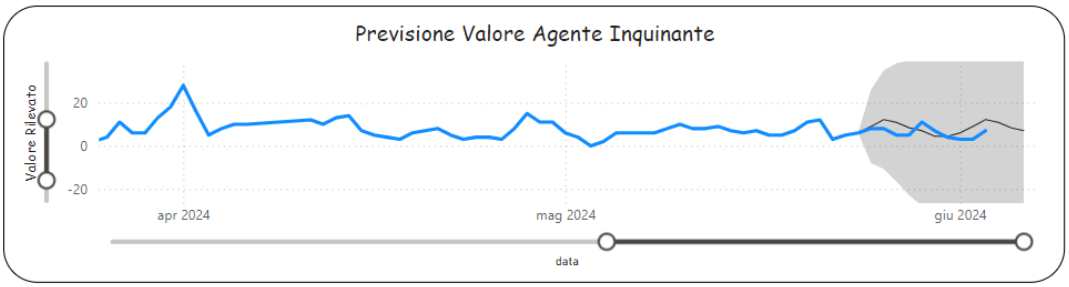
Eventuali valori nulli, dovuti da problemi di rilevazione dei dati, vengono sostituiti, tramite la funzione *'na.approx'*, con stime dei valori noti più vicini, al fine di approssimare quanto più possibile il valore mancante.

Figura 3-30 *Modello ARIMA per previsione dei valori futuri di inquinamento*



Il secondo grafico a linee (Figura 3-31), e quindi il secondo modello utilizzato, è stato creato grazie alle funzioni integrate su *Power BI*. Infatti, questo software, una volta selezionati di dati da voler mostrare, offre una serie di strumenti per l'analisi dei dati, tra i quali anche la previsione dei dati futuri tramite il modello *ETS*.

Figura 3-31 *Modello ETS per previsione dei valori futuri di inquinamento*



Entrambi i grafici si aggiornano dinamicamente alla selezione dei filtri 'Data', 'Agente Inquinante' e 'Stazione'. La sostanziale differenza tra i due risiede nella capacità di trasmettere

visivamente sia i valori storici che quelli predetti e soprattutto la possibilità di poter interagire con il grafico. Infatti, il grafico a linee creato tramite gli strumenti già implementati in Power BI (Figura 3-31), permette di poter navigare all'interno del grafico grazie alla barra di scorrimento sia verticale che orizzontale, ma soprattutto permette la visualizzazione grafica dei valori storici e quelli predetti.

In conclusione, l'obiettivo di questa dashboard, è quella di prevedere i valori di inquinamento nei giorni successivi al fine di poter prendere precauzioni in tempo e quindi tutelare la salute della popolazione.

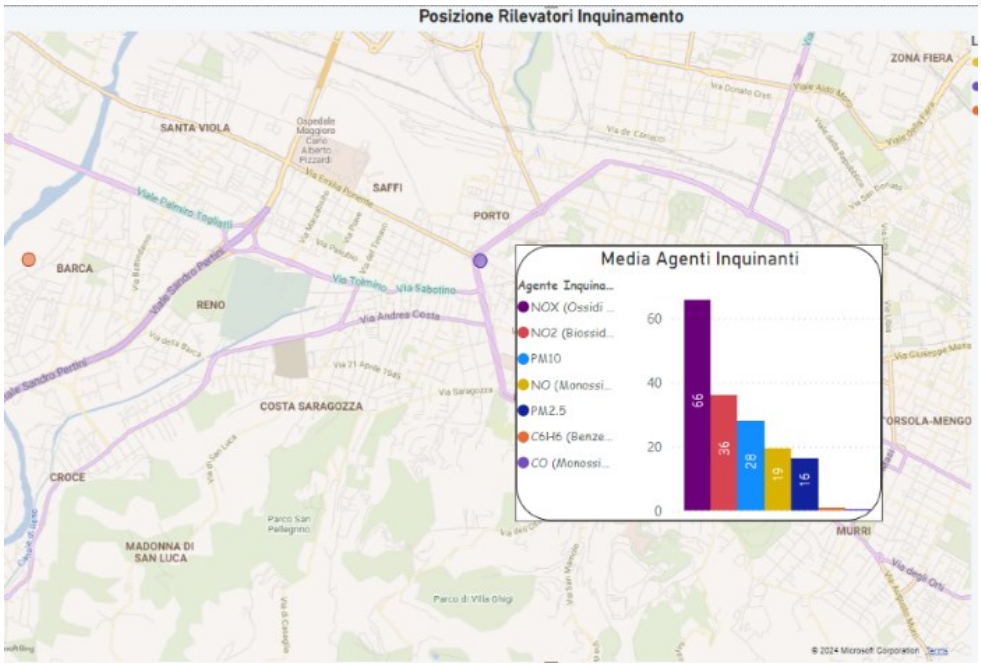
3.3.5 *Tooltips*

I tooltip in Power BI rappresentano uno strumento utile per migliorare l'interattività e l'usabilità delle visualizzazioni. I tooltip, che appaiono quando si passa il mouse su un elemento di un grafico, forniscono ulteriori informazioni che possono aiutare l'utente a comprendere meglio i dati senza dover aggiungere alla dashboard oggetti visivi che la possano appesantire.

