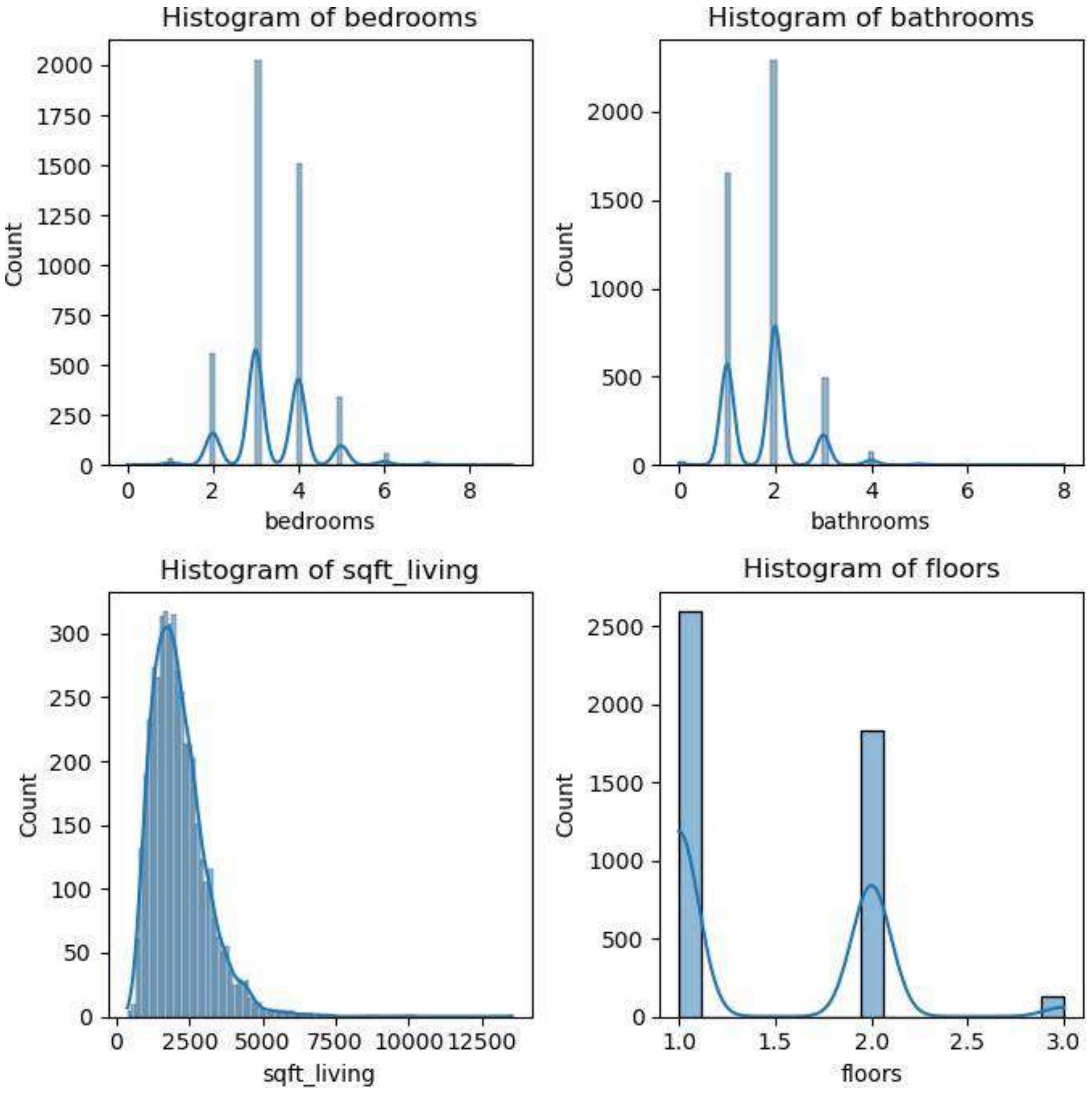
GitHub Repo Link: https://github.com/muhammadtalhajamil/Clustering-and-Fitting_ADS

