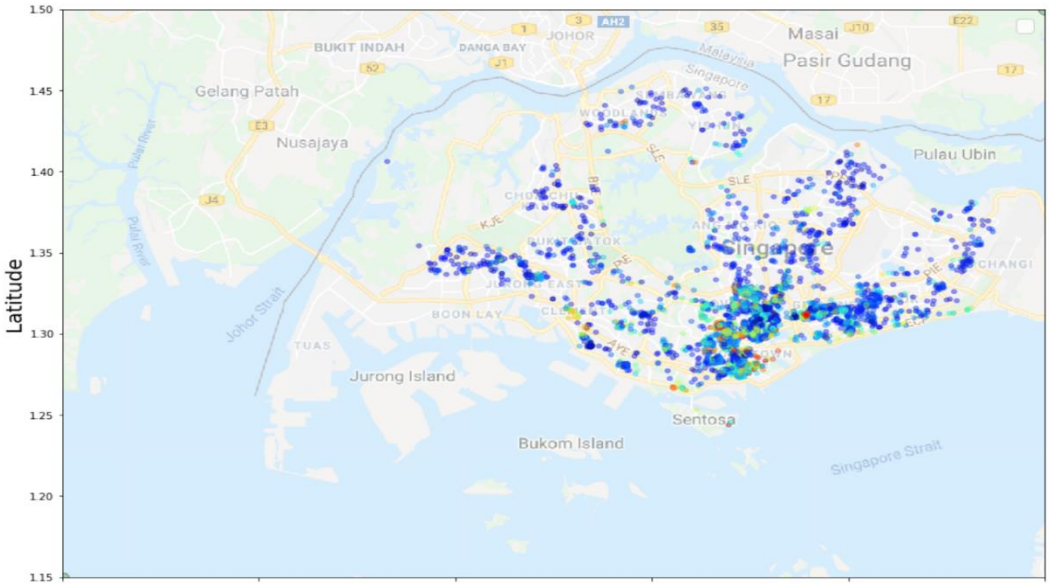
DATA 602:

Assignment 1 – Report

Sana Sharma

Singapore Airbnb Data Prediction



Abstract

Airbnb provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities. Guests can search for specific types of homes, such as bed and breakfasts, unique homes, and vacation homes etc. This dataset allows us to explore how Airbnb's are being used in the different regions of Singapore in 2019.

Introduction

The Airbnb's are growing rapidly in Singapore and generating highly comprehensive data within Southeast Asia. This is an Airbnb Singapore dataset, that was collected on 28 August 2019, according to the website Insiderairbnb and is also available on Kaggle.. It has 7907 rows & 16 columns, and some missing data on some feature/variable. The data is sourced from publicly available information from the Airbnb site.

Motivation

Singapore is reported to be one of “Airbnb’s most penetrated markets globally”, with 1.5 million Singaporeans using Airbnb listings overseas. There are currently around 8,100 listings in Singapore alone. Since the start of Airbnb in 2008, the company has expanded possibilities to travel stay. By exploring this dataset, we can analyze distribution of Airbnb listing in Singapore.

Goal

Hypothesis – I think that the price rage of the Airbnb’s would be somewhere between 50 to 250 and would be using regression analysis to prove my hypothesis. Through exploratory data analysis, the goal is to use the dataset to analyze and answer some of these questions -

1. Regional analysis:
    - Which region has the most Airbnb’s in Singapore?
    - Where are the Airbnb’s located according to the region and neighborhood?
2. Price Analysis:
    - Which region is the priciest and which is the least pricy?
    - Estimate price for minimum nights?
    - Price range for each room type?
3. Which is the most booked room type?
4. Which room has the most and the least reviews? Which region has the most reviews?
5. Which room in which region has the most occupancy?

Proposed method

I will explore and visualize the dataset from Airbnb in Singapore using exploratory data analysis techniques. I will be finding out the distribution of Airbnb listing based on their location, including their price range, room type, listing name, and other related factors. The Objective here is to find different analysis by using this dataset and following these steps –