Power BI permette la personalizzazione dei tooltip, consentendo di mostrare non solo informazioni di base come valori e categorie, ma anche dettagli aggiuntivi che possono essere cruciali per l'analisi. Questa funzionalità è particolarmente utile per fornire un contesto più ampio ai dati visualizzati, permettendo agli utenti di prendere decisioni più informate.

Per arricchire i cruscotti appena descritti sono stati creati tre tooltips differenti posizionati in altre tante dashboard diverse. Il primo tooltip impiegato è quello aggiunto nella visualizzazione grafica della posizione geografica delle stazioni di rilevamento degli agenti inquinanti. Come è possibile vedere nella Figura 3-32, passando il mouse su una bolla, che rappresenta una stazione di rilevamento, spunta un grafico a barre che riassume sinteticamente il livello degli inquinanti rilevati da quella stazione in un determinato giorno.

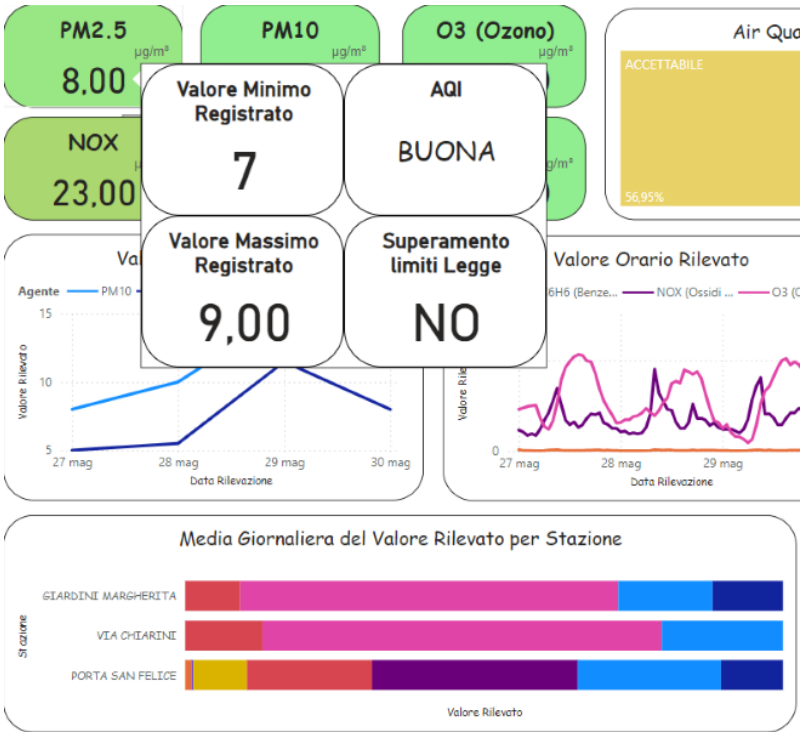
Figura 3-32 *Tooltip mappa stazioni di rilevamento*



questo tooltip si rivela fondamentale per poter desumere intuitivamente e chiaramente le zone della città maggiormente inquinate.

Il secondo tooltip (Figura 3-33), collocato sui KPI relativi ai singoli agenti inquinanti nella dashboard che descrive la situazione attuale delle rilevazioni. Posizionando il cursore su uno dei KPI, si otterranno informazioni aggiuntive sul valore minimo e massimo rilevato dell'agente inquinante all'interno della stessa giornata, il relativo Air Quality Index in forma categorica ed infine se quel determinato valore ha superato o meno i valori massimi di riferimento elencati in Tabella 1-1.

Figura 3-33 *Tooltip KPI agente inquinate*



Il terzo ed ultimo tooltip sviluppato (Figura 3-34) è quello situato nella dashboard che descrive la situazione attuale del traffico veicolare all'interno del grafico che mostra le dieci strade più trafficate della giornata (Figura 3-23).

Selezionando la strada per la quale si vogliono avere più informazioni, verrà mostrato qual è il giorno più trafficato e quello meno trafficato della settimana e infine una piccola mappa con che indica la posizione geografica della strada presa in analisi.

Figura 3-34 *Tooltip informazioni relative alla via selezionata*

In conclusione, i tooltip di Power BI sono uno strumento potente che arricchisce l'esperienza dell'utente, offrendo una comprensione ed un'esplorazione più approfondita dei dati presentati. La loro capacità di mostrare informazioni aggiuntive in modo non invasivo rende le visualizzazioni più informative e dinamiche.