Student ID: 23017847

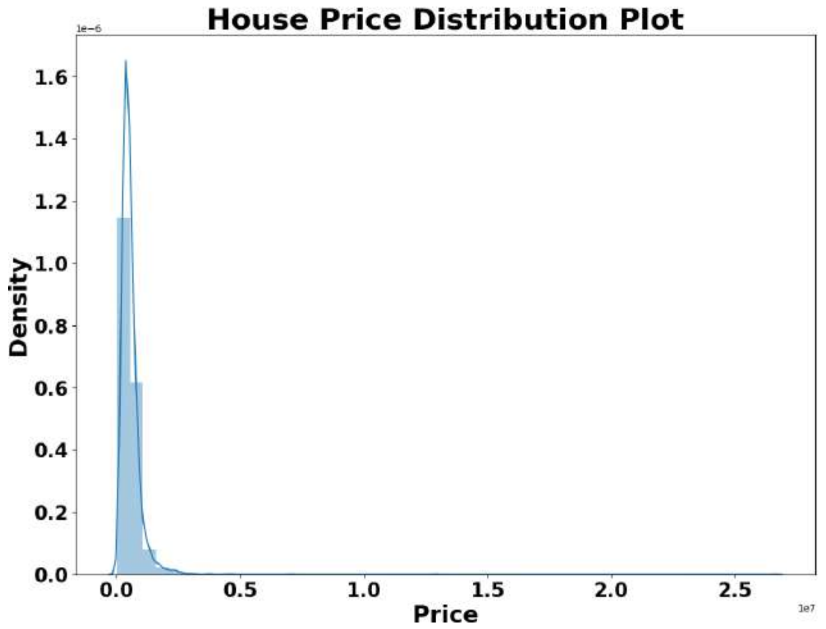
Dataset Link: https://www.kaggle.com/datasets/shree1992/housedata/data

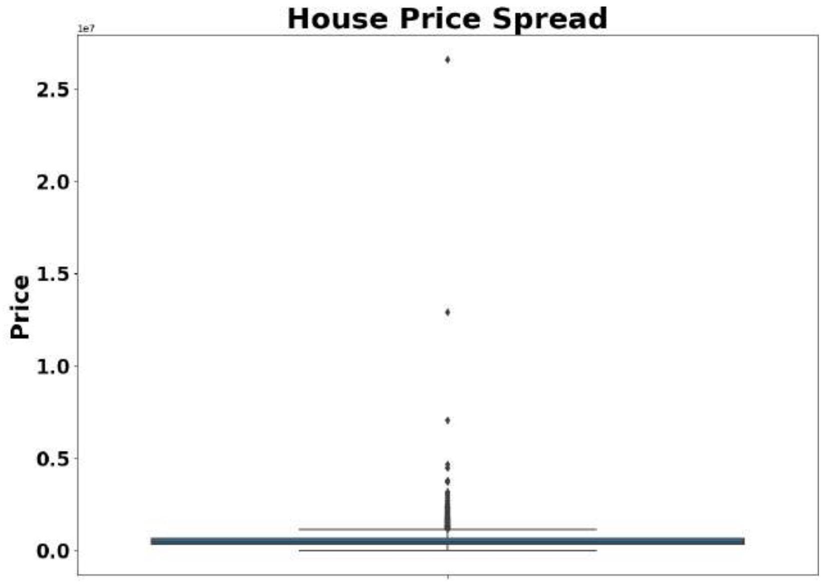
House Price Analysis

We are dealing with data of Sydney and Melbourne market data to predict housing price, we are wrangling large set of property sales records stored in a raw format. We completed data preprocessing by changing data type of different columns and also normalized our data.

