New York Housing Market

Exploratory Data Analysis (EDA)

Hrutheek Reddy Thummala (20018023)

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Contents

Introduction

•**Exploring New York's Real Estate Landscape:** Diving into the intricacies of NYC's real estate market with comprehensive data covering broker titles, property types, bedrooms, bathrooms, and locations, offering a holistic view of the housing scene.

•**Unveiling Stories within the Data:** This dataset enables uncovering narratives by analyzing relationships between square footage and pricing, property types, and regional trends, catering to real estate professionals, data enthusiasts, and curious observers alike.

•**Navigating Complexity:** The world of real estate is multifaceted and ever-evolving, especially in a city as diverse and dynamic as New York. Through this journey, we aim to unravel the nuances, spot emerging patterns, and navigate the complexity of the "New York Housing Market," offering valuable insights for anyone seeking to understand or engage with this fascinating domain.

Data cleaning

- **Handling Missing Values:** Address any missing data points by either imputing values or removing incomplete entries to maintain dataset integrity.

- **Removing Duplicates:** Identify and eliminate duplicate records to prevent redundancy and ensure accurate analysis results.

- **Standardizing Data Formats:** Ensure consistency in data formats across columns, such as dates, numerical values, and categorical variables, to facilitate analysis.

- **Dealing with Outliers:** Detect and manage outliers within the dataset to mitigate their impact on analysis outcomes, employing techniques like outlier removal.

- **Feature Engineering and Encoding:** Enhance dataset quality by creating new features based on domain knowledge and encoding categorical variables into numerical representations for analysis readiness.

Data loading and Initial Exploration

•Importing necessary libraries and reading the dataset into a Data Frame.

•Printing basic information about the dataset such as column names, data types, and non-null counts.

•Performing initial data cleaning tasks, such as converting data types or handling missing values (in this case, converting the 'BATH' column to integer type).

Data loading and Initial Exploration

	BROKERTITLE	TYPE	PRICE	BEDS	BATH	PROPERTYSQFT	ADDRESS	STATE	MAIN_ADDRESS	ADMINISTRATIVE_AREA_LEVEL_2	LOCALITY	SUBLOCALITY	STREET_NAME	LONG_NAME	FORMATTED_ADDRESS	LATITUDE	LONGITUDE
0	Brokered by Douglas Elliman -111 Fifth Ave	Condo for sale	315000	2	2	1400.0	2 E 55th St Unit 803	New York, NY 10022	2 E 55th St Unit 803New York, NY 10022	New York County	New York	Manhattan	East 55th Street	Regis Residence	Regis Residence, 2 E 55th St #803, New York, N...	40.761255	-73.974483
1	Brokered by Serhant	Condo for sale	195000000	7	10	17545.0	Central Park Tower Penthouse-217 W 57th New Yo...	New York, NY 10019	Central Park Tower Penthouse-217 W 57th New Yo...	United States	New York	New York County	New York	West 57th Street	217 W 57th St, New York, NY 10019, USA	40.766393	-73.980991
2	Brokered by Sowae Corp	House for sale	260000	4	2	2015.0	620 Sinclair Ave	Staten Island, NY 10312	620 Sinclair AveStaten Island, NY 10312	United States	New York	Richmond County	Staten Island	Sinclair Avenue	620 Sinclair Ave, Staten Island, NY 10312, USA	40.541805	-74.196109
3	Brokered by COMPASS	Condo for sale	69000	3	1	445.0	2 E 55th St Unit 908W33	Manhattan, NY 10022	2 E 55th St Unit 908W33Manhattan, NY 10022	United States	New York	New York County	New York	East 55th Street	2 E 55th St, New York, NY 10022, USA	40.761398	-73.974613
4	Brokered by Sotheby's International Realty - E...	Townhouse for sale	55000000	7	2	14175.0	5 E 64th St	New York, NY 10065	5 E 64th StNew York, NY 10065	United States	New York	New York County	New York	East 64th Street	5 E 64th St, New York, NY 10065, USA	40.767224	-73.969856

Descriptive Analysis

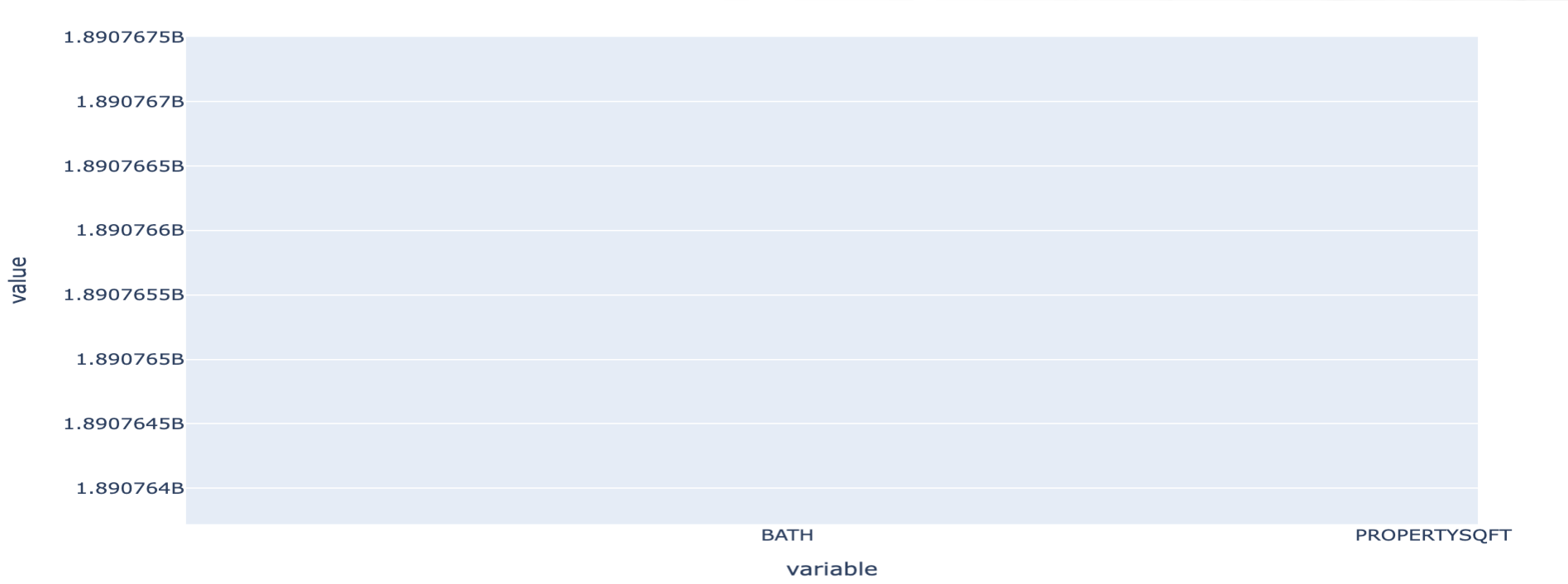
- Calculating summary statistics for numerical columns using 'describe()'.
- Computing the number of unique values in each column using 'nunique()'.

