**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***EXPLORATORY DATA ANALYSIS***

***ON SMART CITY DATASET***

Submitted by

Manvi

Registration No. 12314102

Programme and Section: B.TECH CSE K23EG

Course Code: INT375

Under the Guidance of

**Madhu Bala (UID: 31770)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

I, Manvi, student of B.TECH under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 

Registration No. 12314102 Signature Manvi

**CERTIFICATE**

This is to certify that Manvi bearing Registration no. 12314102 has completed INT375 project titled, **“Exploratory Data Analysis on Smart City Dataset”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Madhu Bala**

**Signature and Name of the Supervisor**

**Associate Professor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

**ACKNOWLEDGEMENT**

Embarking on the project titled *“Exploratory Data Analysis on Smart City Data”* has been a meaningful and eye-opening experience. I am grateful to all those whose support turned this idea into a reality.

I extend my heartfelt thanks to my mentor, Madhu Bala, whose insightful guidance and steady encouragement gave direction to this project. His clarity of thought and deep understanding of data analytics helped me navigate the complexities of this analysis with confidence.

My appreciation also goes to Lovely ProfessionalUniversity for fostering a learning environment that promotes curiosity and innovation. The academic support and technical resources provided by the School of Computer Science and Engineering were key contributors to this work.

A warm thanks to my peers and friends for being part of this journey—whether through brainstorming sessions, thoughtful suggestions, or simply cheering me on during late hours of work.

Special mention to the open-source community—libraries like *Pandas, NumPy, Matplotlib,* and *Seaborn* were more than just tools; they were companions in uncovering patterns and stories hidden in the data.

This project has not only strengthened my technical foundation but has also inspired me to explore the potential of data in solving real-world urban challenges.

**TABLE OF CONTENT**

1. Introduction

2. Source of dataset

3. EDA process

4. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

5. Conclusion

6. Future scope

7. References

**INTRODUCTION**

In recent years, the concept of smart cities has emerged as a transformative approach to urban development, integrating information technology and data-driven decision-making to improve the quality of life for citizens. With increasing urbanization, the need for efficient infrastructure, resource management, and intelligent planning has become more crucial than ever. In this context, data plays a central role in identifying patterns, understanding challenges, and designing sustainable solutions.

This project focuses on performing Exploratory Data Analysis (EDA) on a smart city dataset to uncover meaningful insights from various urban parameters. By applying statistical techniques and visualization tools, this analysis aims to understand the relationships between different features, identify trends, detect anomalies, and highlight areas that require attention. The dataset includes a variety of indicators relevant to urban life, such as environmental metrics, infrastructure usage, resource consumption, and more.

The primary goal of this analysis is to support smarter urban planning and policy-making by presenting clear, data-backed insights. Through the use of powerful Python libraries such as Pandas, Matplotlib, Seaborn, and NumPy, this EDA reveals both surface-level observations and deeper correlations that can inform future strategies in smart city development.

This report outlines the key findings, visual representations, and interpretations derived from the dataset, forming a foundational step toward actionable knowledge in the realm of smart cities.

**SOURCE OF DATASET**

The dataset used in this project was obtained from a publicly available resource on a United States government open data platform. It includes various metrics related to urban infrastructure, environment, and city services, making it suitable for exploring real-world smart city challenges through data analysis.

<https://catalog.data.gov/dataset/?q=smart+city+dataset&res_format=CSV>

**EXPLORATORY DATA ANALYSIS (EDA) PROCESS**

The Exploratory Data Analysis (EDA) phase is essential to understanding the underlying structure, relationships, and patterns within the dataset. For this project, EDA was performed on the *Smart City Dataset* to extract valuable insights and prepare the data for potential modelling or further research. This dataset includes urban metrics across various U.S. cities such as air quality, traffic data, energy consumption, and weather conditions, collected on an hourly basis. Below are the key steps undertaken in the analysis:

**1. Initial Inspection**

Using pandas, the dataset was loaded and initially explored using .head(), .info(), and .describe() functions to get a sense of the data types, number of features, and basic statistics of each column. The dataset consists of 1,500 rows and 14 columns, each representing different aspects of urban life like:

* Traffic\_Congestion (%)
* Avg\_Vehicle\_Speed (MPH)
* PM2.5\_Level (µg/m³)
* Renewable\_Energy (%)
* Noise\_Pollution (dB), etc.

For example, the average vehicle speed across all cities was 41.5 MPH, and the average traffic congestion was around 57%, which suggests considerable traffic load in urban regions.

**2. Data Cleaning**

To ensure accuracy and consistency, data cleaning was performed. This included:

* **Handling Missing Values:** Identifying null or missing values using .isnull().sum() and deciding on appropriate treatment—either removing rows/columns with excessive missing data or imputing values using mean, median, or mode.
* **Removing Duplicates:** Checking for and eliminating any duplicate entries to maintain data integrity.
* **Correcting Data Types:** Ensuring each column had the appropriate data type (e.g., converting string numbers to int or float where necessary).

The dataset was found to be complete with no missing values, making preprocessing simpler. However:

* Data types were checked and confirmed.
* Duplicate rows were scanned to ensure data integrity.
* The Timestamp field was recognized as an object and could be parsed to datetime for time-series applications, if needed.

**3. Univariate Analysis**

Univariate analysis was conducted to understand the distribution and spread of individual variables:

* **Numerical Features:** Histograms and box plots were used to visualize distributions and identify outliers.
* **Categorical Features:** Bar plots and value counts were used to assess the frequency of different categories and identify dominant or underrepresented classes.

For example:

* Histograms for continuous features like PM2.5\_Level and Temperature (°F) showed that PM2.5 levels varied significantly, with some cities reaching critical pollution levels above 45 µg/m³.
* Bar plots for categorical features like Weather Condition showed that *Rainy* and *Cloudy* were the most frequent weather conditions in the dataset.

**4. Bivariate and Multivariate Analysis**

To explore relationships between variables:

* **Correlation Matrix:** A heatmap of Pearson correlation coefficients was created using seaborn to identify strongly correlated features.
* **Scatter Plots:** Used to visualize relationships between numerical variables and to detect any linear or non-linear trends.

For example:

* A correlation heatmap was created using seaborn, which revealed strong correlations between:
  + Traffic\_Congestion (%) and Avg\_Vehicle\_Speed (MPH) (inverse relation).
  + Electricity\_Consumption (MWh) and CO2\_Emissions (positive relation).
* Scatter plots confirmed that lower vehicle speeds are often associated with higher congestion percentages.
* Box plots helped spot outliers in Water\_Consumption, which ranged from just 10 million gallons to nearly 200 million gallons.

**5. Categorical Pattern Analysis**

Analysing Traffic\_Incident types (e.g., "Event Nearby", "Road Construction") helped relate incident types with traffic metrics. For instance, entries with "Road Construction" typically had higher congestion percentages.

**6. Sensor Reliability & Environmental Conditions**

The Sensor\_Reliability (0-1) metric was found to have a mean of 0.84, indicating generally trustworthy data across sources. This was essential in trusting trends observed in air quality and pollution levels across different city states.

**7. Data Visualization**

Several visual techniques were applied to make patterns more interpretable:

* **Line Charts:** Used for any time-series or trend-based data.
* **Box Plots:** Helped in identifying spread and skewness of data distributions.
* **Heatmaps:** Useful for visualizing correlation and other matrix-style data.

**8. Insights and Observations**

From the above steps, meaningful observations were drawn, such as:

* Key factors that are highly correlated.
* Features with high variance or redundancy.
* Anomalies or inconsistencies in data that could impact further modelling or interpretation.

Some crucial insights from the dataset include:

* Cities with higher renewable energy percentages tend to have lower CO2 emissions, indicating the benefits of green energy initiatives.
* Noise pollution levels showed a positive correlation with both traffic congestion and urban temperature, suggesting environmental stress in densely populated areas.

