**Singapore AIRBNB**

**Team Details:**

|  |  |  |
| --- | --- | --- |
| Name | SRN | Section |
| S S Priya | PES1UG19CS404 | G |
| Sadhvi Sushravya H S | PES1UG19CS410 | G |
| Sathi Sreyahsri | PES1UG19CS434 | G |
| Sejal Priya | PES1UG19CS441 | G |

**Table of Contents**

[Abstract](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.2et92p0) 2

[Introduction](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.26in1rg) 2

[Dataset Description](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.3dy6vkm) 2

[Preprocessing and Data cleaning](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.1t3h5sf) 3

[Exploratory Data Analysis](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.4d34og8) 3

Normalization and Standardization 7

[Hypothesis Testing](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.17dp8vu) 7

Correlation 9

[Results and Conclusion](https://docs.google.com/document/d/13P7uv2s6-kIHqwuuWIhC2rFOqPePxCLn/edit" \l "heading=h.3rdcrjn) 10

**Abstract:**

The dataset under consideration contains data like the name of the Airbnb, the name and ID of the host etc. It also has the information about the geographical location of the lodging. This project is about combining programming skills, knowledge of mathematics and statistics to extract meaningful insights out of the raw data. Our first task was to clean the raw data to obtain a clean dataset without empty cells or any outliers which we performed using different libraries from Python. Then we visualized our data using different techniques like boxplots, bar charts and histograms. We then normalized the data whose goal is to change the values of numeric columns to a common state without distorting differences in the range of values. And then the Hypothesis testing was performed (Using Z-scores) to check if the Null Hypothesis is rejected or not where we got to know that the Null Hypothesis is plausible. And as the final step, the variables in the data are checked for mutual relationship using Correlation.

**Introduction:**

Airbnb is unique for its name and also the facilities, which matches hosts with guests, as part of the sharing economy. It first started in San Francisco and is now making great progress by also having branches in cities like New York, Tokyo, Singapore, Seattle, etc. Our dataset is on the data collected from different tourist spots about the accommodation facilities present in that area.

**Dataset:**

Source: The dataset is taken from Kaggle.

(https://www.kaggle.com/jojoker/singapore-airbnb)

The dataset consists of 12 columns, each attribute representing a feature of a course

This dataset is from Singapore that contains data like the name of the Airbnb i.e., the lodging or the homestay, the name of the host, the ID that is given to the host, etc. The data from Airbnb also gives information about the neighborhood and also the latitude and longitude of the place, which reduces the burden of worrying about the directions to the place from the shoulders of the customer. Our dataset contains a total of 16 columns where each of the columns represents different attributes like name, ID, cost, etc. It has information about 7907 places of accommodation which shows the growth of a 12-year-old company into a successful one. There are a few empty cells where we can see the absence of any data regarding that attribute. There are 5518 empty cells for a total of 126528 cells which makes the percentage of empty cells 4.361%.

Column wise description :

In total there are six columns of categorical values, and 10 columns of numerical values.

* **ID** - A unique code of the location of the Airbnb
* **HOST\_ID** - An ID provided to each host of Airbnb
* **LATITUDE** - Latitude of the Airbnb Location
* **LONGITUDE** - Longitude of the Airbnb Location
* **PRICE** - Cost per night in the room
* **MINIMUM\_NIGHTS** - Minimum duration of stay
* **NUMBER\_OF\_REVIEWS** - Reviews available for a particular location
* **LAST\_REVIEW** - Date of the last review of the Airbnb
* **REVIEWS\_PER\_MONTH** - Number of user reviews per month
* **CLACULATED\_HOST\_LISTING\_COUNT** - Number of total customers who have stayed in a particular Airbnb
* **AVAILABLITY\_365** - Number of days available in the year
* **NAME**- Name of the Airbnb
* **HOST\_NAME**- The owner of the Airbnb
* **NEIGHBOURHOOD\_GROUP** - The region the Airbnb is located
* **NEIGHBOURHOOD** - Name of the locality of the Airbnb
* **ROOM\_TYPE** - Type of room ( private / shared / whole apartment or house)

**Data Cleaning:**

Techniques used to clean our dataset:

1. Handling missing data by dropping and imputing

2. Removing unwanted observations

3. Fixing the typos and inconsistent capitalization.

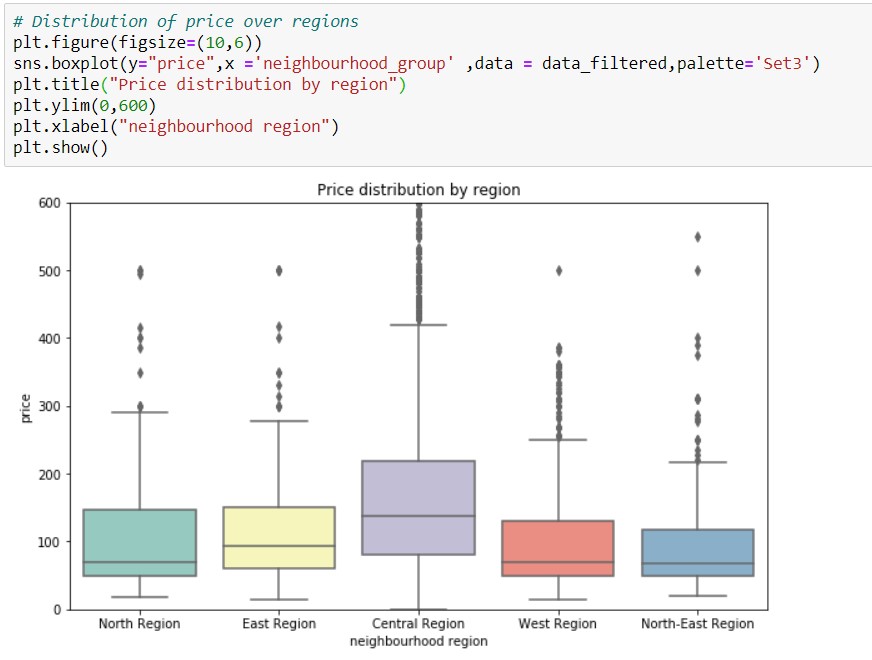
Why is cleaning required in our data set?

As stated earlier 4.361% of our dataset contains empty cells so we need to implement the above methods to clean our dataset. We find the empty cells using .isnull() function. If we observe in ‘least\_reviews’ and ‘reviews\_per\_month’, we find that there are many null values, so we decided to drop the whole column. We now fill the empty categorical column with null. We decided to capitalize all the categorical data. Finally to remove outliers we used a box plot.

**Exploratory Data Analysis:**

We used four graph visualization techniques for our Exploratory Data Analysis.

1. **Boxplot**: It represents a box the middle line represents median, left or the bottom -1st quartile and right or top-3rd quartile. Left whisker contains min value. Right whisker contains max value. They also shows any outliers if present.



Here to visualise distribution of

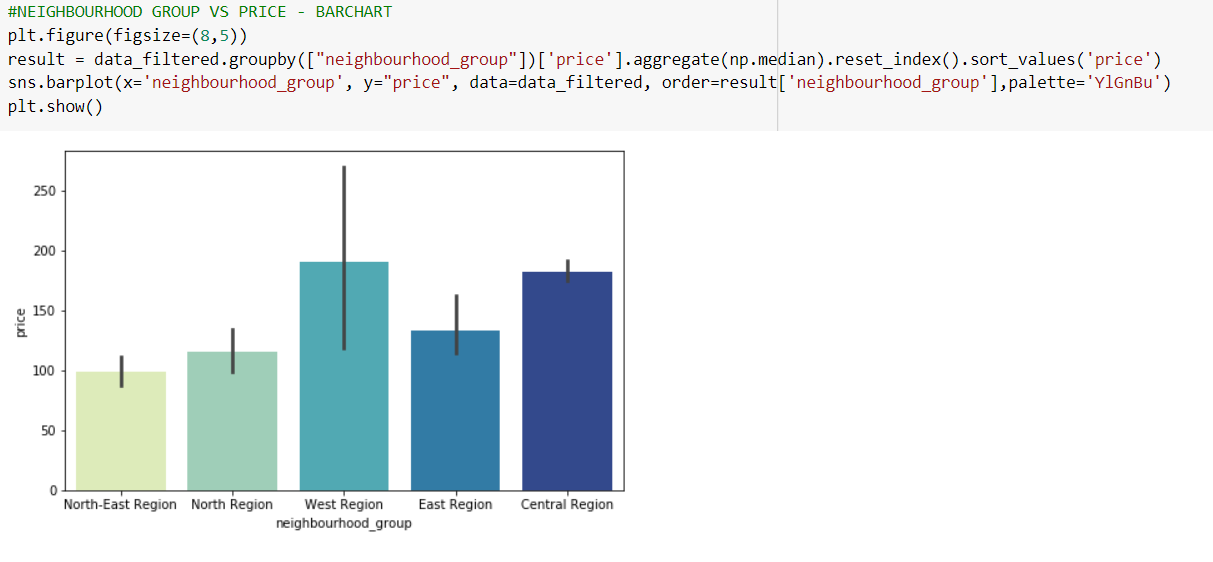
->The price feature in general is positively skewed over all regions.

