**HOUSE PRICE PREDICTION**

**Problem Statement: To Predict the housing price based on certain factors like house area, bedrooms, furnished, nearness to main road, etc.**

**Dataset Overview:**

The dataset consists of 545 entries and 13 features.

**Price:** This typically represents the price of the property in the local currency (e.g., dollars, rupees).

**Area**: Refers to the total area of the property in square feet or square meters.

**Bedrooms**: Indicates the number of bedrooms in the property.

**Bathrooms:** Specifies the number of bathrooms in the property.

**Stories:** Represents the number of floors or stories in the building.

**Main road:** A binary (0 or 1) indicator showing if the property has direct access to a main road.

**Guestroom:** A binary indicator showing if the property has a guest room.

**Basement:** A binary indicator showing if the property has a basement.

**Hot water heating:** A binary indicator showing if the property has hot water heating.

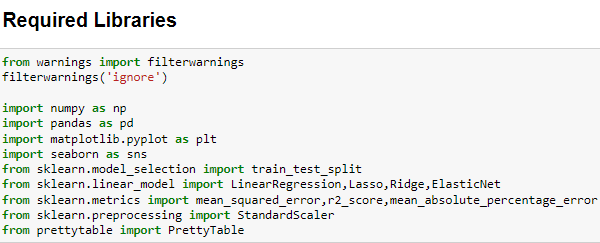
**Air conditioning:** A binary indicator showing if the property has air conditioning.

**Parking**: Specifies the number of parking spots available on the property.

**Prefarea:** A binary indicator showing if the property is located in a preferred location.

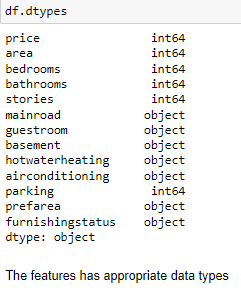
**Furnishingstatus:** Describes the furnishing status of the property (e.g., furnished, semi-furnished, unfurnished).

**Imported Necessary Libraries:**

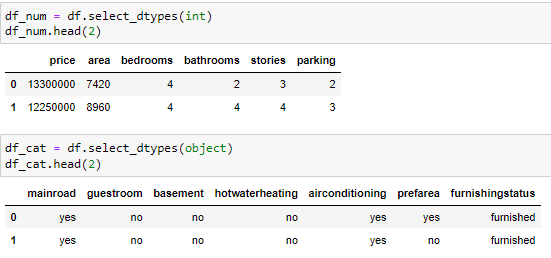


**Exploratory Data Analysis**

**Check for the data types:**

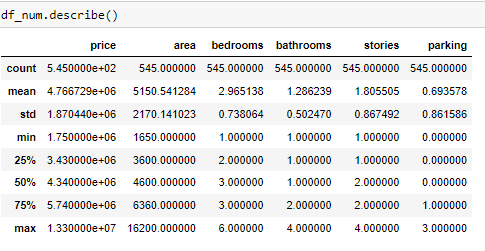


**Distinguishing between categorical and numerical data:**



**Summary Statistics:**

The statistical summary for the numerical data.



Based on the summary, the average price is $4.77, the average house area is 5,150 square feet, and typically, there are 2 bedrooms, 1 bathroom, and 1 story, with no parking available.

**Encoding:**

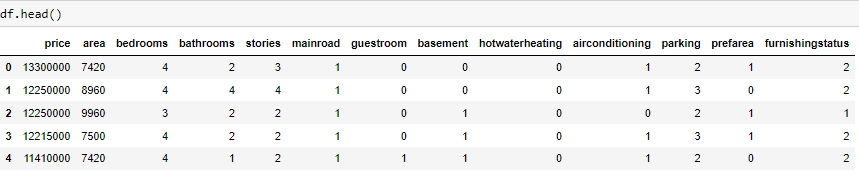
Encoding is a technique used to convert categorical data into numerical data used in machine learning models.

Note: Instead of using (n-1) dummy encoding, I'm opting for the `replace` function.

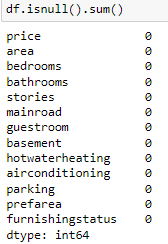
It's computationally efficient and works well with lower-dimensional data.



**Outcomes after applying the replace function.**



**Checking for missing values:**

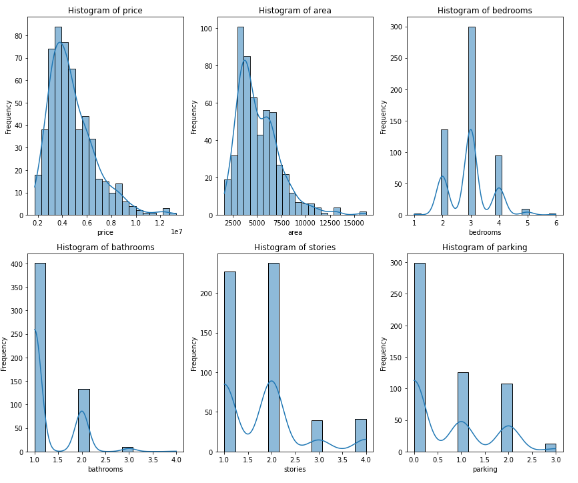


The data is very clean, there are no missing values present in the data.

