**Data Science Toolbox: Python Programming  
PROJECT REPORT**  
Analysis of Commercial Building Dataset in the USA  
  
Submitted by  
Prince Kumar  
Registration No: 12309518  
B.Tech CSE - Section K23GX  
Course Code: INT375

Discipline of CSE  
Lovely School of Computer Science and Engineering  
Lovely Professional University, Phagwara

# DECLARATION

I, Prince Kumar, student of B.Tech CSE under CSE 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: 12-4-2025  
  
Signature: Prince  
Registration No. 12309518  
Name: Prince Kumar

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guide, Madam Gargi Sharma, for her invaluable support, guidance and encouragement throughout the project. I would also like to thank the faculty members and staff of the Department of CSE for their support and help.

# TABLE OF CONTENT

1. Introduction  
2. Source of dataset  
3. EDA process  
4. Analysis on dataset  
5. Conclusion  
6. Future scope  
7. References

# 1. INTRODUCTION

The purpose of this project is to perform exploratory data analysis (EDA) on a dataset containing information about commercial buildings in the United States. The analysis covers aspects such as rentable area, geographical distribution, availability, correlation between variables, and outlier detection.

# 2. SOURCE OF DATASET

The dataset used for this project is provided in the file named '2025-3-28-iolp-buildings.xlsx'. It includes details about commercial buildings across different states in the USA.

Dataset Link: https://catalog.data.gov/dataset/inventory-of-owned-and-leased-properties-iolp

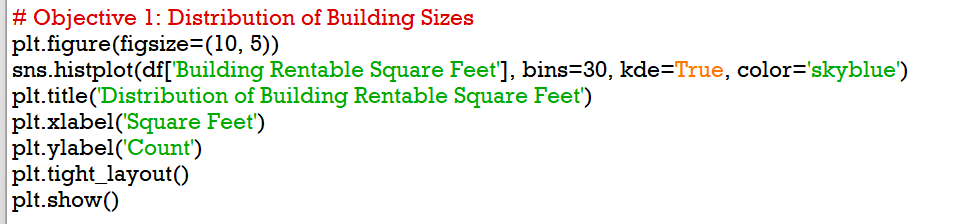
# 3. EDA PROCESS

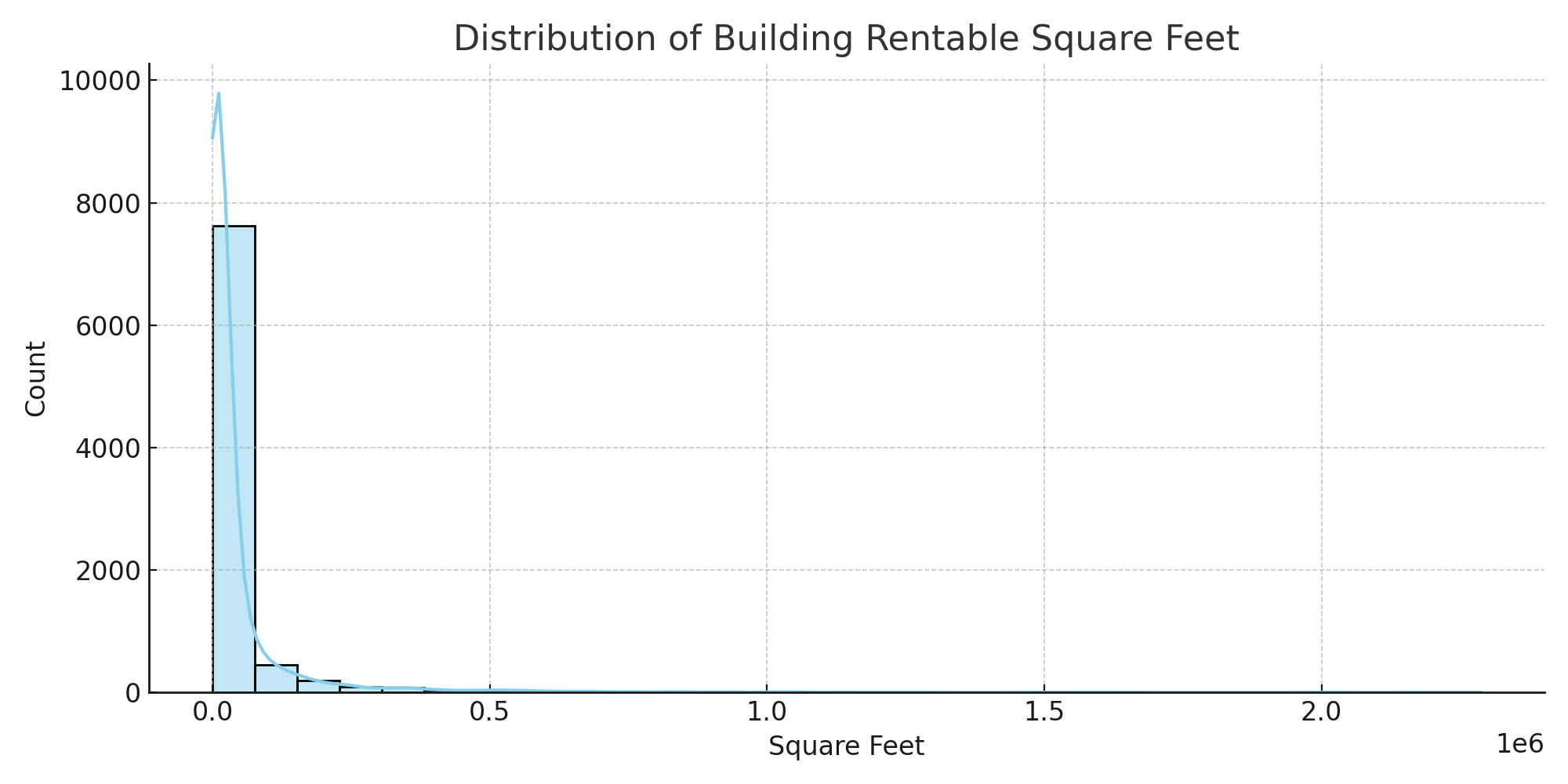
Exploratory Data Analysis (EDA) involves summarizing the main characteristics of the dataset using statistical graphics, plots, and information tables. The analysis was done using Python with pandas, matplotlib, and seaborn libraries.