->Minimum price of rooms is almost same for all regions

-> The central region also shows the maximum range in prices while the north east region shows the least.

1. **Barchart**: Presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

a)



The lines represent error intervals. Which represents confidence/ standard deviation. Seaborn takes the avg by default.

The bars here indicate the the neighbourhood groups.

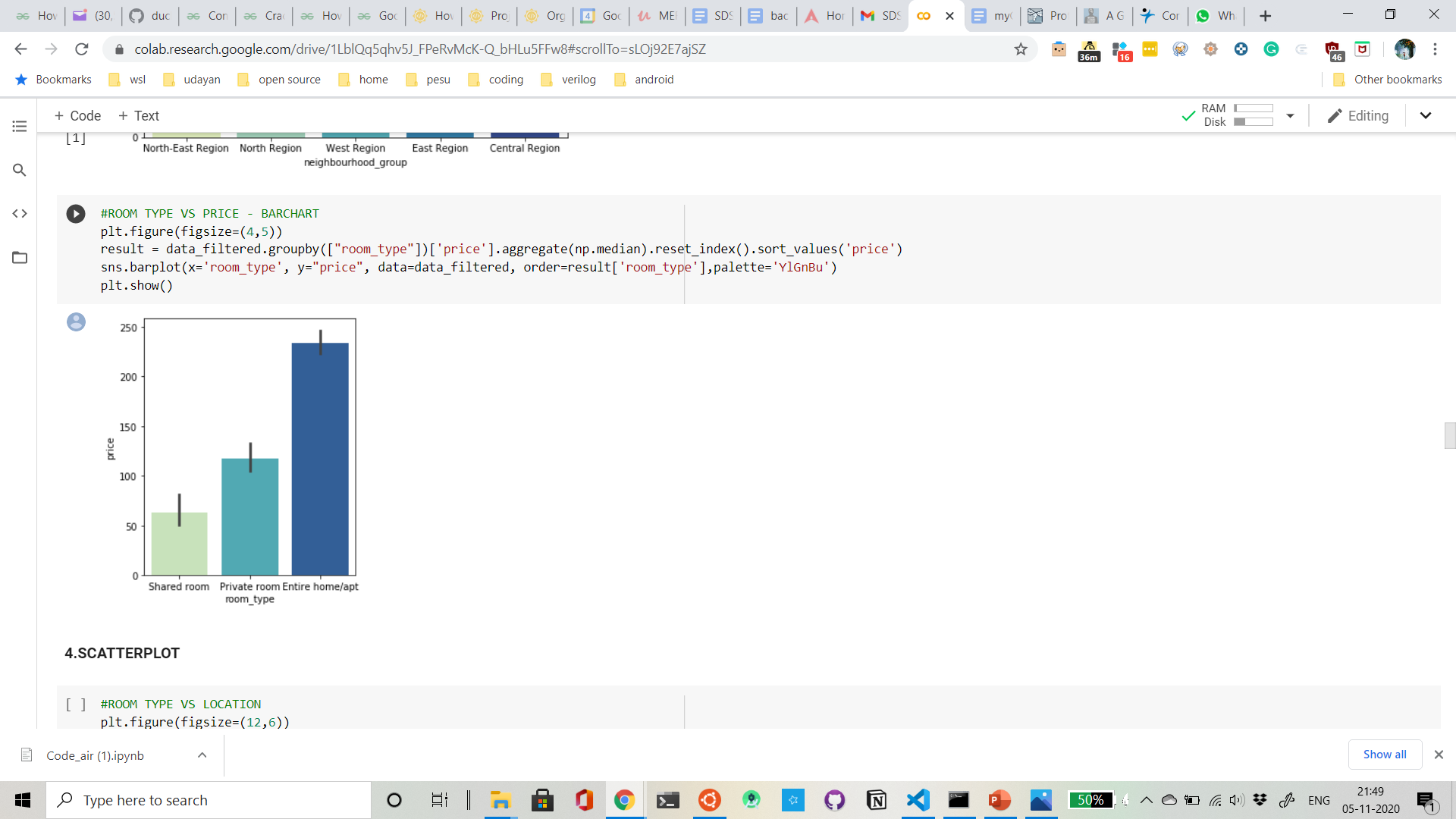
Though west region might have the maximum number

