Statistics for Data Science (UE19CS203) Report

Studying patterns in Singaporean AirBnB

Team Bare Minimum

Section : A

-Anagha H M (PES1UG19CS057)
-Ananya Uppal(PES1UG19CS058)
-Anchal Sharma(PES1UG19CS059)
-Anisha Ghosh(PES1UG19CS067)

Abstract

In this project, our aim is to determine relationships between variables such as location, room type and reviews and how they impact the price of an AirBnB in Singapore. A comparison was also made between the price of the rooms in the west and east region of the country with that of the total population. We started off by examining the dataset to clean it for any structural errors and remove outliers. This was followed by an extensive analysis of all the parameters to come up with fruitful conclusions based on the patterns observed. However, our null hypothesis was rejected proving that despite our assumptions, the prices of the areas were different from what was expected.

Introduction

Our assumption was that the mean price of rooms in the east and the west regions of Singapore would be approximately equal to the mean price of all rooms in Singapore. The dataset included a variety of variables for each airbnb in Singapore including price, location, availability, etc.

The theory behind this assumption was that airbnb rooms would be scattered evenly across the country, and the prices would also vary evenly according to the location. Therefore the price of Airbnbs in the east and west regions would be an accurate representation of the whole country. We also assumed that location would be the only variable affecting the price of an Airbnb.

Dataset

Dataset consisting of 7907 rows and 16 columns from the Singapore AirBnb listings.
Source: kaggle.com

- **ID** - A unique code given to the user.
- **HOST ID** - An ID provided to each AirBnb location
- **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

- **CALCULATED_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)

Preprocessing or Data Cleaning

To begin with we calculated the no. of null values in each column and the results were as follows, 2 columns with 35% null values (last_reviews & reviews_per_month) were **dropped** as they had data which wasn't relevant to our analysis.

name	:	2
last_review	:	2758
reviews_per_month	:	2758

Missing Names In 'Name' column: To remove rows with no values in column 'name' they were replaced with the value 'NULL', as name was a column of not much relevance, instead of losing data and dropping those rows, we decided to replace those with 'NULL'

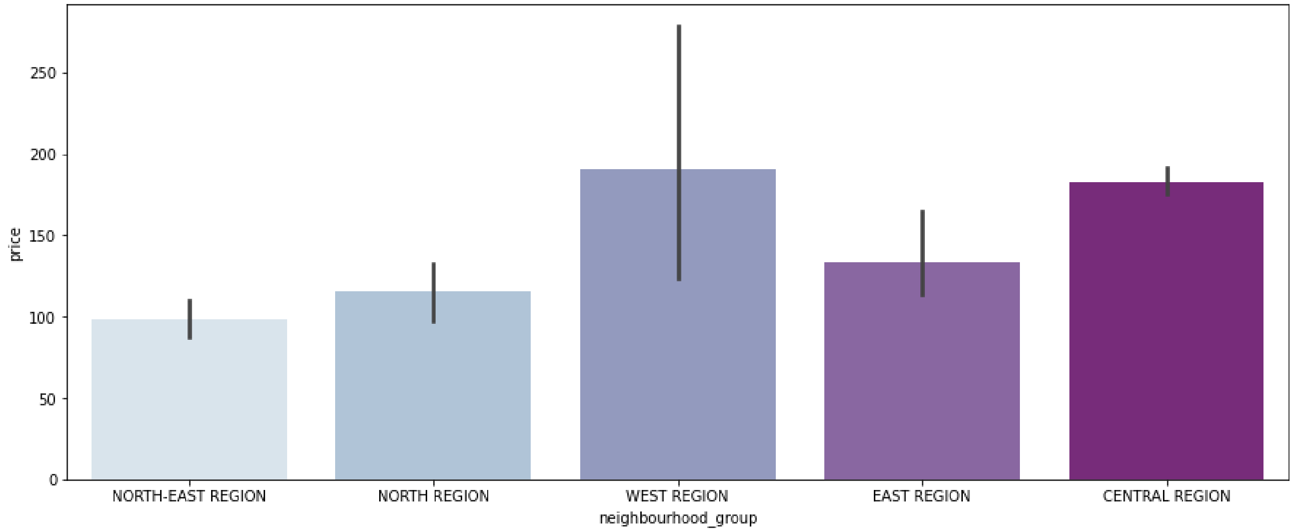
Capitalising Categorical variables: To remove any structural error(inconsistent capitalisation). Inconsistency capitalisation can lead to formation of repeated categories under varied names causing errors in analysis, hence all the categorical variables were capitalised.