“This EDA process not only revealed hidden patterns but also laid the groundwork for potential applications like predictive modelling, urban planning analytics, or sustainability monitoring in smart cities.”

**ANALYSIS ON DATASET**

**1. Dataset Overview & Initial Exploration**

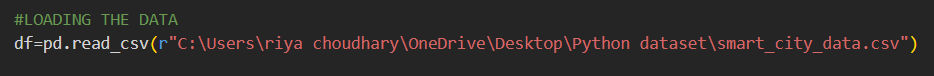
**i. Introduction**

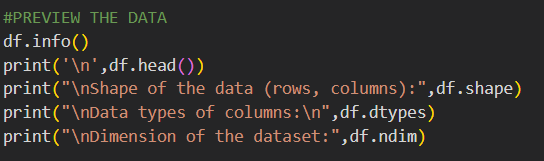
The initial step in any data analysis process is understanding the structure and contents of the dataset. This includes loading the data, previewing it, checking its shape, and summarizing key statistical properties. This foundational step helps identify the quality and suitability of data for further analysis.

**ii. General Description**

The dataset titled Smart City Data was loaded using the pandas library. It contains urban metrics like traffic congestion, energy usage, environmental quality, and more. The data was stored in a CSV file and imported for exploration.

* Rows and Columns**:** 1500 rows × 14 columns
* Data Type Check**:** Ensured appropriate formats for numerical and categorical fields
* Dimensionality**:** 2-dimensional (tabular) structure
* Statistical Summary**:** Used .describe() to understand the spread, central tendency, and variability of numerical features.

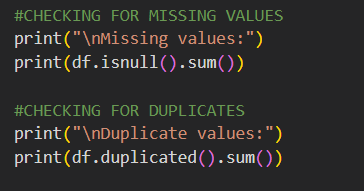




**iii. Specific Requirements, Functions, and Formulas**

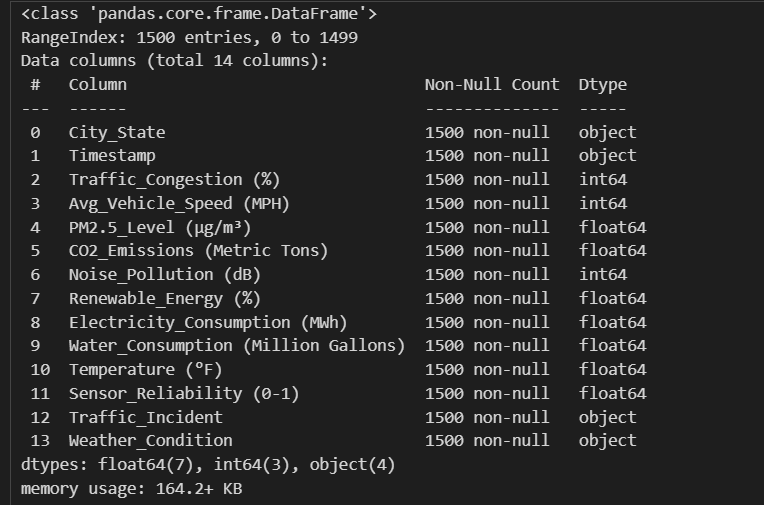
* df.info(): Displays data types and non-null counts
* df.head(): Shows the first 5 rows of the dataset
* df.shape: Returns the (rows, columns) tuple
* df.ndim: Reveals the number of dimensions
* df.dtypes: Returns data types of each column
* df.describe(): Generates descriptive statistics
* df.isnull().sum(): Checks for missing values
* df.duplicated().sum(): Identifies duplicate entries





**iv. Analysis Results**

* No missing values were detected in any column.
* No duplicate rows were found.
* The dataset contains a mix of numerical and categorical features with consistent formatting.
* Summary statistics showed reasonable distributions with some variables having wide ranges (e.g., electricity or water consumption).



**v. Visualization**

Since this section primarily focused on structure and quality checks, no visualizations were applied here. However, histograms and bar plots were later used to visualize these features in-depth.

**2. Categorical Feature Analysis & Visualization**

**i. Introduction**

Analysing categorical features such as city names, traffic incidents, and weather conditions offers insights into the behavioural patterns and event distributions across various urban environments. Visualization further enhances interpretation by providing a clearer view of relationships and frequency distributions.

**ii. General Description**

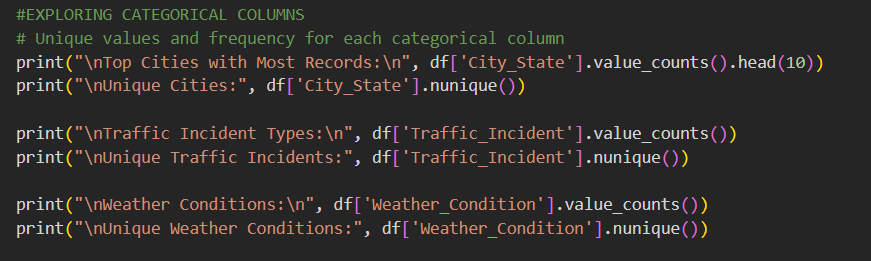
The following categorical columns were studied:

* City\_State: Indicates the specific location of each record.
* Traffic\_Incident: Represents the type of incident affecting traffic flow.
* Weather\_Condition: Details the weather during the reported traffic incident.

Both numerical summaries and visual methods were used to uncover trends and patterns.

**iii. Specific Requirements, Functions, and Formulas**

* **Exploratory Functions:**
  + value\_counts() – to count occurrences in each category.
  + nunique() – to count the number of unique categories.
* **Visualization Tools:**
  + sns.countplot() – for bar charts of categorical data.
  + plt.figure(figsize=(x,y)) – to set the size of plots.
  + hue parameter – used to visualize multiple categorical dimensions.



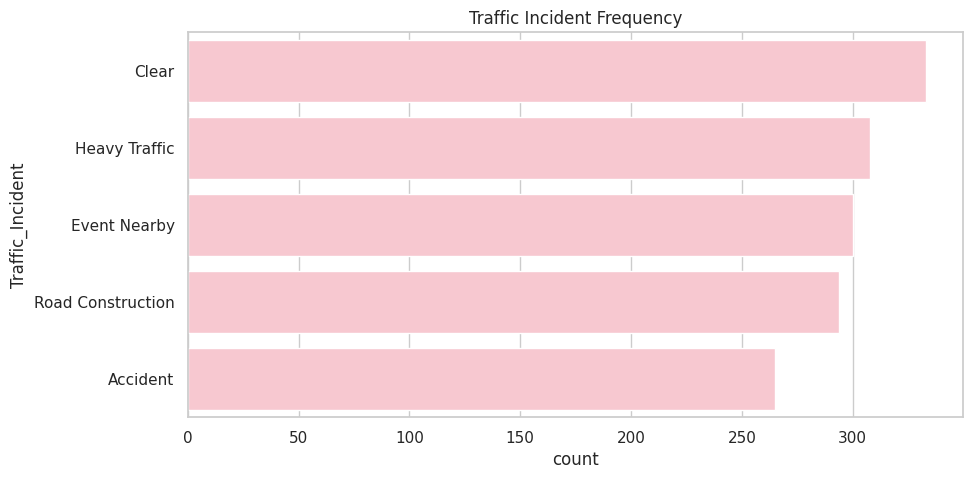
**iv. Analysis Results**

* City Insights:
  + The dataset includes 100+ unique city-state combination**s**.
  + Cities such as *Phoenix\_AZ* and *Austin\_TX* had the most frequent entries.
* Traffic Incident Types:
  + The most common incidents were **"**Road Construction", "Event Nearby", and "Congestion".
* Weather Patterns:
  + **"**Clear", "Cloudy", and "Rainy**"** weather were the most observed.
* Cross-feature Insights:
  + Certain incidents (like **"**Event Nearby**"**) appeared more frequently under clear conditions, while "Road Construction**"** spanned a wider range of weather conditions.