	PRICE	BEDS	BATH	PROPERTYSQFT	LATITUDE	LONGITUDE
count	4.582000e+03	4582.000000	4582.000000	4582.000000	4582.000000	4582.000000
mean	2.381709e+06	3.360978	2.352030	2176.471515	40.713996	-73.941501
std	3.208579e+07	2.624625	1.969932	2283.751257	0.087692	0.101214
min	2.494000e+03	1.000000	0.000000	230.000000	40.499546	-74.253033
25%	4.990000e+05	2.000000	1.000000	1200.000000	40.638651	-73.987387
50%	8.250000e+05	3.000000	2.000000	2184.207862	40.726691	-73.949116
75%	1.495000e+06	4.000000	3.000000	2184.207862	40.771913	-73.869418
max	2.147484e+09	50.000000	50.000000	65535.000000	40.912729	-73.702450

Outlier Detection:

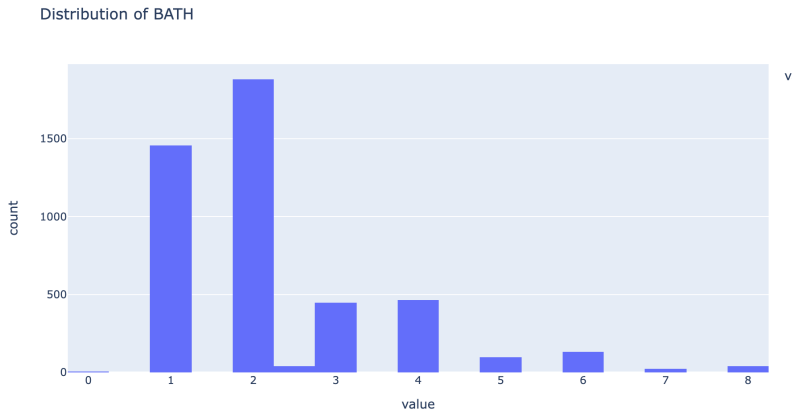
•Visualizing the distribution of data and detecting outliers using box plots.

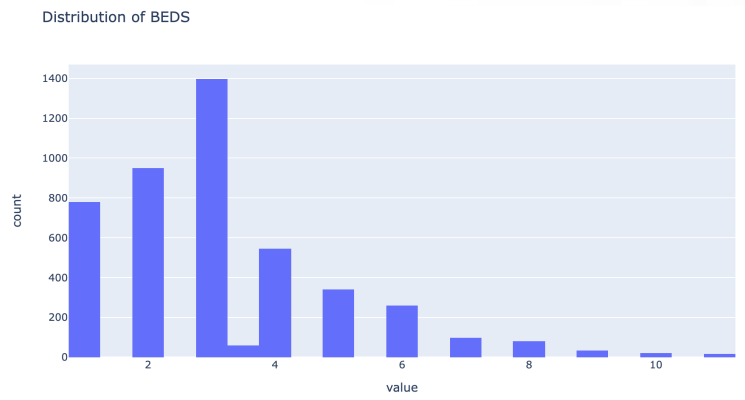
•Handling outliers by replacing them with appropriate values, such as the mean or median.

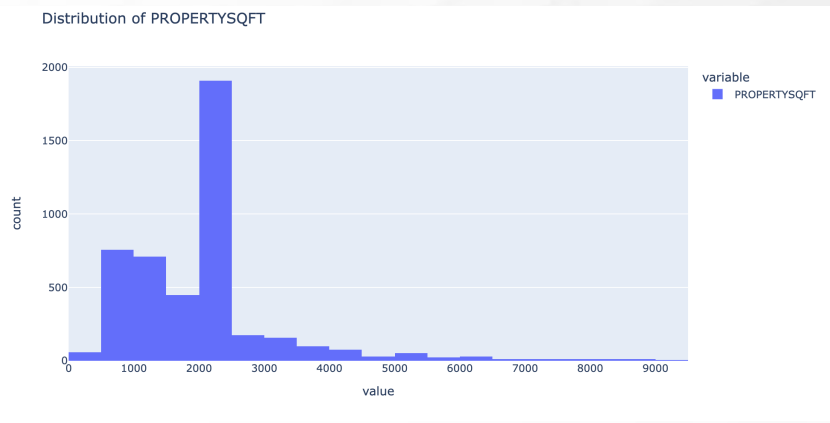


Plotting Distributions:

- Generating histograms and box plots to visually explore the distribution of numerical columns.



