New york HOUSING DATA 2022

Programming for Data Science Project Report Under Guidance of Prof. Jason Parker

Abstract

Analysis of New York Real Estate Sales data of 2022   
Proposal  
Data source  
 Exploratory data analysis  
 Model variety  
Accuracy/MSE

Rushad Dhanore

rxd220036@utdallas.edu

NAME: RUSHAD DHANORE

NET ID: RXD220036

New York Housing Data 2022

This dataset contains prices of New York houses, providing valuable insights into the real estate market in the region. It includes information such as broker titles, house types, prices, number of bedrooms and bathrooms, property square footage, addresses, state, administrative and local areas, street names, and geographical coordinates.

**Key Features:**

* **BROKERTITLE**: *Title of the broker*
* **TYPE**: *Type of the house*
* **PRICE**: *Price of the house*
* **BEDS**: *Number of bedrooms*
* **BATH**: *Number of bathrooms*
* **PROPERTYSQFT**: *Square footage of the property*
* **ADDRESS**: *Full address of the house*
* **STATE**: *State of the house*
* **MAIN\_ADDRESS**: *Main address information*
* **ADMINISTRATIVE\_AREA\_LEVEL\_2**: *Administrative area level 2 information*
* **LOCALITY**: *Locality information*
* **SUBLOCALITY**: *Sublocality information*
* **STREET\_NAME**: *Street name*
* **LONG\_NAME**: *Long name*
* **FORMATTED\_ADDRESS**: *Formatted address*
* **LATITUDE**: *Latitude coordinate of the house*
* **LONGITUDE**: *Longitude coordinate of the house*

This comprehensive dataset serves as a rich resource for analyzing and understanding the dynamics of the New York real estate market. Encompassing a diverse range of information, it delves into the intricacies of property transactions and showcases a myriad of factors influencing housing prices.

In addition to broker titles, house types, prices, and the number of bedrooms and bathrooms, the dataset furnishes detailed insights into the physical attributes of the properties. This includes information on property square footage, enabling a nuanced exploration of size-based pricing trends. The dataset goes even further by providing essential details such as addresses, state, administrative and local areas, street names, and geographical coordinates. This holistic approach allows for a comprehensive spatial analysis, identifying correlations between location, amenities, and property values. The objective of this project is to develop predictive models for estimating house prices in New York using a comprehensive dataset containing various attributes such as house types, square footage, number of bedrooms and bathrooms, and geographical coordinates. By leveraging machine learning techniques, we aim to provide valuable insights into the factors influencing housing prices in the region.

Researchers and analysts can leverage this dataset to uncover patterns and trends in the New York housing market, facilitating informed decision-making for various stakeholders. Whether it be prospective buyers seeking to understand market trends or real estate professionals aiming to refine pricing strategies, this dataset offers a multifaceted perspective on the dynamic real estate landscape in New York. Its depth and breadth make it an invaluable resource for anyone interested in gaining a thorough understanding of the factors shaping the housing market in this vibrant and diverse region.