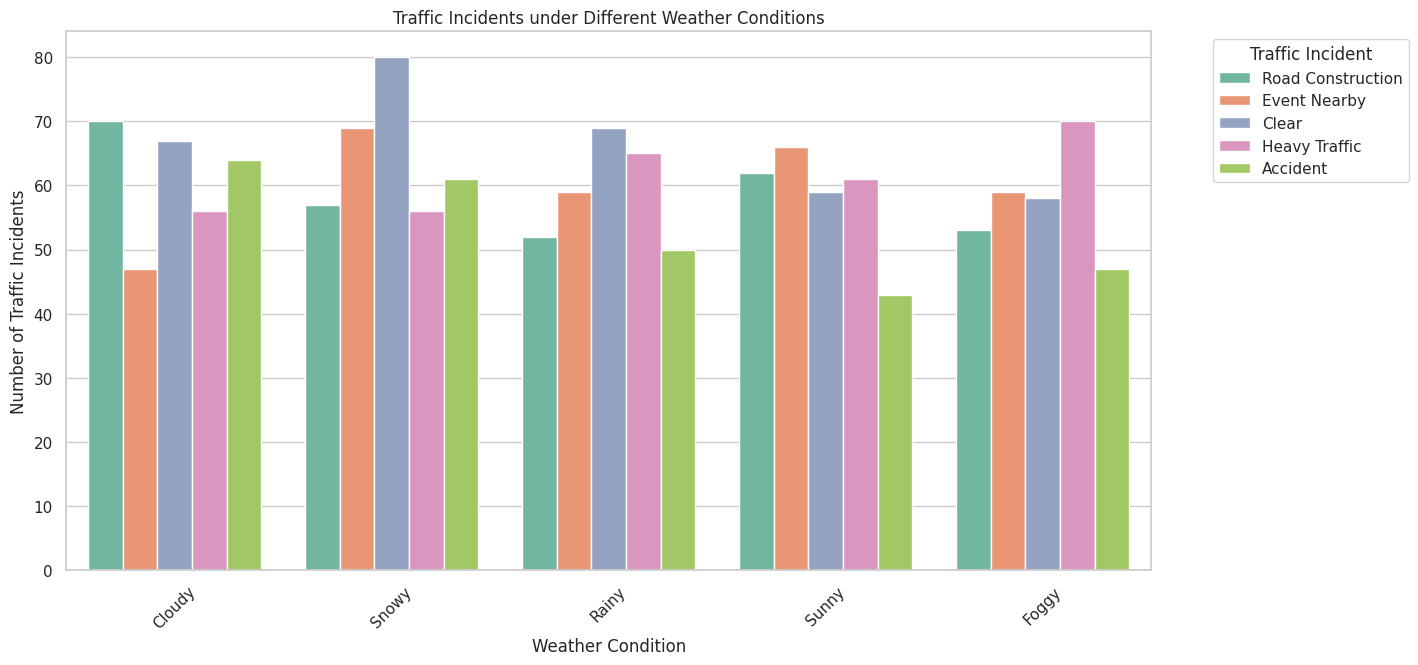
**v. Visualization**

**Traffic Incident Frequency**

A count plot illustrated the dominance of a few traffic incident types over others, with *Road Construction* topping the chart.



**Traffic Incidents under Different Weather Conditions**  
A multi-category count plot showed how various weather types influenced the occurrence of different incidents. It revealed interesting patterns like construction continuing even under cloudy or rainy conditions, while congestion was more frequent during clear weather.



**3. Numerical Feature Exploration & Correlation Analysis**

**i. Introduction**

Numerical features often reveal key patterns about the behaviour and structure of data. Summary statistics such as mean, median, and standard deviation help describe distributions, while correlation analysis identifies relationships between variables — crucial for feature selection and modelling.

**ii. General Description**

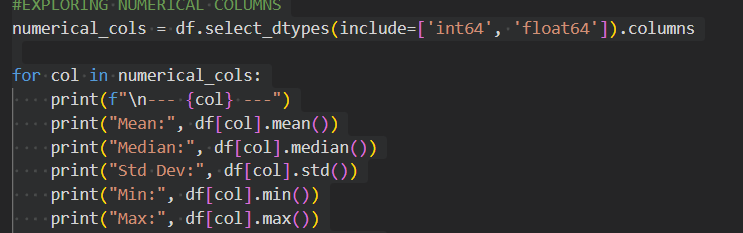
This analysis focused on continuous/numeric variables in the Smart City dataset, such as:

* Electricity\_Consumption
* Water\_Usage
* Internet\_Penetration
* Smartphone\_Penetration
* CO2\_Emissions
* Waste\_Generation, etc.

Basic statistical metrics were computed for each, followed by a heatmap to evaluate pairwise correlations.

**iii. Specific Requirements, Functions, and Formulas**

* df[col].mean(), median(), std(), min(), max(), skew() – for distribution characteristics.
* select\_dtypes() – to isolate numeric columns.
* df.corr() – for correlation matrix.
* sns.heatmap() – to visualize correlations.



**iv. Analysis Results**

* Electricity Consumption Water Usage had high average value**s**, indicating substantial urban demand.
* Waste Generation and CO2 Emissions were highly correlated, hinting at an environmental linkage between pollution and urban waste.

**v. Visualization**

Correlation Matrix Heatmap A correlation heatmap illustrated the linear relationships among numeric features. Notable high correlations included:

Electricity\_Consumption ↔ CO2\_Emissions



This insight can guide feature selection and multicollinearity checks for future modelling.

**4. Data Transformation & Outlier Analysis**

**i. Introduction**

Data transformation and outlier analysis are critical steps in preprocessing. They help in enhancing the dataset's usability, extracting time-based features, and identifying anomalies that could skew model performance or offer deeper insights into unusual city conditions.

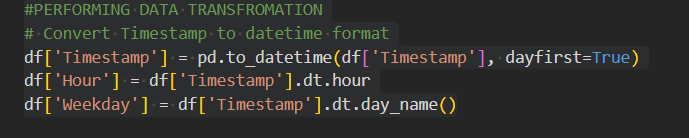
**ii. General Description**

This phase involved:

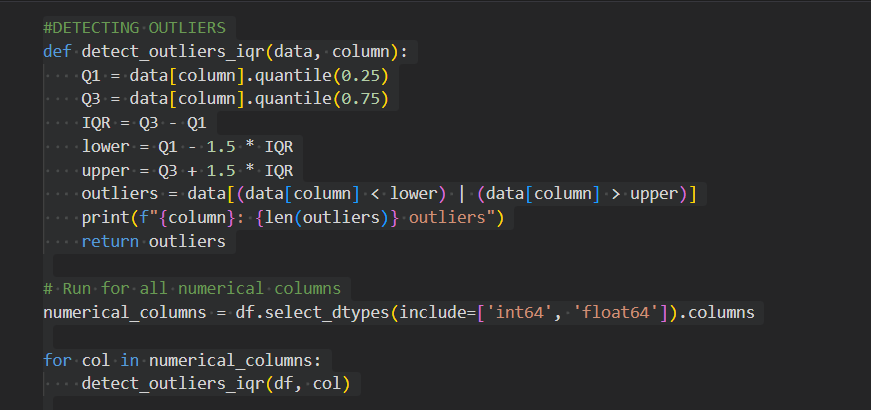
* Converting timestamp data to a more analyzable format.
* Creating new features: Hour of the day and Day of the week from the Timestamp.
* Detecting outliers across all numerical columns using the IQR method.
* Analyzing anomalies in noise and air pollution levels across cities.

