Multiple Linear Regression

Shayan Anwar

27027

Python Code:

<https://colab.research.google.com/drive/1xlQc8ECAjzN60fdU6pzi5a3uqKp1Dyfu?usp=sharing>

(feature Scaling + outlier treatment by z-score) (For this pdf)

# Aim and Objective:

→ Apply Multiple Linear Regression Model on Housing dataset and train the model on the dataset to eventually predict results on testing data

# All the Processes Involved:

## → Data Preprocessing:

* What’s there in the data?
* Looking for Duplicate and Null values and treating them.
* Converting Categorical features to Numerical features to be used in Regression
* Outlier Detection and Treatment, especially for Area and Price
* Feature scaling
* Data Description
* Correlation and VIF

## → Exploratory Data Analysis with Visualizations:

* Distribution Plots
* ScatterPlots
* RelationalPlots
* SwarmPlots
* CountPlots

## → Linear Regression:

* Statistical Regression
* Regression by Sklearn
* Model Testing
* Residual Analysis
* Model Evaluation

## 

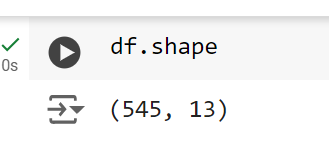
## 

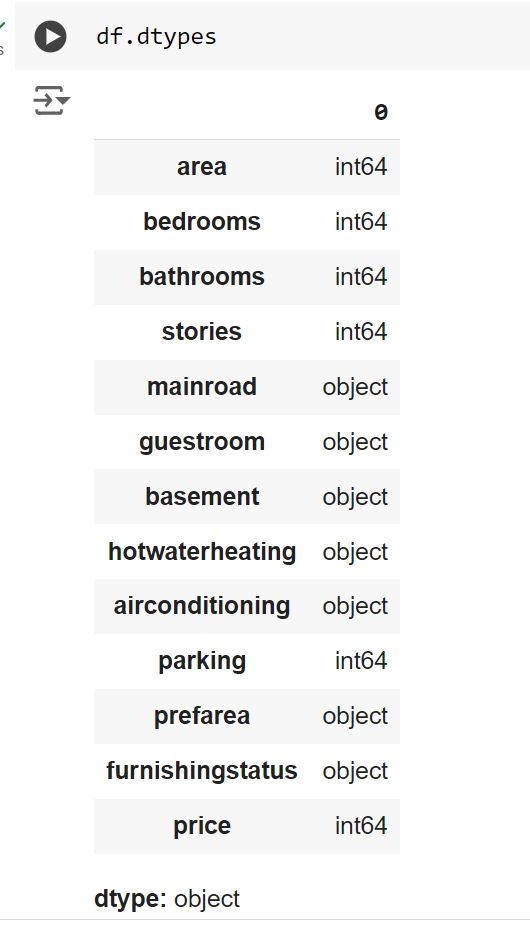
## 

## **Data Preprocessing**

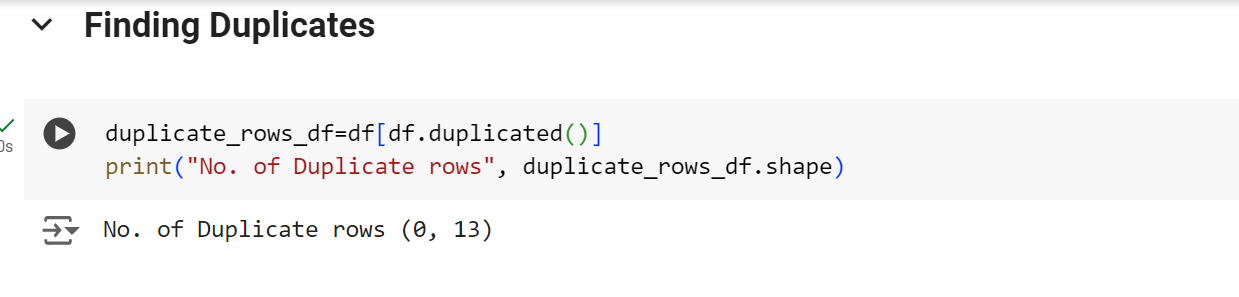
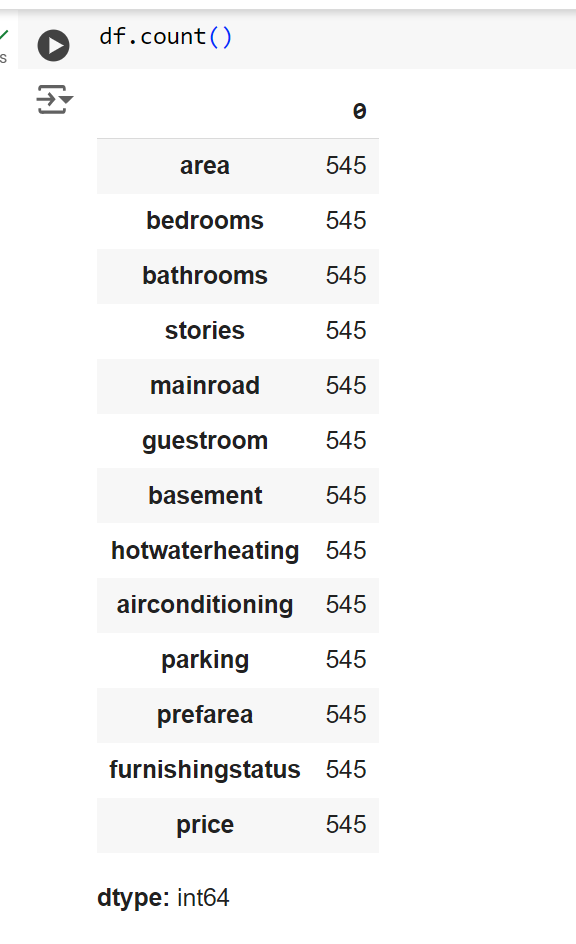
### Exploring Data:

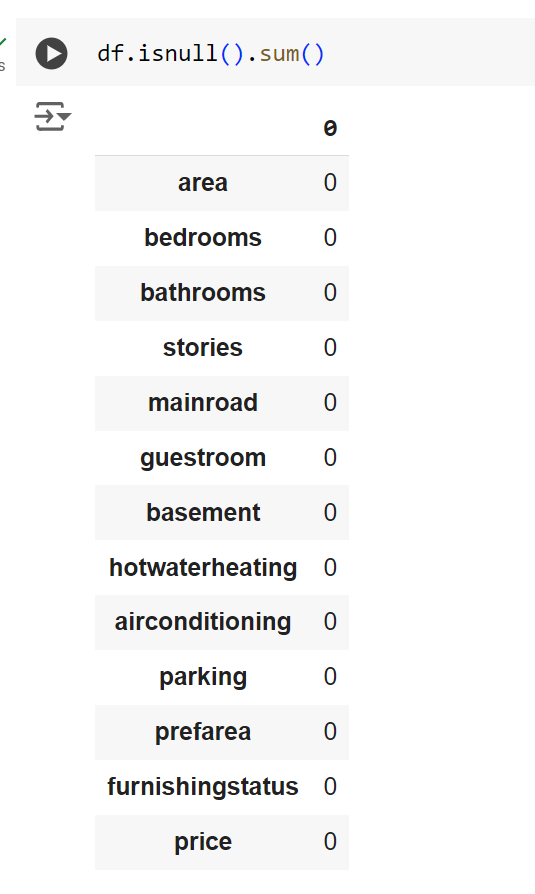






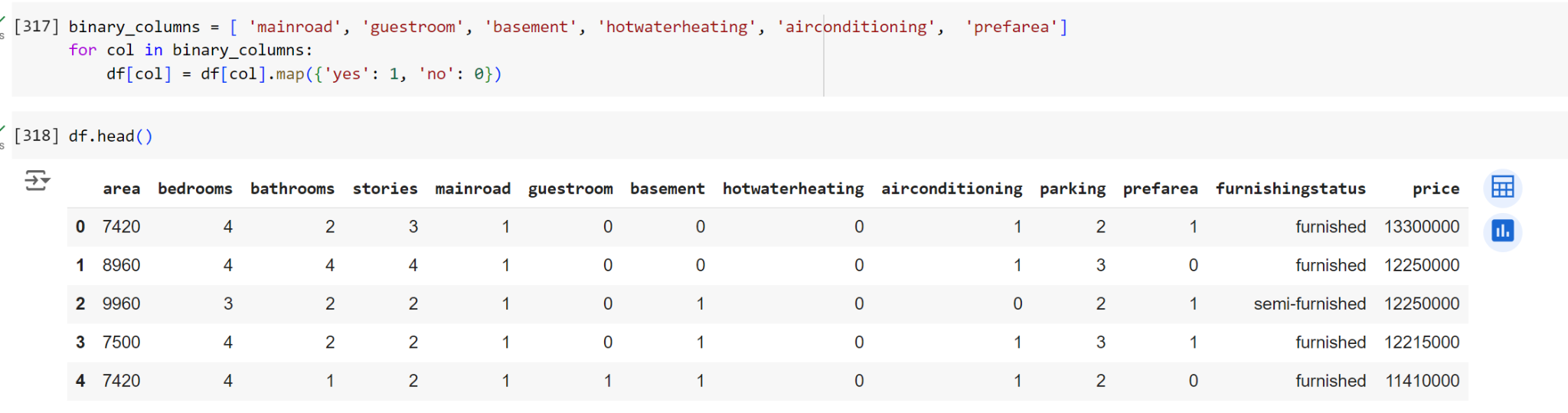
### Duplicates and Null Values

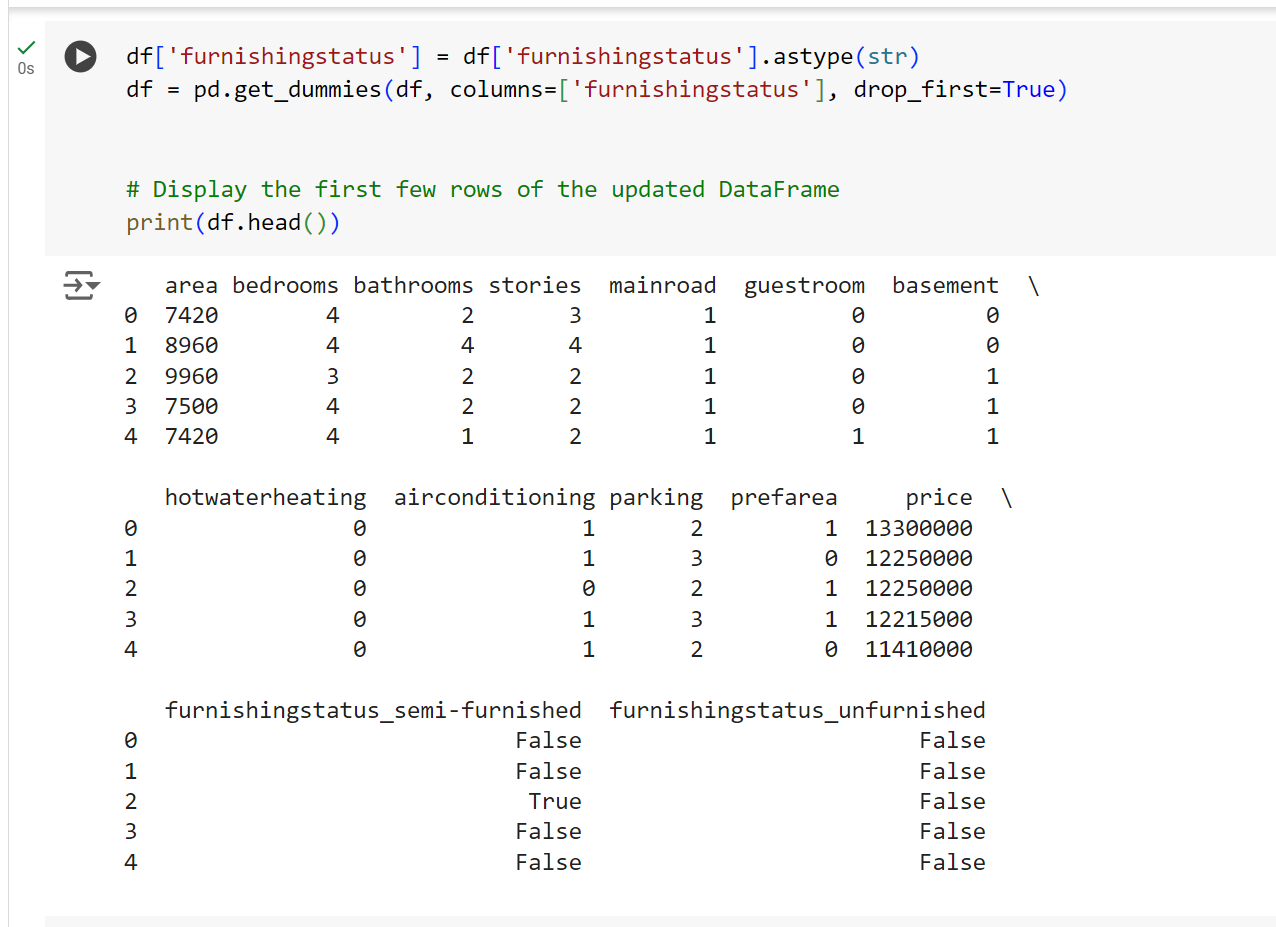


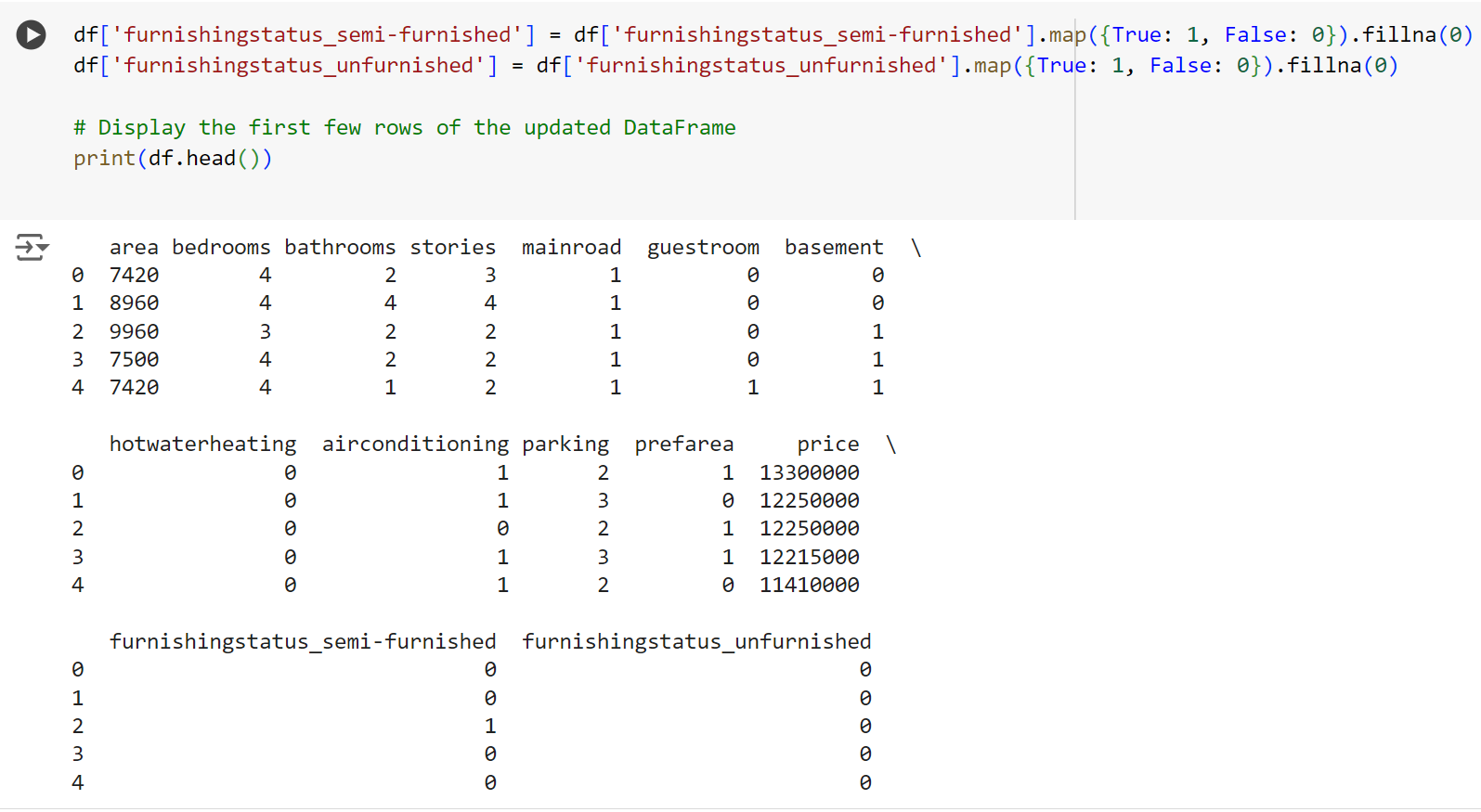


No null and Duplicate Values available in data so no need to clear them or replace them with some other appropriate values

### Working with Categorical Features and preparing for Linear Regression:





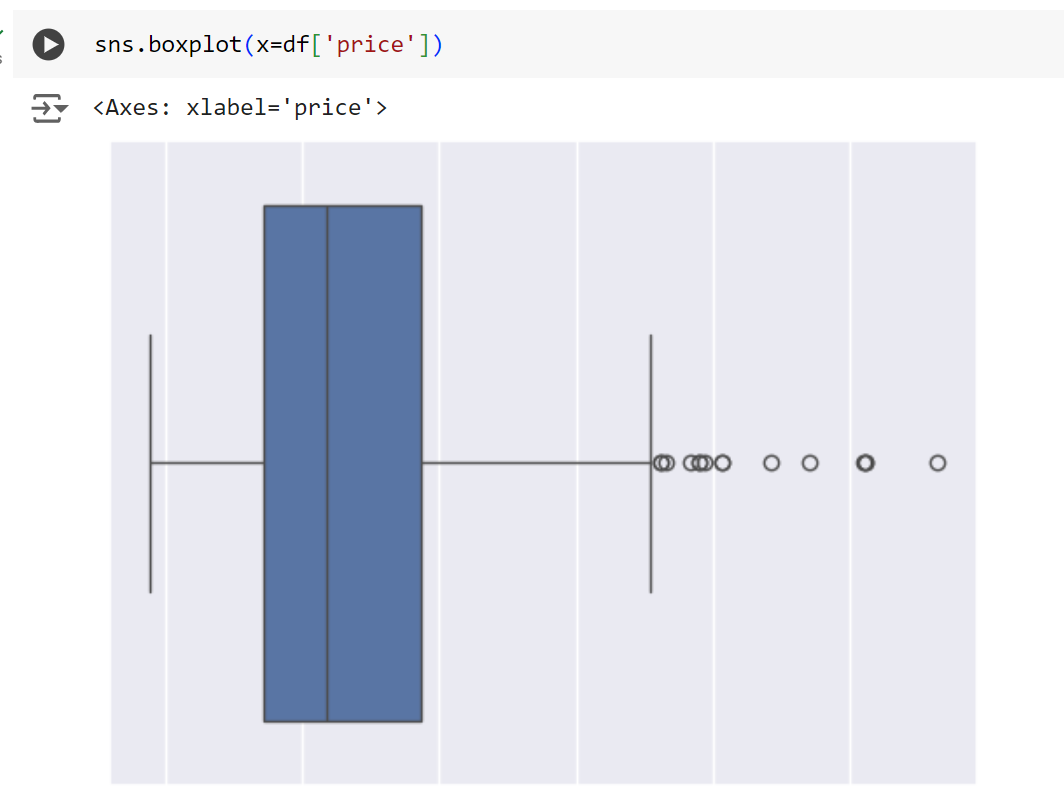


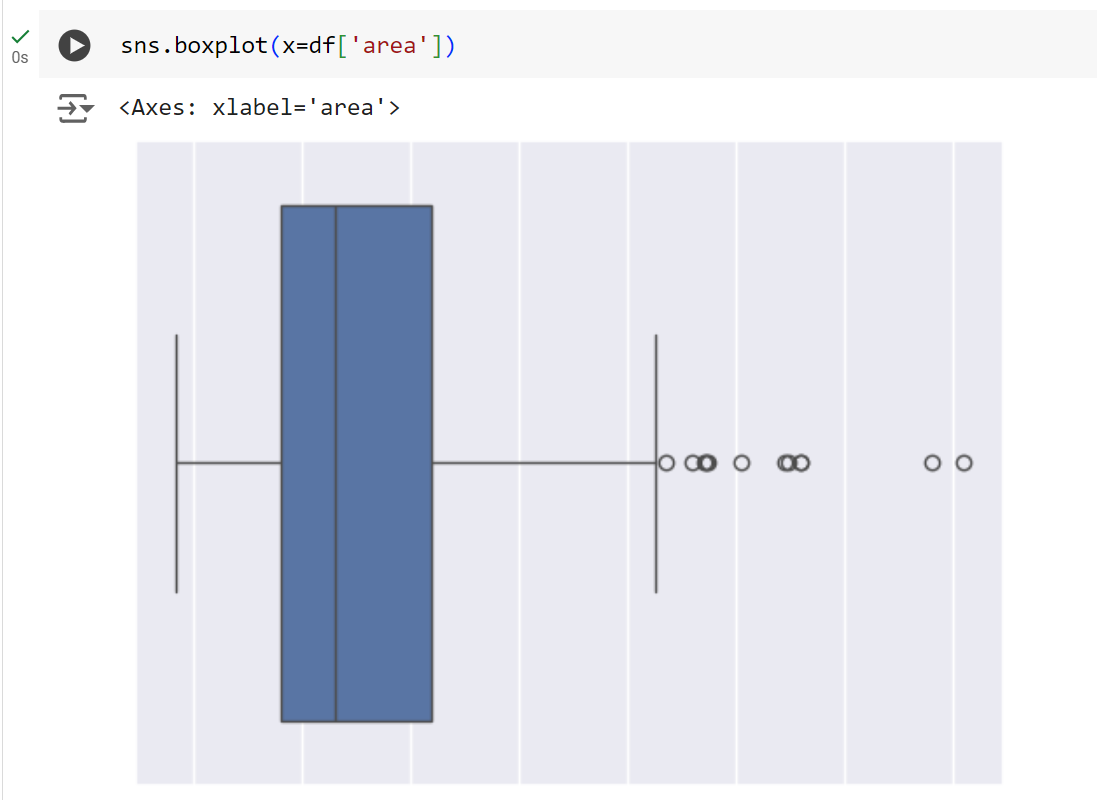
The categorical features with yes and no have been converted to binary outputs like either 1 or 0

For Furnishingstatus, it had 3 results: unfurnished, semi-furnished, and furnished. **Furnished** has been taken as a dummy/reference, while the other two are mapped to 1 and 0. If both of the included ones are 0, then it means we have furnished

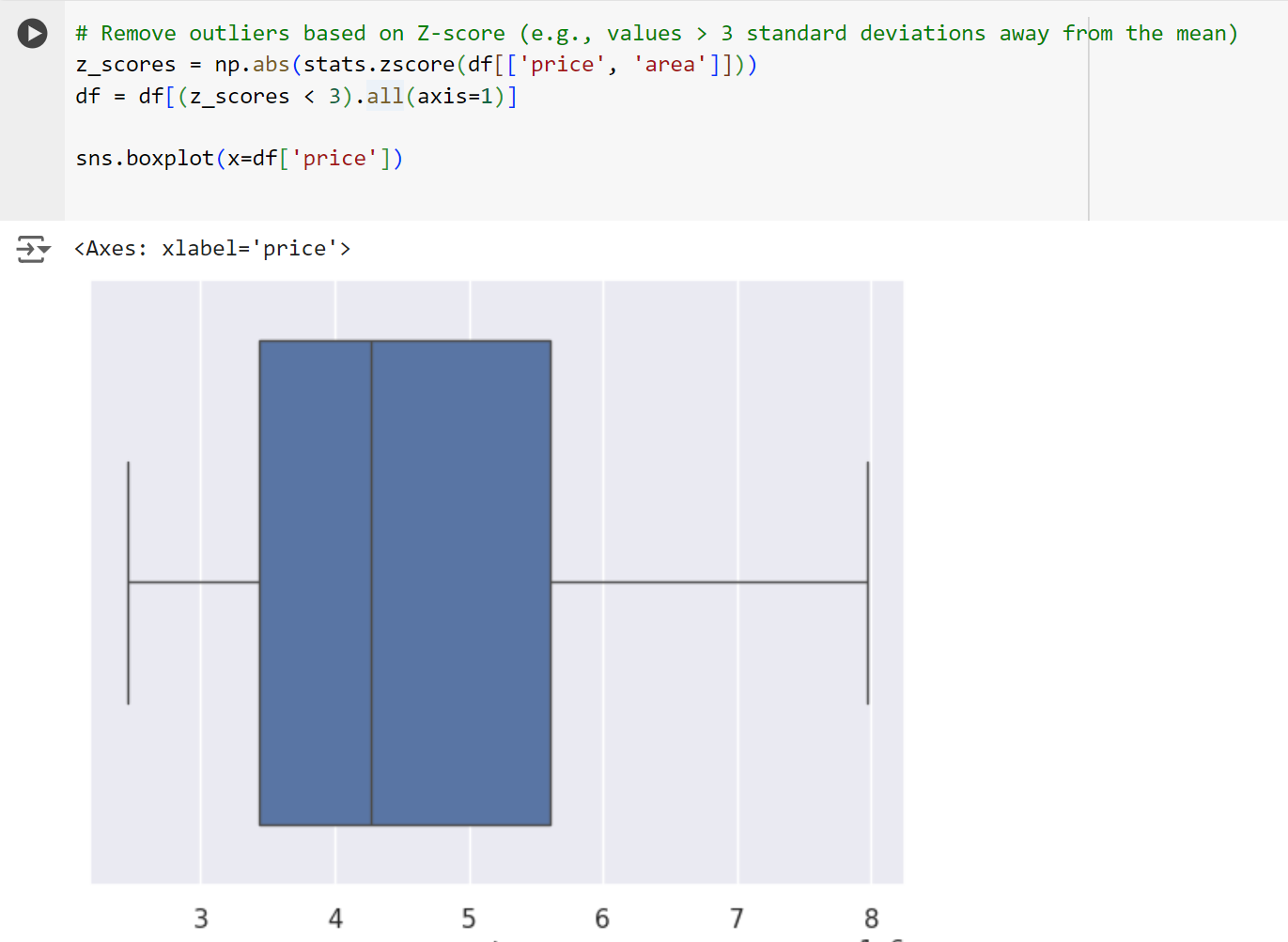
### Outlier Detection and Treatment:

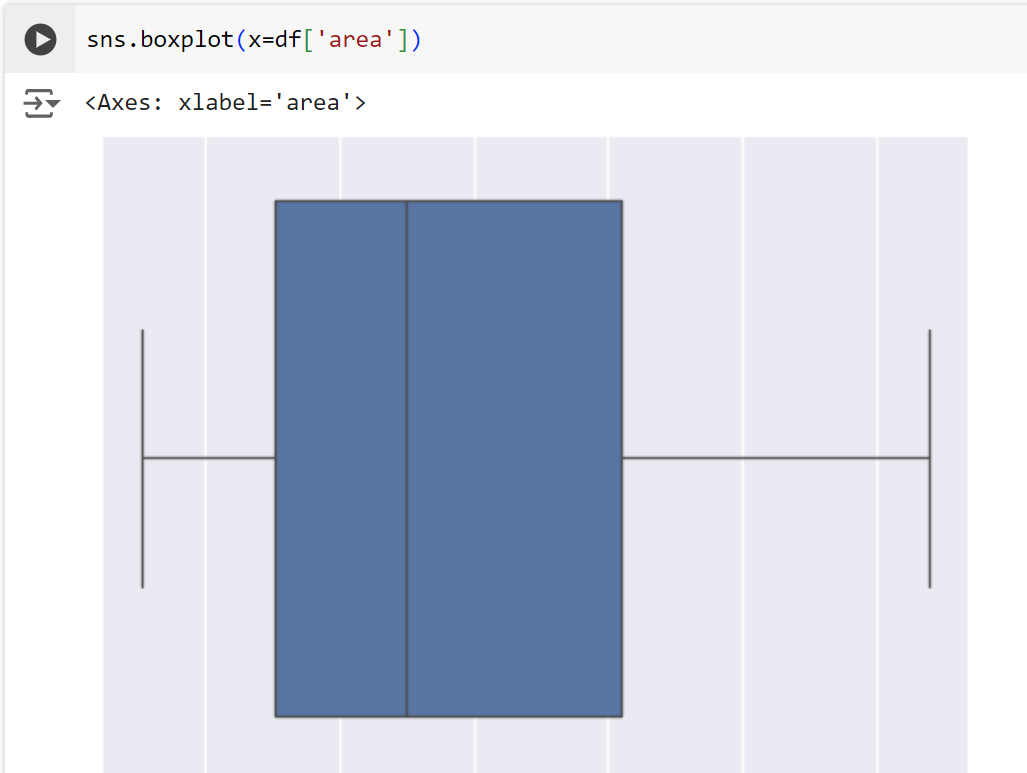
The two most essential features in the dataset are area and price, which are continuous.