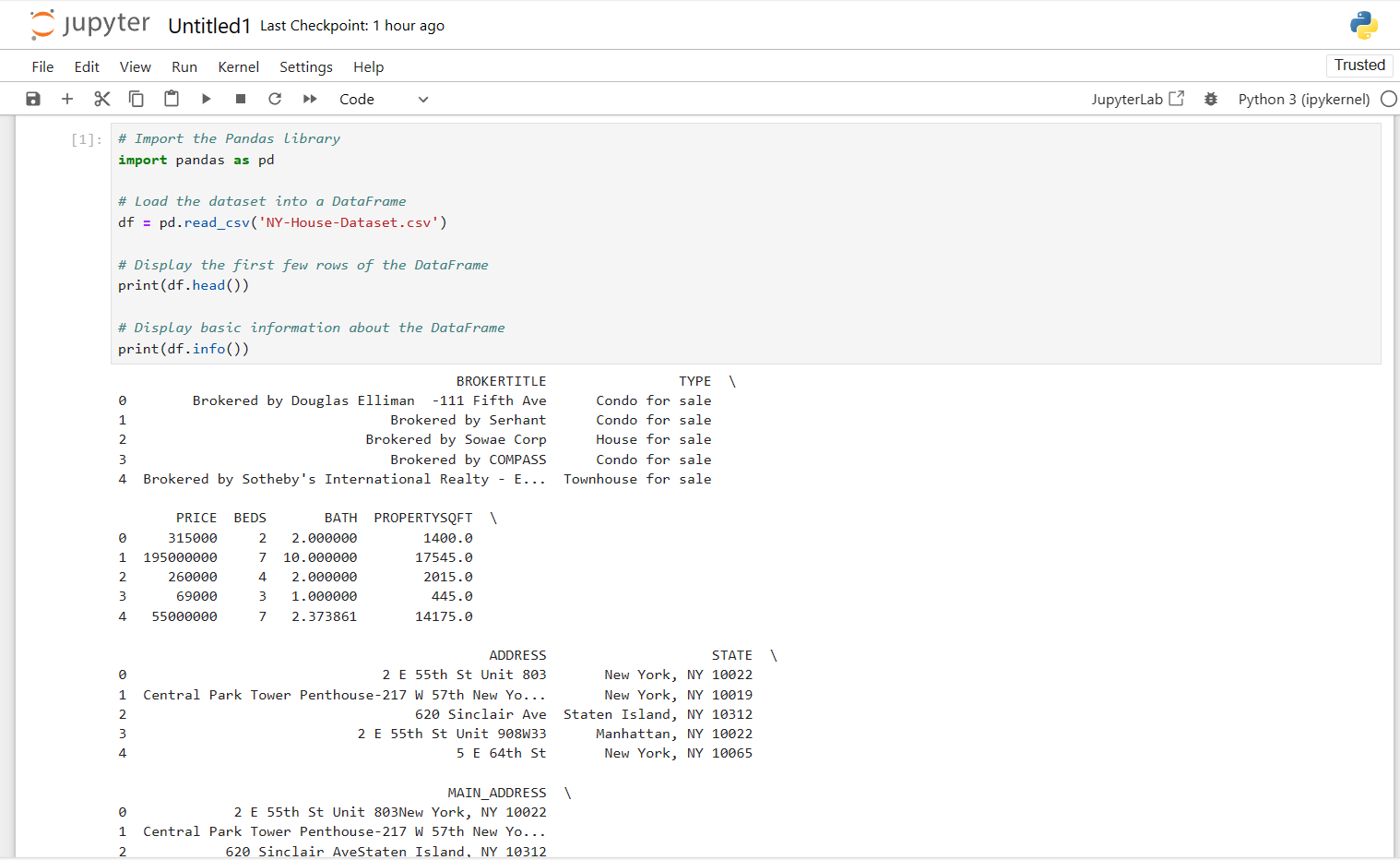
I got this dataset from Kaggle website and here is link to it [New York Housing Market (kaggle.com)](https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market?resource=download).

**Exploratory Data Analysis (EDA)**

EDA involves examining and summarizing the main characteristics of your dataset to gain insights and identify any patterns or anomalies.

**Load the Dataset:**

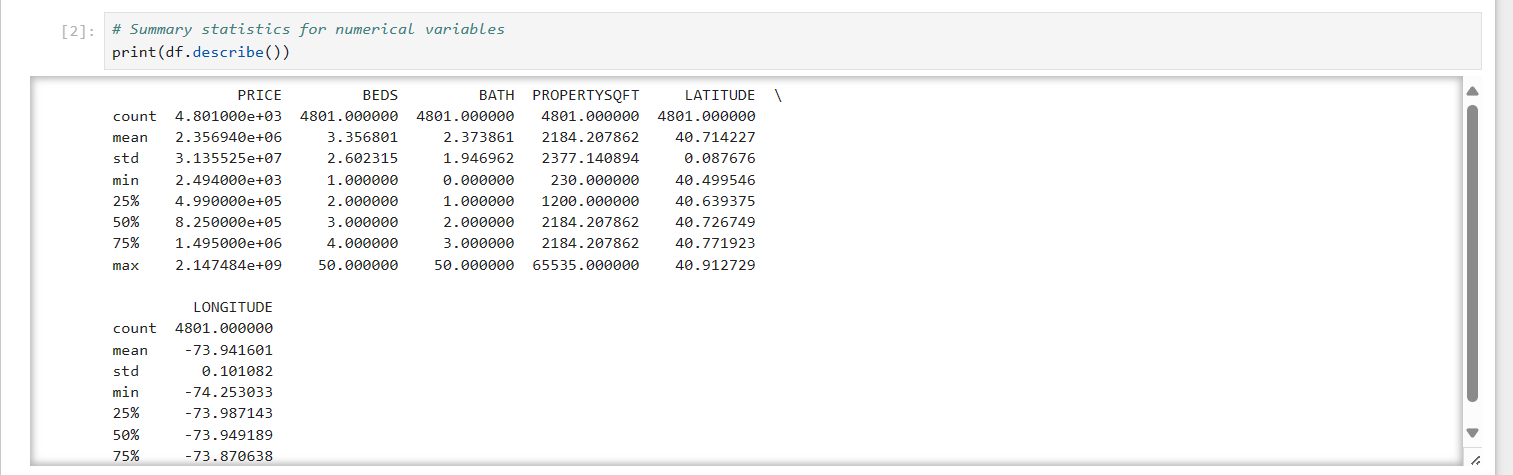
* Begin by loading the dataset into your preferred data analysis tool or programming environment using Python with libraries like Pandas and Matplotlib/Seaborn.