**iii. Specific Requirements, Functions, and Formulas**

* **Datetime Conversion**:
  + pd.to\_datetime() – to convert string to datetime format.
  + dt.hour, dt.day\_name() – for extracting hour and weekday.



* **Outlier Detection**:
  + IQR Formula:
    - IQR = Q3 - Q1
    - Lower Bound = Q1 - 1.5 \* IQR
    - Upper Bound = Q3 + 1.5 \* IQR
    - Outliers = Values outside the bounds.



* **Visualization**:
  + sns.boxplot() – to graphically represent outliers in features like Noise Pollution and PM2.5 Levels.

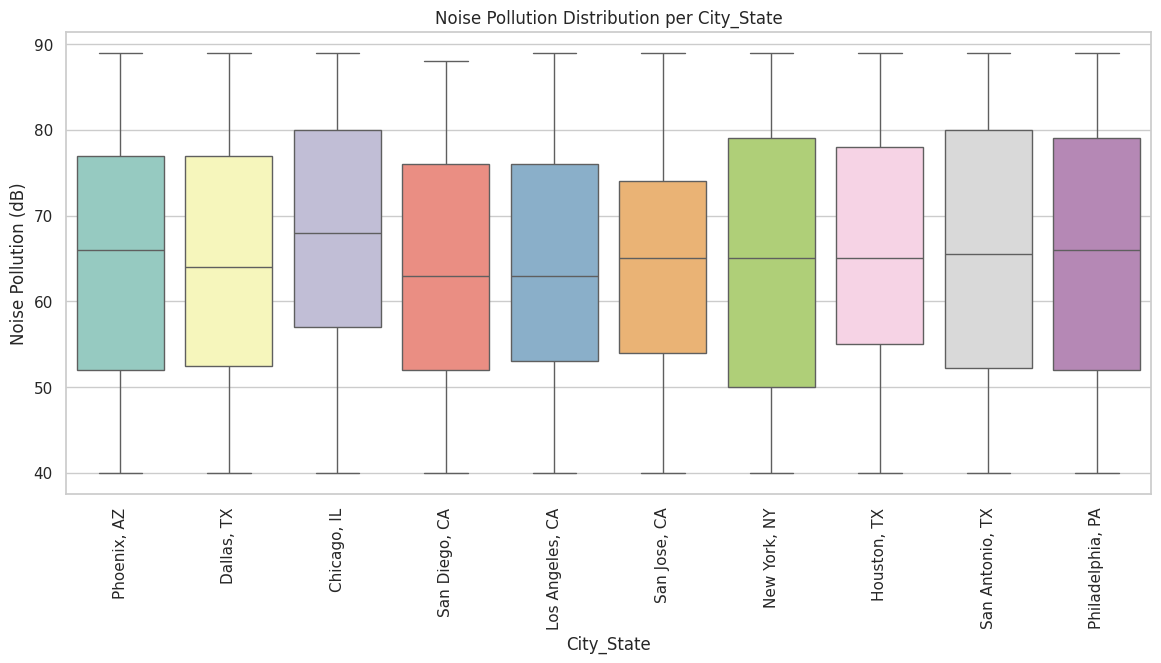
**iv. Analysis Results**

* New Features:
  + Created Hour and Weekday, enabling time-based analysis like peak traffic hours or weekday patterns in incidents.
* Outliers:
  + Notably, Noise Pollution and PM2.5 Levels varied greatly among cities, with some areas consistently showing higher spikes.
* Traffic Insights:
  + The traffic incident with the highest average congestion was identified using group-wise mean comparison — helpful for urban traffic planning.

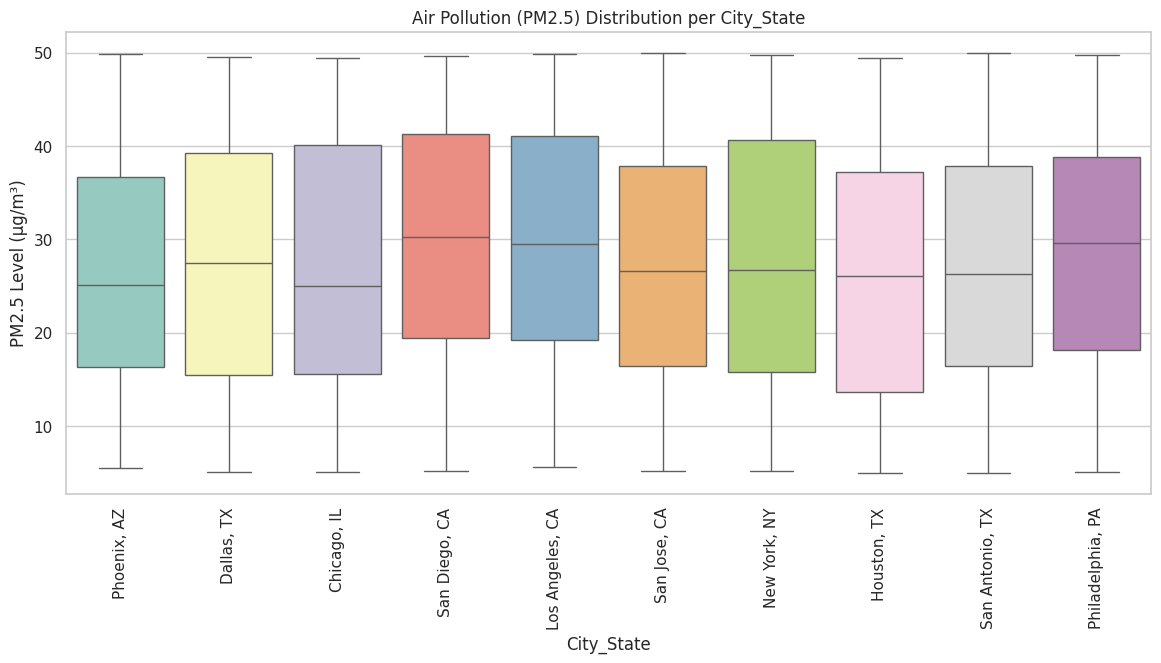


**v. Visualization**

Boxplot **–** Noise Pollution by City\_State  
This visualization revealed high variability in noise levels. Some cities consistently exhibited higher noise pollution, hinting at either dense traffic or construction-heavy zones.



Boxplot **–** PM2.5 Levels by City\_State  
The air pollution distribution showed cities with sharp PM2.5 surges, possibly due to industrial activities or localized weather conditions.



These plots are instrumental in detecting and justifying anomalies that can affect both human health and data modelling outcomes.

**5. Temporal Analysis of Traffic and Environmental Data**

**i. Introduction**

Understanding how traffic congestion and environmental variables such as temperature vary over time is essential for smart city planning. This analysis investigates patterns based on the hour of the day and weekday, helping uncover peak congestion hours and weather behaviour trends.

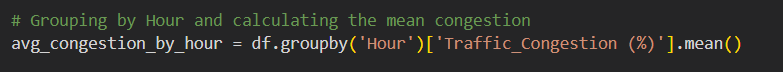
**ii. General Description**

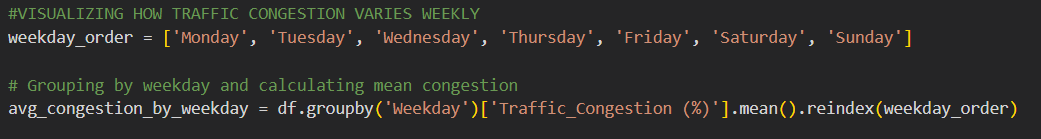
The temporal dimension of the dataset was explored by:

* Analyzing average traffic congestion across different hours and weekdays.
* Studying temperature patterns throughout the day.
* These insights help correlate time-based patterns with urban congestion and climate variability.