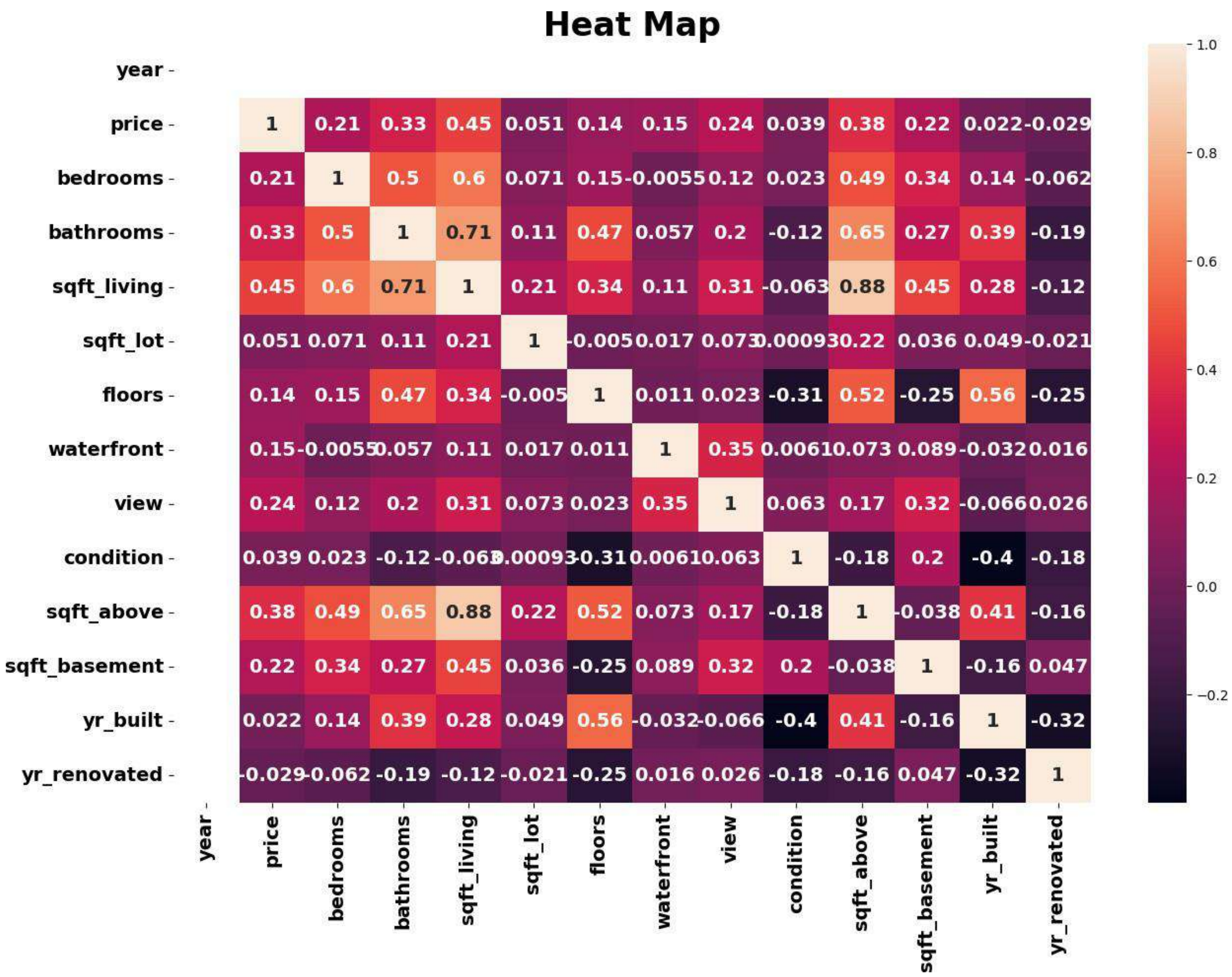


We plotted histogram to show the distribution of 4 features of house pricing dataset we can observe that sqft_living is 0-5000sqft which shows that more towards left and we have more data of 1 floor apartments or houses. We have least data for 3 floors. The distribution of bathrooms is around same.

