**Analysis of Private Residential Property in Singapore**

**1. Introduction**

**1.1 Background**

With a population count of 5,894,228 and a population density of 8358 people per square kilometer as at 2020 (According to Worldometer), Singapore is one of the densest countries in the world in terms of population. Singapore is a global financial hub known for its stability, security and infrastructure. Singapore is also a city that attracts large investments from abroad due to its ease of doing business, strategic location, low corporate tax and investment incentives. In 2020, Singapore attracted $17.2 billion in fixed assets investments despite the Covid-19 pandemic.

If you are looking to move to Singapore be it for business or leisure, the suitable option may be to go for private residential properties. Let’s call them PRPs. You may be a Singaporean looking to go private for the initial low minimum cash payment, long-term investment, the less restrictions, or a single individual looking to own a property before the age of 35 (That is if you can afford to, really?). Private Condominiums/apartments come with facilities of a condominium (Gyms, security, swimming pools). There is no Minimum Occupancy Period (MOP) although you have to pay a Seller’s Stamp Duty (SSD) if you sell within 3 years.

Executive Condominiums (ECs) on the other hand are a hybrid between private apartments (Private condominiums) and public housing (HDB flats) in Singapore. ECs are more suited for housing instead of investments/renting out for income because of the Minimum Occupancy Period (MOP) requirement of 5 years. Housing grants are also available for first time EC buyers. However, they are only available to Singapore Citizens and permanent residents.

This project will be focused on private apartments/condominiums (excluding landed property) although future projects could dive into ECs. The following report will include an overview of the data first with visualization, regression and correlation analyses first followed by the use of the Foursquare API to get the most common values in each ‘street’ in Singapore.

References:

<https://www.straitstimes.com/business/economy/singapore-draws-172-billion-of-investments-in-2020-most-since-2008-despite-covid-19>

<https://www.globenewswire.com/en/news-release/2021/05/19/2232506/28124/en/Singapore-Data-Center-Market-Report-2021-Investment-Analysis-and-Growth-Opportunities-Market-will-Witness-Investments-of-5-Billion-by-2026.html>

<https://www.99.co/singapore/insider/7-reasons-a-private-property-is-more-affordable-than-you-think/>

<https://blog.moneysmart.sg/property/private-properties-hdbs-differences/>

<https://blog.seedly.sg/deciding-between-private-condominium-vs-executive-condominium/>

**1.2 Problem**

A home buyer may have several considerations and preferences unique to each individual, including amenities, facilities and budget. This project aims to find out which area will be the best choice for your needs and budget. I hope that it will help you weigh the characteristics (e.g. public amenities, nearby restaurants) of certain postal districts with the prices.

**1.3 Reasons of Interest**

Homebuyers, real estate agents would be interested in the analysis of data categorized into different streets, each tagged to one private condominium nearby. This could factor into investment strategy or just a quality of living better catered to one’s preferences. Tourist may also use the findings to choose their Airbnb location.

It would be advantageous for new homebuyers, be it if you’re a Singapore Citizen or a permanent resident, to know the mean prices of each postal district to narrow down your choices and make better buying decisions. It may also be advantageous for real estate agents to help their clients make better choices.

**2. Data acquisition & cleaning**

**2.1 Data Sources**

I used datasets of private apartments/condominiums from the Urban Redevelopment Authority website: <https://www.ura.gov.sg/realEstateIIWeb/transaction/search.action>

The datasets were dated from May 2016 to May 2021 and were downloaded 5 postal district numbers at once. There are 28 postal district numbers in total.

I later converted the 6 CSV files downloaded into DataFrames and combined them into one. I then obtained the mean of the prices in each postal district. As the CSV did not contain the coordinates for each street name, I extracted the street names from the dataset of private properties and used the One map API to obtain the longitude and latitude coordinates based on the StreetName variable for use in visualization. The coordinates were converted into a DataFrame and combined using a left outer join to fit all the coordinates into the original df\_merged DataFrame. The mean price of PRPs for each street was then calculated and complied into a new dataframe mapdata and eventually visualized onto a map.

**2.2 Feature Selection**

After data cleaning there were 96,675 samples (rows) and 16 features in the data. After examining the meaning of each feature, I found some redundancy in the features. For examples, the features unit price ($psf) and net price are very closely related features which includes other factors such as area (sqft) and excluding/including real estate fees. This means they are very closely related. The most comprehensive feature ‘Price’ will hence be taken as the main metric for this analysis. No features were dropped so further analysis in the future can be carried out on this set of data.

Besides this, the other features were largely categorical and kept as “master data” to be analyzed against price in an ongoing work-in-progress for this research. The features are ‘ProjectName’, ‘StreetName’, ‘Type’, ‘PostalDistrict’, ‘MarketSegment’, ‘Tenure’, ‘TypeOFSale’, ‘NoUnits’, ‘Price’, ‘Sqft’, ‘TypeOfArea’, ‘FloorLevel’, ‘UnitPrice’, ‘DateOfSale’, ‘latitude’ and ‘longitude’.

**2.3 Data Cleaning**

Upon starting to work on the CSV file raw format imported into Pandas, the columns were not recognized as I reset the index and renamed the columns headers for them to be recognized by Pandas. The DataFrame included duplicated string values spanning several rows in the price column labelled “NotFound”, which were dropped.

The latitude and longitude values were object types which were required to be converted to float types and where the latitude columns with missing values were dropped to mitigate a format type error when visualizing the map using folium.