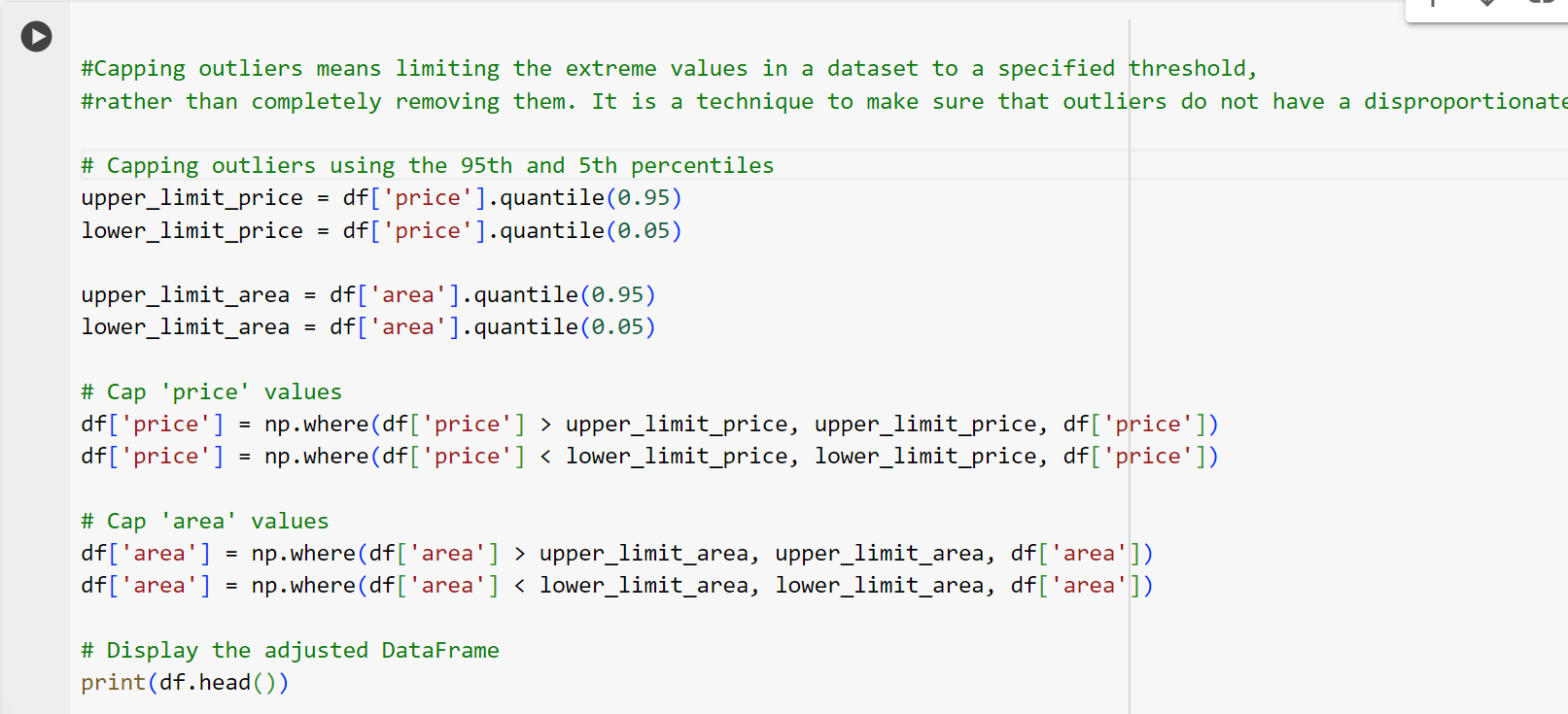




After Treatment of Outliers

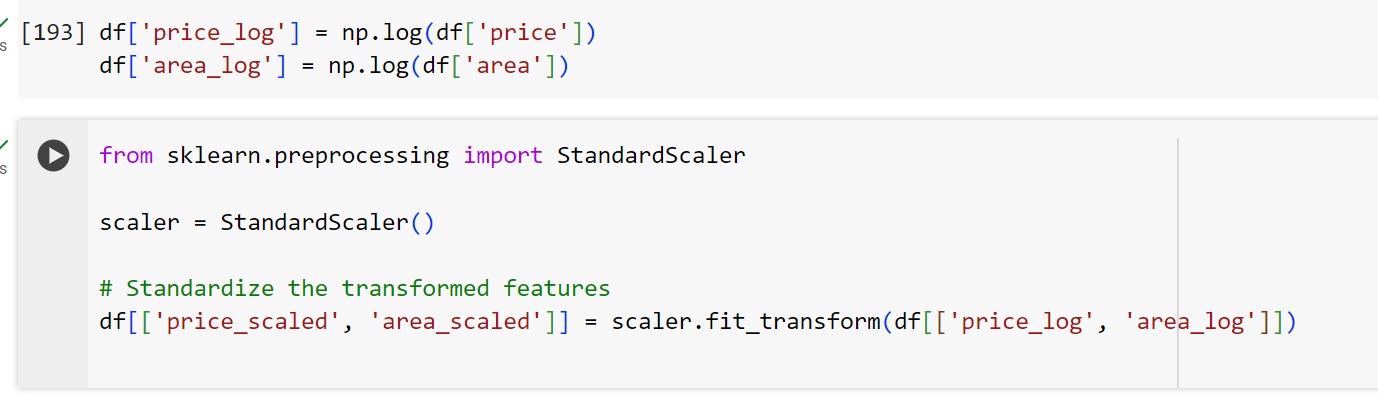






Outliers are treated by z-score(df now contains only values that have a z-score within 3 deviations from the mean) and are then capped at 95th and 5th percentiles to negate the disproportionate effect on the model.

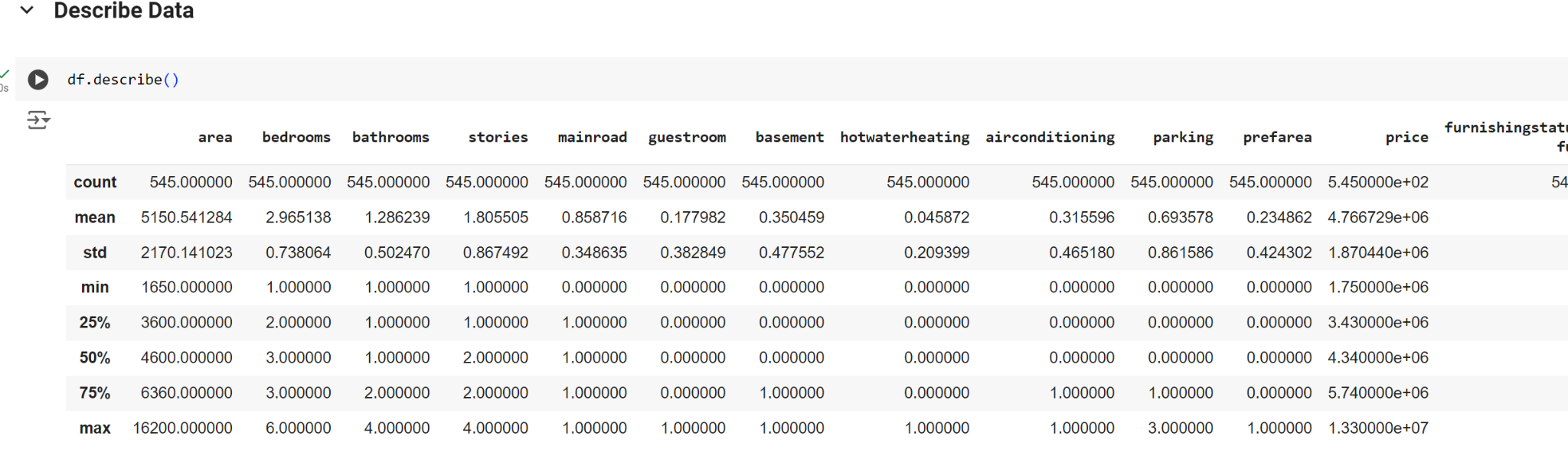
### Feature Scaling:



The Need for Scaling:

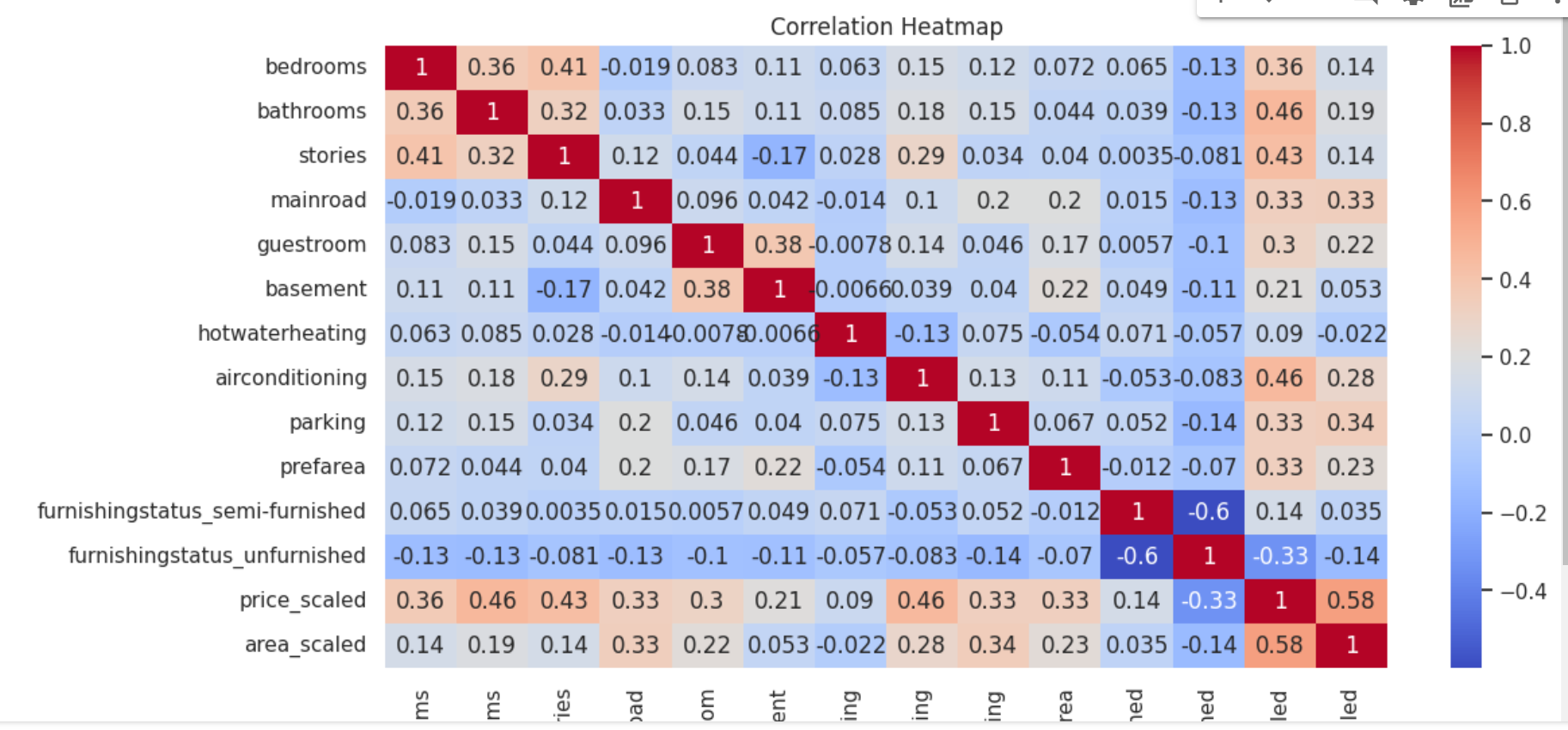
It's essential to scale the area and price to get accurate results. Our original scale might not lead to linearity and accurate multiple-linear regression because the price and area have a different scale.