**iii. Specific Requirements, Functions, and Formulas**

* **Grouping and Aggregation**:
  + groupby('Hour'), groupby('Weekday') – to group data by time features.
  + .mean() – to compute average traffic congestion and temperature.





* **Visualization Techniques**:
  + plt.plot() – for line plots (congestion over hours).
  + sns.barplot() – for weekday-based congestion analysis.
  + sns.lineplot() – for temperature variation over time.
* **Time Features**:
  + Hour: Extracted from timestamp, shows hourly trends.
  + Weekday: Reveals congestion trends across the week.

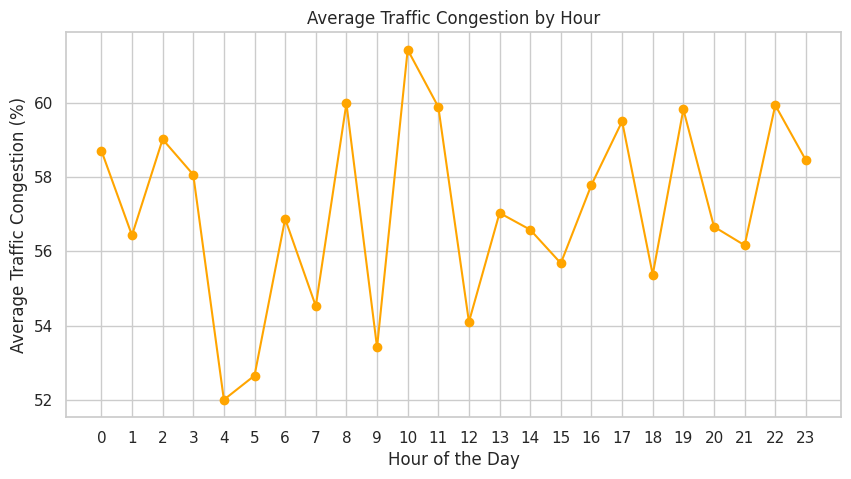


**iv. Analysis Results**

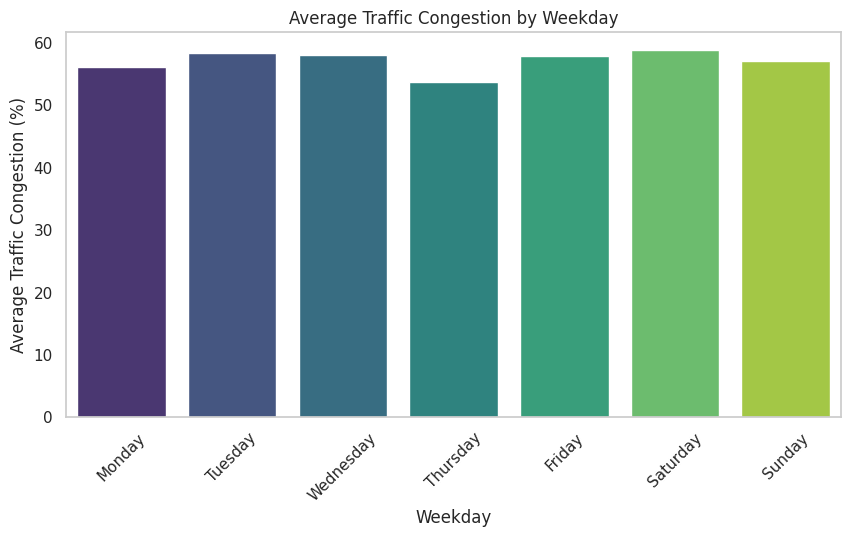
* Hourly Congestion:
  + Traffic congestion peaked during morning and evening rush hours, aligning with typical office commute times (e.g., 8–10 AM, 5–7 PM).
* Weekly Congestion:
  + Weekdays, especially Monday to Friday, showed higher congestion compared to weekends, highlighting commuter-driven traffic.
* Temperature Variation:
  + Temperature typically increased through the day, peaking during early afternoon hours and declining towards evening and night.
  + This trend can be linked with air quality and energy consumption patterns in smart cities.

**v. Visualizations**

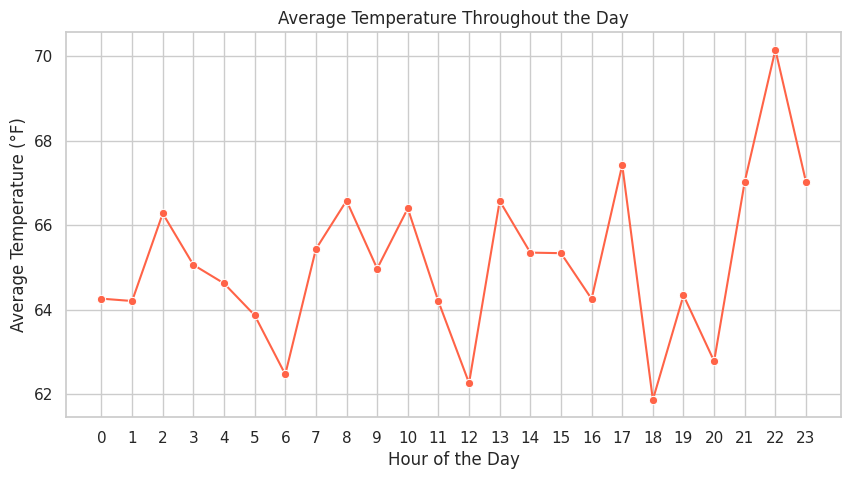
1. Line Plot – Average Traffic Congestion by Hour  
   Clearly depicts congestion peaks during rush hours, valuable for traffic light scheduling and public transport planning.



1. Bar Plot – Average Traffic Congestion by Weekday  
   Reveals how weekdays face heavier traffic, suggesting the need for targeted congestion management strategies on workdays.



1. Line Plot – Average Temperature by Hour  
   Offers insights for environmental planning, HVAC control, and urban heat island analysis.



**6. City-wise Traffic and Emissions Analysis**

**i. Introduction**

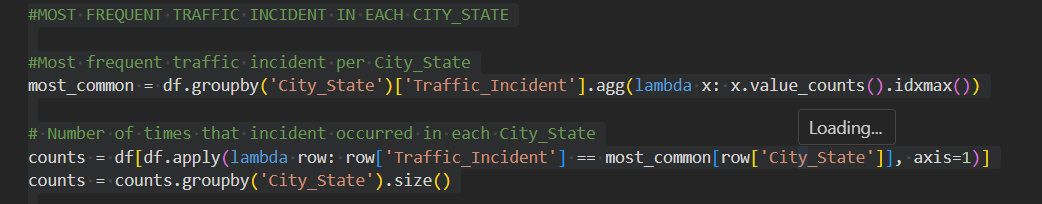
This section focuses on understanding traffic patterns and environmental impact on a city-wise level. It identifies the most frequent traffic incidents in each city and analyzes CO2 emissions to understand air quality variations across different urban areas.

**ii. General Description**

* Most Frequent Traffic Incident: Identifies the most common traffic issue in each city, offering insight into localized urban challenges.
* CO2 vs PM2.5 Analysis: Examines the relationship between CO2 emissions and PM2.5 levels, both of which are critical indicators of air quality.
* City-wise CO2 Averages: Highlights which cities contribute most to carbon emissions, supporting sustainable planning.

**iii. Specific Requirements, Functions, and Formulas**

* groupby('City\_State') with value\_counts().idxmax() – to find the most frequent incident type per city.
* apply() – used to match each city’s incident and count its frequency.
* scatterplot() – for comparing CO2 and PM2.5.
* barplot() – to show average CO2 emissions across cities.