I also realized that the ‘Price’ column was the total prices for the row which may contain more than one unit (up to 436 units for en bloc/collective sale). This may cause the calculated mean of the prices to be inaccurate. To mitigate this inaccuracy, I divided the ‘Price’ column by the ‘NoUnits’ column and assigned it back to the ‘Price’ column.

Those labelled (-) includes 3/4 storey mansionettes and townhouses (such as Kew Green, Bagnall Court, Gilstead Court). This is likely the reason for the high mean price compared to the rest of the floor levels but the data also includes regular condominiums where the floor level is not listed, which may deem it inaccurate. Therefore, we will be excluding this from the trend analysis in the number of floors.

In order to visualize the bar chart when comparing tenure period, I separated string values including categorized ‘Freehold’ data rows to convert them to floats which were sorted by ascending and set to the index. The ‘Freehold’ data was added back to compare with the leasehold data afterwards.

I split the columns and converted the year column into float format and removed their NaN values. This allowed for visualization and regression analyses to be conducted.

**3. Exploratory Data Analysis**

**3.1 Calculation of target variable**

Mean prices based on each ‘StreetName’ was not a feature in the dataset and had to be calculated. ‘StreetName’ was chosen because it is compatible with the OneMap API and also because it is the most specific identifier of a location other than postal district number which has only 28 unique classifications compared to 839 unique street names. As seen below, we can click on the individual points to reveal the location, street name and the mean price of the property on that specific street.

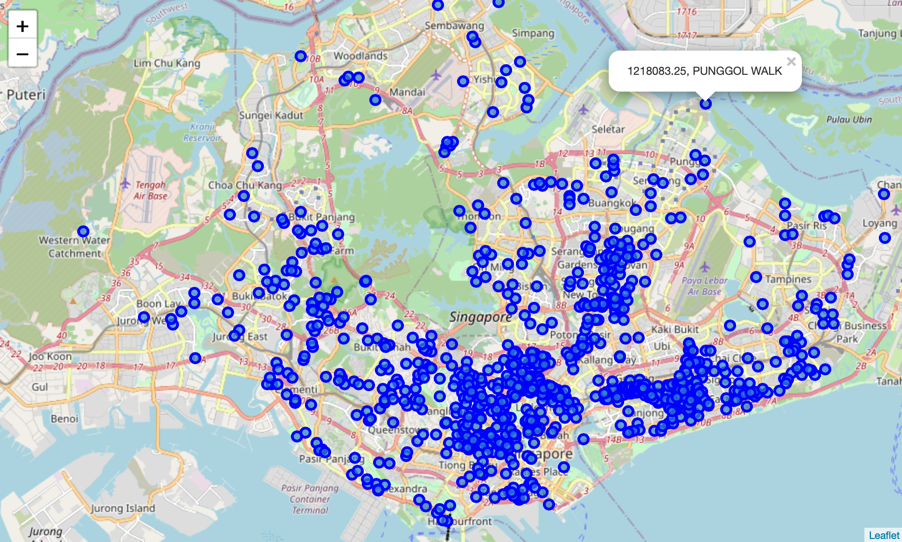


Figure 1(Visualized map of Singapore using OneMap API)

**3.2 Relationship between Floor Level and Price**



Figure 2 (DataFrame named groupfloor)

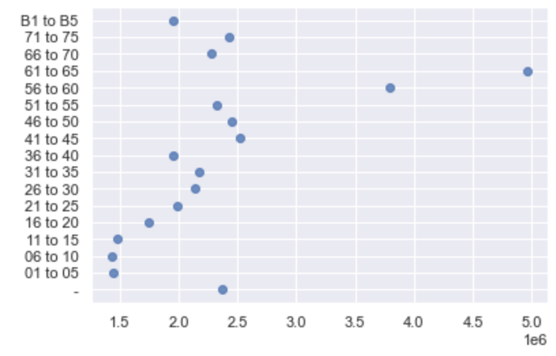


Figure 3 (Visualized using seaborn, Scatterplot of ‘Price’ (Horizontal axis) and ‘Floor Level’(Vertical axis))

As seen from figures 2 and 3, there is a positive relationship between floor level and price especially starting from floor level category ’16 to 20’ and above. This relationship is particularly strong as seen in the significant increases in the floor level ‘56 to 60’ and ‘61 to 65’ ranges.

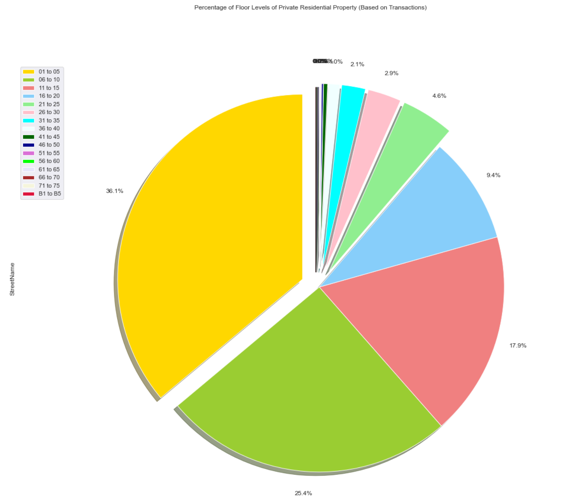


Figure 4 (Percentage of Floor Levels of Private Residential Property(Based on Transactions))

In figure 4 above, we can see the distribution of floor levels of private residential properties in Singapore based on transaction information from our dataset. A large majority of the floors (up to 88%) are 20 levels and below. Comparing to how mean prices begin to rise with the levels ‘16 to 20’ in the analysis above, it could be hypothesized that the rise may be due to a shortage or scarcity of private residential property units for sale of floor levels above 20.

