Applied Data Science Capstone Project

IBM Data Science Specialisation on Coursera

An Analysis of Rental Prices by Towns in Singapore

Contents

[1. Introduction 2](#_Toc18434475)

[1.1 Background 2](#_Toc18434476)

[1.2 Objective 2](#_Toc18434477)

[2. Data 2](#_Toc18434478)

[2.1 Data Collection 2](#_Toc18434479)

[2.2 Data Cleaning 2](#_Toc18434480)

[3. Methodology 3](#_Toc18434481)

[3.1 Calculation of Target Variable 3](#_Toc18434482)

[3.2 Exploring the Relationship between Surrounding Venues and Rental Prices 3](#_Toc18434483)

[4. Results 3](#_Toc18434484)

[4.1 Rental Price Distribution in Singapore 3](#_Toc18434485)

[4.2 Linear Regression Analysis 5](#_Toc18434486)

[4.3 Principal Component Analysis 5](#_Toc18434487)

[6. Conclusion 6](#_Toc18434488)

# 1. Introduction

## 1.1 Background

Singapore is a small and vibrant country located just above the equator in Southeast Asia. It is a global hub for education, finance, healthcare, technology, trade and transport, and has been recognised as the world’s smartest and safest city. Hitting the top rankings in the world in various areas, it’s no wonder why the city-state is always attracting a huge influx of immigrants. Currently, 39% of Singapore’s population are foreign nationals, of which a large portion consist of immigrant workers and foreign expats.

One of the issues when foreigners first move to the country is finding the right place to live. For expats especially, many are just introduced to the most expensive condominium the agent has without knowing much about the other options available.

## 1.2 Objective

The purpose of this study is to provide an overview of the public housing available in Singapore. Public housing comes in the form of government-built flats, known as HDB locally. There are various sizes, ranging from 1-room to 5-room or executive suites. While foreigners are not allow to buy them, they are allowed to rent a flat from the locals. This study will help to provide some insight on where to find the most reasonably priced flats, and explore the types of amenities nearby and their contribution to rental prices.

# 2. Data

## 2.1 Data Collection

For this study, we will need the following data:

* Rental Prices (median rent by town and flat type) – can be downloaded from <https://data.gov.sg/dataset/median-rent-by-town-and-flat-type>
* Amenities in each town – obtained from FourSquare API
* Geojson file of Singapore map – obtained from another user’s project on Github (with some modifications to fit my data) <https://github.com/yinshanyang/singapore/tree/master/maps>

## 2.2 Data Cleaning

Firstly, for the data on rental prices, after loading the csv file into Jupyter Notebook, it needs to be restructured into the desired format. The original file included rental price by quarter all the way back from 2005, but for the sake of simplicity, only the latest price in 2019 Q1 will be used, so all the earlier rows with historical prices are dropped. Next, based on local knowledge, 1-room, 2-room and executive suites are flat types that are unavailable for rent to foreigners. So data on these flat types should be omitted as well. Last but not least, the data frame is pivoted such that each flat type becomes a column of their own, with each row representing a town.

In order to get the choropleth map of Singapore showing the different rental prices by town, I got a geojson file from another user’s past project on github. Some of the location names were renamed to match the town names in my dataset.

# 3. Methodology

## 3.1 Calculation of Target Variable

For the sake of simpler analysis, the median price of 3-room, 4-room and 5-room flats are averaged for comparison. This is fine for the analysis as locations where 3-room flats are cheaper will have cheaper 5-room flats as well. We are only interested in the general comparison of prices in the different regions.

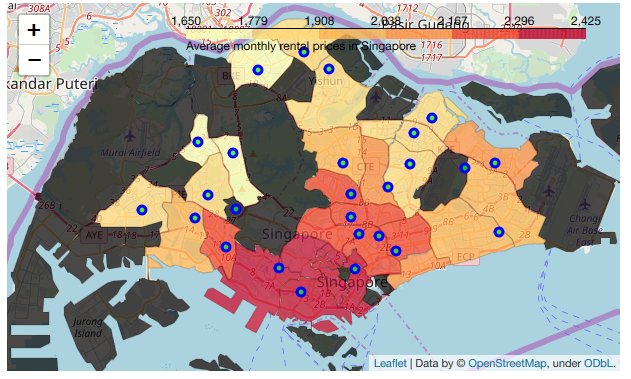
## 3.2 Exploring the Relationship between Surrounding Venues and Rental Prices

With FourSquare API, the surrounding venues in each town are obtained and the count of each type of venue in each town is obtained via one hot encoding. To explore if there is any relationship between the type of venues and rental prices, a linear regression model is used. The venue types with the most positive effect, the most negative effect, as well as the least effect are found using the linear regression coefficient via argsort from numpy library.

If the results from the linear regression model is not conclusive, further analysis can be done with principal component analysis. PCA will help to reduce the dimensions of the feature space and hence end up with less relationships between variables to consider. As such, overfitting is less likely to occur.

