House-Rent Prediction Report

Group F

MEMBERS

|  |  |  |
| --- | --- | --- |
| NAME | STUDENT NUMBER | REGISTRATION  NUMBER |
| SUUBI TREVOR | *2000702045* | 20/U/2045/EVE |
| NAKIMBUGWE RUTH LUCKY | *2000707830* | 20/U/7830/PS |
| NABWIRE ESTHER | *2000702014* | 20/U/2014/EVE |
| WASSANYI KEVIN STEPHEN | *2000707808* | 20/U/7808/PS |
| MULUNGI STEVEN JUNIOR | *2000723507* | 20/U/23507/PS |

Contents

[Introduction 3](#_Toc115265619)

[Objectives 3](#_Toc115265620)

[Data Set Description 3](#_Toc115265621)

[The statistics for the data (data description) in the table is as follows 4](#_Toc115265622)

[Below is the summary about information on the individual columns. 4](#_Toc115265623)

[Data set preparations 5](#_Toc115265624)

[Merging of columns 5](#_Toc115265625)

[Identification Assignments. 5](#_Toc115265626)

[Dropping Columns 5](#_Toc115265627)

[Column Modifications 5](#_Toc115265628)

[Cleaning the Dataset 6](#_Toc115265629)

[o Removing null values 6](#_Toc115265630)

[o Dealing with duplicate values. 6](#_Toc115265631)

[o Removing Outliers 6](#_Toc115265632)

[1. Rent outliers 6](#_Toc115265633)

[2. BHK outliers 6](#_Toc115265634)

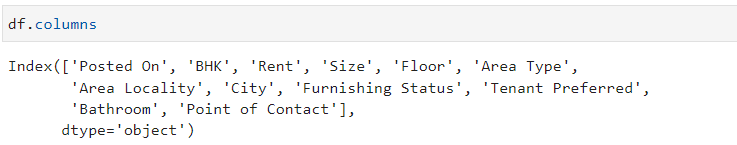
[3. Bathroom outliers 7](#_Toc115265635)

[Data Visualization 7](#_Toc115265636)

[Conclusion 9](#_Toc115265637)

# Introduction

The dataset in question gives detailed information about houses and rent, for different area localities and cities which includes the respective size, bathroom count and other information as represented in the dataset columns.



Columns in the data set

## Objectives

The objectives of the project are as follows;

* The main objective of our data analysis is to predict rent.
* To determine the factors that affect rent.
* Determine which location has the highest or lowest rent.
* Determine why a particular region has the highest rent or lowest rent

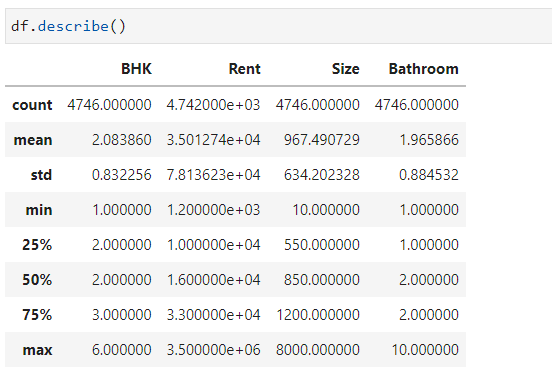
# Data Set Description

The House-Rent dataset is made up of 12 columns and 4746 rows as defined by the shape function.

The columns for the dataset include;