* After running this script, I have successfully loaded your dataset into Python, and can proceed with the next steps of the exploratory data analysis.

**Summary Statistics:**

* Calculate summary statistics for numerical variables such as PRICE, BEDS, BATH, and PROPERTYSQFT. This includes measures like mean, median, standard deviation, minimum, and maximum values.



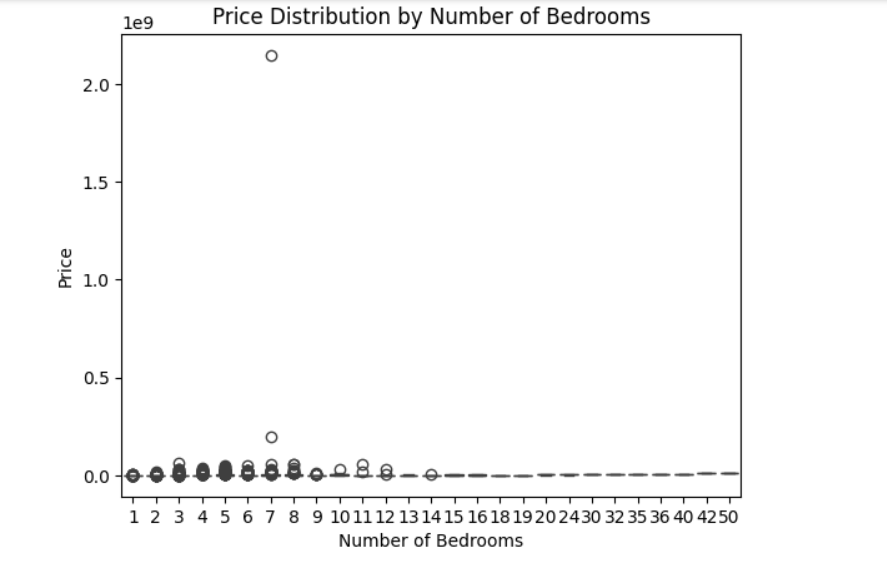
**Data Visualization:**

* Create visualizations to understand the distribution of numerical variables and uncover any patterns or outliers.

A white background with text

Description automatically generated

* **Boxplots**: Construct boxplots to identify outliers and understand the spread of numerical variables.

