## DATA SCIIENCE USING PTHON PROJECT REPORT

(Project Semester January-April 2025)

# *Comprehensive Analysis and Visualization of Government Building Data for Enhanced Urban Planning and Resource Allocation*

### Submitted by

Eswar Bongu

Registration no: 12304639

Section: K23FD

Course Code: INT375

**Under the Guidance of** Dr. Baljinder Kaur

(27952)

### Discipline of CSE/IT

**Lovely School of Computer Science Engineering Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Eswar Bongu** bearing Registration no. **12304639** has completed **INT375** project titled, **“Comprehensive Analysis and Visualization of Government Building Data for Enhanced Urban Planning and Resource Allocation”** 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.

### Dr. Baljinder Kaur

### Assistant Professor

School of Computer Science Engineering

Lovely Professional University Phagwara, Punjab.

Date: 12th April, 2025

**DECLARATION**

I, Eswar Bongu, 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: 12th April, 2025 Signature

Registration No.12304639 Eswar Bongu

**Acknowledgment**

I would like to express my deepest gratitude to Prof Dr. Baljinder Kaur for his exceptional mentorship and unwavering support throughout the duration of this project. His vast knowledge in the fields of data science and machine learning, combined with his patient and thoughtful guidance, played a pivotal role in the successful completion of this work. His insightful suggestions and feedback consistently challenged me to think critically and improve the quality of my research. I am also grateful for the learning environment he fostered, which encouraged exploration and innovation.

In addition, I sincerely thank my peers and classmates for their helpful discussions, encouragement, and collaborative spirit during this project. Their input provided fresh perspectives that contributed meaningfully to the final outcome. I am also thankful to the open-source community for providing the tools, libraries, and resources that made the implementation of this project possible. Lastly, I acknowledge the dataset contributors for making this analysis feasible.

# Introduction

In today's data-driven era, understanding infrastructure trends and property characteristics plays a vital role in decision-making across sectors like urban planning, government asset management, and sustainable development. With the increasing availability of public datasets, it's now feasible to explore the composition, distribution, and condition of government-owned buildings through the lens of data analytics.

This project focuses on exploring and analysing historical data on federal buildings using Python. The dataset includes detailed records with attributes such as construction year, square footage, agency ownership, location (state and city), and property type. By leveraging this dataset, we aim to identify patterns related to building sizes, geographical distribution, and construction timelines.

Through systematic data cleaning and exploratory data analysis (EDA), this project uncovers insights by detecting missing values, identifying outliers, and visualizing the distribution of key variables. One of the primary goals is to evaluate how construction trends have evolved over time and how building sizes vary across different property types and regions.

Powerful data visualization tools such as Seaborn and Matplotlib are used to generate bar charts, pie charts, scatter plots, box plots, and correlation heatmaps. These visualizations not only support data interpretation but also highlight hidden trends and anomalies within the dataset.

Beyond academic enrichment, this analysis serves as a foundational step for practical applications in policy planning, asset optimization, and sustainability studies. Understanding the nature and scope of government infrastructure can empower researchers, urban developers, and data scientists to build predictive models, forecast maintenance needs, and support smarter public infrastructure decisions. The following sections detail the data preprocessing techniques, analytical findings, and visual summaries derived from this federal building dataset.

# Dataset Description

Dataset Description and Preprocessing

The dataset used in this project contains records of federal government buildings across the United States, collected from a centralized public infrastructure registry. Each row in the dataset represents an individual building and includes various physical, administrative, and geographical attributes associated with that property. These features encompass both numerical and categorical variables, enabling a holistic analysis of government-owned assets.

Key features in the dataset include:

Year\_Built: The year in which the building was constructed (extracted from the original Construction\_Date).

Square\_Footage: Usable building area measured in square feet.

Agency: The name of the federal agency that owns or manages the building.

State: The U.S. state where the building is located.

City: The city in which the building resides.

Property\_Type: Classification of the building based on its usage (e.g., "Office", "Warehouse", "Laboratory").

Before conducting the analysis, the dataset underwent a series of preprocessing steps to ensure data quality, consistency, and readiness for visualization. These steps included:

Cleaning column names by removing extra spaces, line breaks, and special characters for uniformity.

Handling missing values using appropriate strategies:

Numerical columns such as Year\_Built and Square\_Footage were filled using the median value.

Categorical columns like Property\_Type, Agency, State, and City were filled with a placeholder value like "Unknown".

Data type conversion to ensure all columns were in suitable formats for analysis (e.g., converting year to float and square footage to numeric).

Removal of irrelevant or redundant columns such as ABA\_Accessibility\_Flag, Historical\_Type, and Owned\_Leased to simplify analysis.