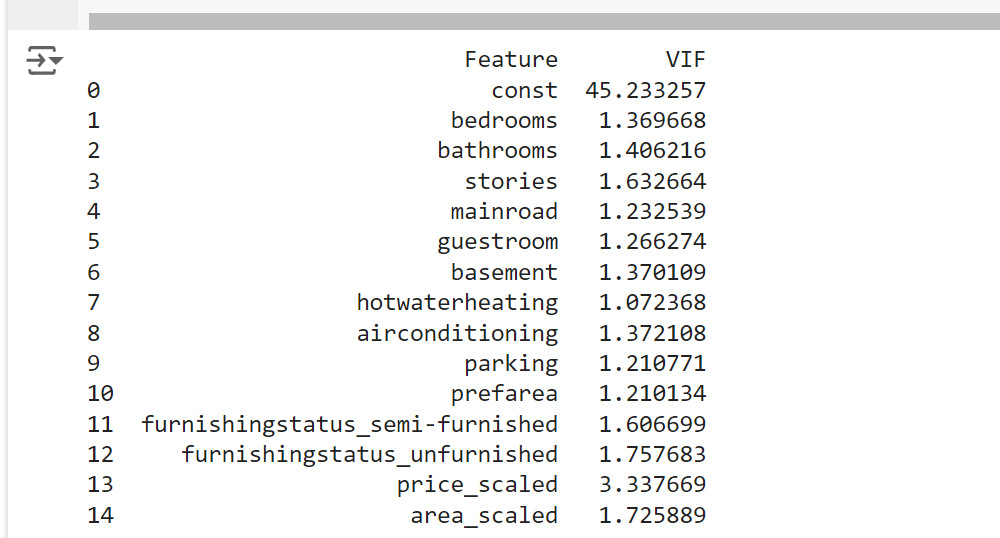
### Data Desciption:



These give 5 different statistics of data, giving an overview of every feature.

### Correlation and VIF:





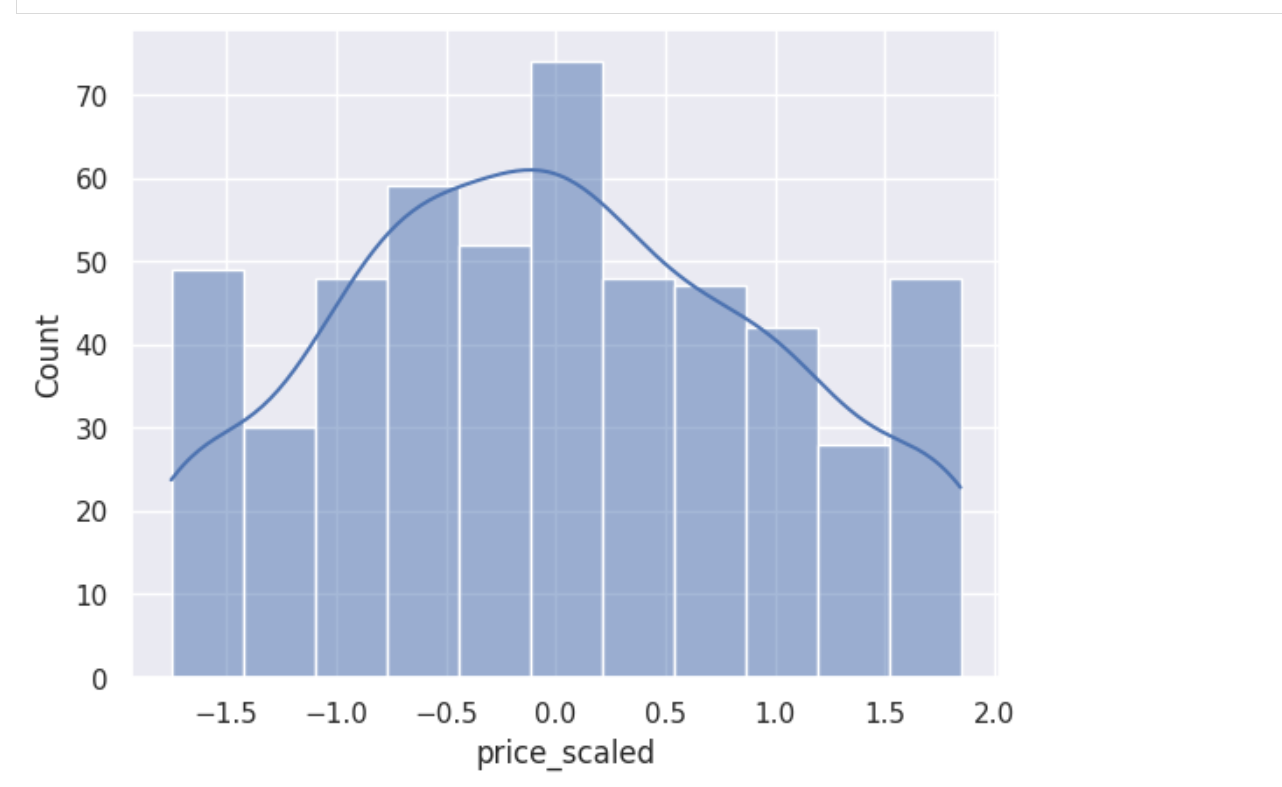
Variance Inflation Factor determines Collinearity between different features in the dataset, and correlation between 1 and 5 is considered moderate. None of the scaled features have high VIF nor does any of them have high Collinearity between each other.

The only price has a VIF of 3.337, which is also moderate and is not problematic.

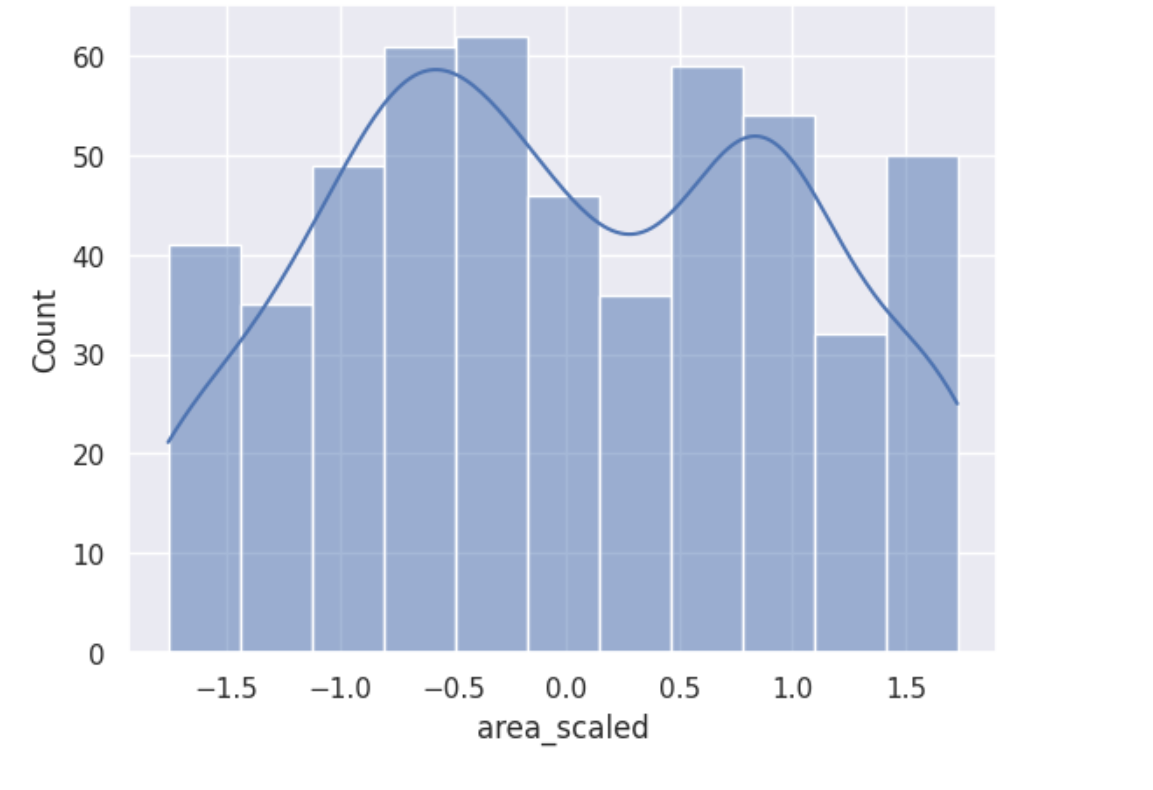
## **Visualizations**

## 

### Distribution Plots of Continous Features:

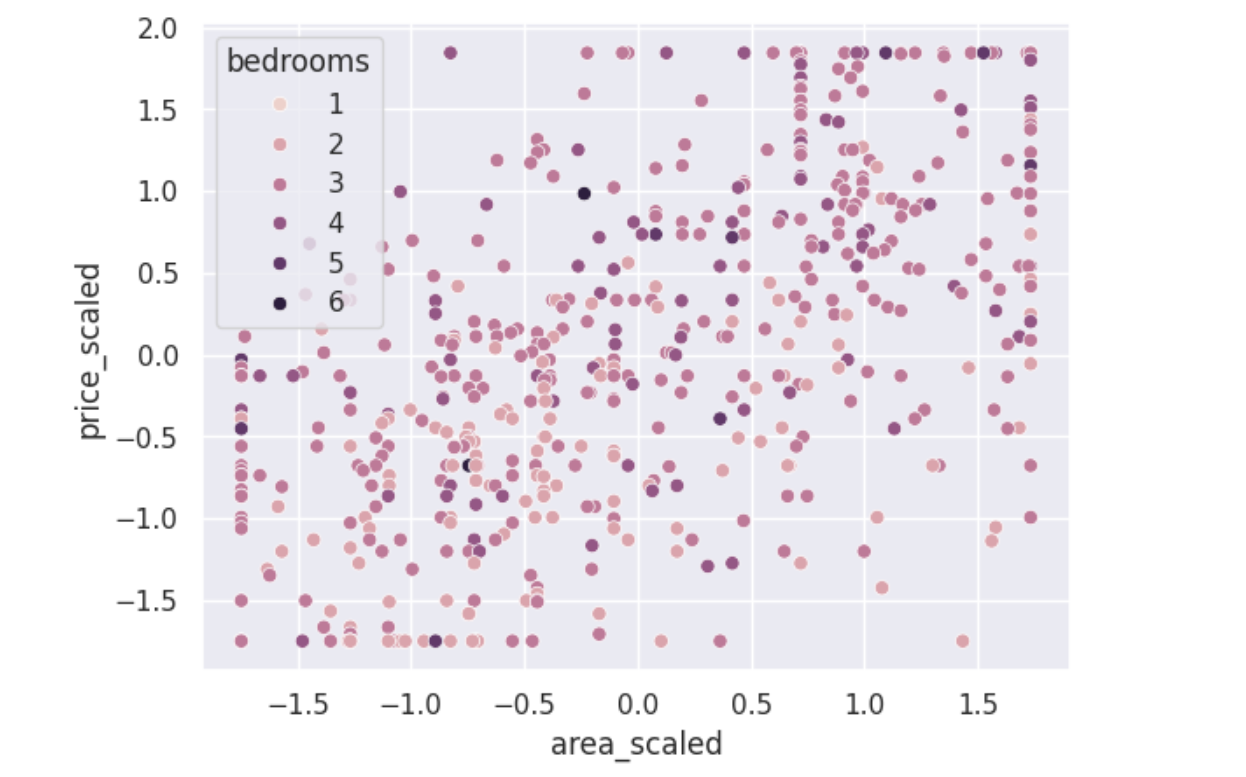


The price has been appropriately scaled, and it is evident that price\_scaled approximately follows the standard Normal distribution/normal curve

Area features have been scaled to follow normal distribution, and we can see that it is a bimodal curve with 0 mean and 1 standard deviation, which suggests Area distribution is not exactly normal.

Peaks suggest that small cottages and large houses are more than midsized houses due to peaks.

### Scatterplot of area and price with respect to the number of bedrooms



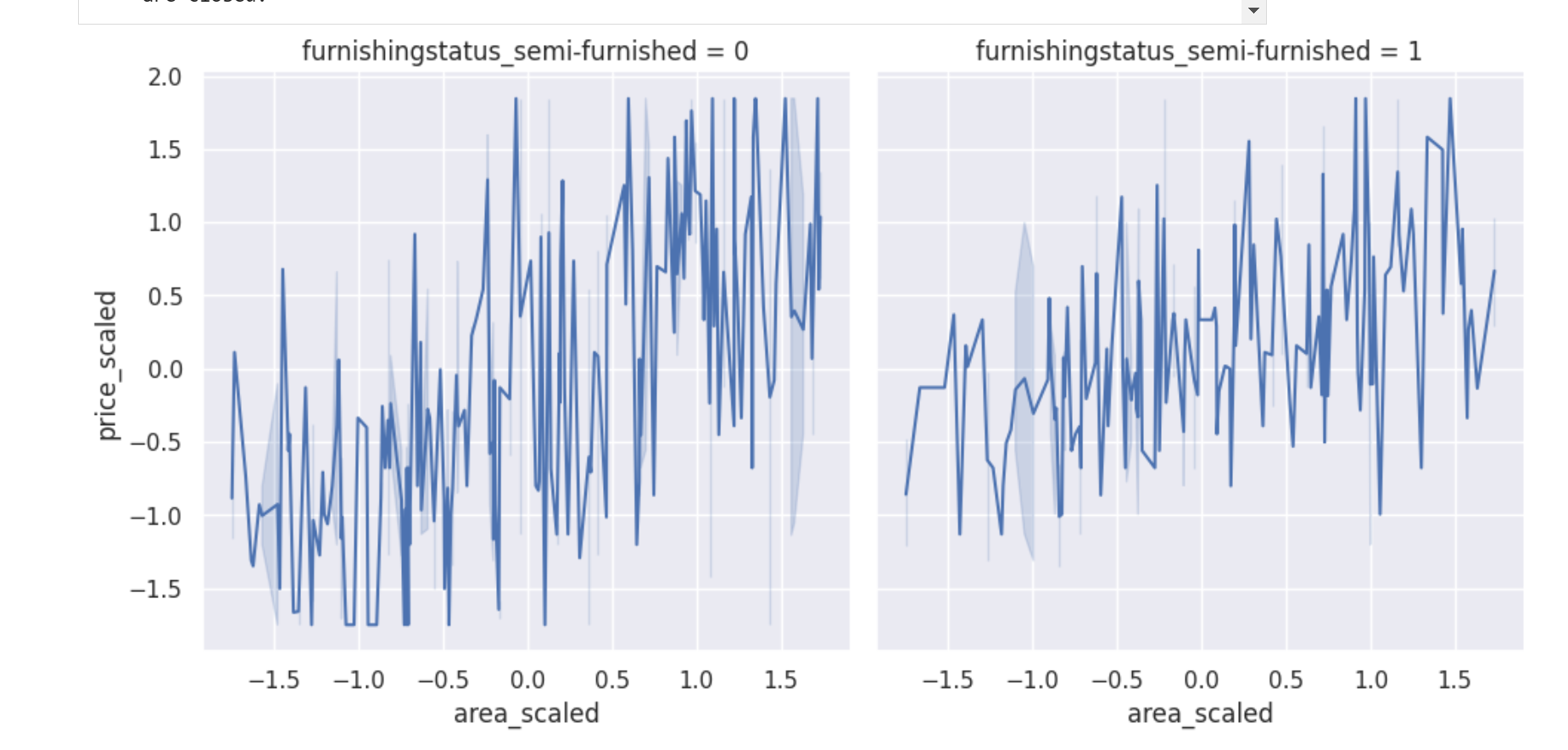
This shows variation in price with respect to area and number of bedrooms.