Dopo la panoramica generale sullo sviluppo, implementazione e significato delle dashboard sulla situazione storica e attuale dell'inquinamento e del traffico veicolare, risulta utile sfruttare ulteriormente le potenzialità di Power BI per predire situazioni future e diminuirne in maniera marginale l'incertezza che le caratterizza.

3.4 **Previsione della qualità dell'aria**

Con l'avvento dell'evoluzione tecnologica, le realtà aziendali possono elaborare una grande mole di dati con un'accurata precisione e rapidità. Tuttavia, oltre l'elaborazione in tempo reale è necessario avere uno sguardo al futuro per poterlo anticipare proattivamente. Conosciuto comunemente come forecasting, concretizza la necessità di anticipare le tendenze di mercato e orientare tempestivamente le scelte aziendali in modo da guadagnare il famoso vantaggio competitivo limitando l'incertezza che caratterizza gli scenari futuri del dataset.
L'obiettivo principale dell'analisi predittiva è quello di identificare pattern e relazioni nei dati per restituire insights in modo da prendere decisioni tempestive. Per fare ciò, in questo paragrafo, si esplorerà l'utilizzo del forecasting in Power BI. Come precedentemente accennato, questo software, offre molteplici feature di analisi avanzata e non si limita all'analisi descrittiva dei dati ma può essere utilizzato anche per anticipare trend, eventi e comportamenti.
Tra le numerose potenzialità offerte da Power BI, una delle più rilevanti è la sua integrazione con il linguaggio di programmazione R, che consente di realizzare analisi e statistiche avanzate. In questo paragrafo analizzeremo più nello specifico i due modelli implementati nella dashboard relativa alle previsioni future del valore dell'inquinamento atmosferico: il modello ARIMA e quello ETS.

3.4.1 *ARIMA*

Utilizzando R, è possibile implementare e visualizzare modelli complessi come i modelli esponenziali e i modelli ARIMA (AutoRegressive Integrated Moving Average)[43]. Questi modelli sono fondamentali per l'analisi predittiva, poiché consentono di analizzare dati storici e prevedere tendenze future con un alto grado di precisione. Le serie storiche sono considerate come una sequenza di dati definita su un intervallo temporale, che può essere data da ore, giorni, settimane o una qualsiasi durata quantificabile. Fare predizioni di serie temporali significa analizzare una sequenza di osservazioni campionate ad intervalli regolari.
Le sue componenti principali sono tre: **l'AutoRegressione (AR), l'Integrazione (I)** e la **Media Mobile (MA).**

[43] https://otexts.com/fppit/arima.html

- ***L'AutoRegressione*** si riferisce a un modello che utilizza la dipendenza tra un'osservazione e un numero specifico di ritardi delle osservazioni precedenti. Il modello AR(p), o modello autoregressivo di ordine p, usa *'p'* lag delle osservazioni precedenti per prevedere il valore attuale. Matematicamente può essere espresso come:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t$$

dove Y_t è il valore attuale, c è una costante, ϕ_i sono i coefficienti autoregressivi e ϵ_t è il termine di errore.

- ***L'Integrazione*** rappresenta la differenziazione delle osservazioni per rendere la serie temporale stazionaria, cioè senza trend o stagionalità. Il parametro *'d'* indica il numero di differenziazioni necessarie per ottenere una serie stazionaria. Se una serie temporale non è stazionaria, può essere differenziata, una o più volte, fino a quando diventa stazionaria. La differenziazione può essere espressa come:

$$Y_t' = Y_t - Y_{t-1}$$

dove $Y\_t\text{^}'$ è la serie temporale differenziata al tempo *t*.

- La ***Media Mobile***, infine, utilizza la dipendenza tra un'osservazione e un residuo di errore proveniente da un modello di media mobile applicato ai ritardi. Il modello MA(q) può essere espresso come:

$$Yt = c + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Dove θ_i sono i coefficienti di media mobile e ϵ_t è il termine di errore.

Il modello ARIMA richiede la definizione di tre parametri: p, d e q, che rappresentano rispettivamente i gradi dell'autoregressione, il numero di differenziazioni e i gradi della media mobile.

- **p (Ordine dell'Autoregressione)**: Determina il numero di termini autoregressivi da includere nel modello. Viene determinato esaminando il grafico di autocorrelazione parziale (PACF).
- **d (Grado di Differenziazione)**: Indica il numero di differenziazioni necessarie per rendere la serie temporale stazionaria. La serie differenziata dovrebbe essere priva di tendenza e stagionalità.

- **q (Ordine della Media Mobile)**: Indica il numero di termini di media mobile da includere nel modello. Viene determinato esaminando il grafico di autocorrelazione (ACF).

