In these R code snippets, we begin by reading a dataset named 'house\_prices.csv' using the 'read\_csv' function. We then proceed to inspect and understand the dataset's structure and content. First, we retrieve and display the column names using 'colnames(HousePrices).' Following that, we continue to display the column names once more, emphasizing the dataset's variable names. Finally, we inspect the initial rows of the dataset using 'head(HousePrices),' providing a brief preview of the data's content

A screenshot of a computer

Description automatically generated

A close-up of a computer screen

Description automatically generated

In this series of R code snippets, we begin by examining a dataset named 'HousePrices.' We calculate the number of missing values in each column and display the results. Next, we explore the structure of the dataset, including its column and row counts. We set a threshold for missing data at 50%. We then calculate the percentage of missing values in each column and filter the columns based on the threshold, resulting in a new dataset called 'filtered\_Housing\_Prices.'

A screenshot of a computer program

Description automatically generated

This filtered dataset is explored, and we display its structure and content.

Following this, we proceed to handle missing values. We remove rows with missing values in the 'Price (in rupees)' and 'Carpet Area' columns, and we eliminate the 'Dimensions' and 'Plot Area' columns. Furthermore, we convert the 'Price (in rupees)' values from Lakhs to Crores for consistency.

A screenshot of a computer program

Description automatically generated

Afterward, we display the converted values and perform a final step of removing any remaining rows with missing values, resulting in the 'Cleaned\_House\_Prices' dataset.

Additionally, we explore the 'Cleaned\_House\_Prices' dataset, check its structure, and convert categorical values in the 'Furnishing' variable to numeric representations. The code sequence is designed to clean, preprocess, and transform the data, making it suitable for further analysis or modeling.

A screenshot of a computer program

Description automatically generated

Top of Form

Bottom of Form

Top of Form

Bottom of Form

Creating a Word Cloud to show most frequent cities occurring in Data Set:

A computer screen shot of a computer code

Description automatically generated

A close up of words

Description automatically generated

There were 16 variables total but unfortunately majority of these variables were not numerical. Instead, we decided to focus on those that were already numerical, converted the Carpet Area, and converted the Furnishing from a categorical variable to a numerical variable.

A white and black text on a white background

Description automatically generated

Below we plotted the independent variables we chose in order to try to visualize a relationship. Unfortunately, it was not very clear as to where the relationship was.

A screenshot of a computer

Description automatically generated

As you can see, although there are statistically significant variables, R^2 is too low. It is basically saying our independent variables only explain .8% of the change in pricing.

A computer screen shot of a computer program

Description automatically generated

Process of cleaning up housing data for multiple regression analysis:

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A white rectangular sign with black text

Description automatically generated

We decided to narrow down to the capital of India, New Delhi, and plot this data with the 5 independent variables.

A screenshot of a graph

Description automatically generated

You can start to see a relationship between Carpet Area and Pricing so we decided to test this with New Delhi.

A computer code with black text

Description automatically generated

A graph and chart with text

Description automatically generated with medium confidence

We continued to test the larger cities in India to see if this theory worked and it was successful for Mumbai, Jaipur, and Kolkota.

A screenshot of a computer program

Description automatically generated

This led us to believe that combining all the cities into one data set was **not** a proper way to view the data given Carpet Area and Pricing vary from city to city. For example, a 750 sq foot apt in Nob Hill could be the same cost as a 2000 sq foot apartment in Tenderloin. Additionally, if there were further investigation into the data, our suggestion would be to potentially convert other categorical variables to numerical to test.

Lastly, after researching reasons as to why our independent variables were not coming out as expected we found the below potential reasons for a weak MLR:

* + The relationship is indeed insignificant
  + The larger data set could be causing a loss in degrees of freedom.
  + Independent variables could be interacting with each other (multicollinearity)
  + There are potentially missing variables

Further analysis could include confirming no multicollinearity and investigating potential additional variables not captured in this data set.