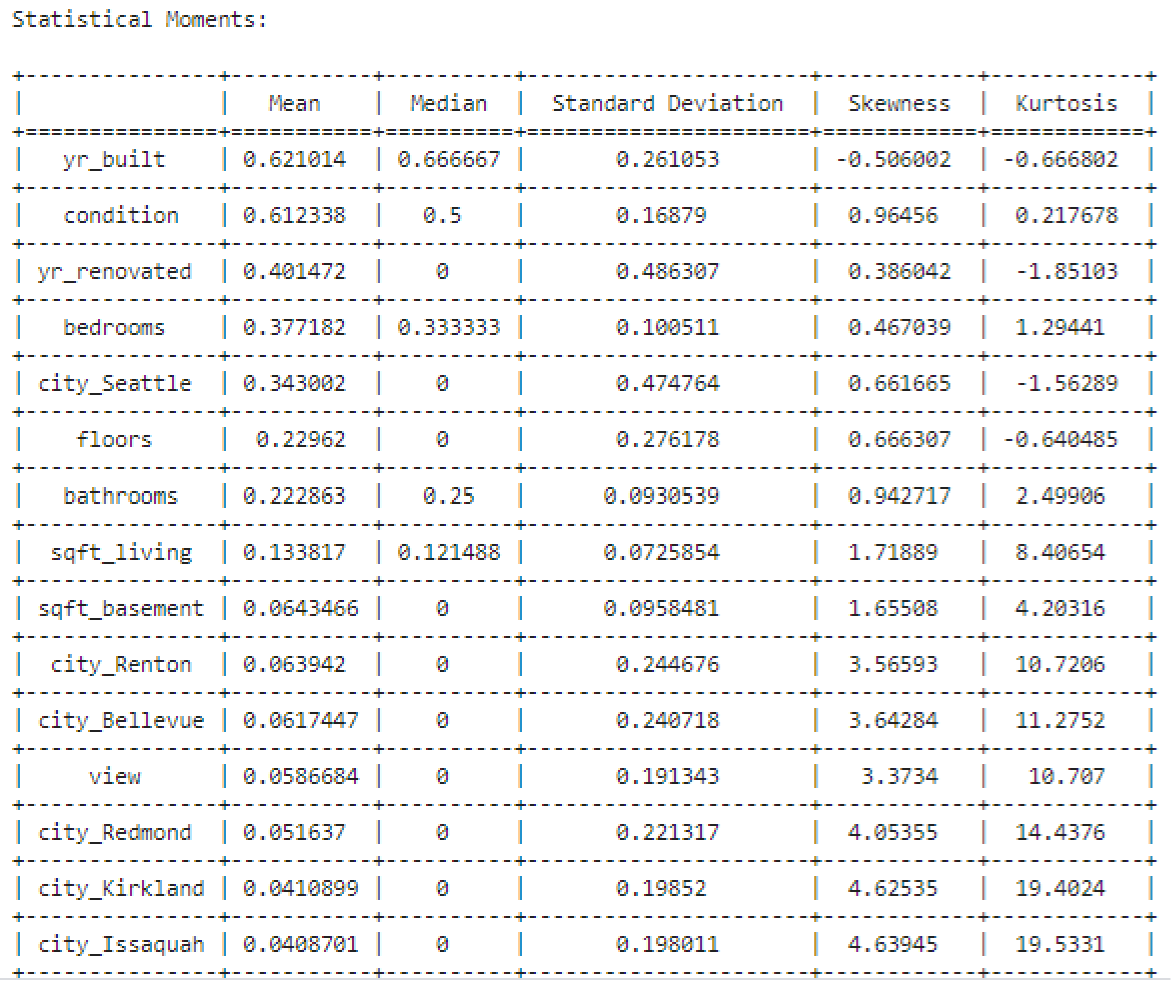




We also plotted density and spread of price we can observe that there are very few outliers present in our dataset. However price range density is uplifting.