Minimum price.

But as the standard deviation is more compared to central region.

so we can conclude that it is maximum for the central region and not west region.

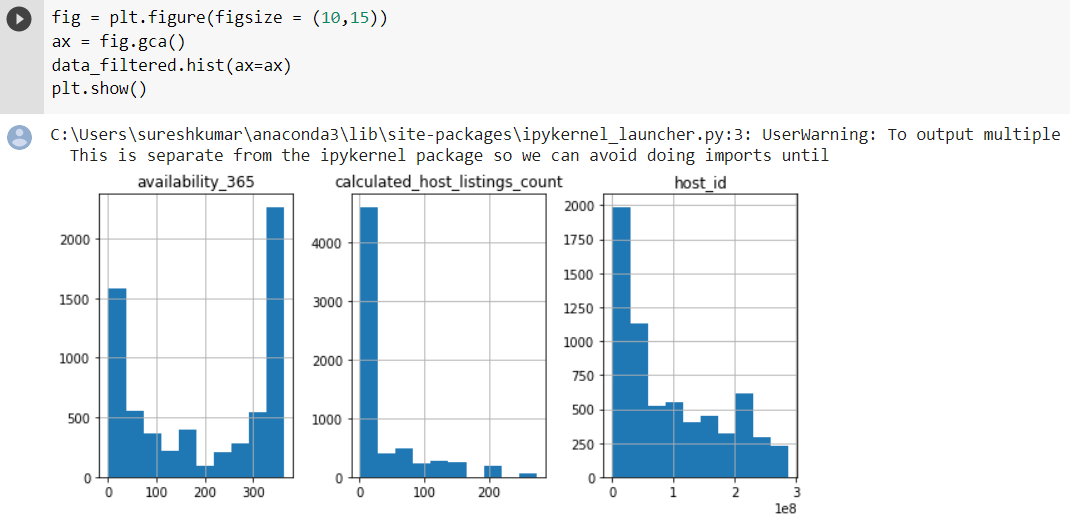
b)