1. Data Cleaning

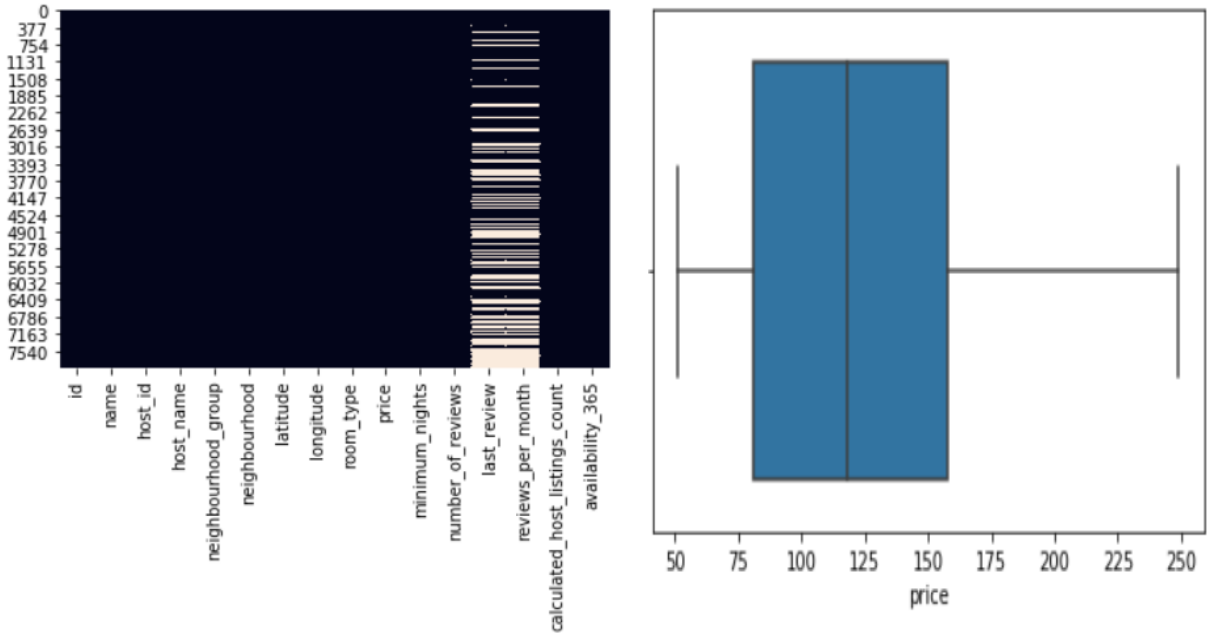
2. Regression Modelling

5. Visual Analysis

Experiments

1. Data Cleaning –

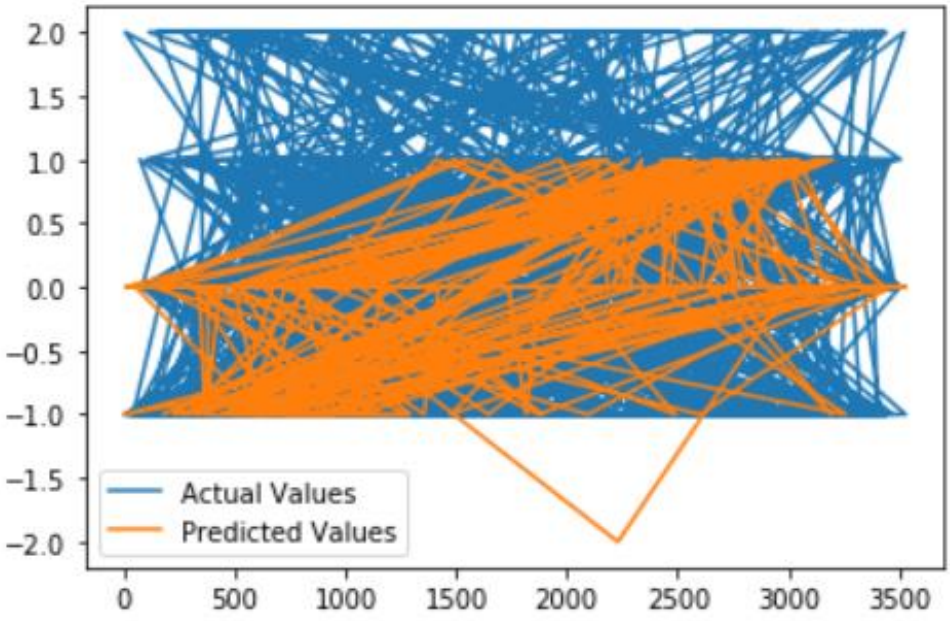
For this project I am using jupyter notebook with python programming language to write the script. After importing the libraries and loading the dataset: listings.csv, I used a heatmap to observe the missing values and then dropped them. After dropping the Nan values, the dataset was down from 7907 rows to 5148 values. Then I extracted important variables and identified outliers, any other missing values or human error.