Il modello ARIMA si basa su diverse fondamenta statistiche:

- ***Stazionarietà***: Una serie temporale è stazionaria se la sua media e la varianza, sono costanti nel tempo. La stazionarietà è una condizione importante per l'analisi delle serie temporali poiché molti metodi statistici assumono che i dati siano stazionari.

- ***Autocorrelazione*****:** L'autocorrelazione misura la relazione tra un'osservazione e lag delle osservazioni precedenti. La funzione di autocorrelazione (ACF) e la funzione di autocorrelazione parziale (PACF) sono strumenti fondamentali per identificare il grado di AR e MA nel modello.

- ***Rumore Bianco (White Noise)*****:** I residui di un modello ARIMA ideale dovrebbero essere 'rumore bianco', ovvero dovrebbero avere una media zero, una varianza costante e nessuna autocorrelazione significativa.

L'abilità del modello ARIMA di fondere autoregressione, differenziazione e media mobile, consente di modellare con precisione un ampio assortimento di dati temporali rendendolo uno strumento versatile e potente per l'analisi di previsioni di breve periodo.

3.4.2 *ETS*

Come già accennato in precedenza, Power BI mette a disposizione una serie di strumenti per la previsione dei valori futuri. Per il calcolo dei valori futuri viene utilizzato il modello ETS. I modelli ETS (Error, Trend, Seasonal)[44] rientrano nella famiglia di modelli noti come modelli di *Exponential Smoothing*, i quali rappresentano un importante insieme di modelli statistici utilizzati per la previsione dei valori all'interno di serie temporali (Hyndman & Khandakar, 2008). Questi modelli sono ampiamente utilizzati per la loro semplicità e capacità di adattarsi rapidamente ai cambiamenti nei dati storici in quanto hanno la capacità di sfruttare errori, tendenze e stagionalità per fornire previsioni accurate. Queste tre componenti vengono combinate insieme e ognuno di esse può essere modellata in modo additivo, moltiplicativo o nullo, dando la possibilità all'utente di poter creare numerosi modelli.
L'Exponential Smoothing applica pesi decrescenti ai dati storici dando, così, maggiore rilevanza ai dati più recenti. Ne consegue che questo metodo è utile quando i dati recenti sono considerati più rilevanti per la previsione rispetto ai dati storici. Questi tipi di modelli utilizzano un approccio iterativo per aggiornare continuamente le stime delle componenti, ad ogni nuovo punto, rendendo il modello altamente reattivo ai cambiamenti nelle serie storiche. L'equazione di base per il *Basic Exponential Smoothing* può essere espressa come:

$$S_t = \propto Y_t + (1-\propto)S_{t-1}$$

dove S_t è il valore smussato al tempo t, Y_t è il valore osservato al tempo t, e α è il parametro di smoothing che varia tra 0 e 1. Un valore di α vicino a 1 dà più peso ai valori recenti, mentre un valore vicino a 0 dà più peso ai valori passati.

Il modello *ETS* si basa su tre componenti principali:

- **Errore (E)**: Questa componente cattura il rumore o le deviazioni casuali nei dati. Può essere additiva (A) o moltiplicativa (M).

[44] https://en.wikipedia.org/wiki/Exponential_smoothing

- **Tendenza (T)**: Questa componente rappresenta la direzione a lungo termine della serie temporale. Può essere nulla (N), additiva (A), moltiplicativa (M) o additiva smorzata (Ad).

- **Stagionalità (S)**: Questa componente cattura i pattern stagionali nei dati. Può essere nulla (N), additiva (A) o moltiplicativa (M).

Solitamente queste tre componenti vengono combinate in diverse configurazioni per avere diversi modelli, ad esempio *ETS(A,A,N)* indica un modello con errore additivo, tendenza additiva e nessuna stagionalità.
L'implementazione di un modello ETS richiede un'analisi preliminare dei dati per identificare la presenza di tendenze e stagionalità. Successivamente, il modello viene affinato attraverso algoritmi di ottimizzazione che determinano i parametri di Smoothing ottimali. I parametri da impostare sono tre e includono: $\propto$ per il livello di Smoothing (quanto rapidamente il modello si adatta ai cambiamenti nei dati), β lo Smoothing relativo alla tendenza (quanto rapidamente il modello si adatta alle tendenze dei dati) e infine, γ lo Smoothing della stagionalità (quanto rapidamente il modello si adatta ai pattern stagionali). Una volta configurato, il modello *ETS* può essere utilizzato per generare previsioni future, fornendo stime che si adattano dinamicamente ai nuovi dati.
I modelli *ETS* rappresentano dei potenti strumenti per la previsione di dati nelle serie temporali, soprattutto se questi presentano forti componenti di trend e stagionalità. La loro capacità di adattarsi rapidamente ai cambiamenti nei dati li rende ideali per contesti dove le condizioni possono variare nel tempo. Nel prossimo sottoparagrafo, verranno messi a confronto i due modelli appena descritti sia dal punto di vista teorico che dal punto di vista di metriche di errore sulle previsioni dei dati dell'inquinamento atmosferico presente nell'ultima dashboard.