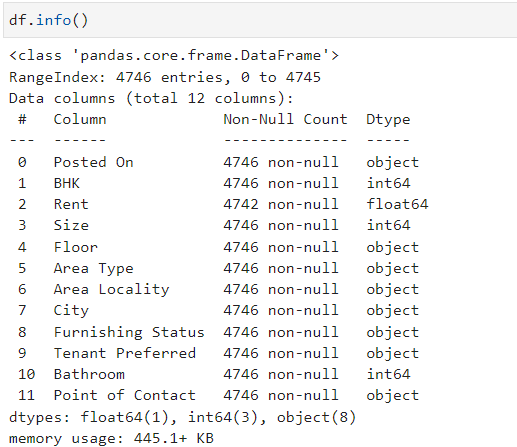
* **Posted On**: The column has the dates on which the data about the houses and the rent were recorded in to the dataset.
* **BHK**: This includes the different number of Bedrooms, Hall and Kitchen in the house. The house can 1, 2, 3, 4, 5 and 6 as defined by the unique method.
* **Rent**: The column contains the different rents for the houses.
* **Size**: Includes the sizes for the different houses.
* **Floor**: Shows or represents what floor the house is at.
* **Area Type**: shows the type of the area that the house is located in. A house can be in a super, Carpet or Built area.
* **Area Locality**: This represents the location of the houses. A house can be located in Kikoni, Kikumikumi, Wandegeya and Nakulubya.
* **City**: Includes the different cities where the houses are located at. There six cities with each city having four different area localities as defined by the unique function. The cities include Kolkata, Mumbai, Bangalore, Delhi, Chennai and Hyerabad.
* **Furnishing Status**: This includes the furnishing status of the house. The values are unfurnished, semi-furnished and furnished as described by the unique function.
* **Tenant Preferred**: This column describes the kind of tenant preferred in that particular house.
* **Bathroom:** This shows the number of bathrooms that every house has.
* **Point of Contact:** represents the person who can be talked to about the house.

## The statistics for the data (data description) in the table is as follows



Summary of data set description

## Below is the summary about information on the individual columns.



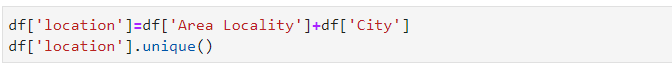
Information about the data set

# Data set preparations

This step involves preparing the data for data operations. This involves the following steps

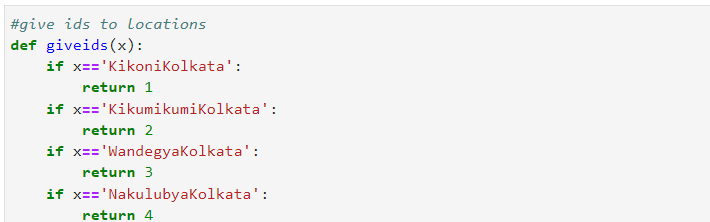
## Merging of columns

In order to simplify data operations, Area Locality and City were merged together to make Location



## Identification Assignments.

In order to be able to include columns that were strings in model training, numbers were assigned to the different values in the columns by using different functions as shown in the image below.



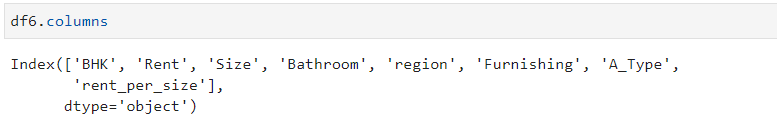
A function assigning numbers to the different values in the location column.

The function created is called and assigned to a new column created named region. This was applied to different columns including Area type and Furnishing Status where their ids were assigned to A\_type and Furnishing respectively.

## Dropping Columns

Some columns that were considered unnecessary after Id assignments were dropped and also columns like Posted on and Floor that seemed not to affect the data set were also dropped.

This left us with the following columns.



Columns that were retained for data analysis.

## Column Modifications

It was necessary for some modifications to be done on some columns in the data set for better data operations. That is to say rent\_per\_size column was created which had the rent per size to assist in data cleaning.

Cleaning the Dataset

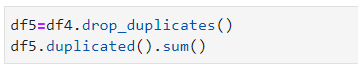
This involves detecting and correcting corrupt or inaccurate records from a given table or data frame. This process involves the following steps

## Removing null values

We checked for missing and duplicated data using the isnull (). From our analysis we discovered that they were 4 null fields under the Rent column which were removed from the data set.

## Dealing with duplicate values.

Duplicate values are values that are repeated in the dataset. Only Three values were found and removed in the dataset.

