Final Project Report

- **Hrutheek Reddy Thummala**

1. Dataset Introduction

The dataset comprises comprehensive real estate information from New York, encompassing diverse attributes such as property prices, the number of bedrooms and bathrooms, property square footage, geographical coordinates (latitude and longitude), and address details. It serves as a valuable resource for understanding the dynamics of the New York real estate market.

2. Problem Statement

The primary objective of this analysis is to extract actionable insights from the New York real estate dataset to facilitate informed decision-making for various stakeholders, including investors, buyers, and real estate professionals. Key areas of focus include identifying patterns, trends, and clusters within the dataset to gain a deeper understanding of the market dynamics.

3. Choices of Methods, Method Introduction, and Rationale of Choosing Them Data Cleaning and Preprocessing

Handling Missing Values: Initial data inspection revealed missing values, which were addressed by implementing appropriate strategies such as imputation or removal, depending on the significance of the missing data.

Outlier Detection and Treatment: Outliers, which could potentially skew the analysis results, were identified using statistical methods such as Z-score analysis. Outliers were replaced with the mean of the respective numerical columns to ensure data integrity and accuracy.

Descriptive Analysis

Summary Statistics: Descriptive statistics, including measures of central tendency (mean, median), dispersion (standard deviation, range), and distribution (skewness, kurtosis), were computed for numerical features to provide a comprehensive overview of the dataset.

Visualization: Histograms and box plots were employed to visually explore the distribution of numerical features, allowing for a more intuitive understanding of the data distribution and identifying any potential patterns or anomalies.

Clustering Techniques

K-Means Clustering: A popular unsupervised learning algorithm, K-Means clustering partitions the dataset into K clusters based on similarity, with each data point assigned to the nearest cluster centroid. This method was chosen for its simplicity, efficiency, and scalability, making it suitable for large datasets.

Hierarchical Clustering: Hierarchical clustering organizes data into a hierarchical tree-like structure, known as a dendrogram, by iteratively merging or splitting clusters based on their similarity or dissimilarity. This approach provides insights into the hierarchical relationships among data points and is particularly useful for understanding complex data structures.

Model Evaluation

Elbow Method: The Elbow Method was utilized to determine the optimal number of clusters for K-Means clustering by evaluating the within-cluster sum of squares (WCSS) for different values of K. The point at which the rate of decrease in WCSS starts to slow down signifies the optimal number of clusters.

Silhouette Score: Silhouette Score measures the quality of clustering solutions by assessing the compactness and separation of clusters. A higher Silhouette Score indicates better-defined clusters, helping to validate the clustering results and guide the selection of the optimal number of clusters.

4. Analytical Outcome

ANOVA Analysis

ANOVA (Analysis of Variance) was performed to assess the significance of numerical features (PRICE, BEDS, BATH, PROPERTYSQFT) in explaining the variance in the target variable (PRICE). The ANOVA table provides F-values and p-values for each feature.

ANOVA Table

	df	sum_sq	mean_sq	F	PR(>F)
PRICE	1.0	3.658803e+11	3.658803e+11	8071.553362	0.0
BEDS	1.0	3.658758e+11	3.658758e+11	8071.476349	0.0
BATH	1.0	3.650028e+11	3.650028e+11	8052.895053	0.0
PROPERTYSQFT	1.0	3.654164e+11	3.654164e+11	8062.407572	0.0
Residual	1207.0	2.474455e+10	2.048727e+07	NaN	NaN

In the ANOVA table, all features have extremely low p-values (close to 0), indicating a statistically significant relationship with the target variable (PRICE). Additionally, the F-values are notably high, further confirming the significance of these features in explaining the variance in house prices.

Data Cleaning: Missing values were appropriately handled, and outliers were mitigated to ensure the integrity and reliability of the dataset for subsequent analysis.

Descriptive Analysis: Summary statistics and visualizations provided valuable insights into the distribution and characteristics of numerical features, enabling a better understanding of the dataset's structure and properties.

Clustering Results: K-Means and hierarchical clustering techniques revealed distinct clusters within the dataset, allowing for the identification of similar groups of properties based on their features such as price, size, and location.

Model Evaluation: The Elbow Method and Silhouette Score aided in determining the optimal number of clusters, ensuring the robustness and effectiveness of the clustering solutions.

5. Conclusion

In conclusion, the analysis of the New York real estate dataset yielded valuable insights into the market dynamics, enabling stakeholders to make data-driven decisions. By employing a combination of data cleaning, descriptive analysis, and clustering techniques, we uncovered meaningful patterns, trends, and clusters within the dataset, providing actionable insights for investment, property valuation, and market analysis purposes. Moving forward, continued exploration and refinement of analytical techniques will further enhance our understanding of the New York real estate market and support informed decision-making in this dynamic and competitive industry.