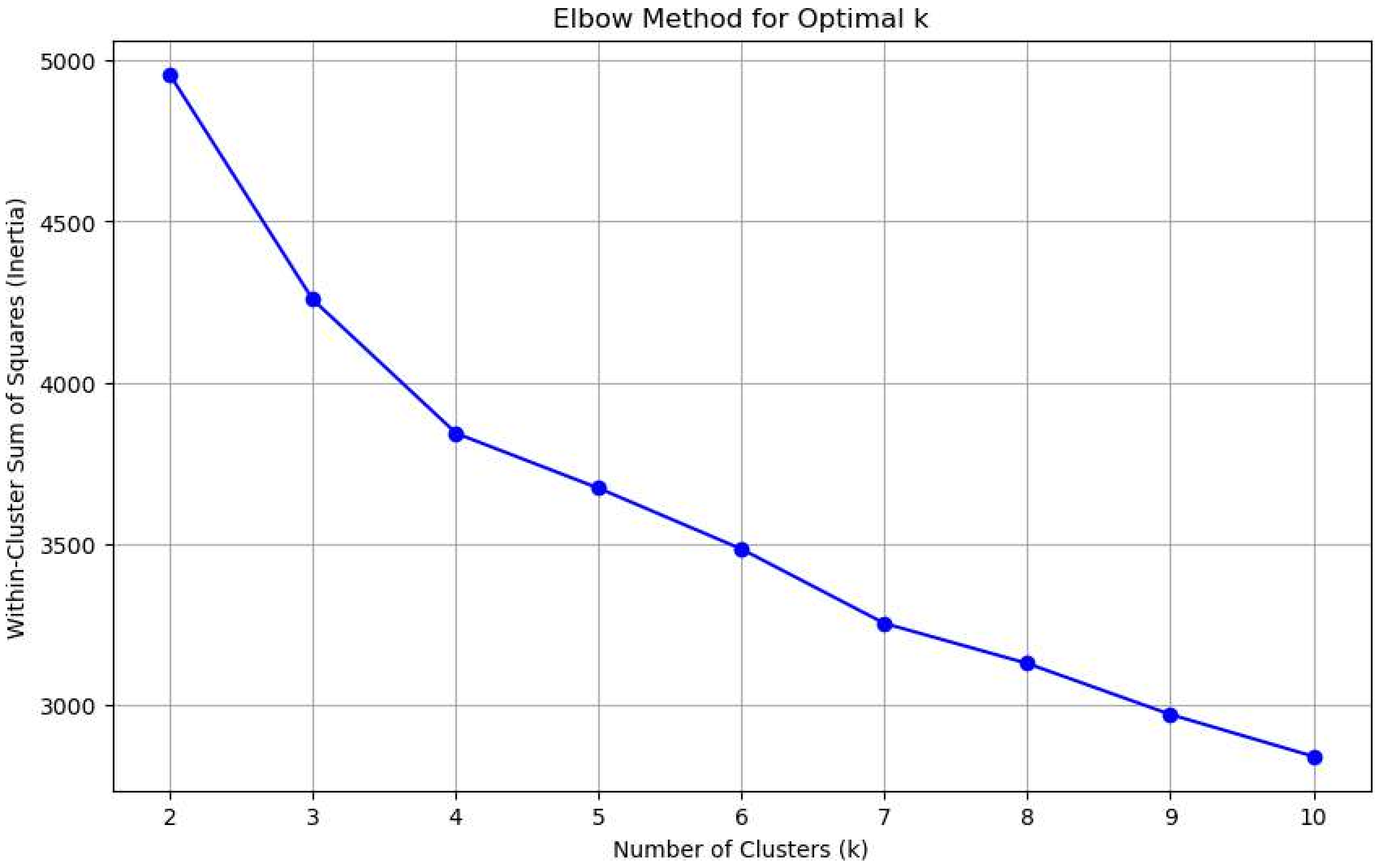
The heatmap shows the positive and negative correlation between numerical features there are some features which are strongly positively correlated like bathrooms and sqft_living and there are also negative correlation sqft_basement and floors it seems floor are less when sqft_basement value is high. Some parameters are moderately correlated.