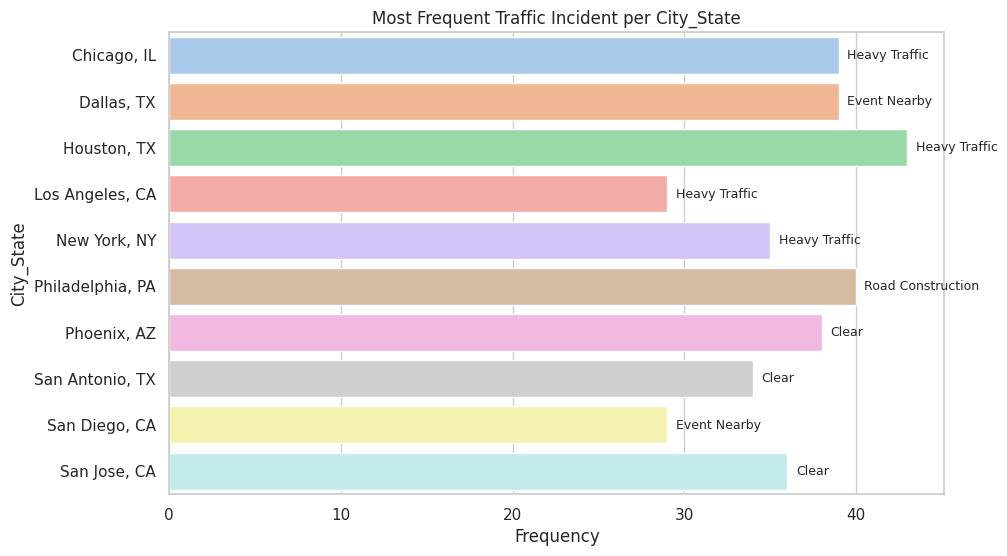


**iv. Analysis Results**

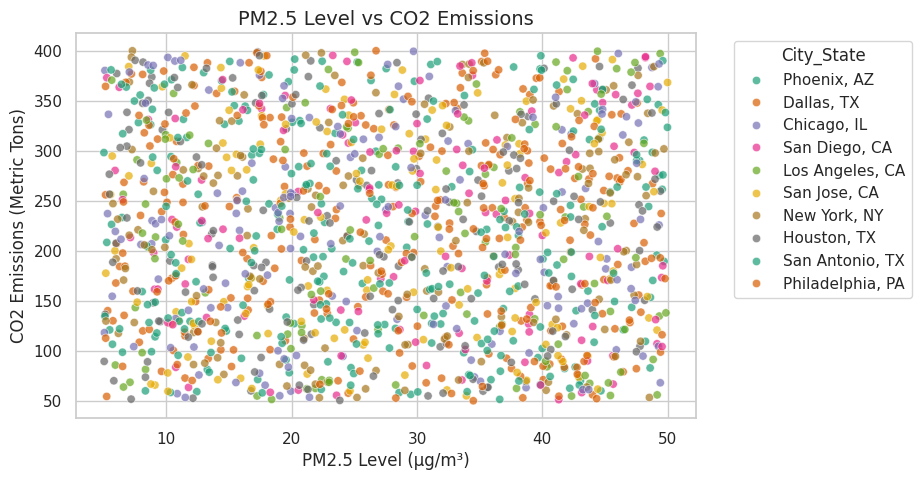
* Traffic Incidents by City:
  + Each city tends to have a dominant traffic issue. For instance, one city might experience more accidents, while another faces road closures frequently.
  + This can guide city-specific safety or infrastructure measures.
* CO2 Emissions vs. Air Quality:
  + A positive correlation was observed between PM2.5 levels and CO2 emissions in some cities, suggesting that higher emissions may contribute to poor air quality.
* Average CO2 Emissions:
  + Cities varied significantly in average emissions.
  + The top emitting cities could be targeted for stricter environmental policies or green initiatives.

**v. Visualizations**

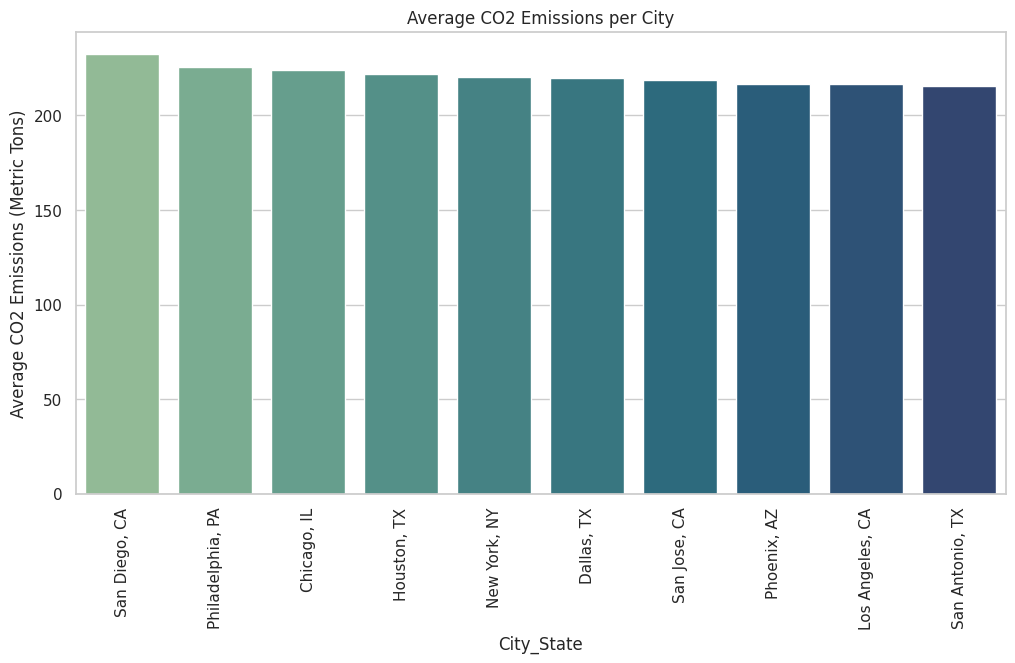
1. Bar Plot – Most Frequent Traffic Incident per City  
   Annotated to show which incidents dominate each region.



1. Scatter Plot – CO2 Emissions vs PM2.5 Levels  
   Displays how pollution metrics correlate city-wise.



1. Bar Plot – Average CO2 Emissions per City  
   Helps pinpoint cities with higher environmental impact.



**7. Environmental and Infrastructure Metrics Analysis (City-wise)**

**i. Introduction**

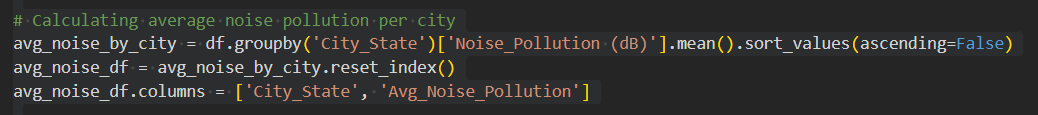
This analysis explores environmental stress indicators and resource consumption patterns across cities. By examining metrics like noise pollution, water usage, and electricity demand, we gain insight into how urban environments are coping with population and infrastructure pressures.

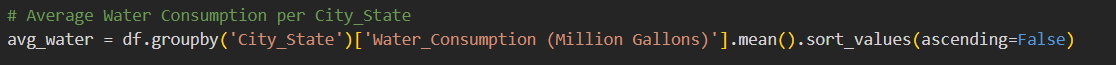
**ii. General Description**

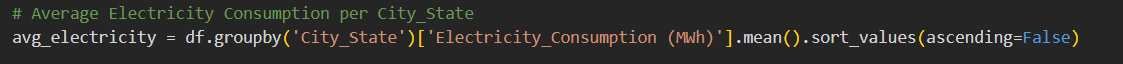
* Noise Pollution: Indicates ambient sound levels, often linked to traffic, construction, and industrial activity.
* Traffic Incident Composition: Shows what types of incidents dominate per city.
* Water and Electricity Consumption: Helps understand utility demand per city.