**3.3 Relationship between Tenure Type and Price**

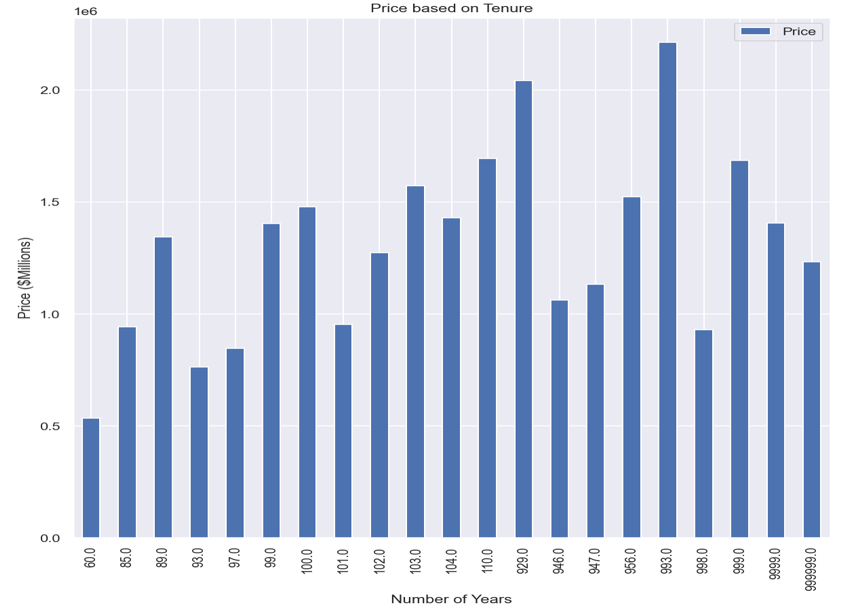


Figure 5

As seen in figure 5 above, the number of years in the x-axis is in a very slight ascending order. If observed closely, the mean prices for the years increases subtly towards 100 to 900 years, but slightly diminishes above 993 years. It does not seem that the tenure is closely correlated with price.

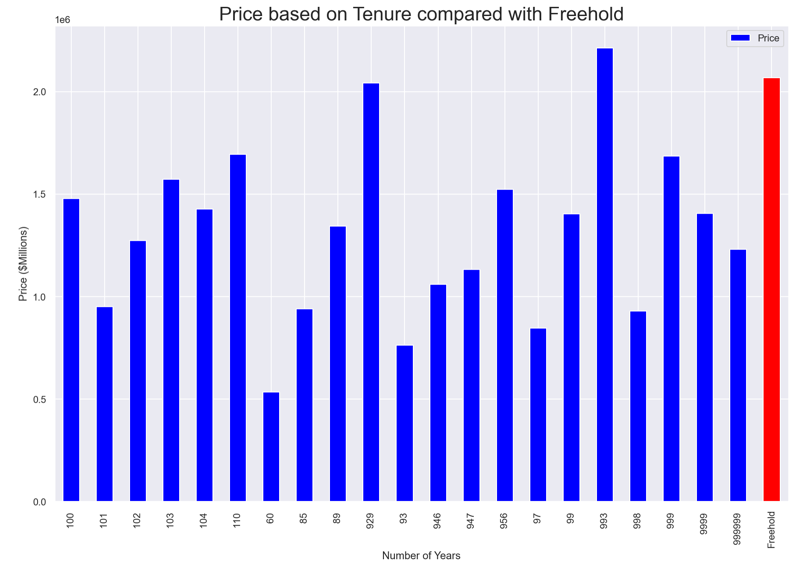


Figure 6 (Prices compared to Freehold Residential Properties)

Even when compared with freehold prices in figure 6, it can be seen that freehold residential properties are priced relatively higher than average when compared to Thus, we can conclude that there are many other factors at play here determining the price of the property besides leasehold vs freehold. The calculated Pearson Correlation Coefficient is -0.040608 which further proves a very weak negative relationship between the tenure years and price.

**3.3 Relationship between Type of Sale and Price**

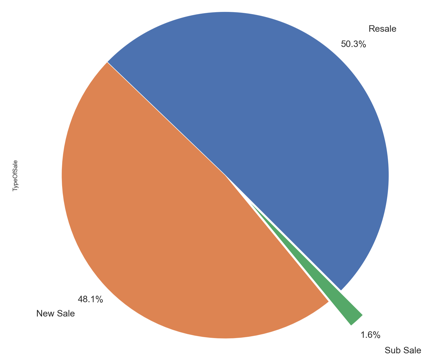


Figure 7 (Proportion of PRPs in dataset based on Type of Sale)

As we can see from figure 7, a large proportion of PRPs are either a new sale or resale. Only 1.6% of PRP transactions are sub-sales.