As the area of houses increases, prices tend to grow as well, and it's practically true as well.

Smaller houses tend to have fewer bedrooms, and lighter shades of dots are more on the left side of the graph, while larger houses have more bedrooms available and are more expensive.

Most houses have just 2 bedrooms, with prices falling below the average rates.

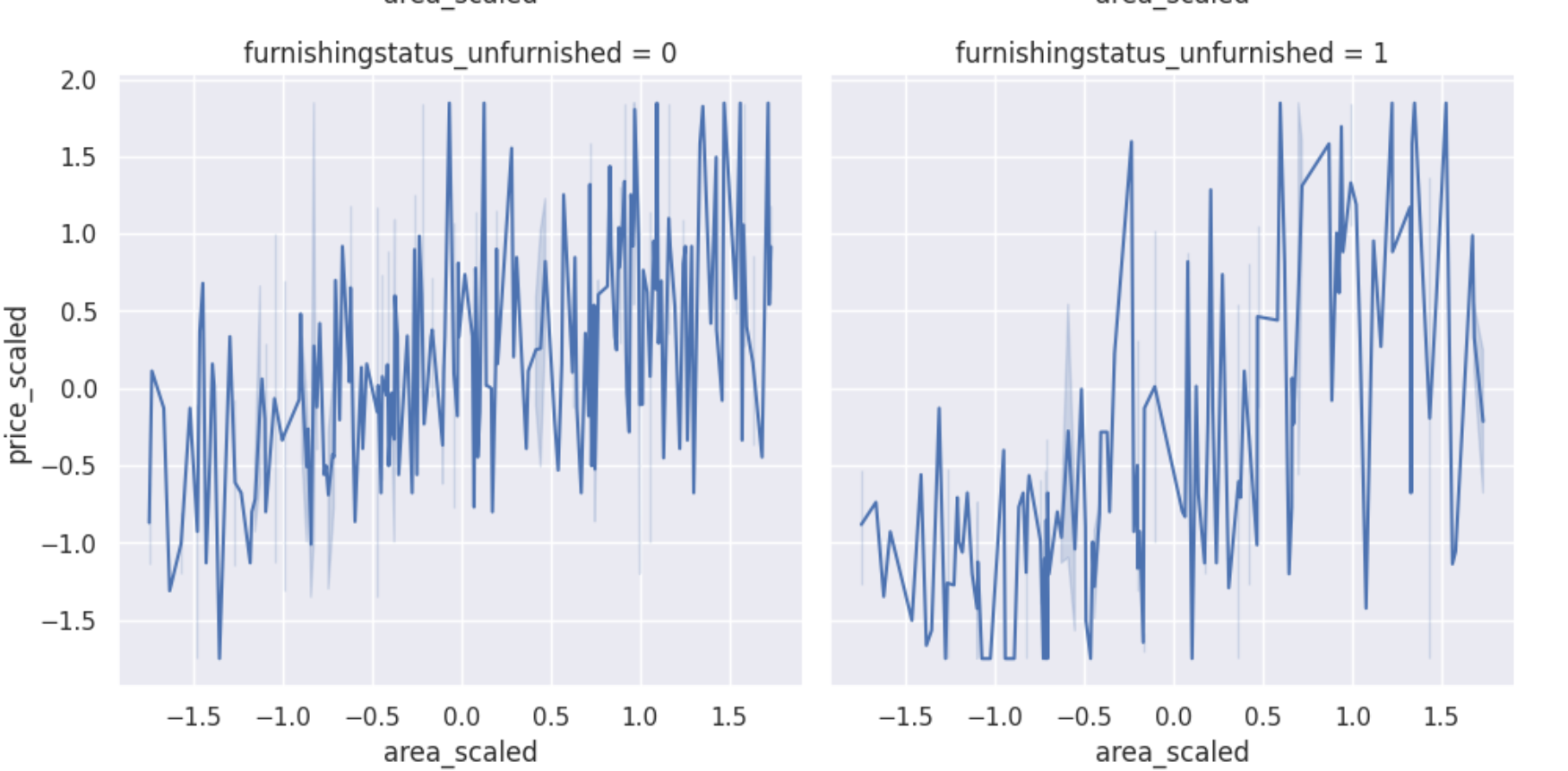
### Relational Plot of Price vs Area based on furnishing status:



**Both plots** indicate a **positive correlation** between area and price, meaning as the area increases, the price generally increases too, regardless of furnishing status.

The data is quite **spread out**, indicating **high variability** in price for given area values, especially for larger homes. For houses that are semi-furnished or furnished (plot 1)

The prices are more stable for the second one as there’s **less variability** in the prices of houses that are unfurnished for lower areas. It shows that as the area of semi-furnished houses increases, so do its rates. (plot 2)



**Both plots** indicate a **positive correlation** between area and price, meaning as the area increases, the price generally increases too, regardless of furnishing status.

The data is quite **spread out**, indicating **moderate variability** in price for given area values. For houses that are either furnished or unfurnished (plot 1)

The prices are more stable for the second one(semi-furnished) for **smaller unfurnished houses** as there’s **less variability** in prices of homes that are unfurnished; however, as the area of the house increases, there’s a high fluctuation in values, showing large houses can vary from low rates to higher rates. (plot 2)



**Semi-furnished and unfurnished houses (plot 1)** have higher price fluctuation as the area increases.

On the other hand, **furnished Houses(plot 2)** tend to have less price variability as the area increases. And practically it’s prominent larger furnished houses have higher rates.

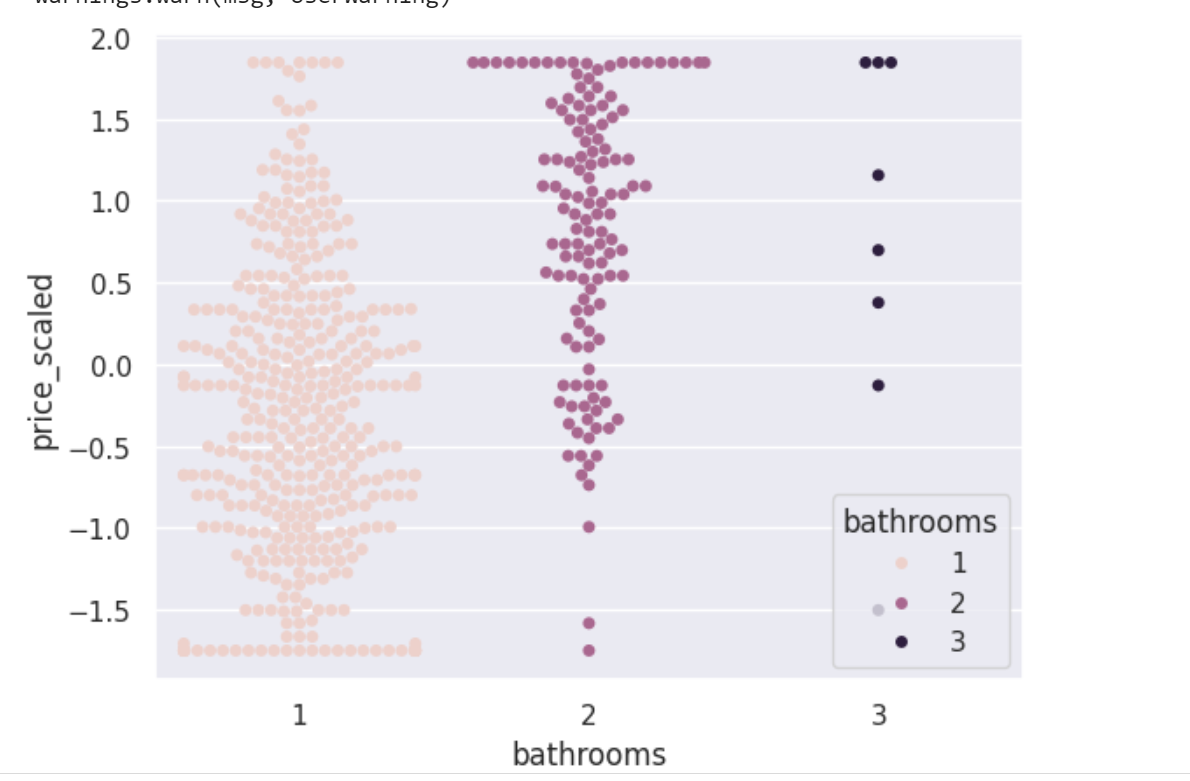
### Swarmplot of price vs other features:



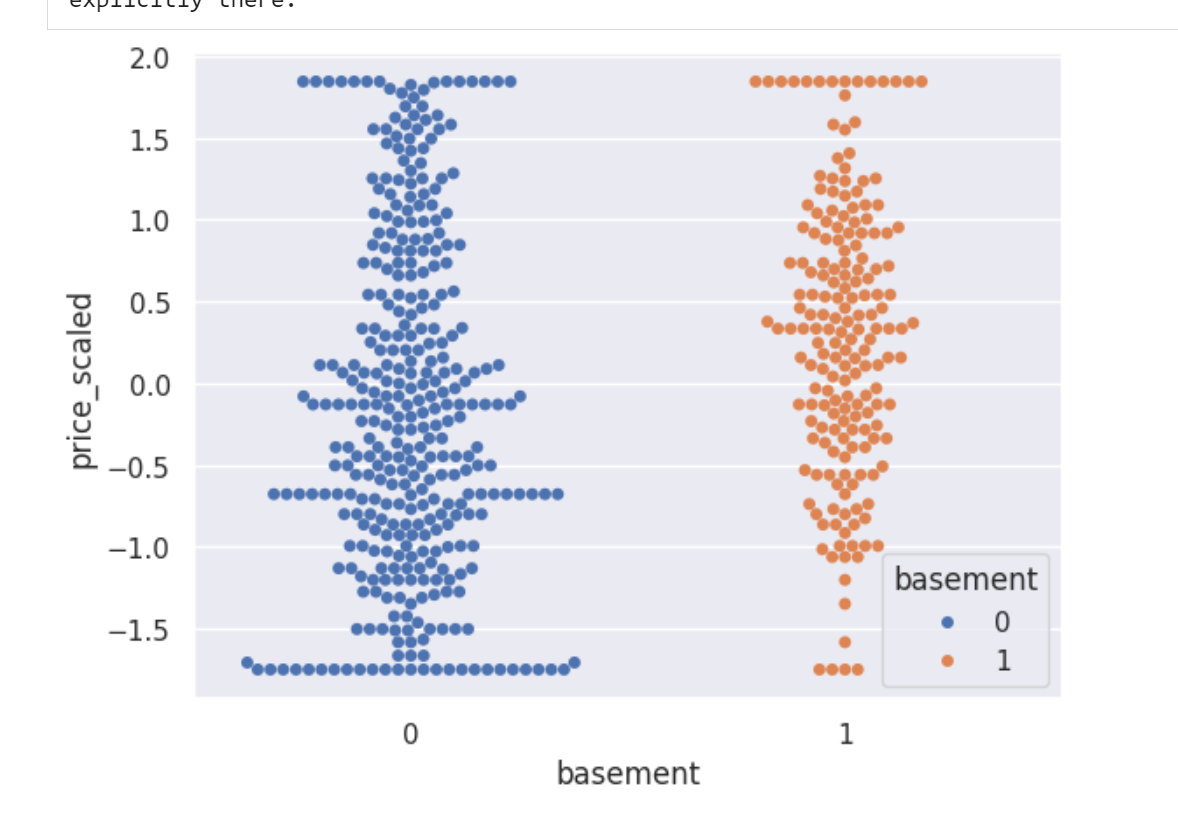
Houses not located on mainroad tend to have lower rates than houses on main roads.

More houses on mainroad have an even distribution of rates around the mean. While a significant number of houses on mainroad have higher than average rates.