**iii. Specific Requirements, Functions, and Formulas**

* groupby('City\_State').mean() – to calculate average values for various environmental metrics.







**iv. Analysis Results**

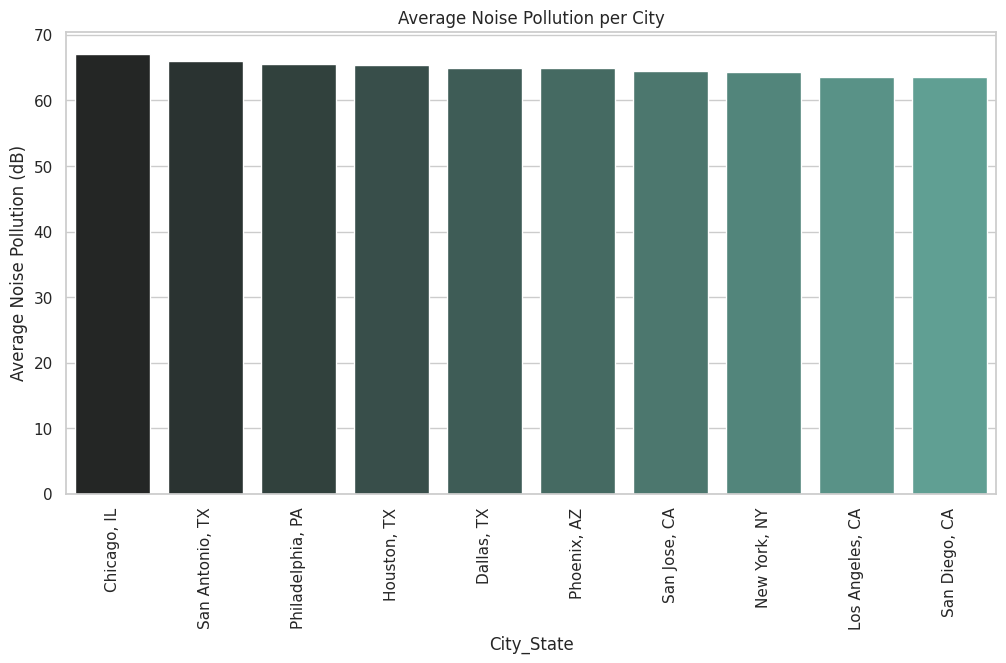
* Noise Pollution: Some cities showed consistently higher average decibel levels, indicating they might be more affected by traffic or industrial zones.
* Traffic Incident Breakdown per City: Stacked bar charts revealed varied incident profiles—some cities had higher counts of accidents, others showed more road closures or constructions.
* Resource Consumption:

Water Consumption🡪 Urban centers with larger populations or industry had higher average water usage.

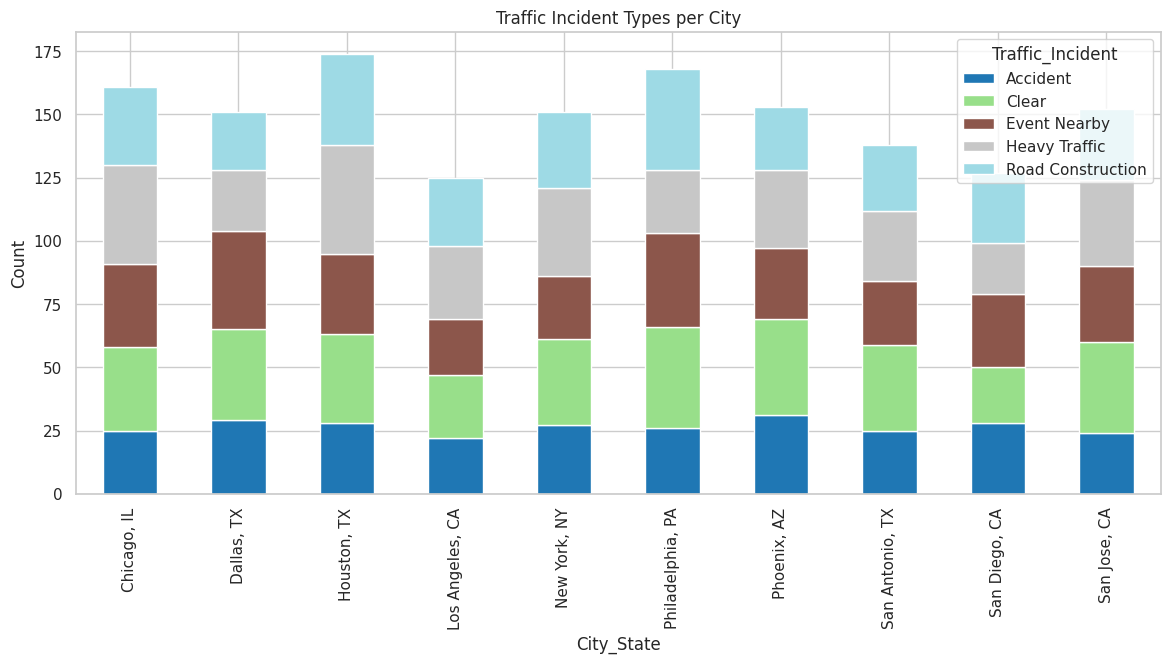
Electricity Consumption🡪Similar trends were seen for electricity, possibly driven by climate (cooling needs), population density, or urban infrastructure.

**v. Visualizations**

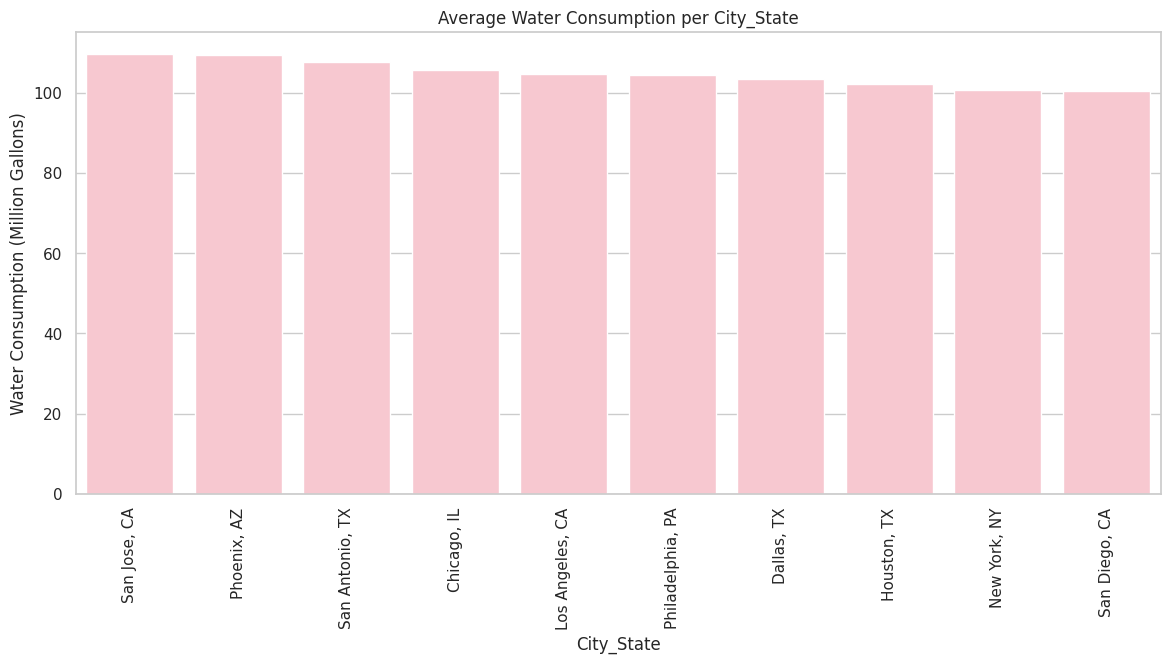
1. Bar Plot – Average Noise Pollution per City