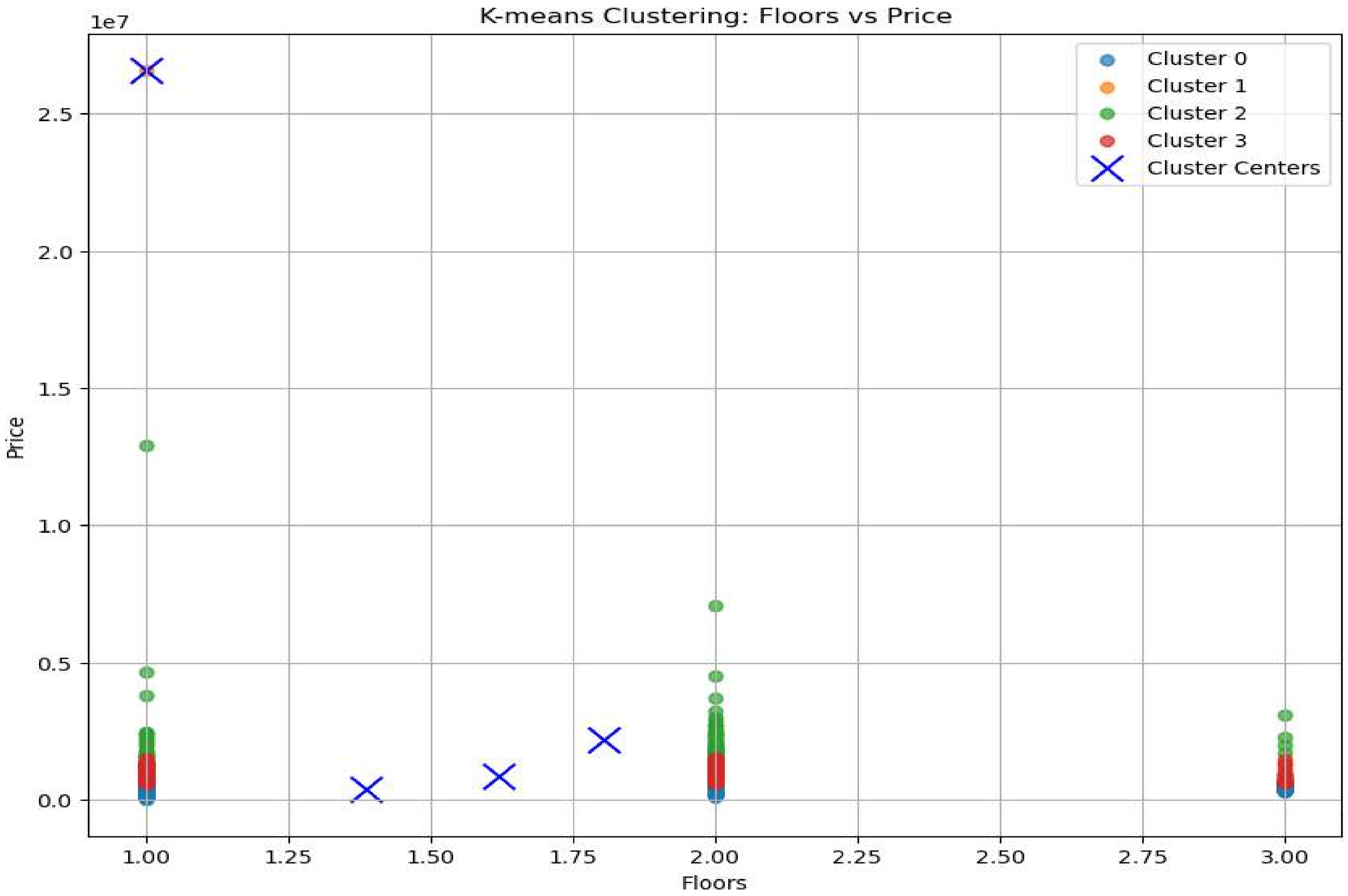
```
Statistical Moments:

+---------------+-----------+----------+----------------------+------------+------------+
|               |   Mean    |  Median  |  Standard Deviation  |  Skewness  |  Kurtosis  |
+===============+===========+==========+======================+============+============+
|   yr_built    | 0.621014  | 0.666667 |       0.261053       | -0.506002  | -0.666802  |
+---------------+-----------+----------+----------------------+------------+------------+
|   condition   | 0.612338  |   0.5    |       0.16879        |  0.96456   |  0.217678  |
+---------------+-----------+----------+----------------------+------------+------------+
| yr_renovated  | 0.401472  |    0     |       0.486307       |  0.386042  |  -1.85103  |
+---------------+-----------+----------+----------------------+------------+------------+
|   bedrooms    | 0.377182  | 0.333333 |       0.100511       |  0.467039  |  1.29441   |
+---------------+-----------+----------+----------------------+------------+------------+
| city_Seattle  | 0.343002  |    0     |       0.474764       |  0.661665  |  -1.56289  |
+---------------+-----------+----------+----------------------+------------+------------+
|    floors     |  0.22962  |    0     |       0.276178       |  0.666307  | -0.640485  |
+---------------+-----------+----------+----------------------+------------+------------+
|   bathrooms   | 0.222863  |   0.25   |      0.0930539       |  0.942717  |  2.49906   |
+---------------+-----------+----------+----------------------+------------+------------+
|  sqft_living  | 0.133817  | 0.121488 |      0.0725854       |  1.71889   |  8.40654   |
+---------------+-----------+----------+----------------------+------------+------------+
| sqft_basement | 0.0643466 |    0     |      0.0958481       |  1.65508   |  4.20316   |
+---------------+-----------+----------+----------------------+------------+------------+
|  city_Renton  | 0.063942  |    0     |       0.244676       |  3.56593   |  10.7206   |
+---------------+-----------+----------+----------------------+------------+------------+
| city_Bellevue | 0.0617447 |    0     |       0.240718       |  3.64284   |  11.2752   |
+---------------+-----------+----------+----------------------+------------+------------+
|     view      | 0.0586684 |    0     |       0.191343       |   3.3734   |   10.707   |
+---------------+-----------+----------+----------------------+------------+------------+
| city_Redmond  | 0.051637  |    0     |       0.221317       |  4.05355   |  14.4376   |
+---------------+-----------+----------+----------------------+------------+------------+
| city_Kirkland | 0.0410899 |    0     |       0.19852        |  4.62535   |  19.4024   |
+---------------+-----------+----------+----------------------+------------+------------+
| city_Issaquah | 0.0408701 |    0     |       0.198011       |  4.63945   |  19.5331   |
+---------------+-----------+----------+----------------------+------------+------------+
```