Through experiments and error, I found out that there are many outliers in the price column that I was using for the regression analysis. After understanding the data by exploring the price column, I set the price to less than 350 for better regression analysis purposes. And lastly, saved the clean dataset into a new csv file: listings_clean.csv

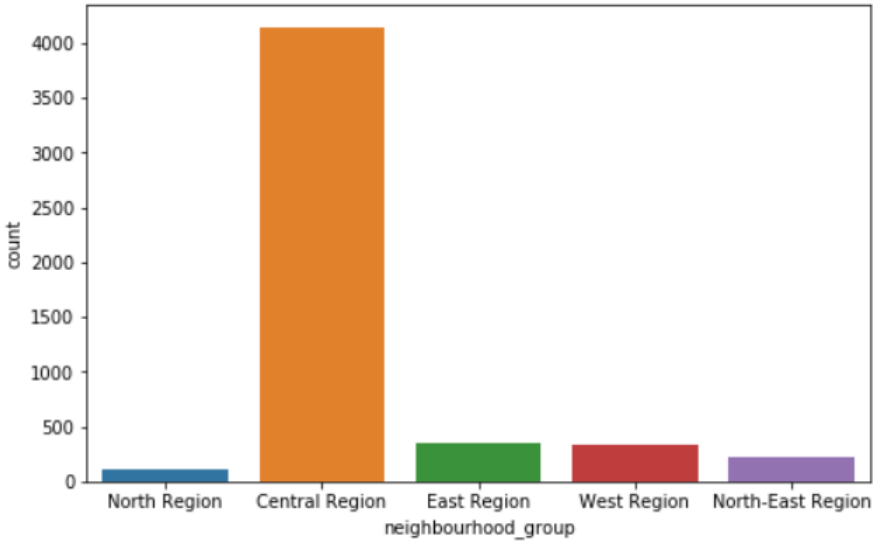
2. Regression Modelling –

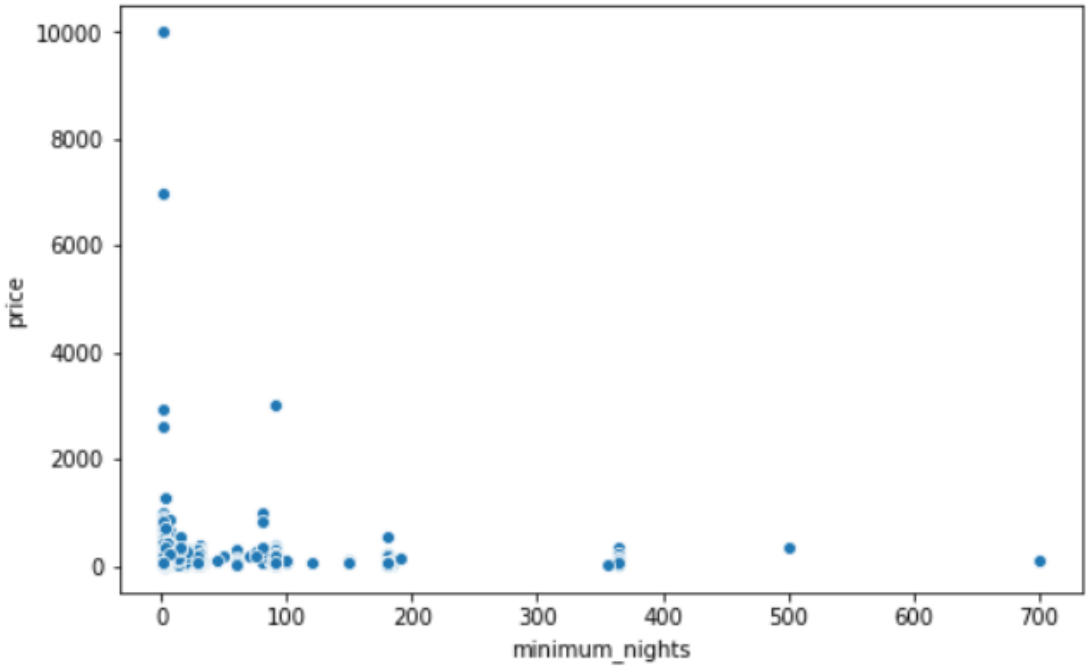
There are 16 columns in the dataset, but we only needed numerical values for our model. I dropped the columns with string values and since we set price to less than 350, the dataset was now 3534 rows and 8 columns. Then I used standard scaler to scale the values, therefore mean is 0 and standard deviation is 1, and replaced the price and reviews column in the dataset with the scalar values. After which I split the data into 70:30, i.e. 70% is the training data and the rest is testing data. I ran the regression using sklearn and with a test-train split to predict the value of prices. Actual verses predicted values were plotted for comparison.



3. Visual Analysis –

For the visual analysis I used seaborn. Through the analysis I explore that Singapore has 5 region area's where the Airbnb's are listed. The Airbnb's have 4 room types i.e. Private room, Entire home/Apt, Shared room, and Hotel room. I tried to answer all the goal questions through my visual analysis i.e. biggest value, room type, most popular listing, price predictions and visualized the longitude and latitude to display the map and the listing on the map.