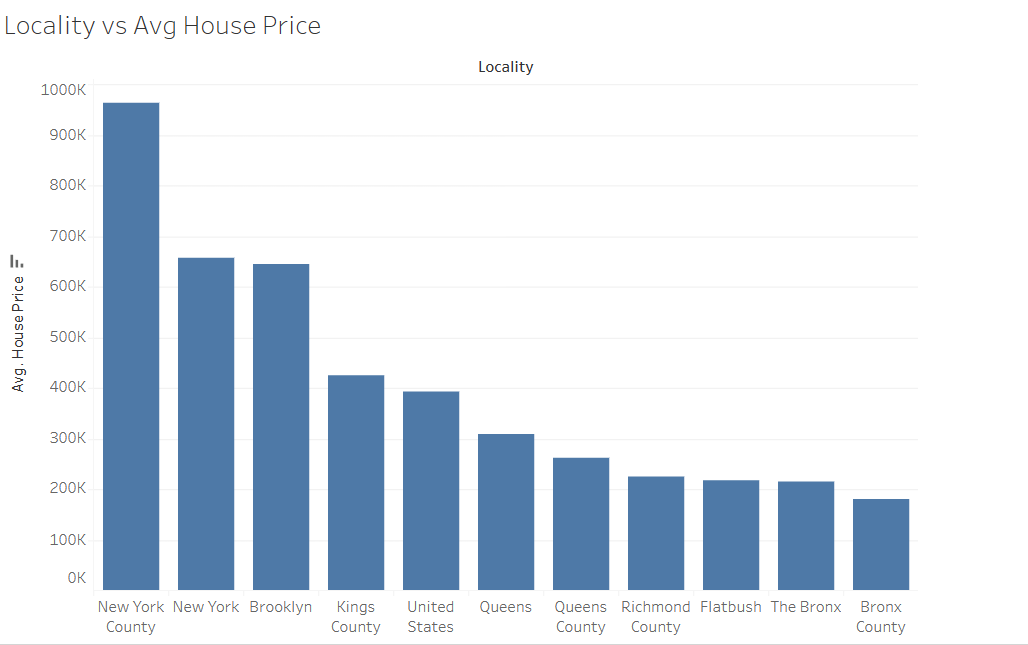


* **Scatter** **plots**: Explore relationships between variables by creating scatter plots. For example, you could plot PRICE against PROPERTYSQFT to see if there's a correlation between house size and price.

A graph with green dots

Description automatically generated

* Histograms: Plot histograms to visualize the distribution of continuous variables like PRICE, BEDS, BATH, and PROPERTYSQFT.



**Handling Missing Values:**

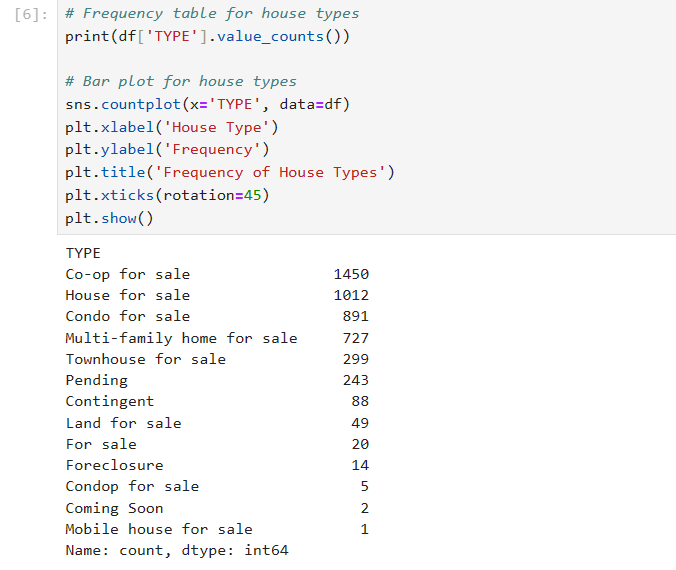
* Check for missing values in the dataset and decide on an appropriate strategy for handling them.

A screenshot of a computer

Description automatically generated

**Categorical Variables:**

* Explore categorical variables like TYPE, STATE, and LOCALITY using frequency tables or bar plots.



A graph of a house type

Description automatically generated

**Correlation Analysis:**

* Calculate correlation coefficients between numerical variables to identify any linear relationships.

A screenshot of a computer program

Description automatically generated

A diagram of a heatmap

Description automatically generated

**Model Variety**

* Explore a variety of models suitable for dataset. Since this is a housing dataset, regression models might be particularly useful for predicting house prices. I have also consider other types of models such as decision trees, random forests, support vector machines, or neural networks.
* For each model, I have split dataset into training and testing sets to evaluate performance. Exploring model variety involves trying out different types of models to see which ones perform best for your specific dataset.

**Selecting Models:**

* Since I am working with a housing dataset, regression models are a natural choice for predicting house prices. Some common regression models I considered include:

Linear Regression

Decision Trees

Random Forests

Support Vector Machines (SVM)

**Data Preprocessing:**

* Before fitting the models, pre-processed the data . This included handling missing values, encoding categorical variables, scaling numerical features, and splitting the dataset into training and testing sets.

A screenshot of a computer program

Description automatically generated

**Model Evaluation (Accuracy/MSE)**

To Calculate evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared (R2) for each model and compare the performance of different models based on these metrics.

A screenshot of a computer program

Description automatically generated

A white screen with black text

Description automatically generated

**Key Insights:** Through our analysis, several key insights were uncovered:

1. Square footage and the number of bedrooms are strong predictors of house prices, indicating that larger properties tend to command higher prices.
2. Location also plays a significant role in determining house prices, with properties in certain areas of New York commanding premium prices.
3. Certain house types, such as single-family homes, tend to have higher average prices compared to other types such as condominiums or townhouses.
4. The choice of broker may also influence house prices, with properties listed by certain brokers fetching higher prices on average.