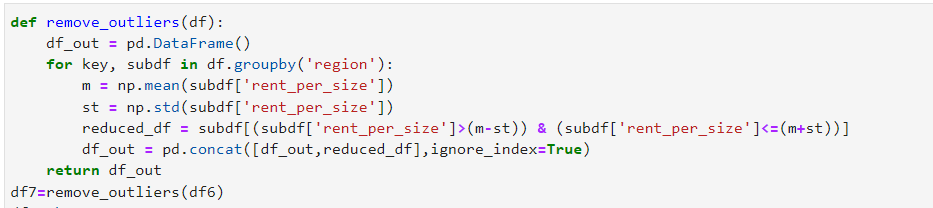


## Removing Outliers

Outliers are values in the data set that don’t make sense. Basing on our objective, we noted that there were some irrelevant fields that weren’t going to help us in further analysis and our prediction.

### Rent outliers

It is always advisable to work with normally distributed data since naturally most data is normally distributed and any extremities are in most cases brought about by human errors. In order to remove rent outliers, all values that had rent that did not exist between the mean and the standard deviation were removed so as to remain with normally distributed data.



Function to remove rent outliers

### BHK outliers

For BHK outliers, a function that removed all values in the dataset with big BHKs that had a rent smaller than Smaller BHKs were removed. For example, A 2 BHK whose rent is cheaper than a 1 BHK.



Function to remove BHK outliers

### Bathroom outliers

For bathroom outliers, Data that had Bathrooms that had bathrooms that were twice the BHKs were removed.



Removing bathroom outliers

### Posted On outliers.

In the column posted on. It was also necessary to change two values that were not in their correct format. We chose to change them since the values were valid but not in their correct formats.

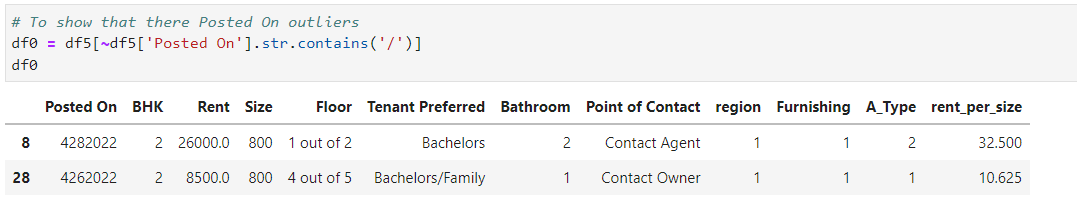


Image showing the values in the posted-on columns that were changed



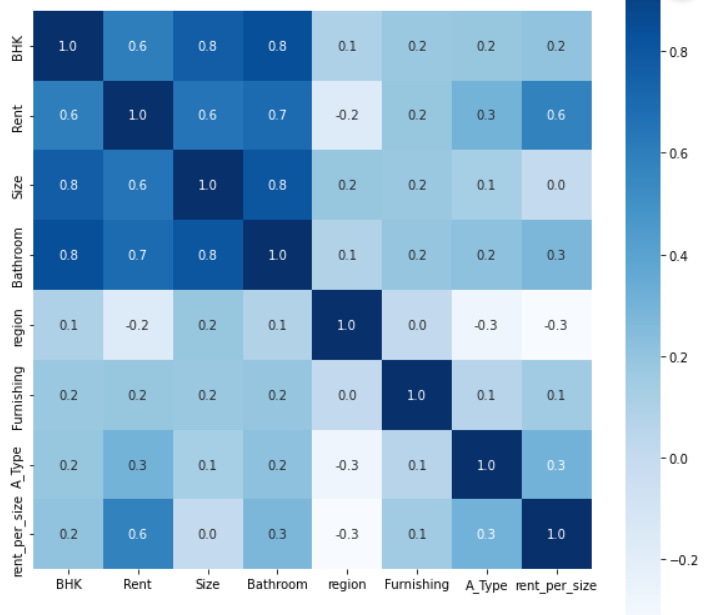
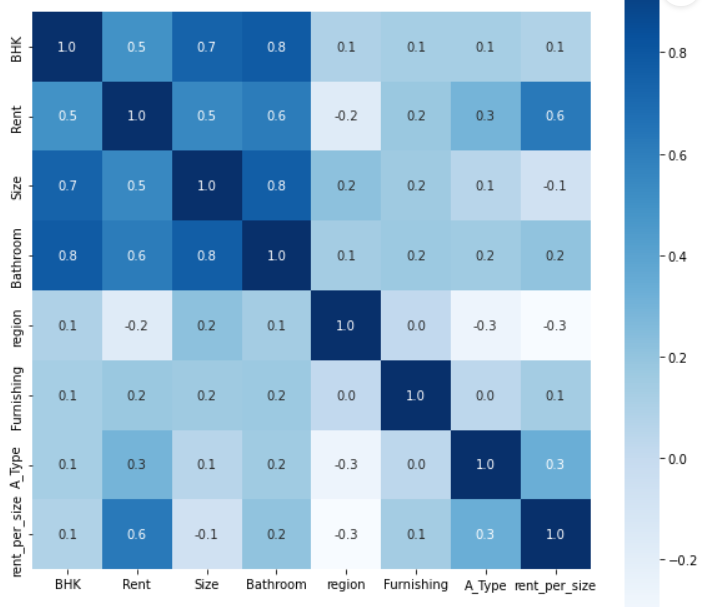
Correcting date outliers