# 4. ANALYSIS ON DATASET

## 4.1 Distribution of Building Sizes

This analysis shows how rentable square footage is distributed across properties. A histogram reveals concentration ranges and outliers.

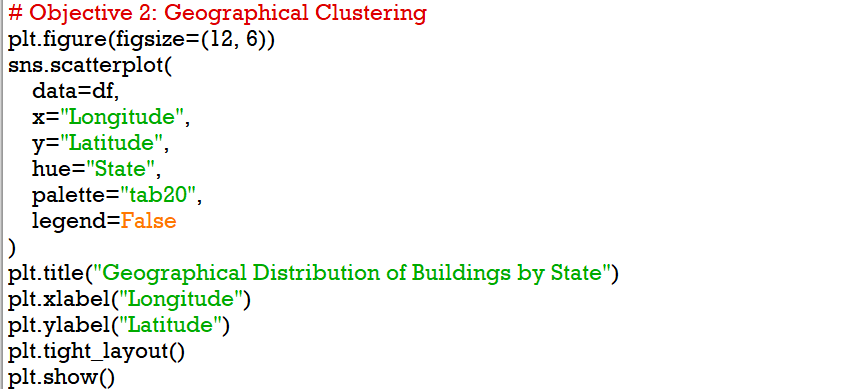


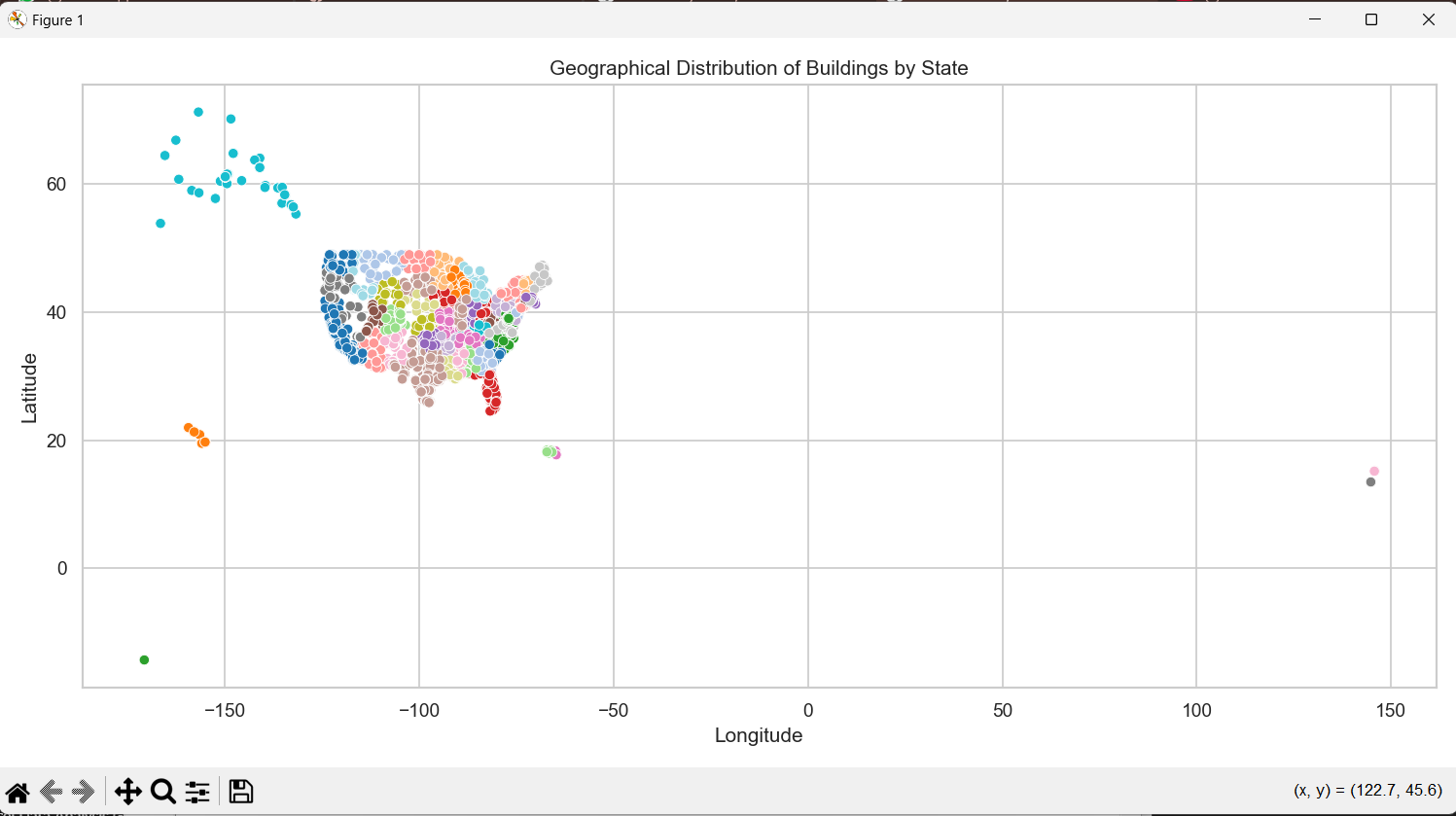


*Figure 1: Distribution of Building Rentable Square Feet*

## 4.2 Geographical Clustering

The scatter plot shows how buildings are geographically spread across the USA based on their latitude and longitude. Each state is color-coded.