Outlier detection and removal using the Interquartile Range (IQR) method for Square\_Footage to improve data reliability and prevent skewed results in visualizations.

Saving a cleaned version of the dataset (cleaned\_rexus\_dataset.csv) for use in further analytical tasks and visual representation.

These preprocessing steps were essential in preparing the dataset for exploratory data analysis (EDA), helping ensure that the resulting insights are both accurate and interpretable. The cleaned dataset now serves as a strong foundation for analyzing building characteristics and identifying meaningful patterns across federal infrastructure.

# Source of Dataset

The dataset used in this project was obtained from data.gov.in, the Government of India’s official open data platform. This portal provides a wide range of datasets collected and maintained by various governmental departments and agencies to promote transparency, innovation, and data-driven research.

The specific dataset, titled "RExUS - Repository of Government Buildings", contains detailed information about federal buildings, including construction year, square footage, location, agency ownership, and property type. It serves as a valuable resource for infrastructure analysis and public asset management.

For this project, a manageable portion of the dataset was used to ensure smooth and efficient data processing. This includes filtering and analysing a representative subset while retaining the diversity and variability of the original data to capture meaningful trends.

Link to original dataset: <https://data.gov.in>

# Exploratory Data Analysis (EDA) Process

Exploratory Data Analysis (EDA) is a crucial step in any data-driven project. It allows me to uncover hidden patterns, identify data quality issues, understand feature distributions, and establish relationships among variables. In this project, EDA played an essential role in cleaning the government building dataset, uncovering trends, and preparing the data for deeper analysis.

The EDA process was carried out through the following steps:

1. Understanding the Structure and Data Types:

I began by examining the structure of the dataset using functions like shape, columns, dtypes, and df.head(). This allowed me to categorize the features into numerical (e.g., square footage, construction year) and categorical (e.g., state, property type) variables. I also identified any inconsistencies in data types and ensured the columns were in the correct format.

2. Handling Missing Values and Data Cleaning:

Using df.isna().sum(), I identified missing values in various columns. I applied appropriate cleaning steps, such as:

* Filling missing values in categorical columns like Property\_Type, State, and City with the mode.
* Imputing missing numerical values like Year\_Built and Square\_Footage with the median.
* Dropping columns with excessive null values, such as ABA\_Accessibility\_Flag, Historical\_Type, and Owned\_Leased.
* Ensuring that Year\_Built and Square\_Footage columns were in correct numerical formats for analysis.

This cleaning process ensured a consistent dataset ready for visualization and analysis.

3. Univariate Analysis:

I analyzed the distribution of individual variables:

* The distribution of Square\_Footage and Year\_Built was visualized using histograms and box plots to understand their spread and detect any outliers.
* For categorical features like Property\_Type, Agency, and State, bar charts and pie charts were used to evaluate the frequency and distribution of different categories.

4. Bivariate Analysis:

I explored relationships between two variables:

* A scatter plot between Year\_Built and Square\_Footage revealed the general relationship between construction year and building size.
* I also examined the distribution of Square\_Footage across different Property\_Type categories using box plots to observe potential differences between them.

5. Correlation and Multivariate Analysis:

A correlation heatmap was generated to visualize the strength of linear relationships among numerical features. The following key findings were observed:

* A strong positive correlation between Square\_Footage and Year\_Built, indicating that larger buildings tend to be newer.
* This correlation is useful for predicting building characteristics and understanding how these factors might interact in future analyses.

6. Outlier Detection:

Using box plots and IQR (Interquartile Range) methods, I identified and handled outliers in Square\_Footage. These extreme values were either addressed through imputation or removed, ensuring the dataset remained clean for analysis and future modeling.

Through these EDA steps, I gained a deeper understanding of the government building dataset, identified key trends, and prepared the data for further analysis, including predictive modeling and anomaly detection.

# Analysis on Dataset

**i. Introduction**

Data analysis is a crucial step in understanding how different building attributes, such as square footage, construction year, and property type, relate to various environmental and infrastructural factors. This section focuses on examining these attributes through statistical and visual techniques. By applying mathematical functions, correlation analysis, and graphical plots, the goal is to extract meaningful insights into how features like square footage, building age, and property type behave and interrelate, providing a deeper understanding of the dataset.

**ii. General Description**

The dataset includes both numerical and categorical attributes related to government buildings. The core objective of this analysis is to:

* Understand the distribution of individual building features like square footage and construction year.
* Identify correlations between variables such as square footage, year built, and property type.
* Analyse how building characteristics vary by state, property type, and agency.
* Use visualization tools to support pattern recognition and outlier detection.