**Univariate Analysis:**

**--Numerical Columns**

* **Histogram**



**Histogram can be useful for-**

Market Insights: Histograms provide insights into market segmentation and pricing trends.

Decision Making: Useful for pricing strategies, market positioning, or identifying investment opportunities.

**Interpretation:**

Price: Most house prices fall within the range of four million dollars.

Area: Most houses range from 2500 sq feet to 5000 sq feet in size.

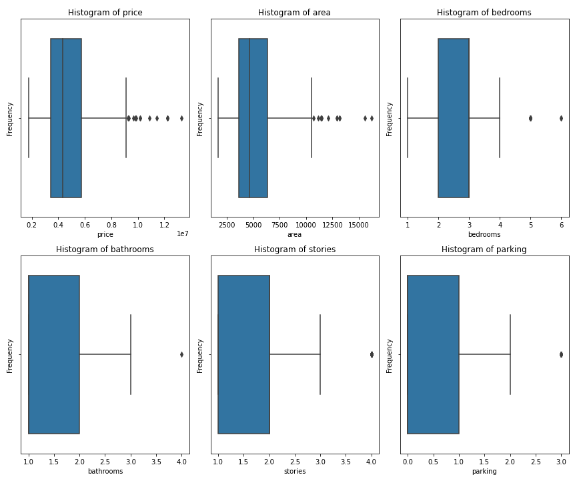
Bedrooms: Most houses have three bedrooms.

Bathrooms: Most houses have one bathroom.

Stories: Most houses are one to two stories tall.

Parking: Most houses do not have a designated parking area.

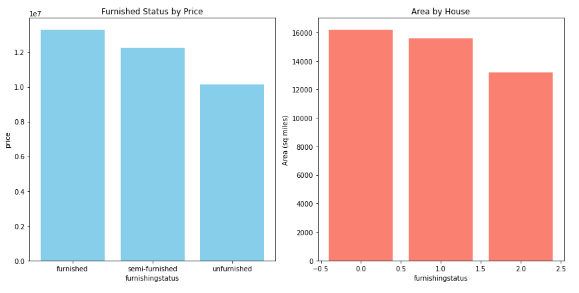
* **Boxplot**



There are outliers present in the data

**--Categorical Columns**

* **Bar plot**

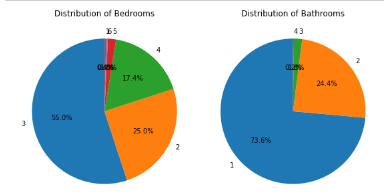


**Interpretation:**

Higher-priced houses are furnished.

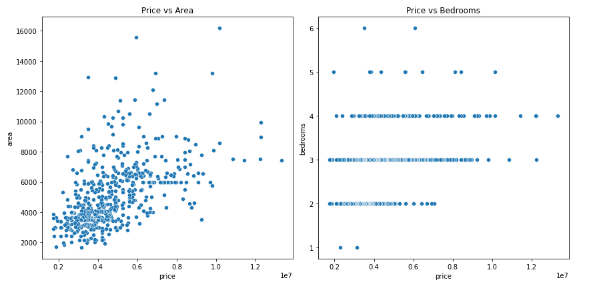
Furnished options are available for larger houses.

* **Pie-Chart**

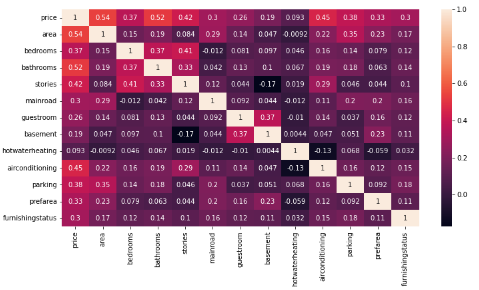


**Bivariate Analysis:**

* **Scatter Plot – Price vs Area and Price vs Bedrooms**



**Multivariate Analysis:**



**Base Model Evaluation**

**Initial Linear Regression Model**

The base model employed for initial evaluation was a linear regression model without any preprocessing steps such as data scaling or outlier removal. This approach aimed to establish a benchmark performance to assess subsequent improvements. The models yielded an R-squared score of 65% of the variance in the target variable, which is moderate. MAPE (Mean absolute percentage error) is 17%, suggesting potential overfitting due to the model's performance on training set.

**Exploration of Data Preprocessing techniques**

To mitigate potential overfitting observed in the base model, various preprocessing techniques were explored. This included feature scaling to enhance the model’s generalization capabilities.

**Ridge and Elastic Net Regression**

There is no change in the metrics. MAPE is less, and accuracy of the model is 65%. An accuracy of 0.659 (or 65.9%) indicates that model's predictions are reasonably close to the actual housing prices, but there is room for improvement. MAPE measures the average percentage difference between model's predicted housing prices and the actual housing prices. A MAPE of 0.1744 (17.44%) suggests that, on average, model's predictions deviate from the actual housing prices by approximately 17.44%.

Note: After performing outlier treatment, we achieved an R-squared of 100%, indicating complete overfitting of the data, so I did not include it in the preprocessing.

**Final Evaluation**

After thorough evaluation, the Ridge regression model scaled input data was chosen as the final model. The decision was based on its ability to maintain strong predictive performance while mitigating overfitting issues observed in earlier stages. The model’s performance metrics and stability make it suitable for eligible house price prediction based on dataset attributes,