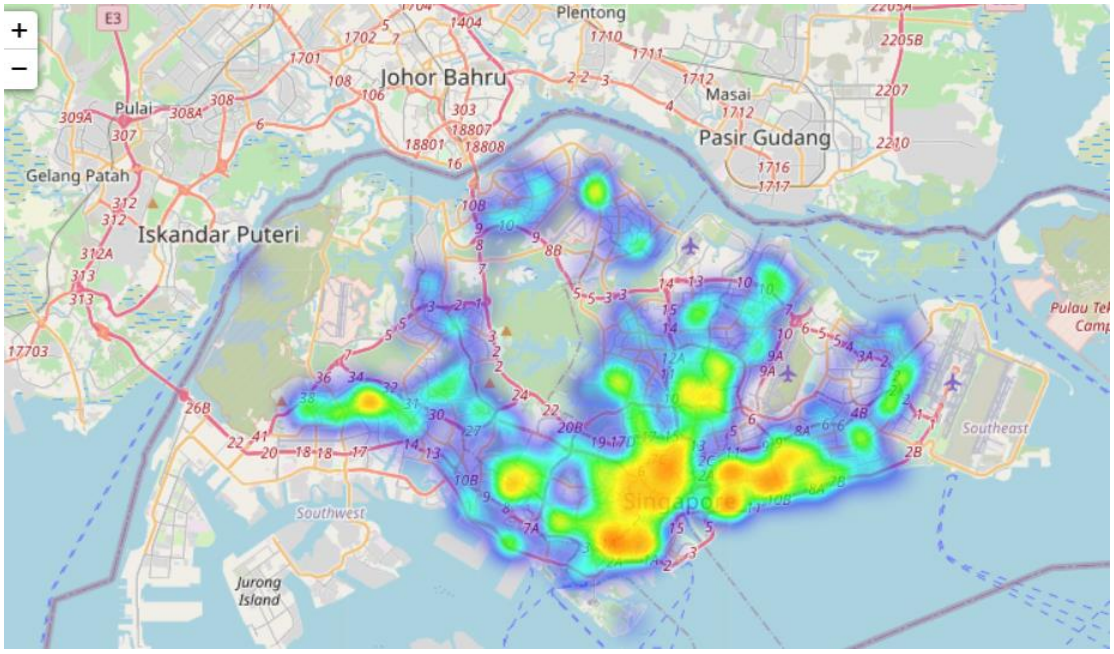




Results and discussion

By performing the data analysis on this dataset, I was able to identify how the Airbnb listings are distributed in Singapore. We now know that the region with the most Airbnb's is the Central Region and with the highest price range. The North-East Region is the least pricey out of the five. There is an estimate price of 1500 or less for minimum nights. Also, the price range for each room type was: Entire room/apt: 200 or above, shared room: 50 or below, private room: 90 or above and the most booked room type is the Private room with the most reviews as well. The most occupancy was in the Central region with the most reviews as well.

For the regression analysis, the mean squared error was '0.9211585253697794'. By plotting the actual and predicted values we can see that they are all over the place. Through the regression and visual analysis, I proved my hypothesis right, which is that the price range was between 50 – 250.



Limitations and later work

I did not investigate my data well enough before fitting it into the model. Initially I was not getting the right prediction and it gave me 1% accuracy and the mean square error was 22,000. With my professor Mr. Murat Guner's guidance, I understood that the distributions of target variable had some outliers. Due to while I was not able to get the right predictions. I needed to do more rigorous cleaning for the regression analysis, which will be my later work now. Once I do the hard-core cleaning, then I will have a better score and better values.

Conclusion and summary

This was still not the most appropriate dataset for regression analysis. It needed a good amount of cleaning to remove the outliers before I work it on my model. But I was able to prove my hypothesis. The most popular region is the Central Region, with prices ranging between 50 – 250 and the highest price range. The visual analysis helped with new insights on this dataset.

For future works, I will do more in depth data cleaning and focus more on the central region and how it dominates over the others.

Documents Submitted

Notebooks –

1. Data Cleaning: airbnb_data_cleaning.ipynb

2. Regression Modelling: airbnb_modelling.ipynb
3. Visual Analysis: airbnb_analysis.ipynb

Dataset –

1. listings.csv
2. listings_clean.csv

Report –

data_602_assignment_1.pdf

References and contributions

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

https://stackoverflow.com/questions/11285613/selecting-multiple-columns-in-a-pandas-dataframe

https://jakevdp.github.io/PythonDataScienceHandbook/04.01-simple-line-plots.html

https://seaborn.pydata.org/generated/seaborn.lineplot.html

http://insideairbnb.com/get-the-data.html