Removal Of Outliers : To remove values which do not lie within the range min and max values calculated, this was one of the major steps in our cleaning process. Over 1000 rows were deleted because they consisted values that fell outside the min,max region

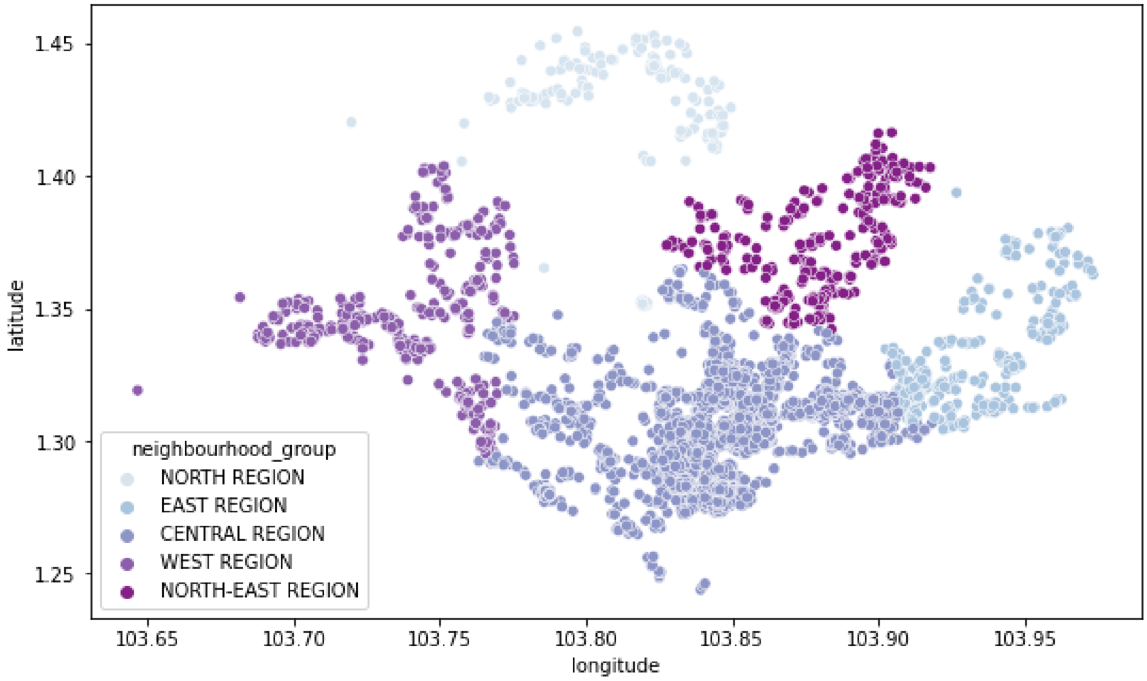
Exploratory Data Analysis

Exploratory Data Analysis is a method of **Data Visualisation** to determine patterns among the various columns in a data set.The graphs plotted help in formulating the null and alternate hypothesis.

The graph given below is a **barchart** plotted for neighborhood group VS price from which we can conclude that the rooms in **western region** have **maximum mean cost**.



The **scatterplot** given below visualizes the distribution of Airbnb's along the latitude and longitude. It closely resembles the map of Singapore. From this, we can conclude that the **maximum concentration** of Airbnb's is in the **Central Region**.



Hypothesis testing

The **Null hypothesis** states that the mean price of the rooms in the east region and the west region is equal to the population mean, the **alternate hypothesis** is that the mean price of the rooms in the east region and the west region is smaller than the population mean
The mean of the sample is equal to 169.3329

$$H_0: \mu = 169.3329$$

$$H_1: \mu < 169.3329$$

This is a left-tailed test, and the method used for hypothesis testing was the z-test, owing to the sample size of 50. We begin by assuming H_0 is true. We calculate the mean of the east region and the west regions.Mean of the east region= 139.4 and mean of the west region=130.84.

The p values are calculated and they are:

- 0.0979 for east region and population mean.
- 0.0988 for west region and the population mean

We considered the significance level at 10%(alpha=0.1), the p value calculated was lesser than 0.1 hence, the null hypothesis was rejected

Results and Discussion

From the z-table distribution, we found the p-value=0.09, we can conclude that

$$p < 0.1$$

Owing to this result, we **reject the null hypothesis, and accept the statement of the alternate hypothesis**. Therefore, we can conclude that the mean price of the rooms in the east region and the west region is smaller than the population mean