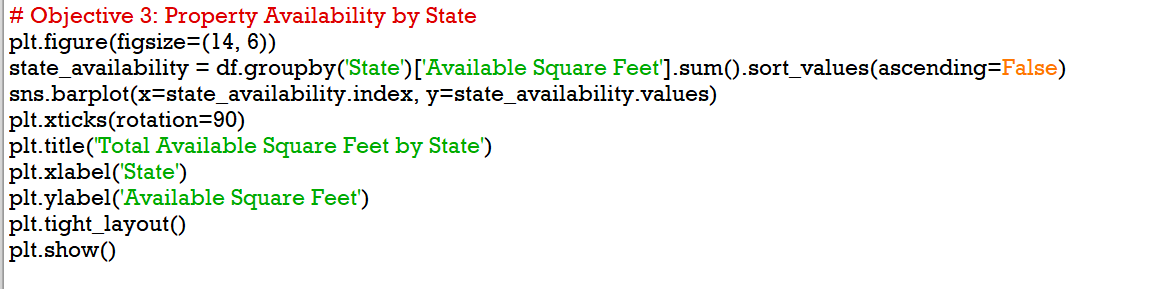


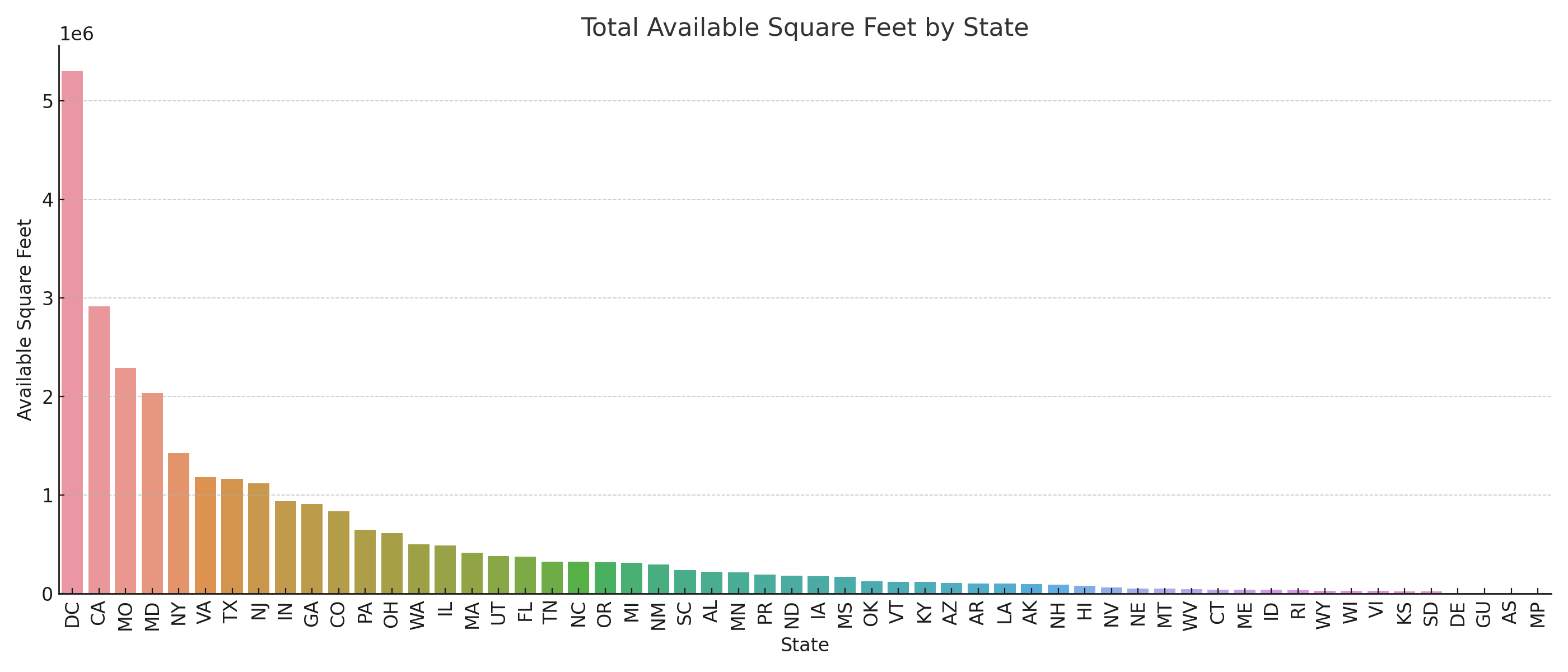


*Figure 2: Geographical Distribution of Buildings by State*

## 4.3 Property Availability by State

This bar chart visualizes the total available square feet of property per state. It helps in identifying which states have the most commercial space.