# Data Visualization

In this section, we include some of the diagrams and graphs we used to visualize our data.

We studied the correlation between all the variables, so as to better understand which fields affect the rent the most, and find out how they each vary with each other. We plotted various graphs and even included heatmaps before and after cleaning our data.

Before cleaning Correlation after cleaning



We concluded that there was an increase in the correlation after the removal of outliers.

We also included a pie chart to show the distribution of rent in the different cities, and a bar chart to show how rent varied in the different locations.

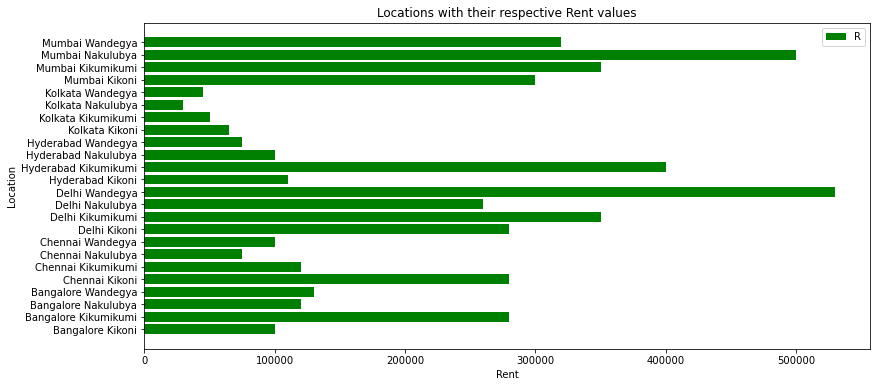
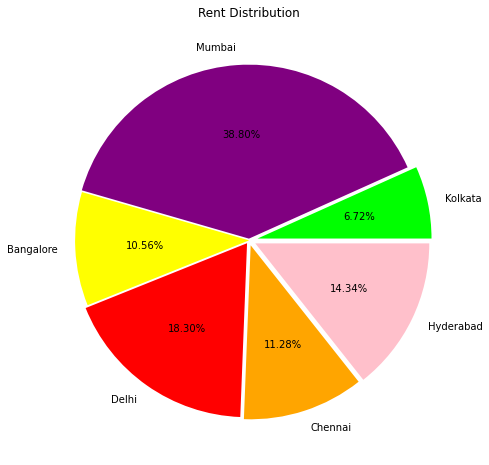
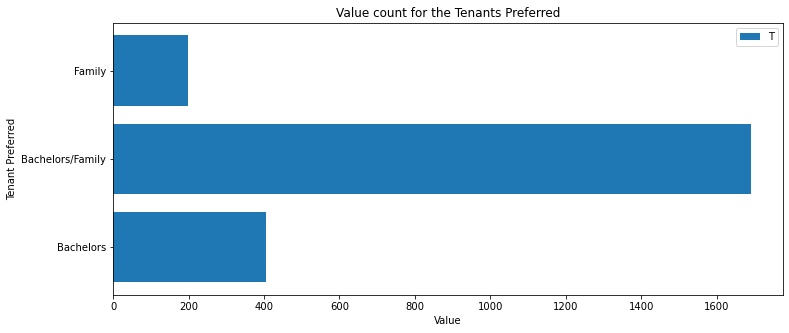