3.4.3 *Confronto tra modelli*

Come abbiamo osservato, i modelli ARIMA e ETS sono entrambi implementati per predire serie temporali. Tuttavia, presentano differenze sostanziali. I modelli ARIMA si basano su tre componenti principali: autoregressione (AR), integrazione (I) e media mobile (MA) e sono idonei per serie temporali stazionarie, dove le proprietà statistiche non cambiano nel tempo. Al contrario, i modelli ETS sono progettati per serie temporali con componenti di trend e stagionalità evidenti.

Un'altra differenza significativa riguarda l'approccio di questi due modelli, alla stagionalità. Mentre nei modelli ARIMA, la stagionalità viene rimossa dalla serie temporale e trattata attraverso la differenziazione stagionale, in quelli di ETS, la stagionalità è integrata direttamente come componente del modello stesso. Consentendo una modellazione più immediata e specifica delle variazioni stagionali nei dati, i modelli ETS vengono considerati più intuitivi e facili da interpretare quando si lavora con dati stagionali.
In termini di configurazione dei parametri, i modelli ARIMA richiedono la selezione degli ordini di autoregressione (p), differenziazione (d) e media mobile (q), che possono essere determinati attraverso l'analisi delle funzioni di autocorrelazione (ACF) e autocorrelazione parziale (PACF). Al contrario, i modelli ETS si basano sull'ottimizzazione dei parametri di smoothing, rendendo il processo di configurazione meno manuale e più automatizzato.

Dopo aver evidenziato teoricamente le principali differenze di questi due modelli, per comprendere al meglio i due diversi approcci, è necessario effettuare una valutazione più approfondita basata sulle metriche di performance.
La valutazione di entrambi i modelli è stata effettuata 'Out of Sample' , ovvero testando il modello su dati che non sono stati utilizzati durante la fase di addestramento del modello. Questo approccio è fondamentale per verificare la capacità del modello di generalizzare e fare previsione su dati nuovi ed indipendenti. Una volta testato il modello possiamo valutarlo sulla base di diverse metriche di performance[45]. Le più comuni, utilizzate per la valutazione di modelli sulla previsione di dati storici sono:

- *Mean Square Error* (*MSE)*
- *Mean Absolut Error* (*MAE)*

[45] https://www.diariodiunanalista.it/posts/valutazione-delle-prestazio9ni-di-un-modello-di-regressione/

- *Mean Absolute Percentage Error* (*MAPE)*
- *Root Mean Square Error* (*RMSE)*

il *Mean Square Error* o *MSE* è definita come:

$$MSE = \frac{1}{n} \sum_{i=1}^{q} (x_i - \hat{x}_i)^2$$

Dove x_i sono i valori effettivi, $\hat{x}_i$ i valori predetti dal modello e n è il numero di osservazioni. Elevando al quadrato si rendono positivi tutti i valore, dando più peso alle differenze maggiori. Questa metrica misura quanta differenza c'è, in media, tra il valore reale e quello previsto dal modello, dunque più è basso l'MSE, minore sarà la differenza media tra il valore reale e quello previsto dal modello. Maggiore sarà l'MSE, maggiore sarà la discrepanza media tra i due valori.

Il *Mean Absolut Error* o *MAE* è definito come:

$$MAE = \frac{1}{n} \sum_{i=1}^{q} |x_i - \hat{x}_i|$$

Dove x_i sono i valori effettivi, $\hat{x}_i$ i valori predetti dal modello e n è il numero di osservazioni. Simile al *MSE,* il MAE non considera la direzione dell'errore, ovvero non si avranno informazioni relativi al segno degli errori. Inoltre, è più robusto alla presenza di outliers grazie alla mancanza dell'elevazione al quadrato. Il MAE, indica quanto le previsioni del modello siano distanti dai valori reali. Analogamente all'MSE, più è basso il MAE, minore sarà la differenza media tra il valore reale e quello previsto dal modello. Maggiore sarà il MAE, maggiore sarà la discrepanza media tra i due valori.