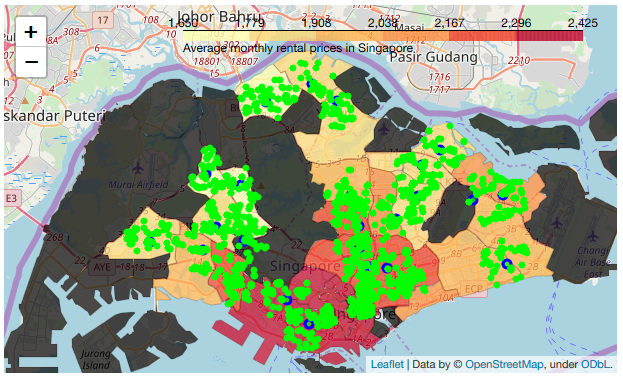
# 4. Results

## 4.1 Rental Price Distribution in Singapore



After cleaning the data we are able to plot a choropleth map of Singapore showing the different average rental prices in different parts of Singapore. We can see that the average rental price is the highest in the central downtown region, and slowly decreases with distance away from there. The prices are the lowest in the north and west regions of Singapore. The black areas are places without housing information. They are either non-residential areas, or areas where houses are mostly not for rental, for example, the Bukit Timah area consists mostly of landed properties that do not fall under public housing category and are also unlikely to be rented out.

With FourSquare API, we get a list of venues in each town. By plotting the venues on the map we can attempt to see if the visualisation can provide any insights on the correlation of these venues with the town rental price.



However, no obvious patterns can be observed from the map above. There are no significant difference in the number of amenities in each town, but perhaps there are some difference in the type of venues. We can study the effect of the different types of venues with further analysis.

## 4.2 Linear Regression Analysis

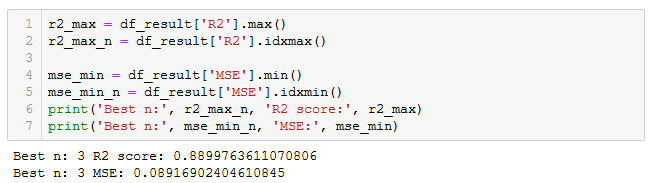
Before fitting the data through a linear regression model, the data was split into train and test set with 20% of the data for testing. The R2 score and mean squared error were calculated to determine how well the model fits.



As seen in the results above, the model has a low R2  score of 0.439 and high mean squared error of 0.293. This suggests that the linear model is not suitable for the dataset. Looking at the results obtained for venue types with most positive impact, most negative impact and least effect, the venue types obtained did not quite make sense as well. It is unlikely for a Pakistani restaurant to have a major positive impact on rental price, and it is also unlikely that a post office will negatively impact the rental price as well.

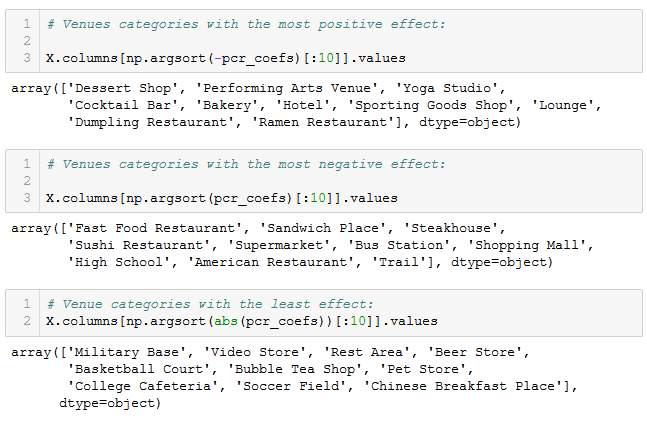
## 4.3 Principal Component Analysis

With the principal component analysis, we first find out the best number of components before redoing the analysis with the linear model again.



It was found that the best R2 score and lowest mean squared error was found when number of components is 3. The results showed a relatively strong correlation as compared to before.

Doing the same analysis earlier, we find the venue categories with the most positive impact, most negative impact, and the least impact as follow:



Despite the better R2 score, the results obtained still do not tally with our daily common sense. However, there are still some insights to be gained. While it does not make sense that the venue categories with the most negative impact are useful amenities that should in no way lower real estate prices, we can observe that these venues are more likely to be found in the heartlands. On the other hand, the venue categories with the most positive impact are places more likely to be found in town. As such, we can conclude that the types of surrounding venues around a residential area are unlikely to be a factor affecting its rental price. Or rather, the rental price is more likely to be due to the location itself – i.e. prices are higher in centre downtown areas whether or not they have useful amenities.

Of course in reality there are a multitude of factors that could affect rental prices. Perhaps the prices in central downtown area are the highest simply because there are less flats available, but that is a whole other topic to analyse. We can see that the hottest and most expensive places to live may not actually have the most to offer, but are expensive simply because of the location. This is an import insight this study can provide for foreigners, that there are plenty of housing options in Singapore and while they might be further away from city central, these areas actually have more to offer with lower rental costs.

# 6. Conclusion

In this study, I attempted to analyse the relationship between rental prices of public housing in different parts of Singapore and the amenities that surround these flats. While a relatively correlation score was obtained with principal component analysis, the results did not match our common sense and we would conclude that there is no relationship between the variables. Or rather, both are results of the location they are in – rental in central region are expensive just because it’s central, and the amenities found there are entertainment related just because it’s central. While there is a pattern, it is not a causal relationship. Foreigners new to Singapore could use the map generated to choose a better location for their stay, perhaps somewhere in the orange zones if their financials allows so that they can enjoy the best of both worlds.

The analysis of relationship between surrounding amenities and rental price of a real estate could probably be more insightful if done on the micro scale – i.e., comparing two actual similar houses in either the same town or a similar town with their actual surrounding venues.