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

The analysis uses descriptive statistics and visualization libraries. The tools and techniques include:

**1. Descriptive Statistics**

* **Mean, Median, Standard Deviation, and Distribution Curves**
* **Formulas:**
  + Mean = (Σ Xi) / N
  + Standard Deviation = sqrt (Σ (Xi - μ) ² / N)

**2. Correlation Analysis**

* Pearson Correlation Coefficient was used to assess linear relationships between numerical variables.
* **Formula:**
  + r = [ Σ (Xi - X̄) (Yi - Ȳ)] / sqrt [ Σ (Xi - X̄) ² \* Σ (Yi - Ȳ) ²]

**3. Data Visualization Techniques**

* **Histograms** to examine distributions of square footage, year built, etc.
* **Scatter Plots** to explore relationships (e.g., Square Footage vs. Year Built)
* **Heatmaps** to visualize the correlation matrix
* **Boxplots** to detect outliers in square footage, construction year, etc.
* **Bar Charts** to compare categorical values like Property Type or Agency frequency

**iv. Analysis Results**

**1. Square Footage and Year Built Relationship**

* A scatter plot revealed a positive correlation: newer buildings tend to be larger in square footage.
* This trend was further supported by a regression line.

**2. Outlier Detection**

* Box plots highlighted a few extreme values in square footage and construction year, which were further investigated and visualized.

**3. Property Type Analysis**

* Bar plots showed that Office buildings were the most common property type, followed by Residential buildings.
* Square footage distributions differed significantly across property types.

**4. Correlation Matrix Insights**

* A heatmap revealed a strong positive correlation between square footage and year built, indicating that newer buildings tend to be larger.
* There was a moderate negative correlation between square footage and the building's age.

**5. Agency-Based Building Trends**

* Bar plots for the "Agency" column showed that buildings managed by the Ministry of Defence were the most prevalent.
* These buildings also tended to have larger square footage compared to other agencies.

**v. Visualization**

Below are the key plots used in the analysis:

1. **Histogram** – Distribution of square footage, year built, and other attributes.
2. **Correlation Heatmap** – Relationships between numerical variables such as square footage, year built, and age.
3. **Scatter Plot** – Square Footage vs. Year Built.
4. **Boxplot** – Outlier detection for square footage and year built.
5. **Bar Chart** – Frequency of property types and agencies.

# Conclusion

This project focused on analysing a government building dataset using Python, combining data preprocessing, visualization, and statistical analysis techniques. Through exploratory data analysis (EDA), we uncovered meaningful patterns such as the positive correlation between square footage and construction year, the frequent occurrence of "Office" as the primary property type, and the distribution of building attributes across various agencies and states.

Key tasks included handling missing values, detecting outliers in square footage and construction year, and using visual tools such as scatter plots, box plots, and heatmaps to understand variable relationships. The correlation matrix provided insights into how different building characteristics interact, aiding in the interpretation of the data and helping to identify potential trends in government building structures.

Overall, the project demonstrated the power of Python in processing and visualizing building data. It lays a strong foundation for future work in building trend forecasting, space utilization analysis, and data-driven decision-making in urban planning and government infrastructure management.

# Future Scope

• While this project successfully provided insights into the characteristics and trends of government buildings, there are several opportunities for further enhancement and expansion:

• Time-Series Forecasting: Implementing time-series forecasting techniques such as ARIMA, Prophet, or LSTM can predict future trends in government building characteristics (e.g., square footage, construction year, or property type) based on historical data.

• Real-Time Data Integration: Integrating real-time building data from various sources (such as public building records or sensors) would allow continuous updates and dynamic tracking of building-related trends, providing more actionable insights for urban planning and resource allocation.

• Geospatial Analysis: Incorporating geospatial data using tools like Folium or GeoPandas can provide deeper insights into the geographical distribution of government buildings, highlighting patterns based on location (e.g., state, agency) and offering a spatial perspective on building characteristics.

• Predictive Modeling: Machine learning models can be trained to predict building-related features, such as the type of property (e.g., office, school) or potential maintenance needs based on historical data, enabling more efficient resource management.

• Interactive Dashboards: Using frameworks like Streamlit or Dash, the analysis can be made interactive, allowing users to filter data, explore different variables, and visualize building trends in real time, making the results more user-friendly and accessible.

# Snapshots:

# 

# 

# 

# 

# 

# 

# Github Link

# h[ttps://github.com/eswar6718/rexus\_data](https://github.com/eswar6718/rexus_data)

# LinkedIn post link

# <https://www.linkedin.com/posts/eswar-bongu-766016289_python-datascience-eda-activity-7317137892659601408-AoCQ?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEXyjb4BUSopLfgLx23EKbfw992QAHVsR2s>