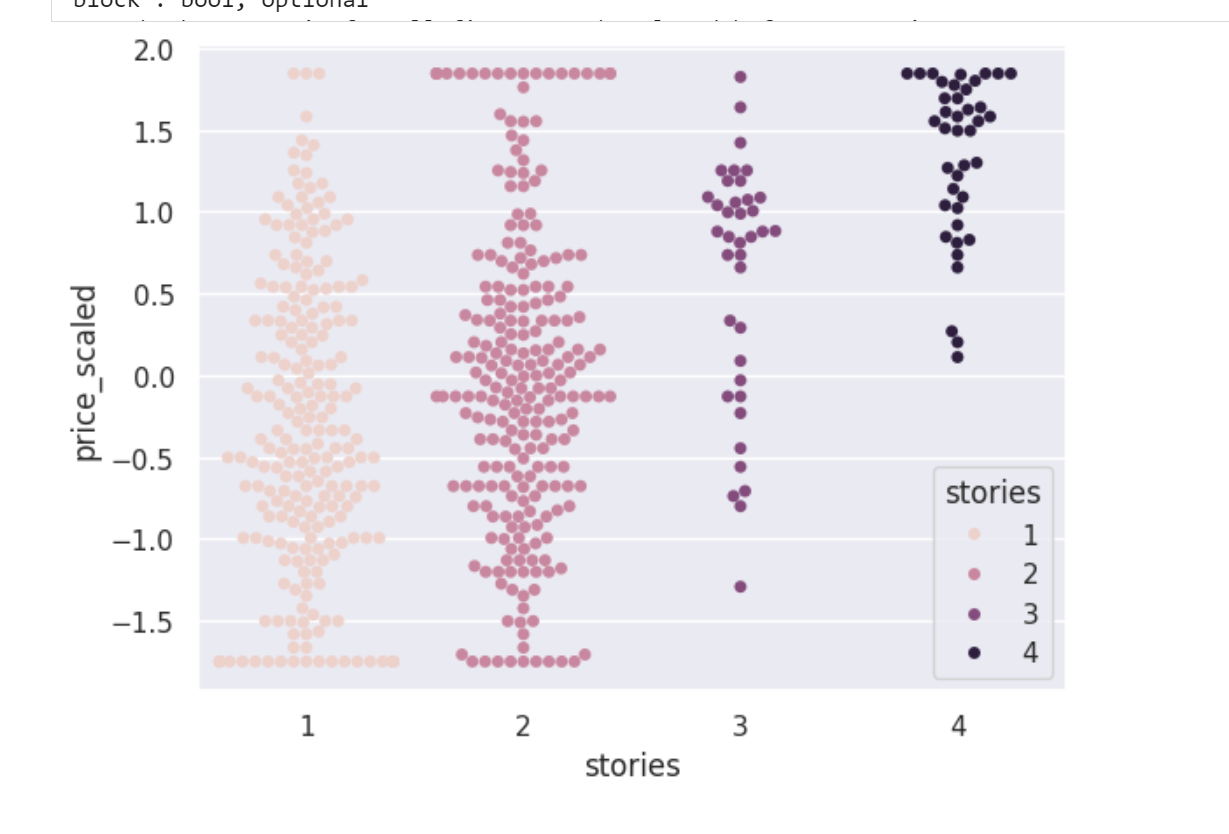
.



Firstly, there are a few houses with 3 bathrooms and many houses with just 1 bathroom. As expected, homes with more bathrooms have higher rates than houses with fewer bathrooms.



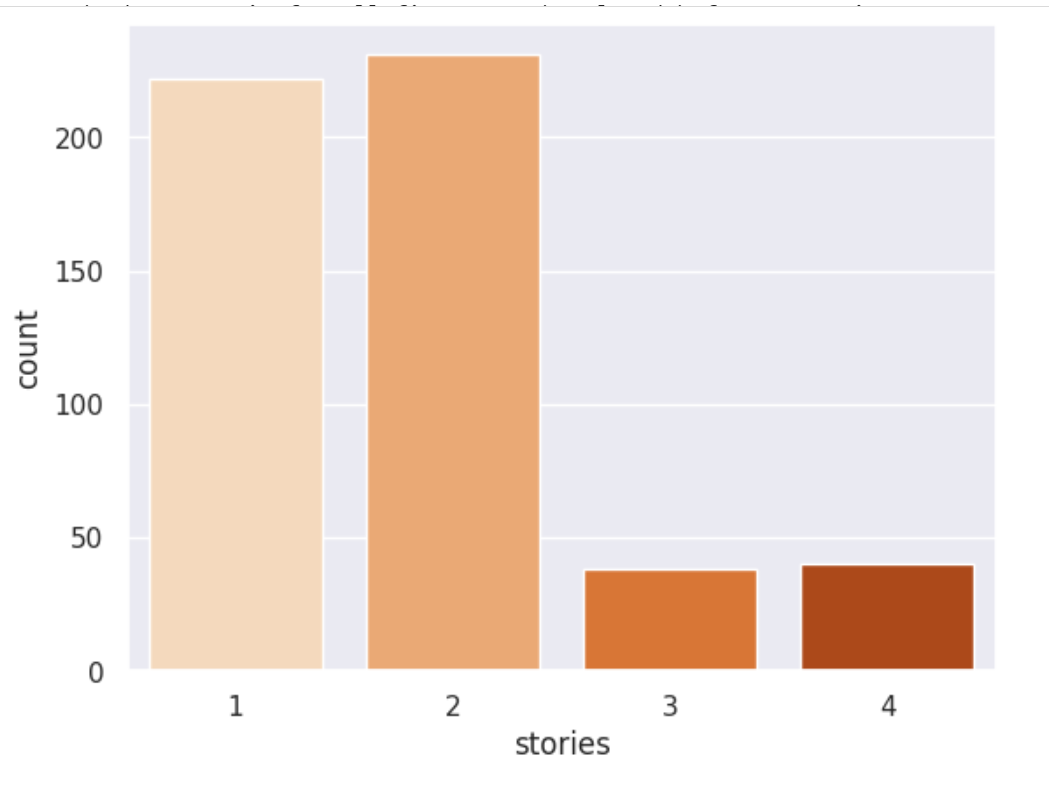
Houses with basements add to the property rates and are thus **valued higher** than houses without basements.



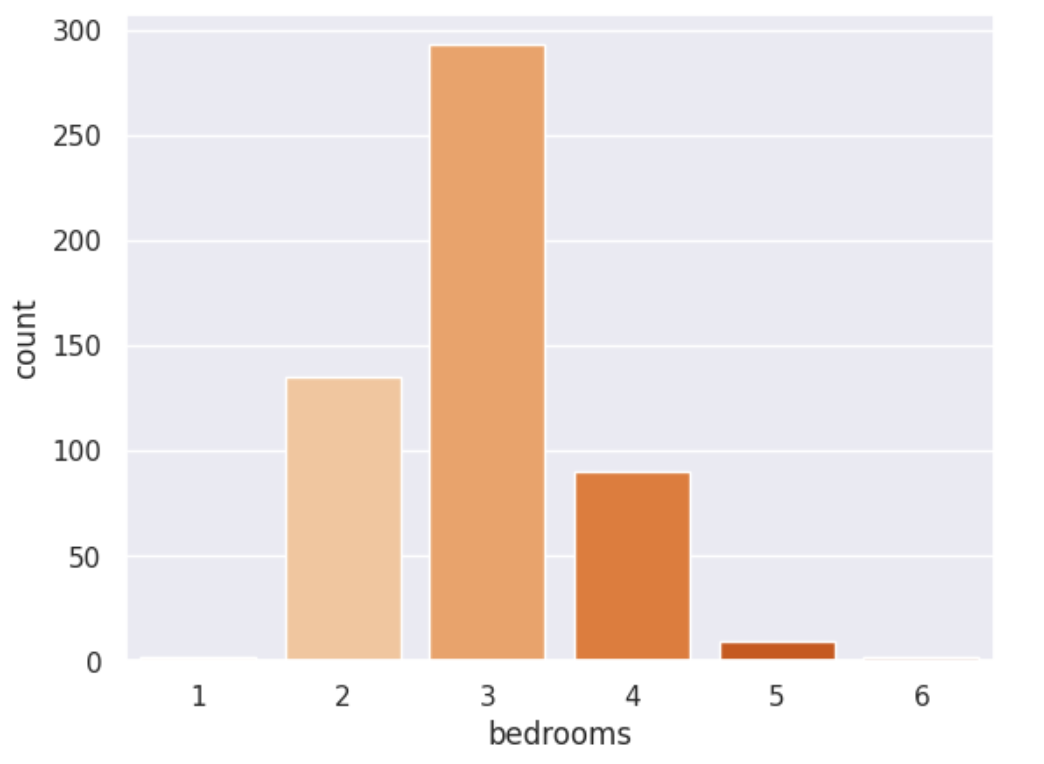
**A lot of houses have either 1 story or 2 stories**. The houses with one 1 story have **average rates lower** than 2 stories. Similarly, houses with 3 stories have **higher average rates** than 2 stories but **lower average rates** than 4 stories.

So, more stories lead to **higher average rates**.

### Countplots:



Houses with 2 stories are more common while houses of 3 and 4 stories are too low.



There are approximately **280 houses with 3 bedrooms,** while we can see that other houses with either 1, 2, 4, 5, or 6 bedrooms can all be added up but will still fall short of the houses with 3 bedrooms.



Most houses have just 1 bathroom.

### Conclusion of Visualizations:

More houses tend to have 1 bathroom, 2 stories, and 3 bedrooms and are located on mainroad. We can also see furnished houses are more costly; houses with more area, more bedrooms, more bathrooms, more stories, and facing main roads are generally more expensive than the others.

### 

## 

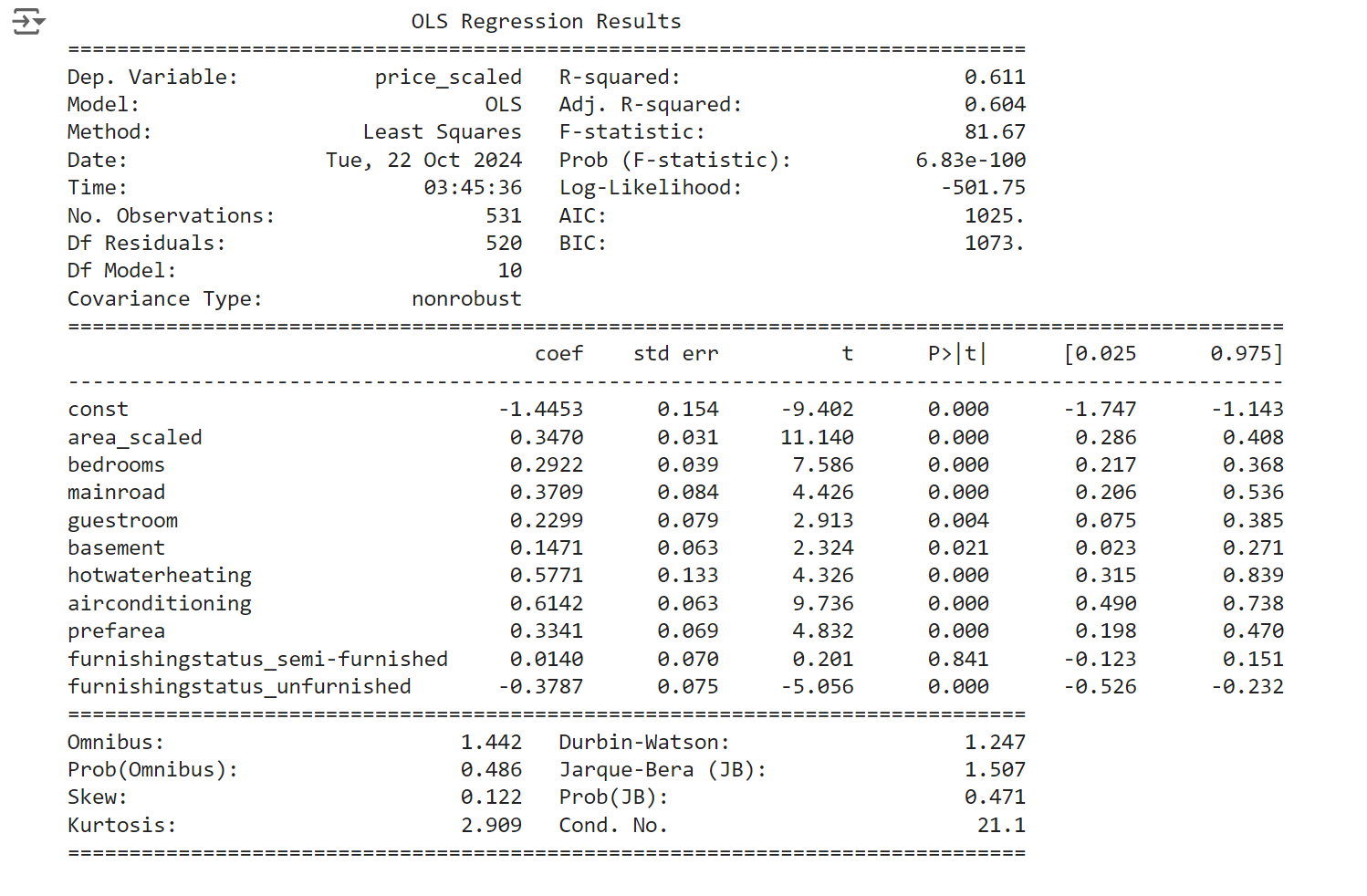
## 

## **Linear Regression**

## 

## 

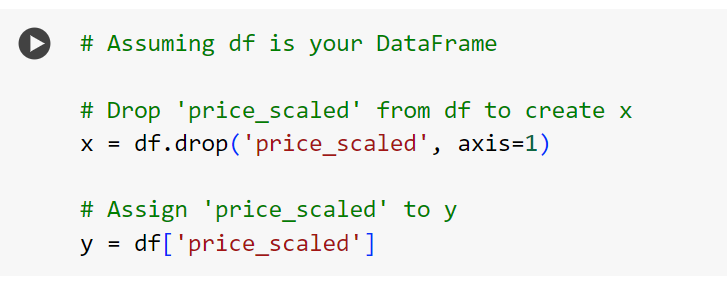
### Statistical Calculations:

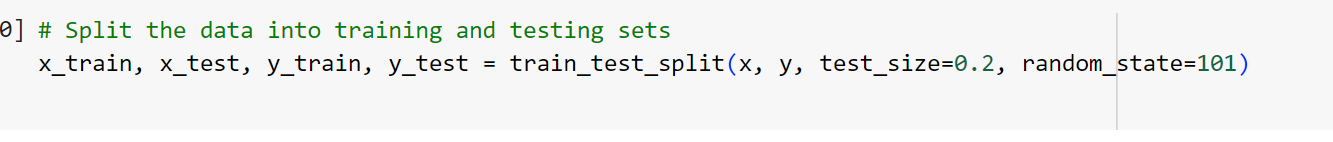


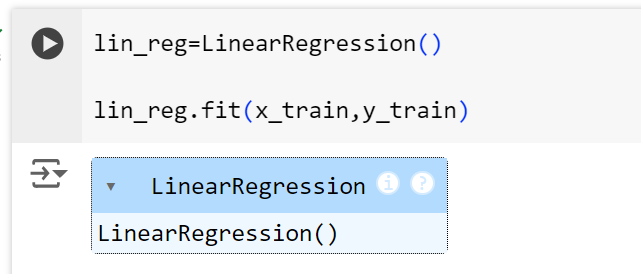
Where

**Price=β0​+β1​⋅Area+β2​⋅Bedrooms+β3​⋅Bathrooms+β4​⋅Stories+β5​⋅Mainroad+β6​⋅Guestroom+β7​⋅Basement+β8​⋅Hotwaterheating+β9​⋅Airconditioning+β10​⋅Parking+β11​⋅Prefarea+β12​⋅FurnishingStatus+ϵ**

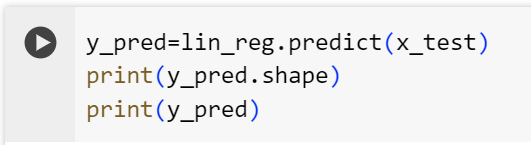
### Linear Regression by sklearn:

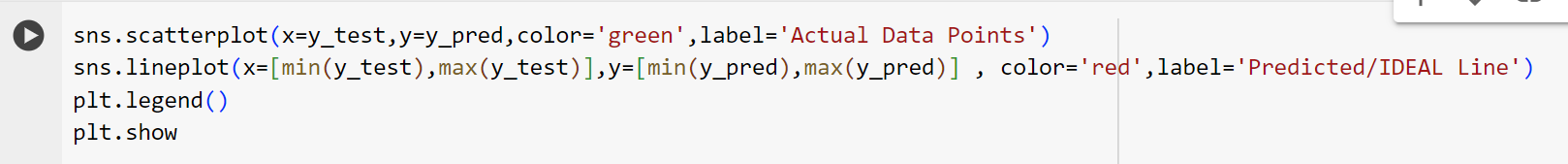


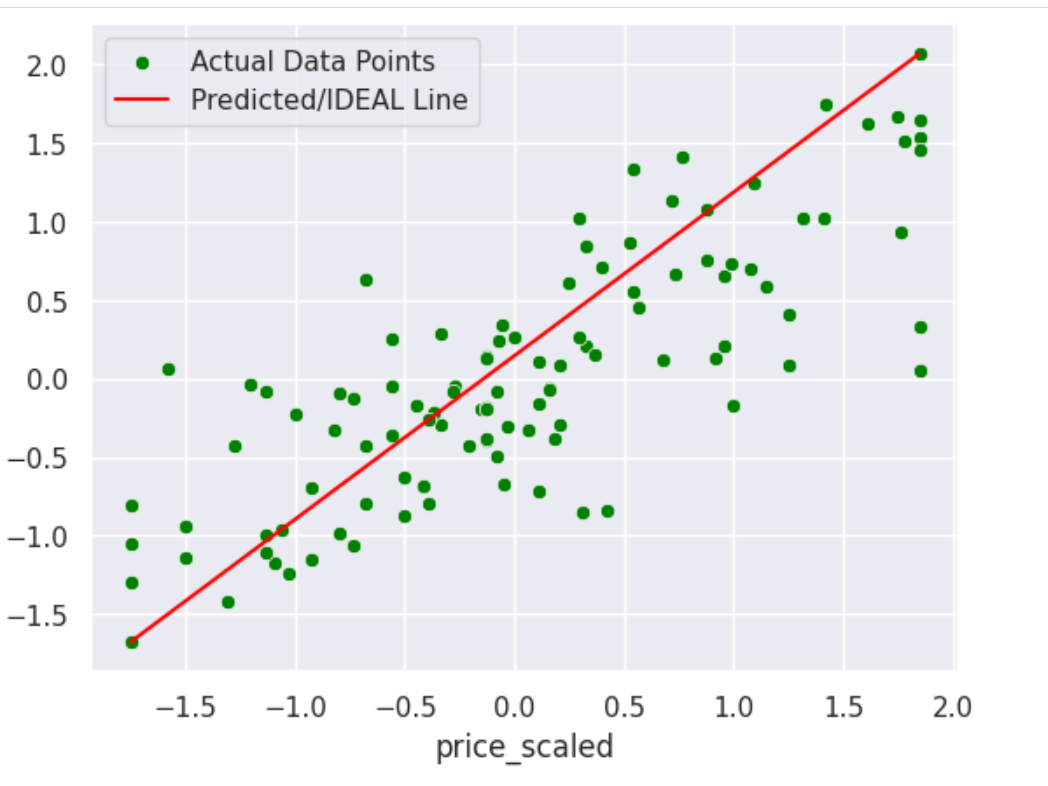




### Model Testing:

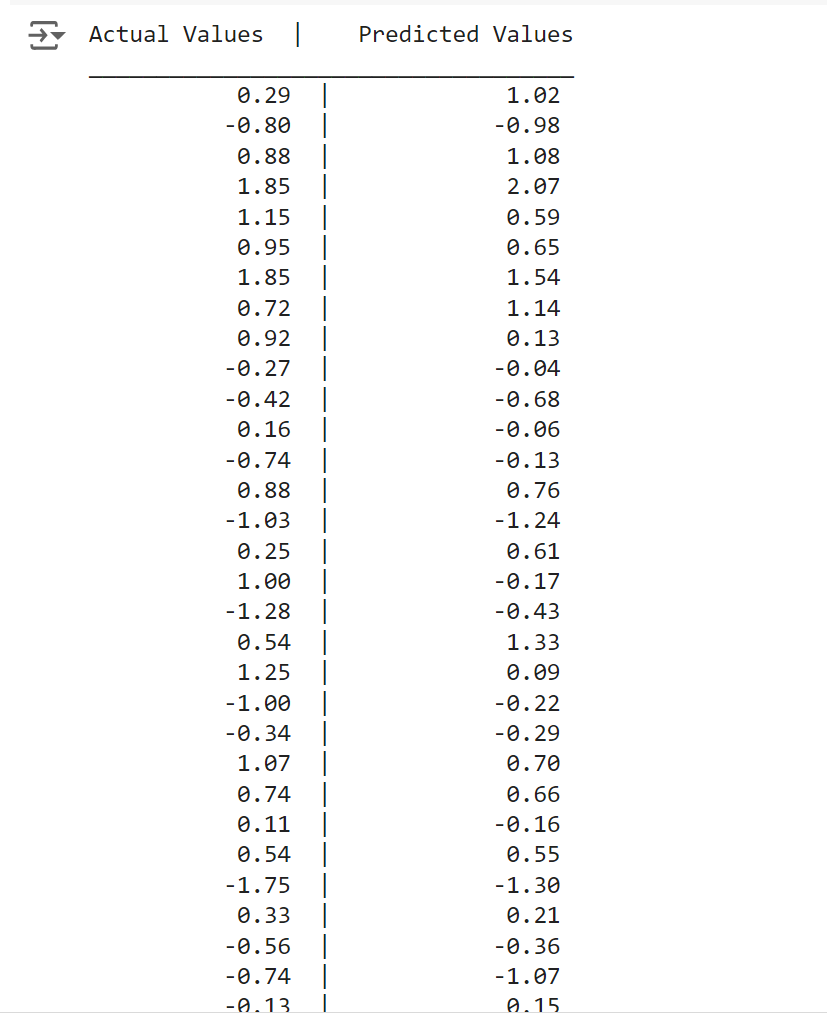






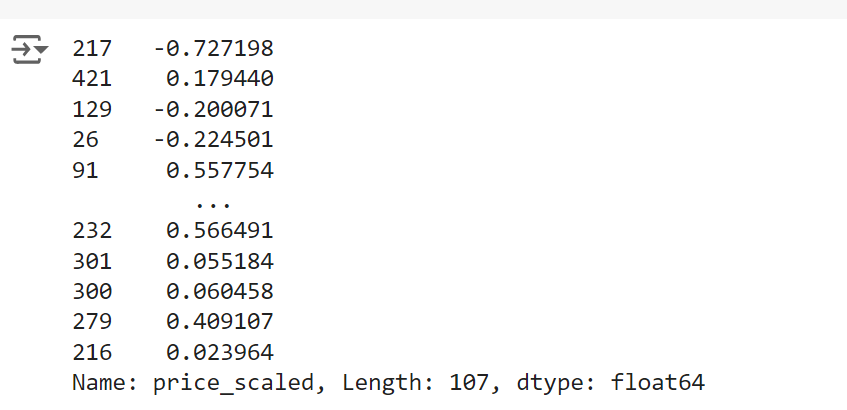
Here we have tested model on our test data and apparently it seems a very good fit through our scatter plot..

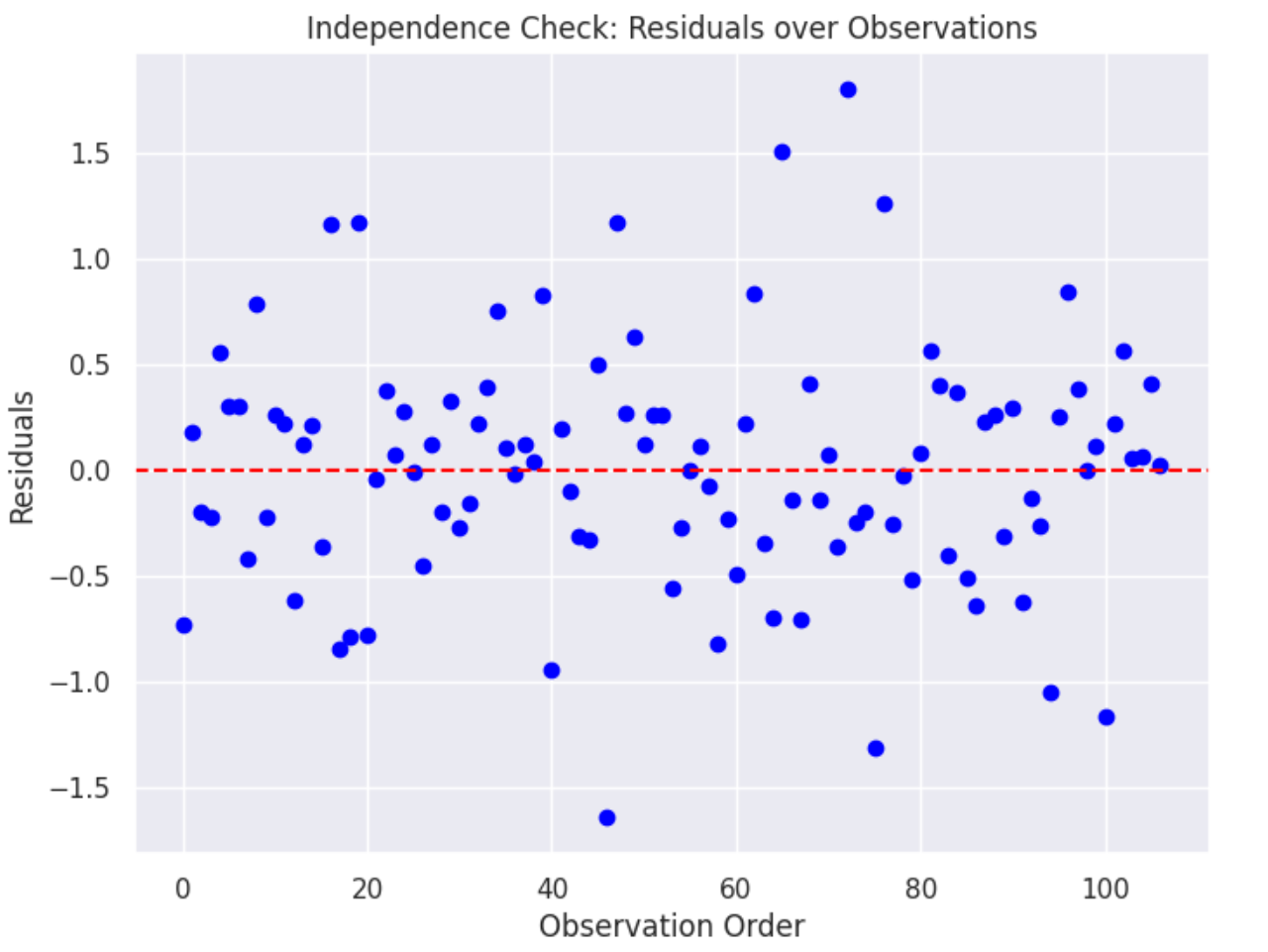
**The observations seem to be very close to the best fit line showing linearity and normality in the model.**



The difference betwen the actual and predicted values are quite low showing multiple linear regression an appropriate fit for the housing dataset.