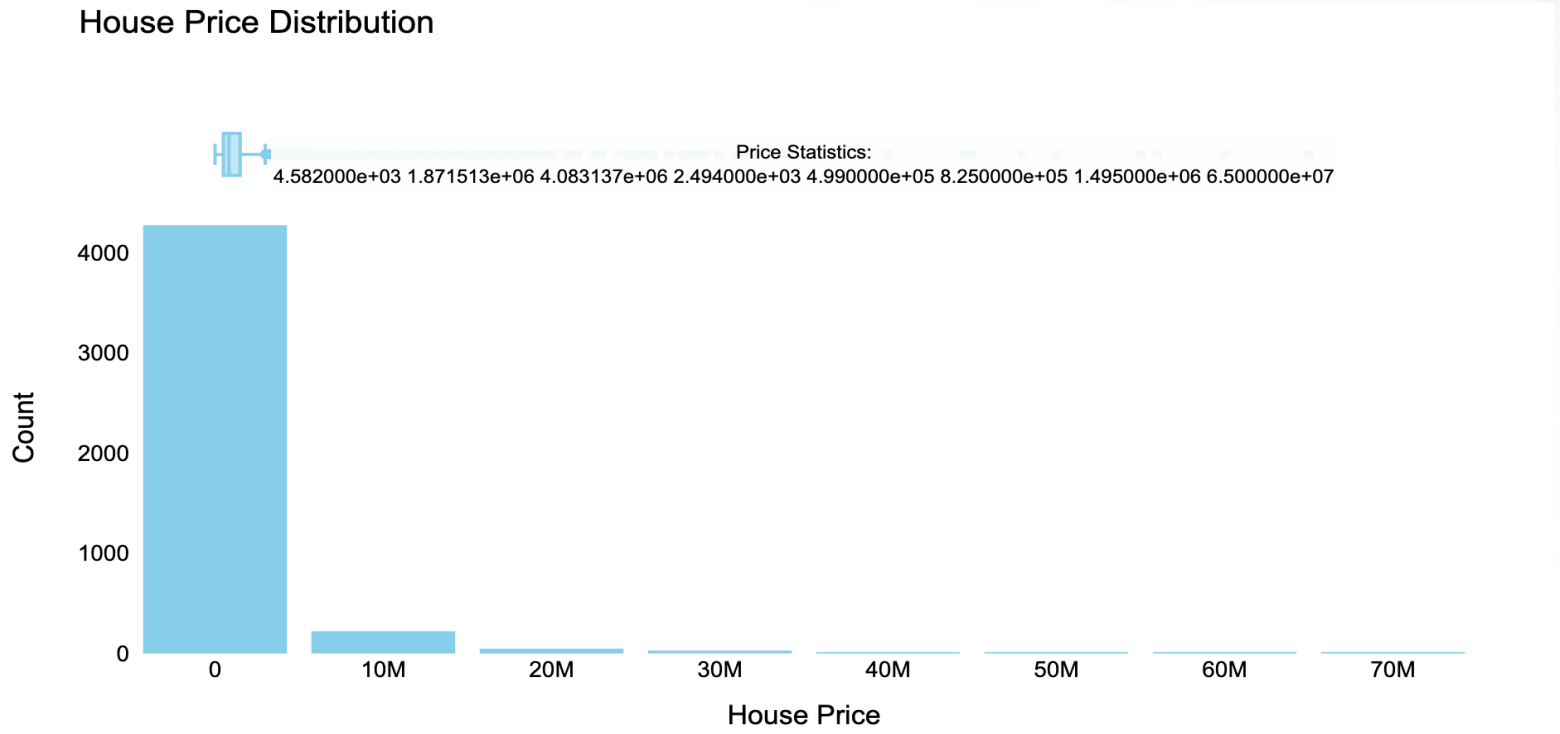




Descriptive Statistics:

•Calculating descriptive statistics, such as mean, median, minimum, maximum, and quartiles, for specific columns (e.g., 'PRICE' column).

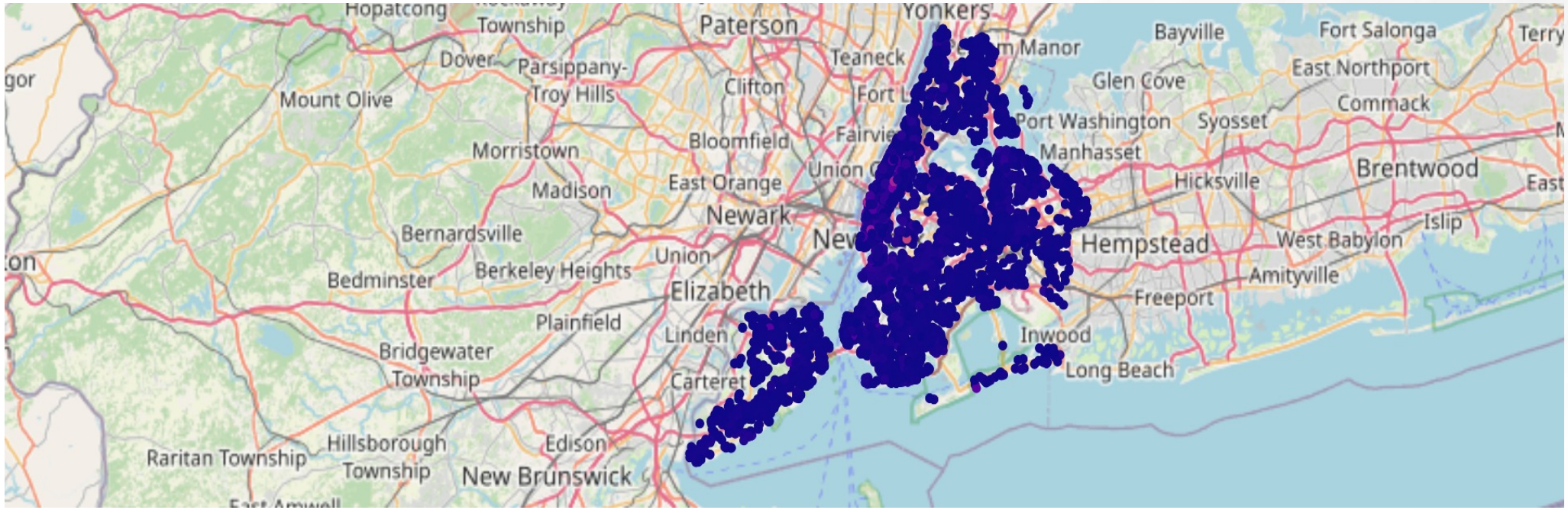
•Visualizing the distribution of house prices using histograms and box plots.



STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

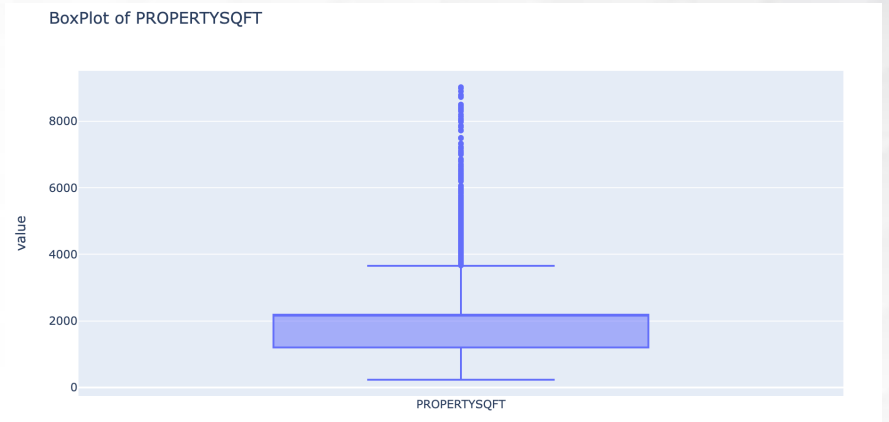
Geographical Distribution Visualization:

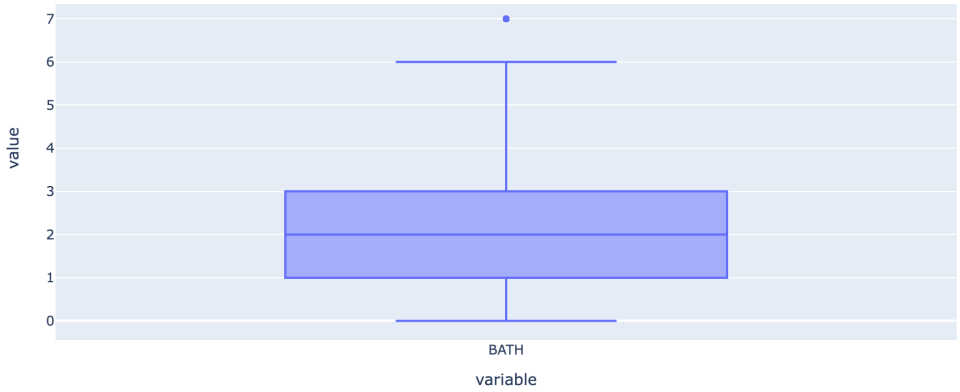
- Visualizing the geographical distribution of real estate properties using scatter plots or maps based on latitude, longitude, and other relevant attributes.

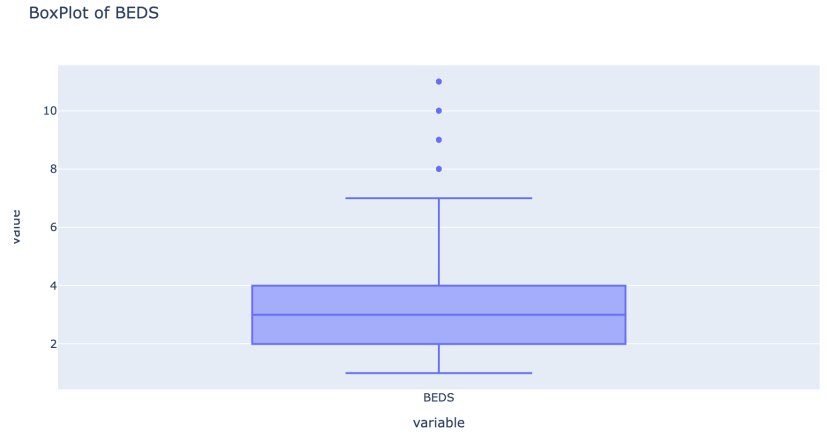


Box plots:

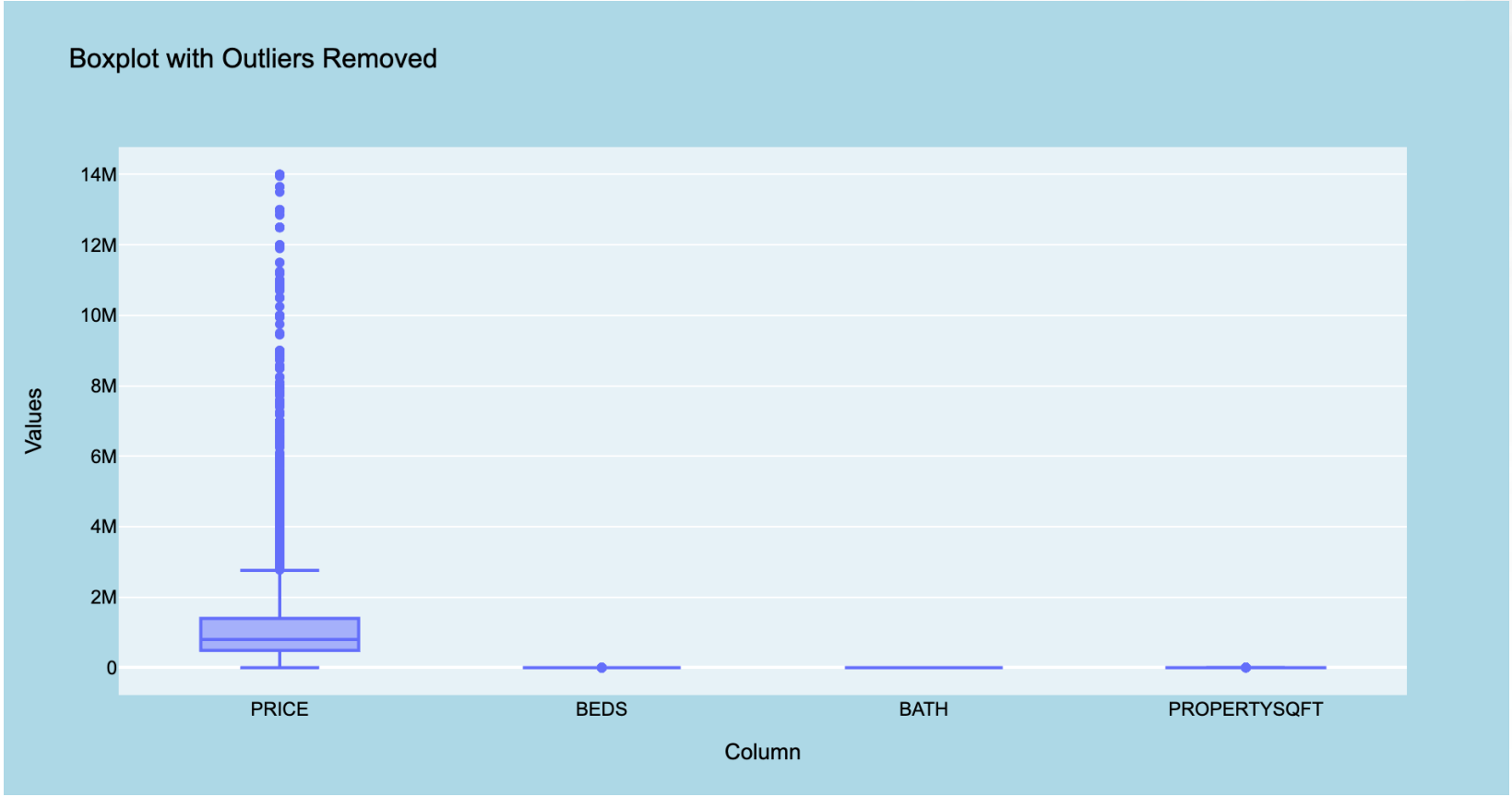

BoxPlot of PRICE
60M
50M
40M
30M
20M
10M
0
value
PRICE
variable


BoxPlot of PROPERTYSQFT
8000
6000
4000
2000
0
value
PROPERTYSQFT


7
6
5
4
3
2
1
0
value
BATH
variable


BoxPlot of BEDS
10
8
6
4
2
BEDS
variable

Box plots with outliers Removed:

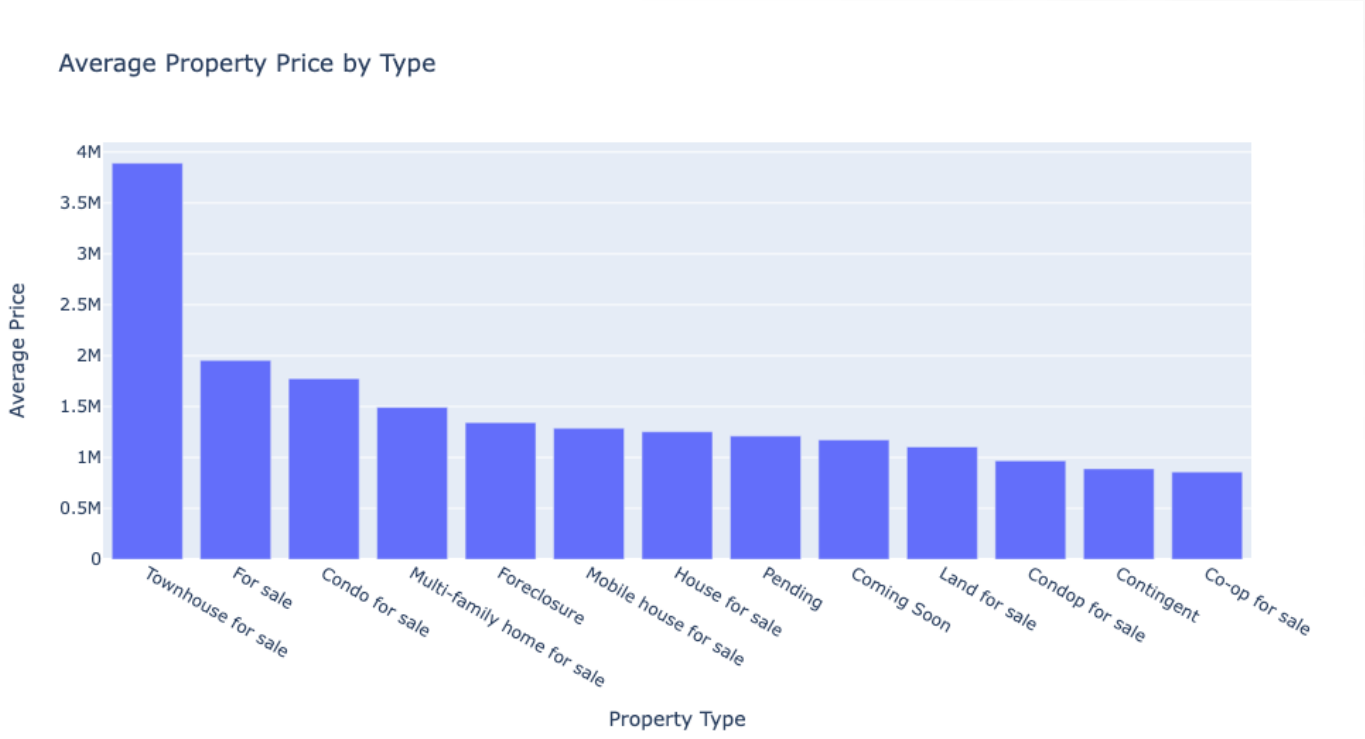


Top Brokers Analysis:

- Analyzing and visualizing insights related to top brokers, such as the average property price associated with each broker.