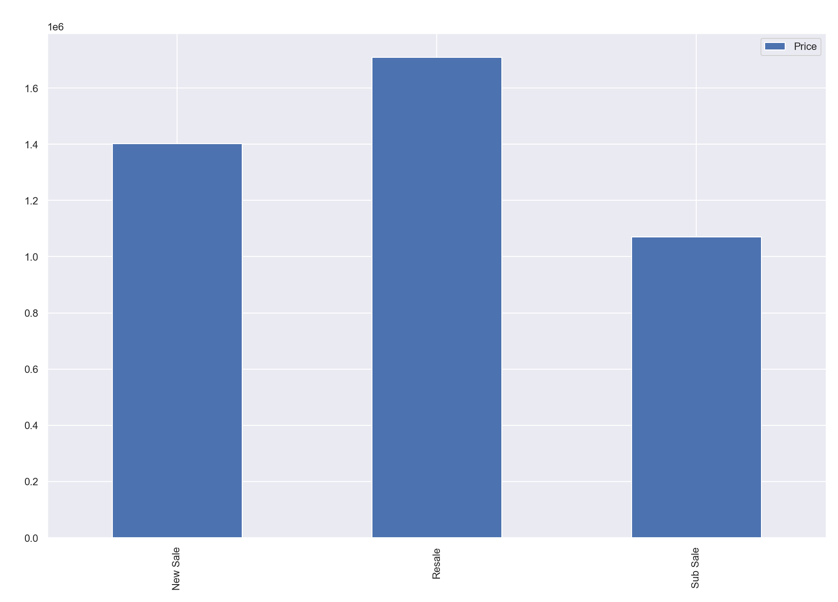


Figure 8 (Price variations between different sale types)

The mean price of resale is highest at $1,709,901 followed by new sale at $1,401,675 and then followed by sub-sale the lowest priced at a mean of $1,069,355. Further research can go into why sub-sale properties are priced considerably lower than resale and new sale properties.

**3.4 Relationship between Date and Price**

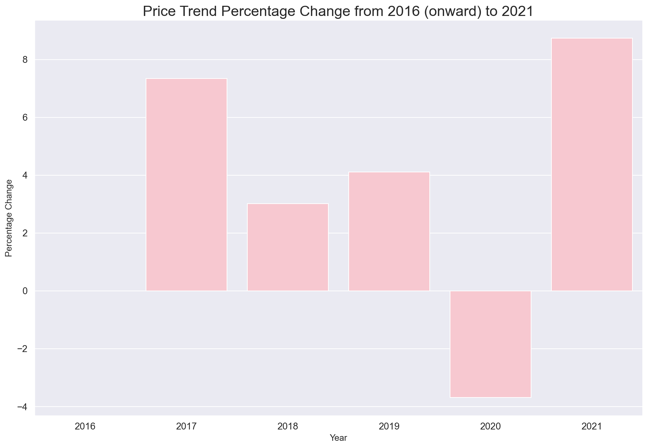


Figure 9 (Price Trend based on percentage change)

The prices of PRPs have seen a trend of slight increase overall, with the exception of 2020. This could be due to significant world events like the Covid-19 Pandemic, which correlation has yet to be proven. A recovery of prices can be seen in 2021, which may indicate a tendency for prices to “regress to the mean”. The standard deviation for the percentage of price change, or price fluctuations is at 4.613%.

**4. Predictive Modeling**

Regression models were developed to help predict prices of a PRP where continuous values are available in the dataset with predictor variables such as floor levels, PRP size, and years of tenure.

**4.1 Regression Models**



Figure 10

Linear regression was applied to further analyze the relationship between floor level and price. The linear model is: Price = 1304377.6 + 23811.402 \* Floor Level

We can see from the regression plot in figure 4.1, that price is slightly positively correlated with floor level. This tells clearer story from our analysis earlier, where the results were affected by outlier unit(s).

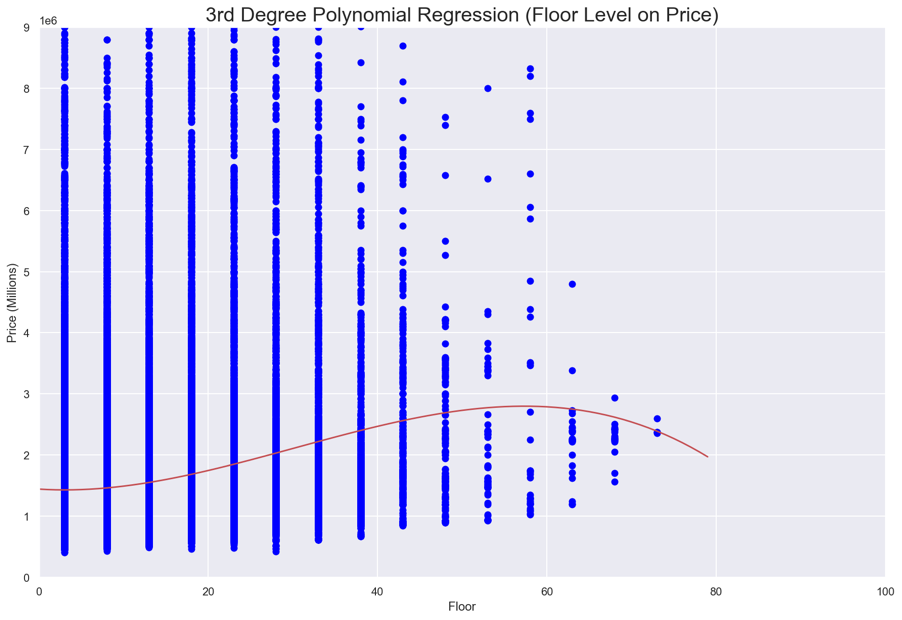


Figure 11

Upon closer analysis of the relationship between floor level and price, the Pearson Correlation Coefficient is 0.1619368 which was lower than the recommended 0.7 that confirms a likely linear relationship. Therefore, using non-linear regression is appropriate. A 3rd- degree polynomial regression trend line showed that the price increased when it approached level 60 but diminished after level 60. However, the R2 score is 0.03 = 3% which shows that floor level is not a good predictor of price. To add on, the mean absolute error is also large at 634442.31.