Il *Mean Absolute Percentage Error* o *MAPE* viene definito come:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \times 100$$

Dove x_i sono i valori effettivi, $\hat{x}_i$ i valori predetti dal modello e n è il numero di osservazioni. È una misura utilizzata per indicare l'errore medio assoluto dei valori predetti espresso in percentuale rispetto ai valori reali. Risulta particolarmente utile nel confronto tra modelli,

anche utilizzando dataset diversi. Anche in questo caso, un valore basso di MAPE indica che le previsioni del modello sono, in media, più vicine ai valori reali, in termini di percentuale. Un MAPE più alto indica che le previsioni del modello sono, in media, più lontane dai valori reali, suggerendo una peggiore accuratezza del modello.

Infine, come ultima metrica di performance troviamo *Root Mean Square Error RMSE* che viene definito come:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{x}_i - x_i)^2}{n}}$$

Dove x_i sono i valori effettivi, $\hat{x}_i$ i valori predetti dal modello e n è il numero di osservazioni. L'errore quadratico medio è una misura dell'errore assoluto in cui gli errori vengono resi positivi per evitare che errori di segno opposto si annullino. Indica quanto le previsioni del modello sono in media distanti dai valori reali. Come le altre misure, un valore basso di RMSE indica che le previsioni del modello sono, in media, più vicine ai valori reali. Un RMSE più alto indica che le previsioni del modello sono, in media, più lontane dai valori reali.

Le metriche appena una comprensione dettagliata dell'accuratezza e dell'efficacia di ciascun modello. Nella tabella 3-3 sono riportati i risultati delle metriche di performance, al fine di poter effettuare un confronto tra i due:

Tabella 3-3 *Risultati metriche di performance dei due modelli*

Modello	**MAPE**	**MAE**	**MSE**	**RMSE**
ETS	70,25%	3,53	15,73	3,97
ARIMA	90,35%	6,26	42,66	6,53

Analizzando il MAPE del modello ETS risulta che, in media, le previsioni si discostano del 70,25% dai valori effettivi. Sebbene questo valore sia piuttosto elevato, suggerendo una differenza significativa tra le previsioni e i dati effettivi, altre metriche forniscono una visione più positiva delle performance del modello di ETS. Il MAE, mostra che l'errore medio assoluto tra le previsioni e i valori effettivi è di circa 3,53 unità. Questo è un risultato relativamente

basso, indicando che le previsioni del modello sono abbastanza vicine ai valori reali in termini assoluti. L'MSE, essendo una misura dell'errore quadratico medio, mette in evidenza che gli errori più grandi sono penalizzati più severamente. Infine, l'RMSE fornisce un'interpretazione dell'errore in scala di misura dei dati originali, confermando la ragionevole accuratezza del modello. D'altra parte, il modello ARIMA presenta un MAPE del 90,35%, che è significativamente più alto rispetto a quello di ETS. Questo valore suggerisce una bassa accuratezza del modello. Il MAE di 6,26 suggerisce che, in termini assoluti, le previsioni del modello ARIMA discostano di 6,26 unità rispetto al valore reale. L'MSE per ARIMA è 42,66, indicando un'alta varianza degli errori. Questo è ulteriormente confermato dall'RMSE di 6,53, che rappresenta un errore medio più elevato in scala con le unità dei dati originali.

Nel complesso, il modello ETS sembra offrire previsioni più accurate rispetto al modello ARIMA, come confermato anche da valori inferiori di MAE, MSE e RMSE. Tuttavia, entrambi i modelli hanno fatto registrare un valore del MAPE elevato, questo ci suggerisce che entrambi i modelli faticano nel prevedere valori affidabili. In ogni caso il modello ETS riesce comunque a mantenere un margine di errore delle previsioni più basso rispetto al modello ARIMA. In conclusione, sebbene nessuno dei due modelli sia perfetto, ETS sembra essere il modello più affidabile per le previsioni basate su queste specifiche metriche di valutazione, probabilmente perché il modello ETS risulta più adatto a catturare le componenti di errore, stagionalità e trend nei dati.