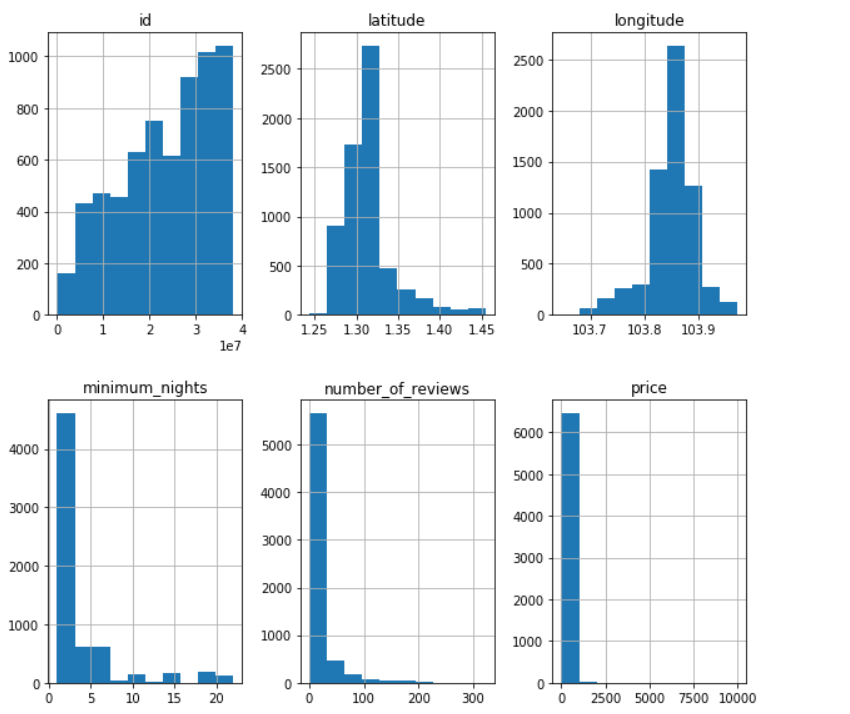
The price for the entire home/apt type is the maximum while that for shared rooms is the least.

1. **Histogram**: Representation of distribution of numerical data, it represents frequency of continuous data.

a)



b)



**Normalization and Standardization:**

The goal of normalization is to change the values of numeric columns to a common state without distorting differences in the range of values. Standardizing the values around the center with mean 0 and a standard deviation of 1 is important when we compare measurements made on different scales or in different units. Without normalization and standardization, our predictions can lead to a biased result where variables with a larger ranger will tend to dominate the prediction.

We used the z-score formula for normalization in this project which is as follows:

Z-score=(x-mu)/sigma.This normalized all numerical columns to mean 0 and standard deviation of 1.

**Hypothesis Testing:**

The sample we considered for the hypothesis test we conducted in the first 50 entries of the eastern region and the first 50 entries of the western region. Can we conclude that the mean price of rooms in the East and the West region is less than the population means?

We will use a 0.05 level of significance (5% statistical significance).

Null Hypothesis(H0): Mean price of rooms in the East and West region is greater than or equal to the population mean.

Research Hypothesis(H1): Mean price of rooms in the East and West region is smaller than the population mean.

We have calculated the Z values and through Z values we have calculated the ‘P’ value.

The value of Z, for Z-score can be known by the formula,

**Z**=(**x**-𝝁)/𝝈

Where **Z**->Z score

**x**->Test Score

𝝁->Population Mean

𝝈->Population Standard Deviation

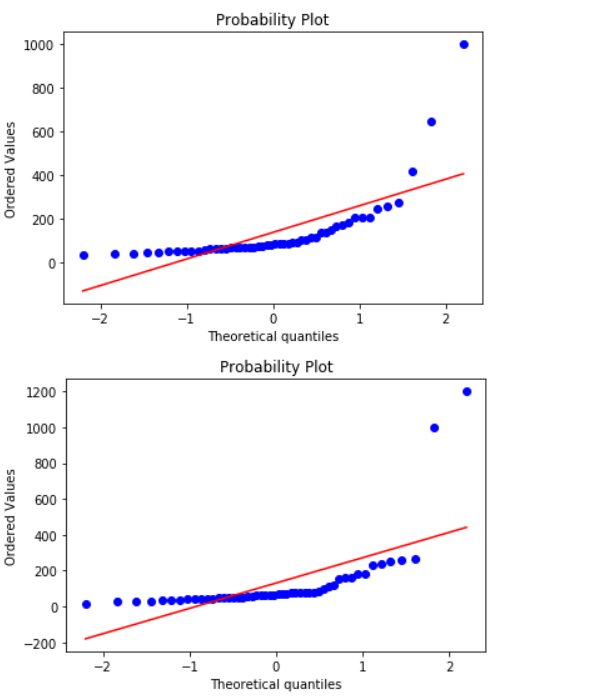
We calculate the value of P using a left tailed test and the result of the test is as follows:

The P-value for the East region will be 0.0979(9.79%) and for the West region, it will be 0.0988(9.88%) which is clearly greater than the level of significance (0.05 or 5%).