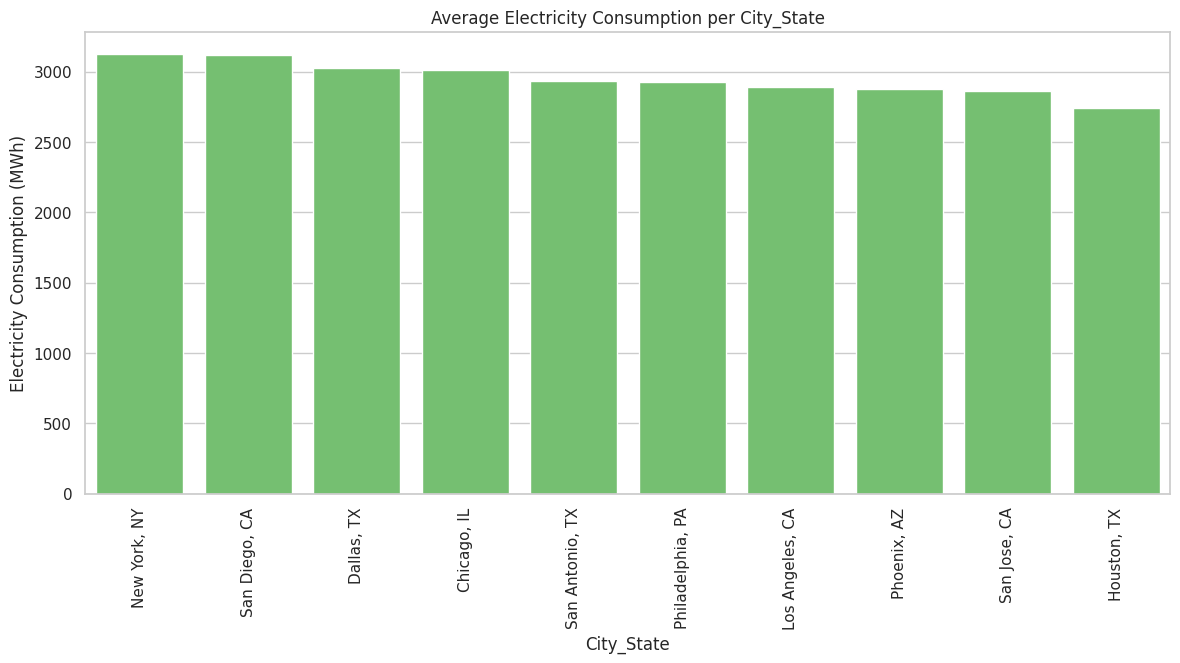
1. Stacked Bar Plot – Types of Traffic Incidents per City



1. Bar Plot – Average Water Consumption per City



1. Bar Plot – Average Electricity Consumption per City



**CONCLUSION**

The Exploratory Data Analysis (EDA) conducted on the Smart City dataset has provided a comprehensive overview of urban infrastructure, environmental conditions, and traffic dynamics across multiple city-states. The analysis enabled structured insights into the relationships among key variables, aiding in the interpretation of urban performance indicators.

Initial data exploration involved data type inspection, missing value assessment, and duplication checks, followed by a descriptive statistical summary. Categorical variables such as City\_State, Traffic\_Incident, and Weather\_Condition were analyzed to understand distributional patterns, while visualizations such as count plots and bar charts were employed to effectively communicate frequency-based insights.

Numerical variables were examined in depth using summary statistics, skewness metrics, and correlation heatmaps. Key findings include notable positive correlations among environmental pollution variables, especially between PM2.5\_Level (µg/m³) and CO2\_Emissions (Metric Tons), which suggests emission-related deterioration in air quality.

Temporal analysis of Traffic\_Congestion (%) revealed peak congestion during early and late working hours, with higher values on weekdays, confirming behavioural traffic trends. The transformation of timestamp data into hour and weekday formats enabled the derivation of time-based insights.

Outlier detection using the IQR method highlighted anomalies in pollution and resource consumption metrics. Box plots provided visual confirmation of such extreme values across cities, particularly for Noise\_Pollution (dB) and PM2.5\_Level (µg/m³).

Further, city-wise aggregations were used to identify cities with the highest average CO2 emissions, noise pollution, water consumption, and electricity usage. A stacked bar plot of traffic incident types per city facilitated comparative analysis across urban zones.

Overall, the EDA process allowed the extraction of meaningful patterns from complex multi-dimensional data. The insights obtained can support evidence-based decision-making in urban planning, environmental monitoring, and resource management. Future work may include time-series forecasting and integration with geospatial data for enhanced predictive modelling and policy optimization.

**FUTURE SCOPE**

This analysis has helped us understand various factors affecting smart cities, such as traffic, pollution, and utility usage. However, there are many more ways this study can be expanded in the future:

1. Prediction Models:  
   In the future, we can use machine learning to predict traffic jams, pollution levels, or energy demand based on patterns seen in the data. This can help city officials plan better.
2. Location-Based Mapping:  
   By adding maps or using tools like GIS, we can visually see which areas face the most traffic or pollution, making it easier to take action in specific locations.
3. Time-Based Forecasting:  
   We can use time-series analysis to predict how things like temperature or electricity use will change over days, weeks, or months.
4. Finding Unusual Patterns:  
   We can build systems that automatically detect anything out of the ordinary, like a sudden increase in noise pollution or traffic, and alert authorities.
5. Adding Citizen Feedback:  
   If we combine this data with feedback from people or smart devices, we can get a more complete view of what’s happening in the city.
6. Measuring Policy Impact:  
   After new rules or policies are made (like reducing traffic or pollution), we can track changes in the data to see if those rules are working.
7. Interactive Dashboards:  
   Creating easy-to-use dashboards can help government departments or the public monitor city data live and make informed decisions.

**REFERENCES**

[1] The Pandas Development Team, “pandas - Python Data Analysis Library,” pandas.pydata.org, 2024. [Online]. Available: <https://pandas.pydata.org/docs/>. [Accessed: Apr. 12, 2025].

[2] NumPy Developers, “NumPy Documentation,” numpy.org, 2024. [Online]. Available: <https://numpy.org/doc/>. [Accessed: Apr. 12, 2025].

[3] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007. [Online]. Available: <https://matplotlib.org/stable/index.html>. [Accessed: Apr. 12, 2025].

[4] M. Waskom, “Seaborn: Statistical Data Visualization,” seaborn.pydata.org, 2024. [Online]. Available: <https://seaborn.pydata.org/>. [Accessed: Apr. 12, 2025].

[5] A. Aggarwal, “Exploratory Data Analysis in Python,” Towards Data Science, 2021. [Online]. Available: <https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce>. [Accessed: Apr. 12, 2025].

[6] A. L. Kromer, “Exploratory Data Analysis and Visualizations in Python,” in Proc. of the 10th International Conference on Data Science, 2019. [Online]. Available: <https://towardsdatascience.com/>. [Accessed: Apr. 12, 2025].

[7] T. M. Mitchell, “Machine Learning,” McGraw-Hill Education, 1997. [Online]. Available: <https://www.mheducation.com/>. [Accessed: Apr. 12, 2025].