### Residual Analysis:





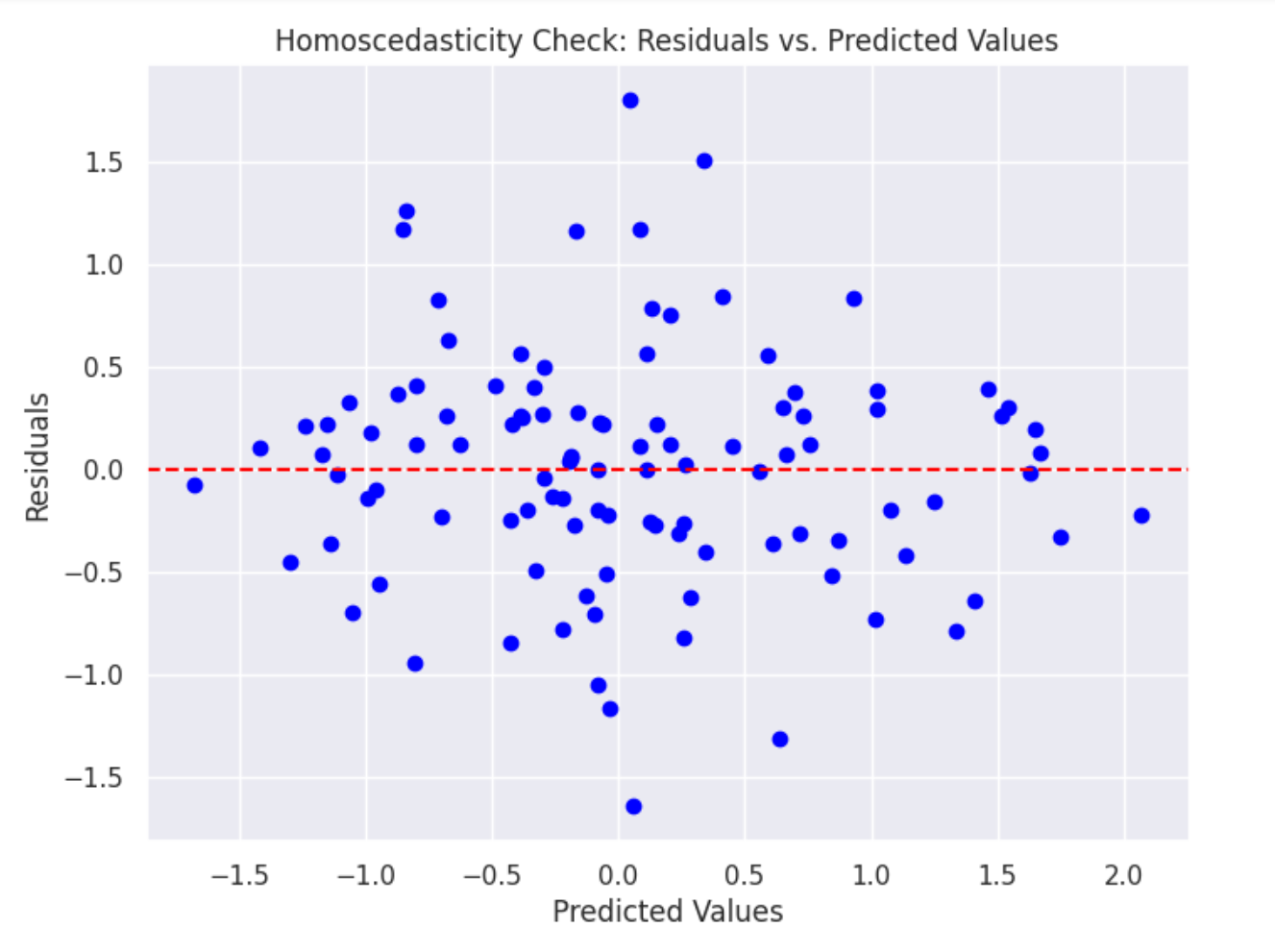
**Residuals**: These are the vertical distances between the actual values and the fitted regression line. Positive residuals indicate that the model **underpredicted** the actual value, while negative residuals indicate that the model **overpredicted**.

**Dashed Red Line (y = 0)**: This reference line represents **perfect model predictions** (where residuals would be 0). Ideally, residuals should be randomly scattered around this line.

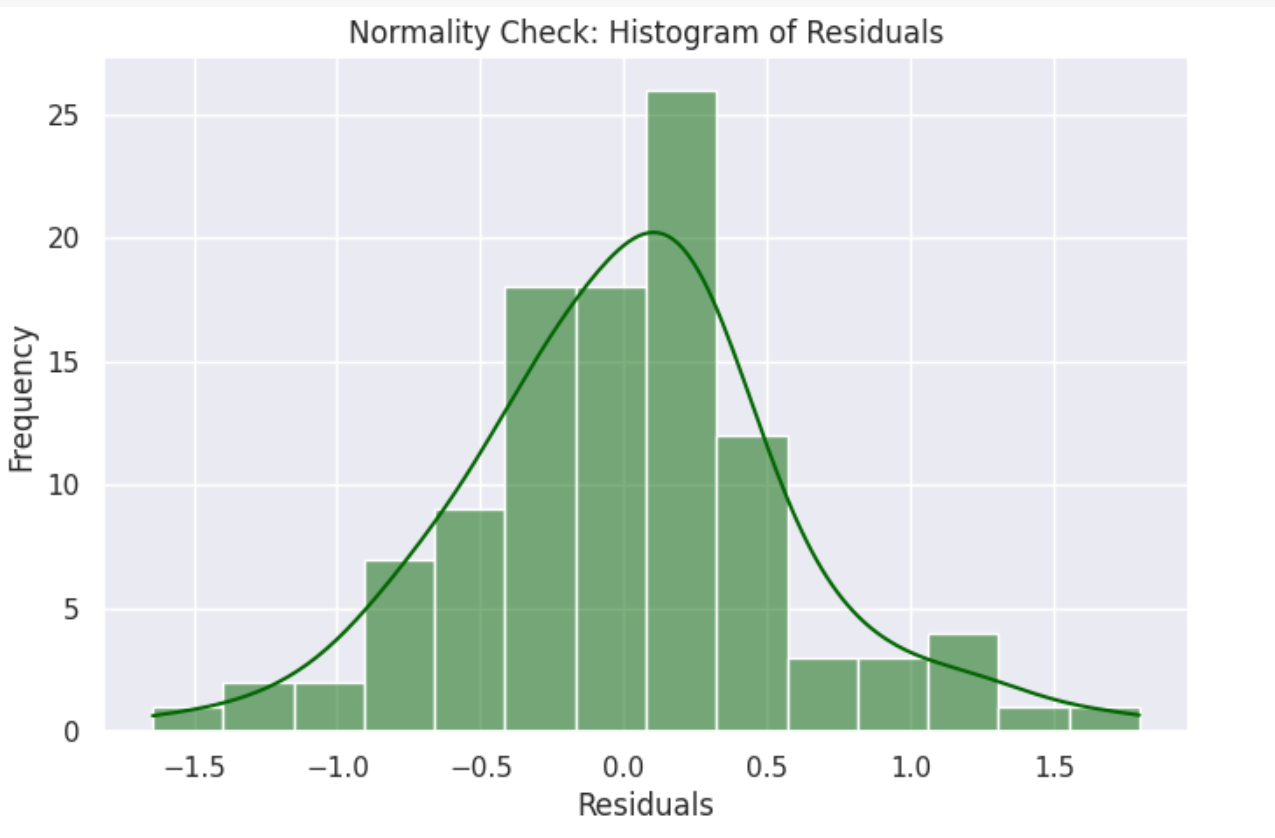
**Independence** : The plot shows no clear pattern or structure (random scatter); the **independence assumption is likely met.**

**Independent residuals** refer to a situation where the residuals (errors) from a regression model are not correlated with each other.

Some of the residuals seems to have high values which exceed 1 or are below -1. These are just a few values which model coudn’t accurately predict.



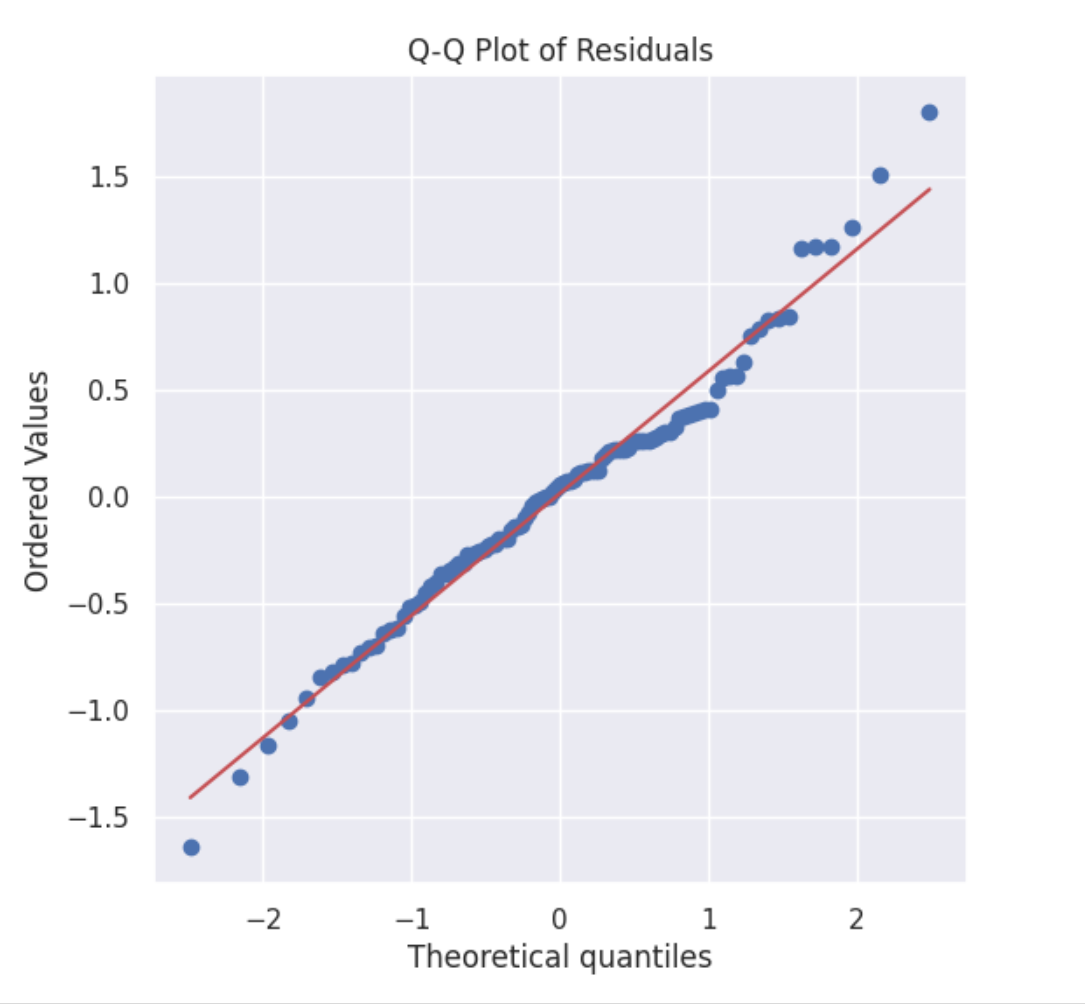
**Homoscedastic →** constant variance on both sides of the mean of residuals(the exact perfect prediction)



The distribution of residuals seems **fairly symmetrical**, with a peak near 0, meaning that most **residuals are small and centered around the mean**. However, there is a **slight skew to the right,** indicating that the residuals tend to have more positive values than negative.

The histogram follows the curve reasonably well, suggesting that the residuals **approximate a normal distribution,** although there are some deviations, particularly in the tail areas (larger residuals).

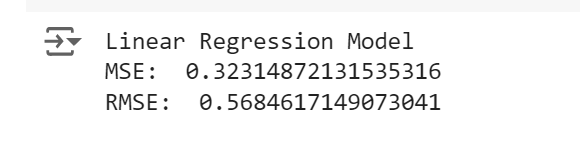
While the residuals appear roughly normal, the slight right skew and the presence of larger positive residuals indicate mild deviations from **perfect normality**.



Most of the points follow the red line where residuals are perfectly distributed. However there’s a slight deviation at the extreme values.

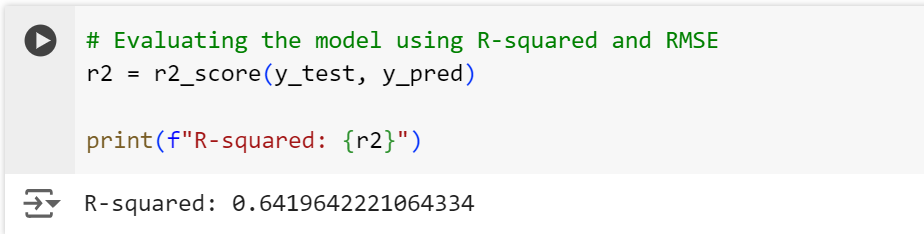
So the Q-Q plot shows **approximate Normal Distribution** with just slight deviation at the tails.

### Model Evaluation:



The more the RMSE and MSE are closer to 0, the more suitable the model is for the Housing dataset.

We can see that the model fits the data **reasonably well.**



64% of the price is expressive by all the features taken in account. There’s moderate level of accuracy in predicting the response feature.

## Conclusion:

The multiple linear regression model shows a generally **linear relationship** and **independent** residuals, suggesting that these assumptions are reasonably met. Similarly, **homoscedasticity** is also available while **normality** is approximately followed.

However, the model is 64% expressive based on the observations in the data. This shows that the model is not an exact fit, and there’s certainly room for improvement that can be further achieved by employing further advanced methods.

<https://docs.google.com/document/d/1wA3xTyXXyVIVwieggNycRhLRrE7BUe2cBhiqS1xEeoM/edit?usp=sharing>

(no scaling + outlier treated with median) (following pdf)

(<https://docs.google.com/document/d/1wA3xTyXXyVIVwieggNycRhLRrE7BUe2cBhiqS1xEeoM/edit?usp=sharing> )