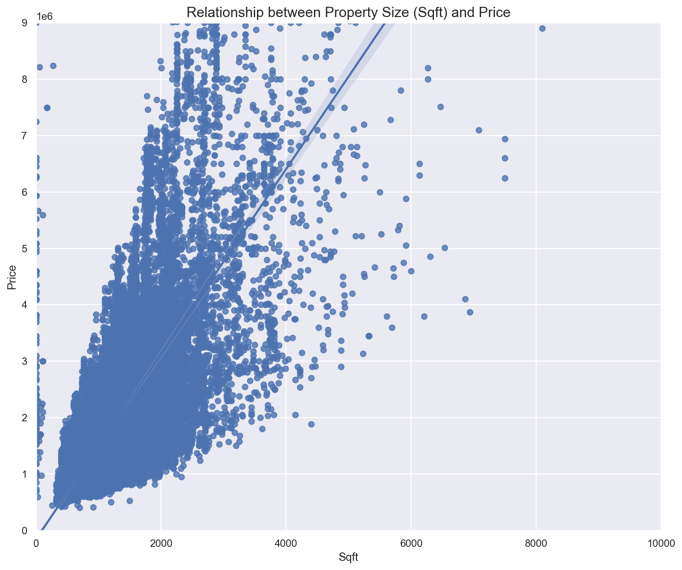


Figure 12

Linear regression was also applied to the relationship between price and property size (Sqft). As seen in figure 11, there is a clear and strong positive correlation between price and property size (Sqft). The linear model is: Price = 484743.2 + 1421.744 \* Sqft

Although this is likely a linear relationship between price and property size with a Pearson Correlation Coefficient of 0.748499, using a polynomial regression function of 3rd order would be useful in churning out more accurate predictions and getting a clearer picture.

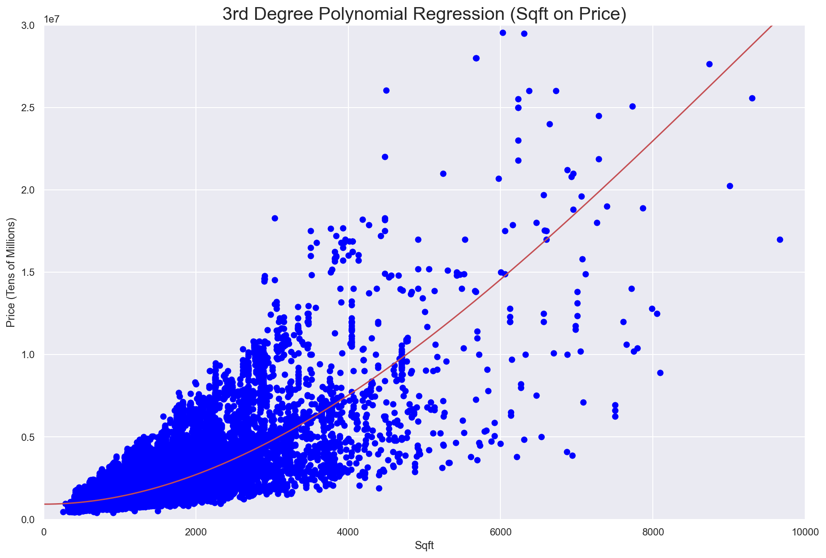


Figure 13

From figure 13 we can see that there is an exponential increase in price especially below 6000 Sqft of property size, continuing with a linear increase above 6000 Sqft. The R2 score is 0.59 (59%). Hence we can say that the size of the PRP (Sqf) is a rather good predictor of price.

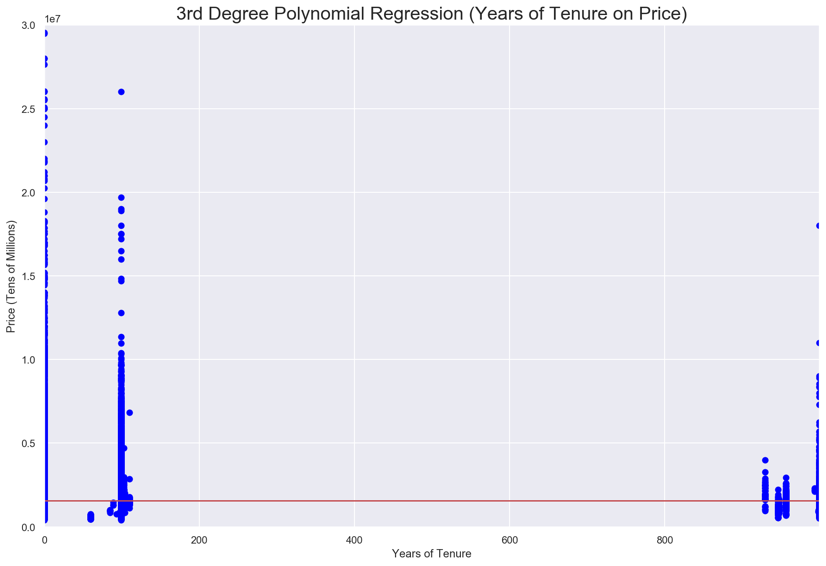


Figure 14

Figure 12 shows that the period of tenure is barely correlated to price with a nearly horizontal line. This is consistent with previous analysis of the effect of tenure years on price. Lastly multiple linear regression was applied using floor, Sqft and tenure period (tenureyears) as predictor variables to create a model in predicting the price of the PRP.