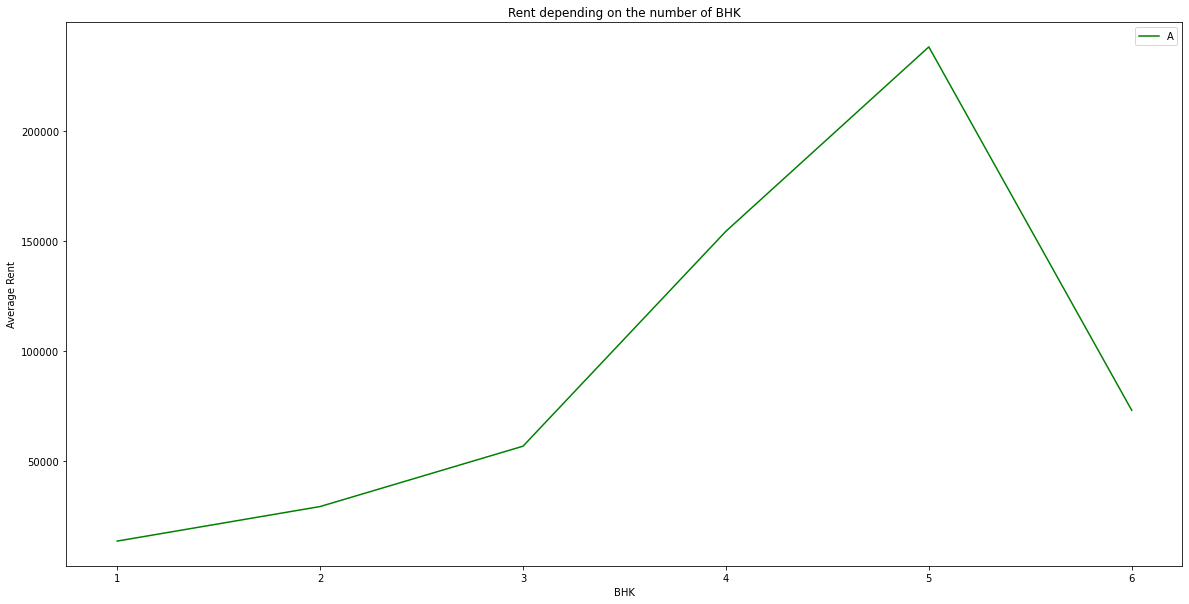
Figure 3.

We can also see from the graph below which tenant was preferred the most according the bar chart.

In this case, it was the Bachelors/Family that were preferred the most.



We also displayed how the rent varies with the BHK (number of bedrooms, halls and kitchen) as shown below:



# Conclusion

After cleaning the data, we created machine learning models using the dataset and trained and tested these models so that we can use them in our predictions. This involved splitting the model into training and testing sets. Then we used a library called Scikit learn or sk-learn to train the model and make our predictions.

The sk-learn library has a number of methods we can use in our machine learning model. They include regression techniques, classifiers and more. Among the classifier techniques we have the DecisionTreeClassifier, using this we got the following results:

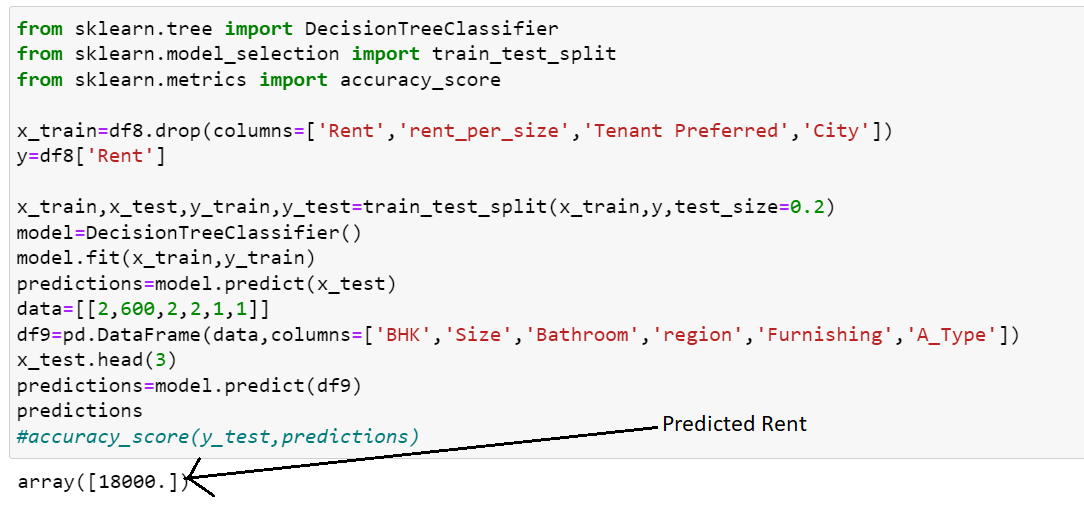
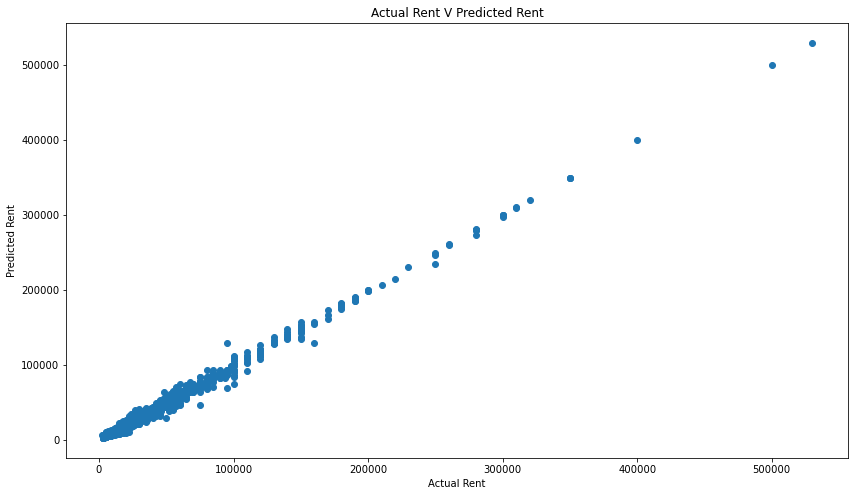
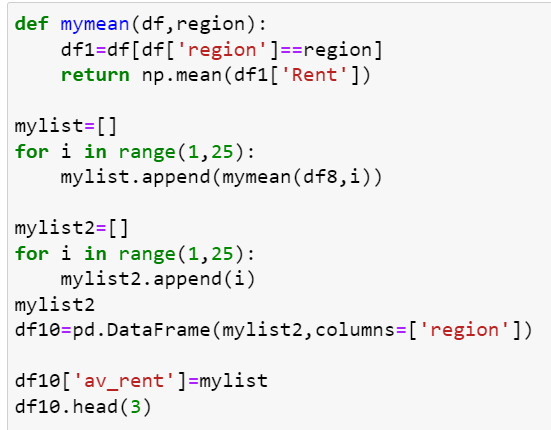