Since the P-value is greater than the level of significance, we cannot reject the Null Hypothesis and confidently conclude that the Null Hypothesis is plausible.

We have also plotted two Probability Plots.

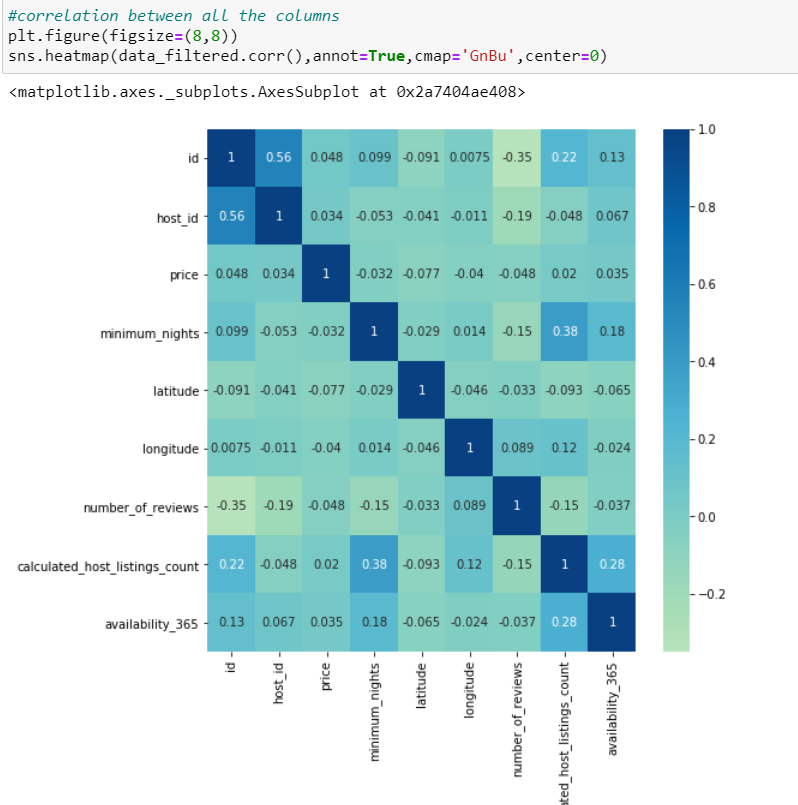
1. Probability Plot of the Eastern region
2. Probability Plot of the Western region



Mean price of rooms in east and west region is greater than or equal to the population mean price.

**Correlation**:

Correlation is a statistical measure that expresses the extent to which two variables are linearly related.



Heatmap showing the correlation between all the columns in the dataset.

We notice that the deeper the color the more is the correlation.

We also calculated the value of the Correlation coefficient between two variables(price and minimum nights) using the python libraries to obtain the Correlation coefficient to be close to zero which infers that the two variables have no relation.

of rooms in east and west region is greater than or equal to the population mean

**Results:**

1.Our dataset is collected from hotels in Singapore commissioned by Airbnb which were last reviewed between 2013 and 2019 and serves as a sample of data from a larger population of all hotels in Singapore (commissioned by Airbnb or otherwise).

2. While cleaning our dataset we found some outliers in price and minimum\_nights column, we removed the outliers when the value is between upper and lower limits for each column.

3.The range of price in our dataset is 0 to 10000 and that of minimum nights is 1 to 1000 and obviously both are measured in different scales so if we would not have normalized or standardized the data, price would have dominated the prediction for our model which involved both variables. And hence the importance of normalization of data is seen.

4. The value of the Null hypothesis being greater than 0.05(0.0979 for the East Region and 0.0989 for the West Region) shows that it cannot be rejected and both Null and Alternate Hypothesis are plausible.

5. The correlation coefficient that is close to zero between price per night and minimum number of nights for which booking can be done was found showing that the price of a room per night and the minimum nights for which a room in a hotel can be booked are not related.

**Conclusion**:

We determined how the price for the hotel varies according to the minimum nights booking required by the hotel and also according to the region of the hotel.