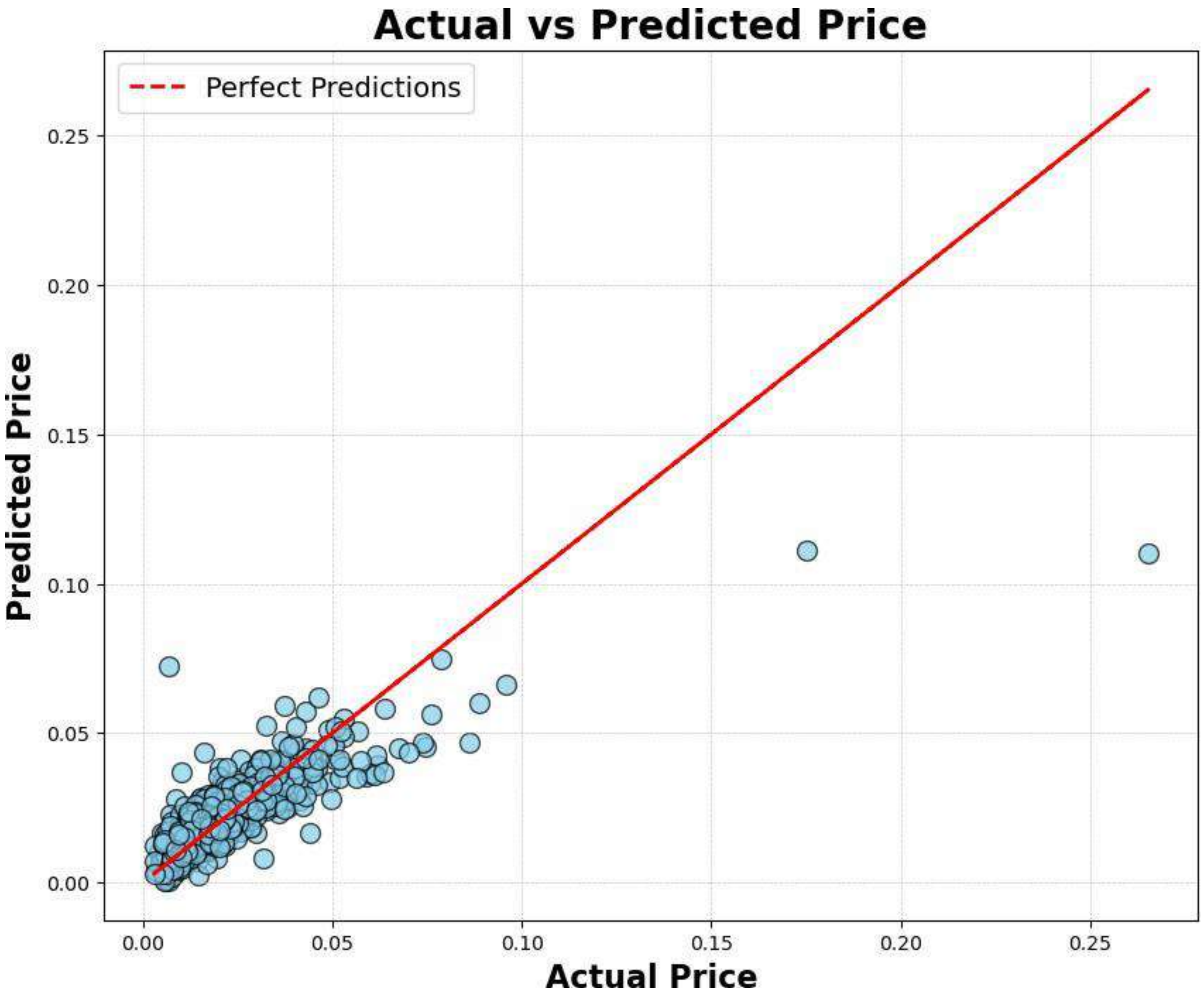
Here’s the statistical measurement of numerical featueres shown in the above table which gives me an important information of how the skewed either left or right skewed or there are values far more mean values getting standard deviation also.