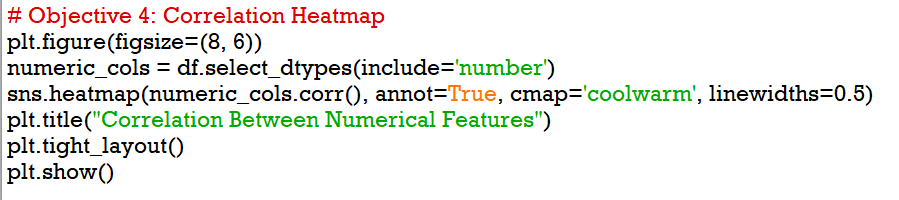


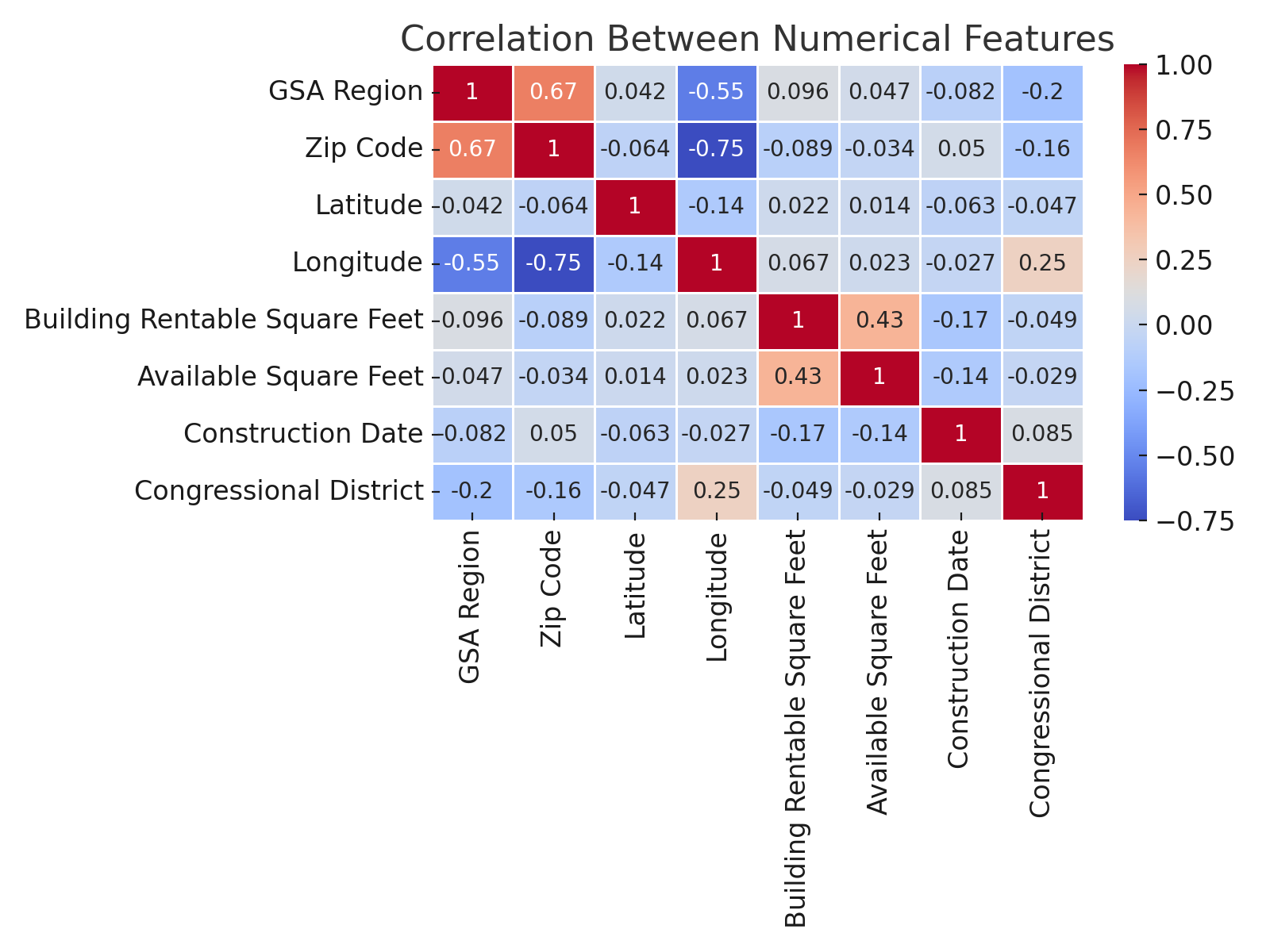


*Figure 3: Total Available Square Feet by State*

## 4.4 Correlation Heatmap

The correlation matrix shows the relationships between different numerical variables in the dataset, highlighting strongly correlated pairs.