DISCUSSIONE

In un'era in cui i dati rappresentano una risorsa fondamentale, la capacità di raccogliere, trasformare e caricare efficientemente ed efficacemente le informazioni è cruciale per il successo e la competitività di qualunque realtà aziendale. Le pipeline di ETL sono essenziali per gestire flussi di dati che alimentano i sistemi di Business Intelligence e analisi aziendali. Nel caso di questa tesi, la pipeline ETL, in combinazione con il software Power BI, ha permesso di ottenere una visione olistica sulla situazione generale dei livelli di inquinamento nella città metropolitana di Bologna, attraverso grafici e KPIs che ne semplificano la visualizzazione. Infatti, ci hanno permesso di estrarre facilmente e dinamicamente i dati che rappresentano la variazione su base mensile della mole del traffico veicolare e dei livelli delle sostanze inquinanti rilevate nell'aria, tra i primi cinque mesi dell'anno 2024 e l'intero anno precedente. Prendendo in considerazione unicamente gli agenti inquinanti PM2.5 e PM10, emessi in quantità maggiore dal traffico di veicoli a combustione, si può facilmente osservare che a fronte di una costante diminuzione del traffico veicolare su base mensile, si sono registrati incrementi dei livelli delle due sostanze nell'aria anche nell'ordine delle due cifre. A fare eccezione ci pensa il mese di maggio in cui si è finalmente registrato una sostanziale diminuzione dei livelli di entrambi gli agenti inquinanti. Questo risultato può essere preso come spunto di riflessione per poter analizzare i fattori che possono aver influito sui livelli di inquinamento. Sebbene non siano stati presi in considerazione i dati relativi alle temperature e alle condizioni meteo relativi ai primi cinque mesi del 2024, è possibile includere un terzo fattore esogeno nell'analisi: l'introduzione del limite di velocità di 30 chilometri orari in tutta la città di Bologna, fatta eccezione di poche vie in cui il limite è rimasto invariato ai 50 km/h. È innegabile che, limiti di velocità così stringenti, se da un lato garantiscono la sicurezza stradale riducendo la probabilità di incidenti e disincentivano l'utilizzo di mezzi propri per i piccoli spostamenti urbani, dall'altro, aumentano inevitabilmente i tempi di percorrenza di tutti i veicoli che transitano nelle strade urbane di Bologna. Tuttavia, nonostante la diminuzione del traffico registrata, possiamo affermare che l'incremento delle tempistiche di percorrenza potrebbe aver contribuito ad un'impennata delle emissioni di due dei principali agenti inquinanti presi in considerazione in questa trattazione, PM2.5 e PM10. Inoltre, l'analisi dei

dati raccolte dalle rilevazioni del traffico veicolare ha fornito informazioni preziose riguardanti la ciclicità e le abitudini di movimento dei veicoli all'interno della città di Bologna. Questo studio ha rilevato che i flussi di traffico presentano picchi ricorrenti durante le fasce orarie 9 – 11 e 17 – 20 , orari che coincidono esattamente con i movimenti di studenti e lavoratori verso le scuole, le università e i luoghi di lavoro al mattino e con il loro ritorno a casa la sera. Queste abitudini di movimento si ripetono in modo costante durante tutti i giorni della settimana lavorativa, dal lunedì al venerdì, evidenziando la forte correlazione tra il traffico veicolare e le attività quotidiane della popolazione.

Durante i fine settimana e i periodi festivi, si osserva una riduzione notevole nel numero di veicoli transitati, indicando un cambiamento significativo nelle abitudini di movimento dei cittadini. In questi periodi, il traffico tende a concentrarsi in altre fasce orarie, spesso legate ad attività ricreative e sociali.

In sintesi, l'analisi dei dati raccolti e le osservazioni fatte suggeriscono che la riduzione del traffico veicolare non è condizione necessaria ma soprattutto sufficiente affinché il livello di inquinamento nell'aria, diminuisca. Si ritiene quindi necessario un approccio olistico che tenga conto di variabili multiple e delle interazioni tra di esse. L'introduzione di limiti di velocità più bassi, pur avendo effetti positivi sulla sicurezza stradale, potrebbe richiedere misure complementari per evitare effetti collaterali indesiderati come l'aumento delle emissioni. Inoltre, le informazioni relative ai veicoli sono fondamentali per la pianificazione urbana e la gestione del traffico, poiché permettono di identificare le ore di punta e le aree più congestionate, consentendo alle autorità di implementare misure mirate per migliorare la fluidità del traffico e ridurre l'inquinamento atmosferico.

Nonostante i risultati ottenuti, l'analisi presenta alcune limitazioni. In primo luogo, come già accennato, non sono stati presi in considerazione i dati che includevano variabili climatiche come la temperatura o le condizioni meteorologiche, che possono influenzare significativamente i livelli di inquinamento presenti nell'aria. Inoltre, non è stato possibile distinguere le varie tipologie di veicoli transitanti sulle strade e soprattutto il tipo di carburante (incluso l'elettrico) utilizzato. Queste limitazioni possono essere prese come spunto per ampliare e aumentare l'accuratezza dell'analisi proposta in questa tesi.

In conclusione, questa riflessione riflette la necessità di una pianificazione più integrata delle politiche ambientali e dei trasporti, basata su un'analisi dati robusta e su un monitoraggio continuo delle misure implementate. Solo attraverso un approccio coordinato e multidisciplinare sarà possibile ottenere miglioramenti significativi e duraturi in termini di qualità, dell'aria e di vita.