The linear model:

Price = 1285002 + 23930.896 \* Floor + 15.740 \* Sqft – 0.306 \* tenureyears

However, the R2 score is 0% and Pearson Correlation Coefficient is -0.003. This means there is close to no relationship between years of tenure and price. Therefore, this linear model should not be used as years of tenure is not a predictor of price.

**4.1 Applying standard algorithms for model evaluation**

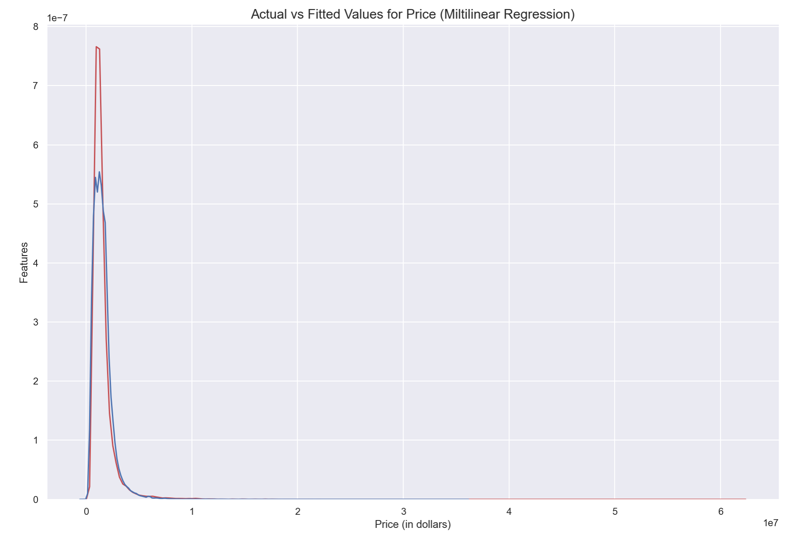


Figure 15

I used mean squared error (MSE) with the 3 continuous variables available in the dataset.

The R-squared (coefficient of determination) for the multiple linear regression is 0.5717667. We can say that about 57.177% of the variation in price can be explained by this multiple linear regression. The mean squared error (MSE) of the price and predicted value using multifit is 648023500000.0 with a Root Mean Squared Error (RMSE) of 804999.07.

Looking at the distribution plot at figure 12, the fitted values are quite close to the actual values although there is still room for improvement. However, the insignificance of the correlation of independent variables such as floor level and years of tenure to price from previous model evaluations calls for consideration that these 2 variables be omitted in future models.

**4.2 Utilizing the Foursquare API and K-Means Clustering**

**4.2.1 Methodology**

For the Foursquare API, I designed with a limit of 100 venues and radius of 500 meters in exploring the streets to get the top 100 venues in each street. After running the data through the function getNearbyVenues, a total of 348 unique categories were retrieved.

**4.2.2 Distribution of Venues**

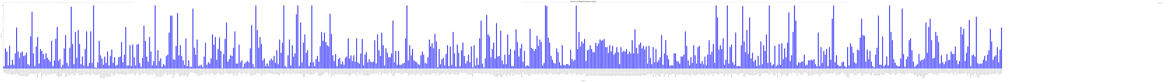


Figure 16 (Shows the variation in the number of categories of the 829 Streets)

In figure 16 above you visually observe that the category count varies widely between different streets of the listed PRPs. This is proven with the standard deviation of 22.59.

The mean number of categories is 26.33. I found the venues with above 80 venues and calculated their mean price which came out to be $1,836,902.33, which is higher than the overall mean price of $1,553,010. This hints at the number of categories being positively correlated with price.