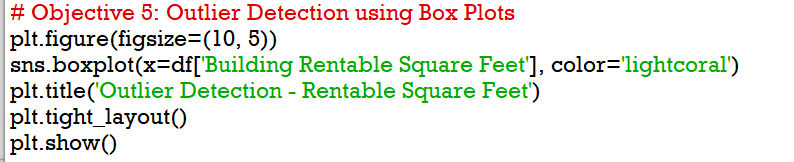


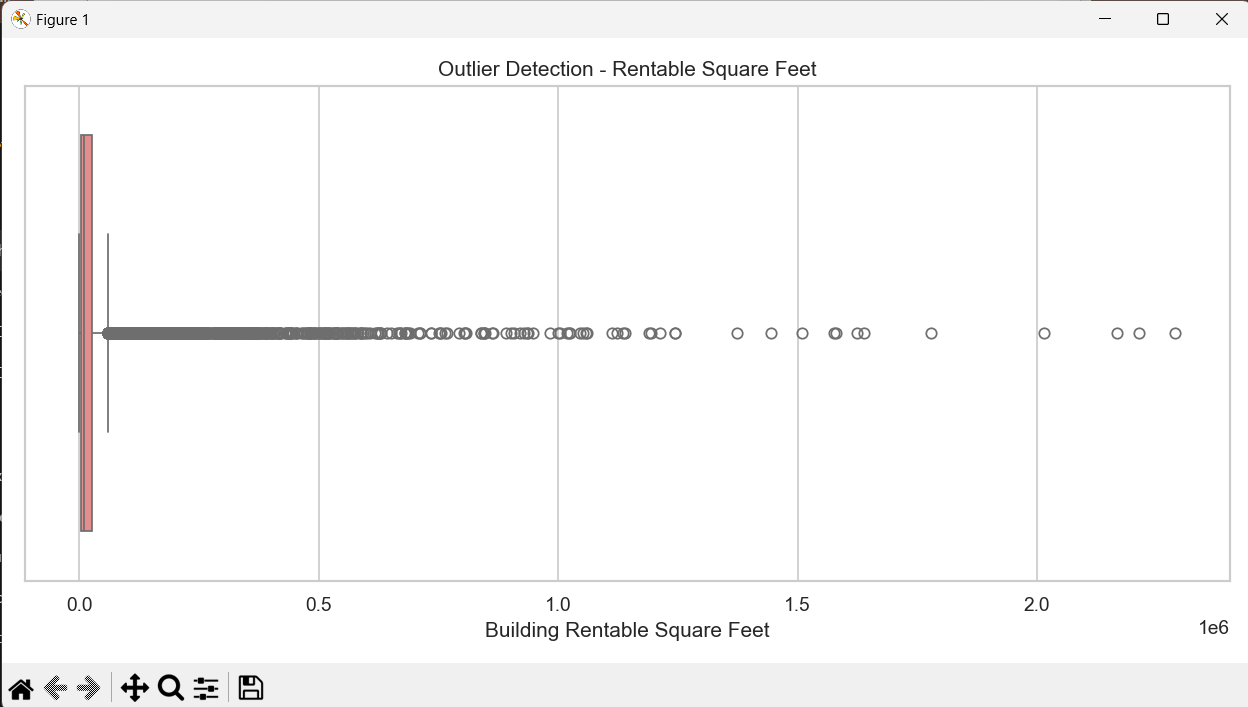


*Figure 4: Correlation Between Numerical Features*

## 4.5 Outlier Detection

A boxplot helps identify outliers in rentable square footage, revealing properties with unusually high or low square footage.





*Figure 5: Outlier Detection - Rentable Square Feet*

# 5. CONCLUSION

The primary objective of this project was to perform a comprehensive exploratory data analysis (EDA) on a dataset containing information about commercial buildings in the United States. Through various visualizations and statistical techniques, valuable insights were extracted regarding the spatial distribution, availability of rentable space, and interrelationships among key numerical features.

The histogram of rentable square footage highlighted a concentration of properties in the mid-range, with a few outliers indicating exceptionally large commercial spaces. Geographical distribution mapping revealed clusters of buildings across specific states, emphasizing regions with higher commercial density. The bar chart showing property availability provided clear evidence that a few states dominate in terms of available rentable area, potentially due to their economic activities or urban development patterns.

The correlation heatmap enabled us to identify meaningful relationships between numerical features. For example, there may be a moderate to strong correlation between rentable square footage and available square footage, indicating that larger buildings generally have more unoccupied space. Outlier detection through boxplots helped in identifying extreme values that might warrant further investigation or cleaning in a data preprocessing stage.

Overall, this project demonstrated how EDA techniques can be effectively used to uncover patterns, detect anomalies, and provide direction for deeper statistical or machine learning models in commercial real estate analysis.

# 6. FUTURE SCOPE

Future studies could incorporate time series data, compare property values over time, or integrate external factors like population or economic growth to enrich the analysis.

# 7. REFERENCES

<https://www.w3schools.com/python/>

<https://data.gov/>

GITHUB LINK: https://github.com/Prince087/Analysis-of-Commercial-Building-Dataset-in-the-USA