CONCLUSIONI

Lo scopo di questa tesi è stato quello di sviluppare e implementare una pipeline ETL per l'analisi dei dati relativi al traffico veicolare e ai livelli di inquinamento atmosferico rilevati da diverse stazioni dislocate nella città di Bologna. Per individuare le possibili variazioni di entrambi i fenomeni è stato scelto l'arco temporale dell'anno corrente, 2024 e del precedente, 2023. Le motivazioni di questa scelta temporale sono state dettate dall'intenzione di rimuovere definitivamente qualunque effetto delle restrizioni introdotte per la pandemia da Covid-19 ancora in vigore nel 2022, e di evidenziare gli effetti e la possibile correlazione di "Zona 30", una normativa introdotta in quasi tutta la città di Bologna a tutela della sicurezza stradale e i livelli degli inquinanti atmosferici.

La fase iniziale del progetto ha riguardato la progettazione e implementazione della pipeline ETL, coinvolgendo l'estrazione dei dati da diverse fonti tramite chiamate API fornite direttamente dal sito web Open Data Bologna, la loro trasformazione in un formato omogeneo, grazie agli script appositamente creati tramite Python, il successivo caricamento all'interno del database MongoDB e infine la loro analisi attraverso il software di Business Intelligence, Power BI. Proprio grazie a questo software, è stata possibile la creazione di dashboard sulle condizioni attuali dei livelli di inquinamento, la visualizzazione su mappa della dislocazione delle stazioni di rilevamento degli agenti inquinanti e dei sensori di rilevamento del traffico, la dashboard sull'andamento e le condizioni di viabilità del traffico e infine la dashboard nella quale sono stati implementati due diversi modelli di previsioni per stimare i futuri livelli di inquinamento atmosferico.

Il progetto realizzato in questa tesi potrebbe avere diverse declinazioni e quindi numerose sono le direzioni in cui poterlo estendere. La prima direzione potrebbe essere quella di un sito web. La creazione apposita di un sito internet permetterebbe la pubblicazione delle Dashboard sviluppate, in modo che i cittadini possano informarsi autonomamente sullo stato di salute dell'aria. Un'altra via percorribile è quella di un ampliamento dello studio integrando i dati relativi alla qualità dell'aria con quelli delle condizioni meteorologiche in tempo reale. L'aggiunta di questa variabile consentirebbe di analizzare l'impatto che le condizioni climatiche hanno sui livelli di inquinamento atmosferico.

Fenomeni come pioggia, temperatura, vento o umidità, influenzano la concentrazione degli agenti inquinanti e aiuterebbero a sviluppare modelli predittivi sempre più affinati e accurati. Infine, anche l'integrazione delle informazioni relative alla direzione della percorrenza ai dati sul traffico giornaliero sulle strade di Bologna potrebbe aiutare nel modulare e ottimizzare la percorrenza urbana.

In conclusione, questa tesi ha dimostrato l'importanza e l'efficacia di una pipeline ETL ben progettata per l'analisi dei dati di traffico e inquinamento. Sebbene ci siano delle limitazioni, le potenziali integrazioni proposte possono aiutare a superarle, fornendo strumenti ancora più potenti per migliorare la qualità dell'aria e la gestione del traffico urbano.

BIBLIOGRAFIA

A. Ferrari, M. R., 2015. *The Definitive Guide to DAX: Business Intelligence with.* s.l.:Microsoft Press.

A. Ferrari, M. R., 2016. *Introducing Microsoft Power BI.* s.l.:Microsoft Press.

El-Sappagh, S. H. A., 2011. A proposed model for data warehouse ETL processes. *Journal of King Saud university - Computer and Information Sciences,* 23(2), pp. 91-104.

F. Marinuzzi, M. L., 2016. *Basi di dati e Big Data: come estrarre valore dai propri.* s.l.:Youcanprint.

Hyndman, R. J. & Khandakar, Y., 2008. Automatic Time Series Forecasting : The forecast Package for R. *Journal of Statistical Software,* 27(3), p. 22.

McNeil, A. J., 2007. Review of Latent Curve Models: A Structural Equation Approach. *Journal of the American Statistical Association,* 102(480), pp. 1479-1481.

Powell, B., 2017. *Microsoft Power BI Cookbook: Creating Business Intelligence Solutions of.* s.l.:Packt Publishing Ltd.

Qaiser, A. a. F. M. a. M. S. M. N. a. A. N., 2023. Comparative Analysis of ETL Tools in Big Data Analytics. *Pakistan Journal of Engineering and Technology,* 6(1), pp. 7-12.

Rezzani, A., 2017. *Big Data Analytics. Il manuale del data scientist.* s.l.:Apogeo Education.

S. Machiraju, S. G., 2018. *Power BI Data Analysis and Visualization.* s.l.:De Gruyter.

Storey, H. C. a. R. H. L. C. a. V. C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly,* 36(4), pp. 1165-1188.

Vassiliadis, P. a. S. A. a. S. S., 2002. *Conceptual modeling for ETL processes.* II a cura di New York, NY, USA: Association for Computing Machinery.