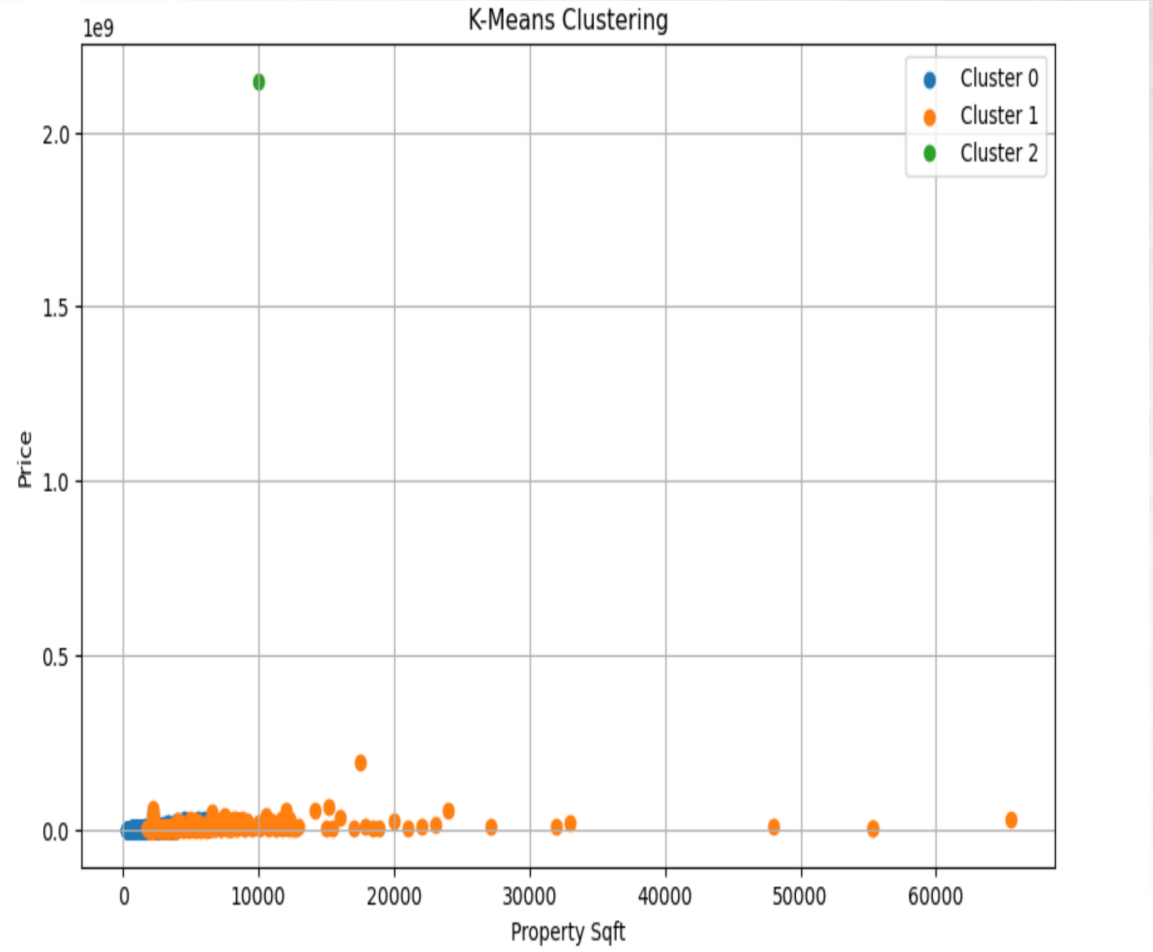
Property Type Analysis:

•Analyzing and visualizing insights related to different property types, such as the average property price for each property



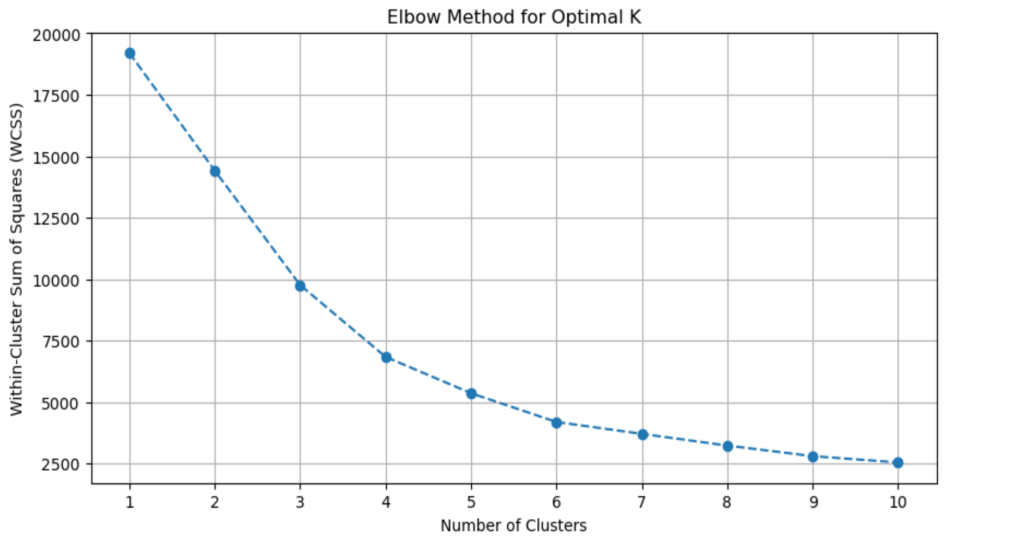
K-Means Clustering:

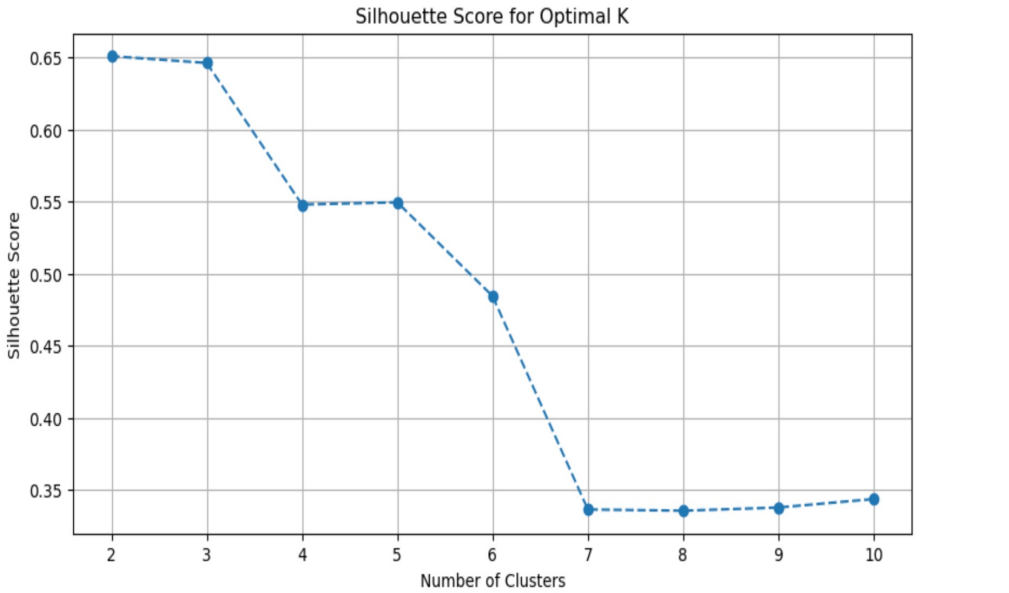
- This graph displays the results of a K-Means clustering analysis on real estate data, where properties are grouped based on square footage and price.
- We used the K-Means algorithm with three clusters, initialized with a random state of 42 for reproducibility. The data was scaled to normalize the scale of square footage and price.
- You can see that most properties are grouped in Cluster 1 as lower-cost properties with smaller sizes. The outlier in Cluster 0 represents a high-value property with significant square footage.
- The orange dots represent Cluster 1, the blue dot is Cluster 0, and the green dot signifies Cluster 2.



Elbow Method and Silhouette Score for Optimal K:

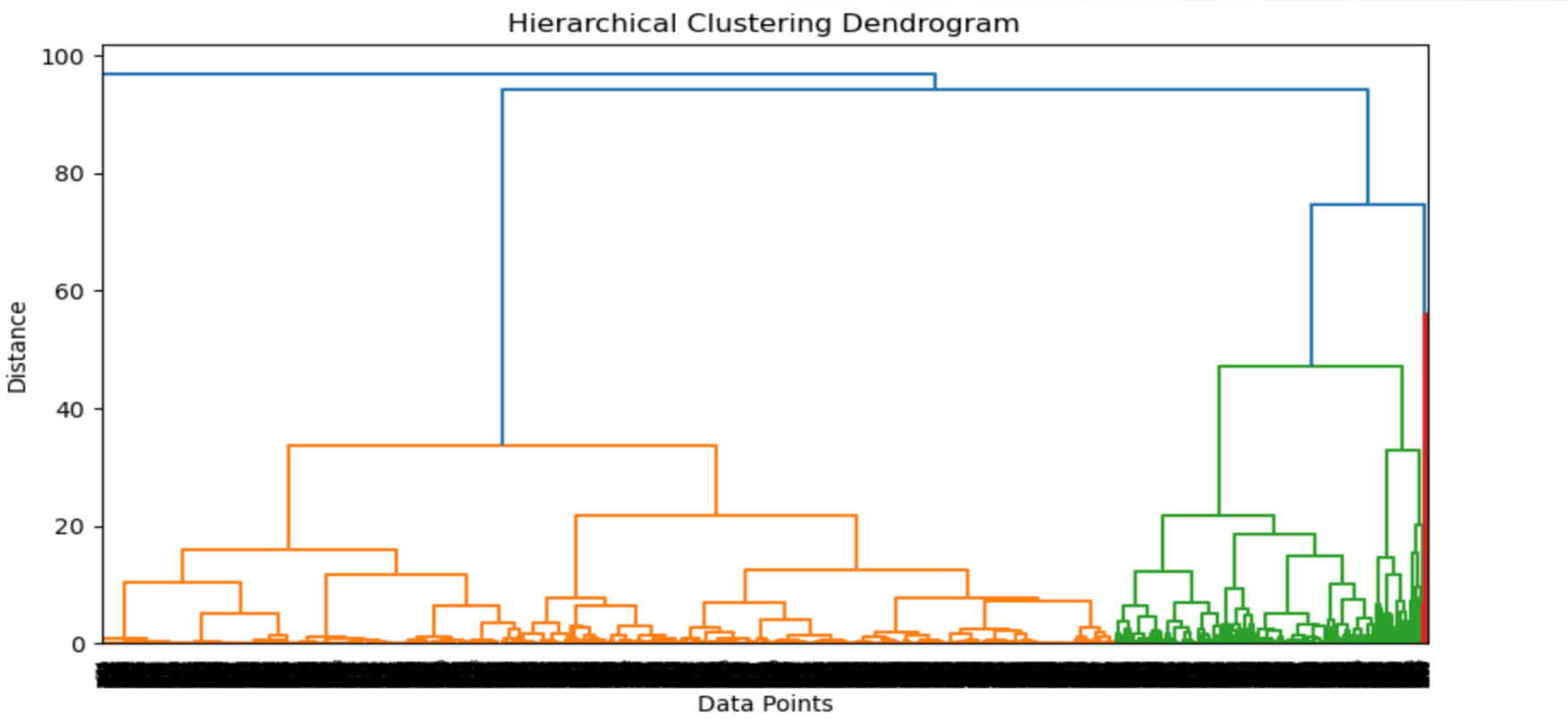
- The Elbow Method graph helps us determine the optimal number of clusters by showing the Within-Cluster Sum of Squares (WCSS).
- Notice the elbow at K=3, indicating that additional clusters beyond three yield diminishing improvements in WCSS, which measures compactness of the clusters.
- Simultaneously, the Silhouette Score provides a measure of how similar an object is within-cluster compared to other clusters.
- A higher silhouette score near K=2 and K=3 suggests good cluster separation and cohesion. The score decreases as more clusters are added, indicating poorer clustering quality





STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Hierarchical Clustering:



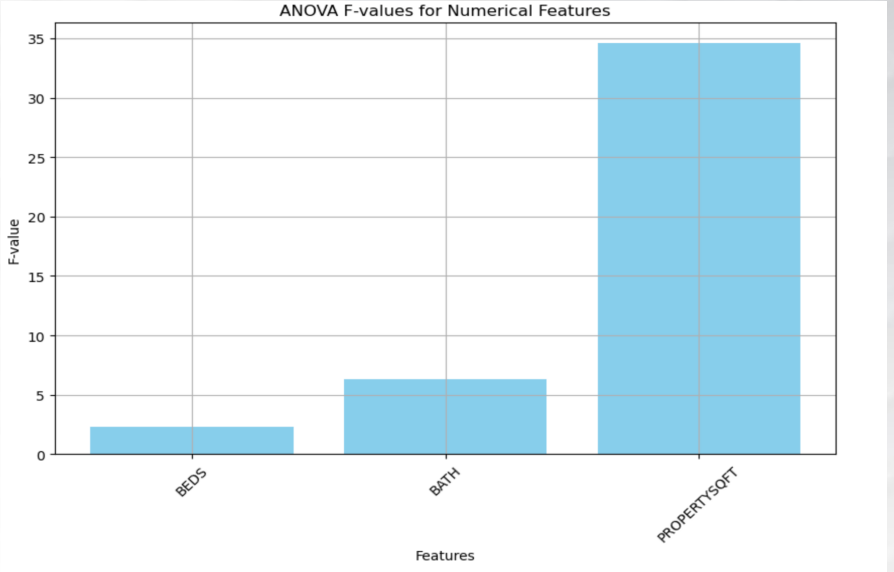
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

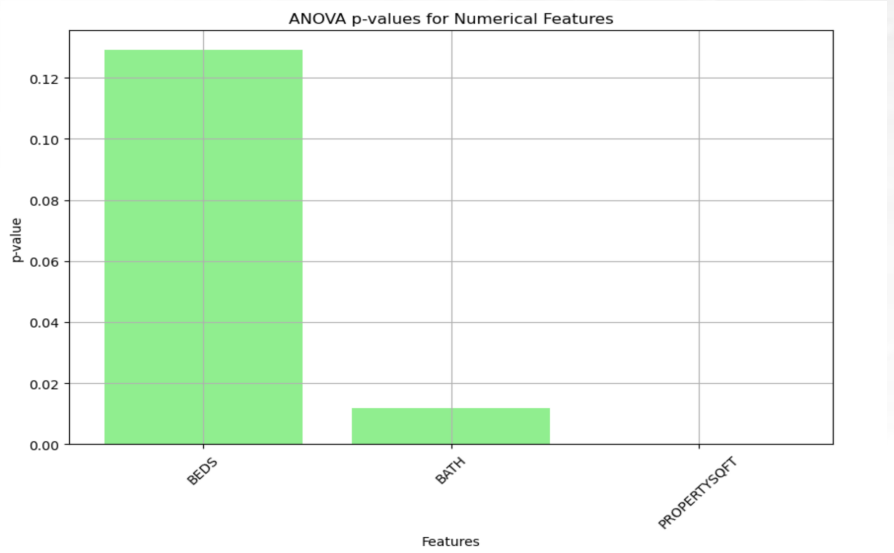
ANOVA Table:

	sum_sq	df	F	PR(>F)
BEDS	2.236459e+15	1.0	2.304786	1.290420e-01
BATH	6.131779e+15	1.0	6.319113	1.197682e-02
PROPERTYSQFT	3.354938e+16	1.0	34.574357	4.380698e-09
Residual	4.654790e+18	4797.0	NaN	NaN

ANOVA Results :

- We performed ANOVA to analyze the impact of various features like beds, baths, and property square footage on property prices.
- The F-values represent the ratio of variance between groups to the variance within groups, where higher values indicate a significant impact on prices.
- Property square footage has the highest F-value, suggesting a strong influence on the price variance among properties.
- Lower p-values indicate statistical significance. For property square footage, the p-value is extremely low, underscoring its significance in price determination.





Conclusion:

Model performance: To evaluate the performance for our model we used root mean squared error which is a metric used for evaluating the performance of trained model on our data set.
The root mean squared error value for our model is **423806.1027852285**

Predicting price: We have five key features in determining the house prices which are latitude, longitude, beds, bath, propertysqft. Using these values we can determine the predicted price of the houses.