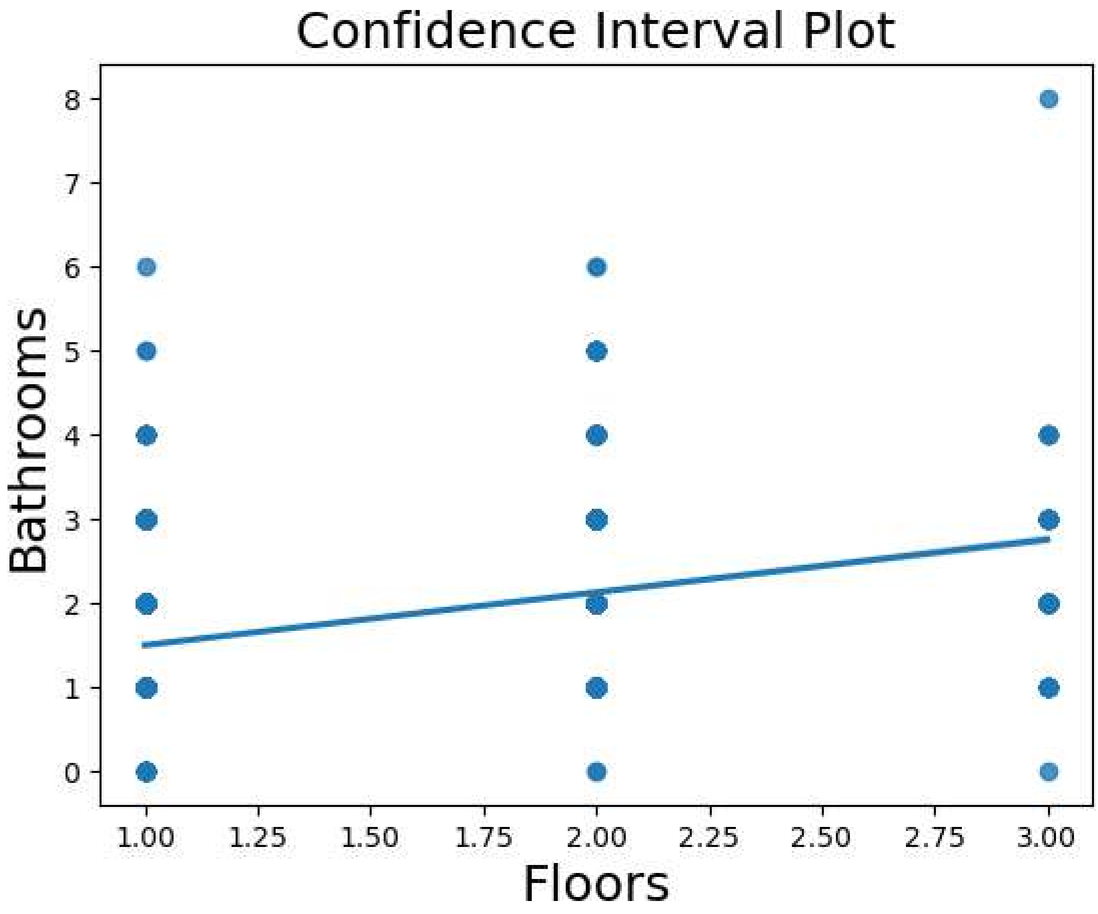
In the elbow plot, we are looking for the elbow point, after which the inertia gets smaller but by only so much. We choose 4 clusters for optimality. This point, where the rate of decrease changes most significantly, is known as the "elbow", and it's typically considered as the optimal number of clusters.



We can observe the result of KMeans clustering, Based on the elbow method, we identified that k = 4 is appropriate. Subsequently, we utilized this value to form clusters based on similarity, as depicted in the plot. We are not getting good results as it’s showing that KMeans is not well fitted.



Here we got linear regression plot results we can see that it's positively increasing it's fitted well but not increasing data. We got around 67% R2 Score which is quite low but pricing of houses is tend to be not increasing much.


Confidence Interval Plot
Bathrooms
Floors