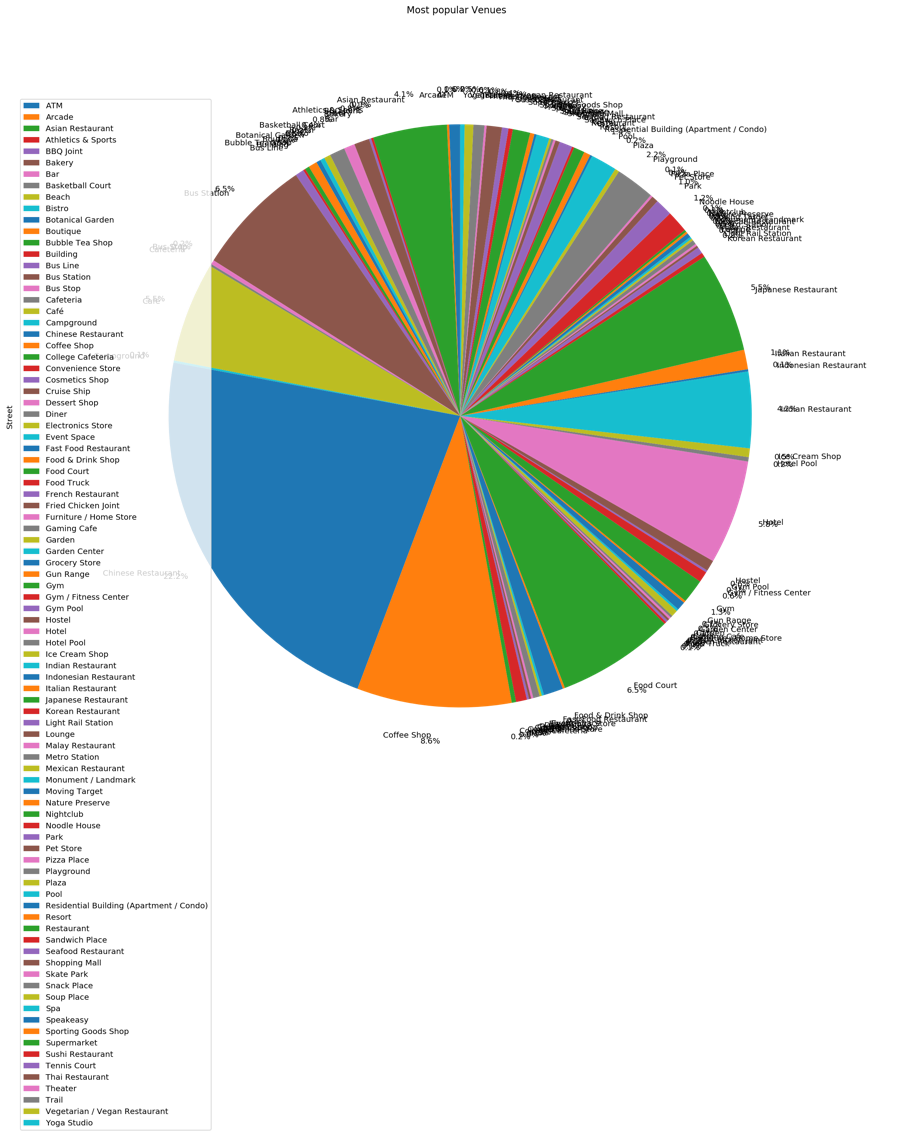


Figure 17 (Proportion of Most Popular Venues)

Upon analysis of the 1st Most Popular/Common/Recommended Venues in each of the streets, it can be seen that F&B establishments such as Chinese, Japanese, Indian and Asian restaurants, Coffee shops, Food Courts and Cafes are observed to take up a large portion of recommended venues by Foursquare.

**4.2.3 K-Means Clustering**

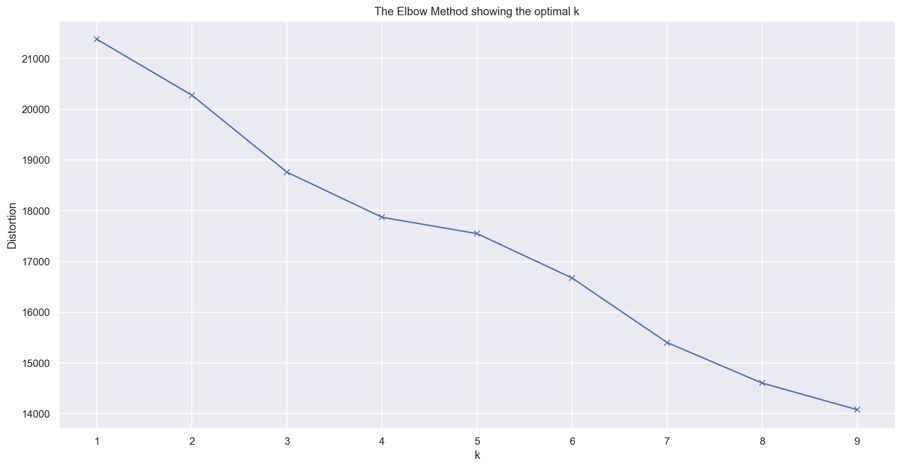


Figure 18 (Elbow point)

Then I used k-means clustering to identify the number of clusters assigned to identify patterns in the category types and separate them. As seen in Figure 18, the elbow method showed that 4 clusters will be optimal.

I then used folium to visualize a map with circle markers, when clicked on reveals information on the street name, cluster, top 5 venues and price. The clusters are separated with colors as seen in figure 19.

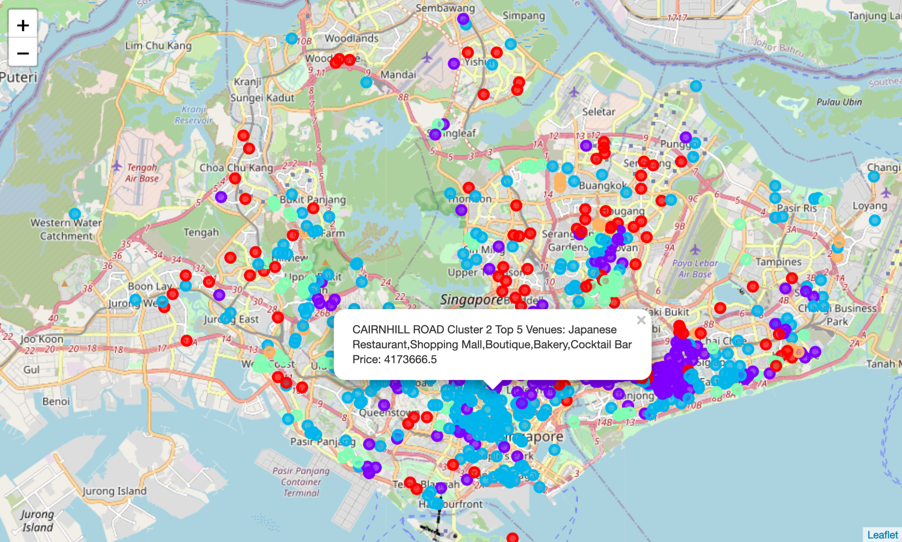


Figure 19 (Clusters visualized using folium)

Let’s take a closer look into clusters.