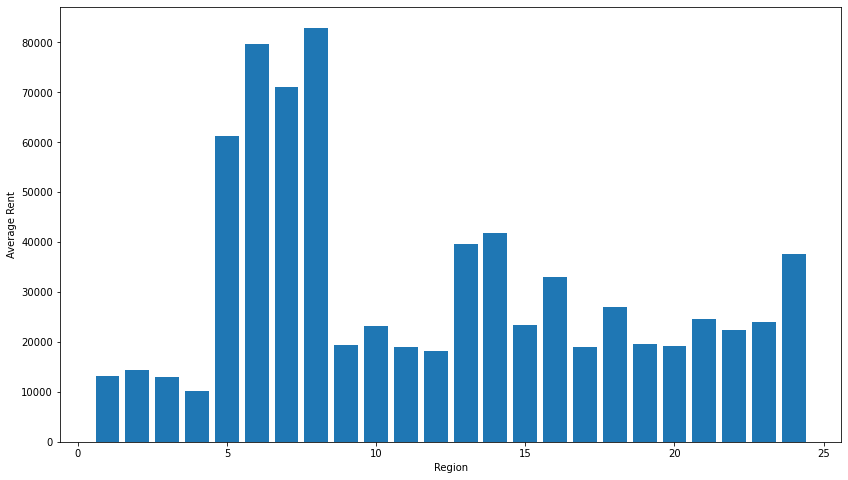
Figure 4.

Lastly, we studied the correlation between the predicted rent and actual rent to further see how accurate our predictions were.



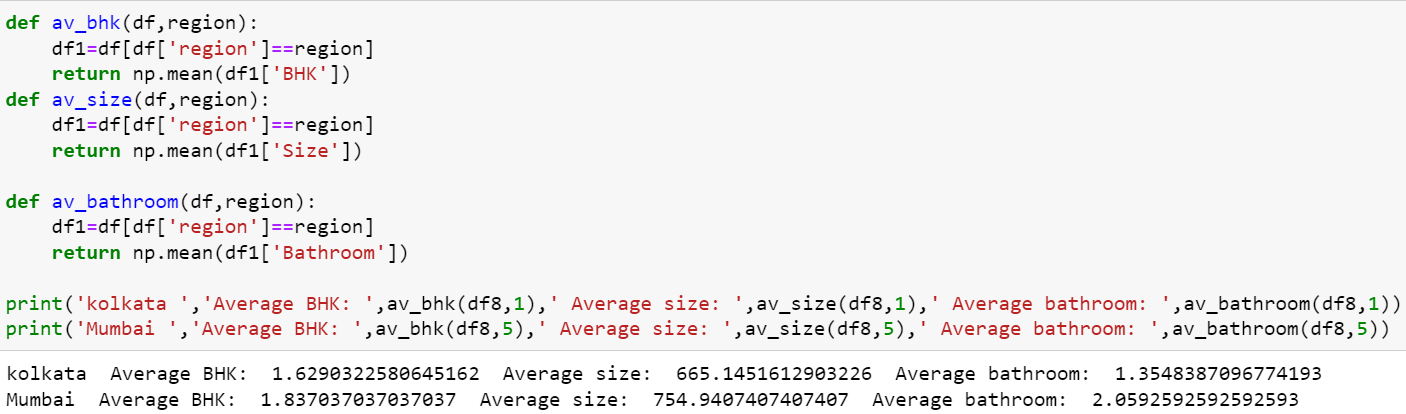
One of our objectives for our analysis, was to find which locations had the highest average rent or lowest, and also the reason why it is so. Our graph alone in Figure 3. 1 wasn’t enough to give us the exact values, so we wrote functions to help us get the average rent for each region.

We used this function on the left to draw a graph for average rent against each region.

**Note:**

* Region 1-4 is Kolkata City
* Region 5-8 is Mumbai City
* Region 9-12 is Bangalore City
* Region 13-16 is Delhi City
* Region 17-20 is Chennai City
* Region 21-24 is Hyderabad City

As we can see Mumbai city has the highest average rent and Kolkata has the lowest.We then wrote a few functions to help us tell why the two cities above had such a distribution of average rent, and we discovered the following:



Each of these functions were basically calculating the average BHK, Size and Bathroom for a particular city, from the results above we concluded that because Mumbai had a higher average BHK, room size, and bathroom count than Kolkata, it had to have a higher average went.

Therefore for any customer willing to rent a house for a cheap price, we’d advise them to visit Kolkata.