Figure 20 (Distribution of categories of recommended venues)

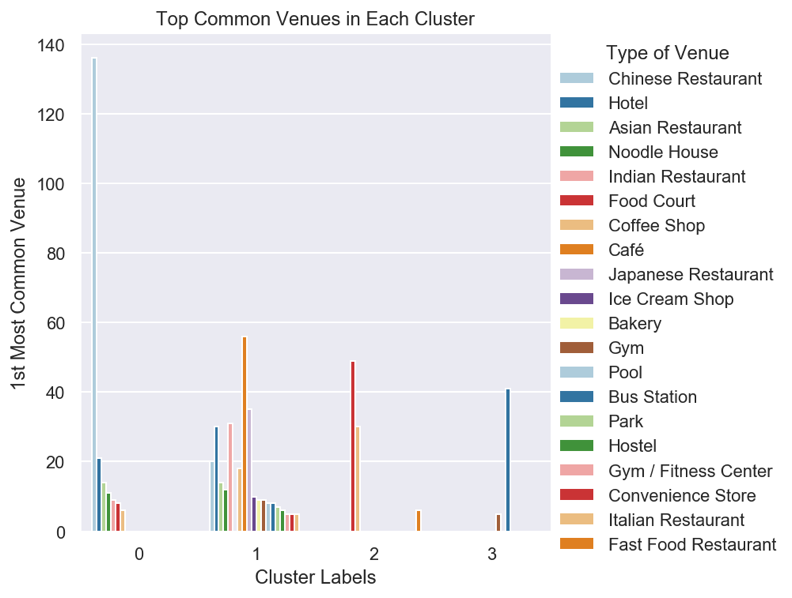


Figure 21 (Visualization of Distribution of categories of recommended venues)

We can see from figure 21 that there are very distinct venue types to each cluster. After examining each cluster we can label each cluster as such below:

§ Cluster 0: “Asian Cuisine F&B/Restaurant Establishments”

§ Cluster 1: “Wide Range of Amenities”

§ Cluster 2: “Budget Food Establishments”

§ Cluster 3: “Transport Accessibility and Fitness”

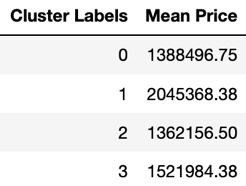


Figure 22 (Mean Prices for Each Cluster)

Looking at the mean prices of properties in these clusters, we can see that the mean price of cluster 1 is significantly higher than the other clusters at $2,045,368. This is followed by cluster 3 at 1,521,984. This result showed that properties with a wide range of amenities are valued higher.

**5. Results and Discussion**

We started this analysis exploring the dataset, visualizing the street name with property price on the map of Singapore using folium and One map API, looking into the relationships between floor levels, tenure type, type of sale and date with the price. We also looked at the proportions/percentages of these categorical features such as floor levels and tenure type to get a better picture of the dataset. One of these categorical variables, floor levels, was transformed into continuous ones for regression analysis and the development of regression models. Property Area (Sqft) already came in a continuous format.

Initial visualizations revealed the floor level to have a positive relationship especially between floor levels 56 to 65. However, regression and coefficient of determination (R squared score) analysis later showed floor level to be a rather insignificant predictor of price.

Similarly, regression analysis also revealed that length of tenure had no relationship with price. Do note that a large proportion of the years of tenure were listed as either 99- or 999-years tenure/lease periods, which may cause inaccuracies in analysis.

Despite this, a strong positive relationship was found between property size/area and price.

Although accessibility to food, amenities and transport in Singapore is generally very high due to its small and dense area, there may be differences in different areas to be explored.

Foursquare API was used with K-Means clustering to identify 4 clusters through one hot encoding and the elbow method. Distinct clusters were identified with cluster 1 properties – Wide Range of Amenities being the most highly priced of all clusters. This strongly implies a strong relationship between the variety of amenities available and price. Other clusters with their unique characteristics were also identified which may cater to a buyer’s preferences.

**6. Conclusion**

The purpose of this project was to identify aspects of a property that affect its price such as floor level, property size, and nearby recommended venues so buyers or investors can better weigh their decisions. There is so much more to consider when moving to a new home or taking up an investment in real estate. This was just an exercise to test out the Supervised (Regression) and Unsupervised (K-Means Clustering) Machine Learning methods and the data that Foursquare API provided.

There are many other aspects home buyers or investors can look at such as view, preinstalled apartment renovations/features, internal condominium amenities (carpark, pools), proximity to key business districts, etc. which should be considered.

**7. Future Directions**

Do note that findings from this analysis are largely deduced out of quantitative data. Certain qualitative aspects such as view from the property and overall design of the property were not considered in this analysis. Classification modelling such as decision trees and logistic regression models can also be applied to the data for better insights.