1.Introduction

Singapore has one of the highest home ownerships in the world, ranking second globally with 91% of its citizens owning a property they call home. About 85% of the population live in Public Housing by the Housing Development Board (HDB), giving these properties its colloquial term “HDB”. Despite this, a report by CBRE Group, the largest commercial real estate services company in the world ranks Singapore as the 2nd most expensive housing market in the world. Singapore not only has one of the most expensive houses, it is also home to the most expensive country to own a car, where a Volkswagen Golf can cost you a cool $75,000 USD and up. As such, few Singaporeans drive, placing emphasis on the convenience of public transport, which translates to the importance of a strategically located home.

1.1 Business Problem

The importance of a conveniently located home in Singapore is ever so important. Therefore, this project will seek to discover the factors that contribute to the price of a resale HDB property in Singapore and whether the additional facilities or business in the vicinity have an effect on the housing prices.

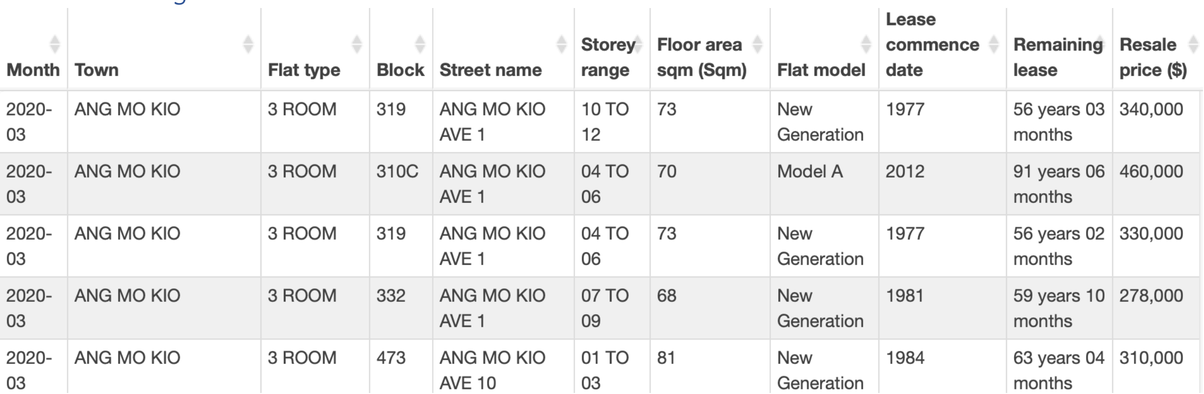
2. Data

2.1 Data Sources

- The resale properties’ information and prices are available on the Singapore Government’s Data website as CSV files. The link is https://data.gov.sg/dataset/resale-flat-prices
- OneMap Singapore, for Geocodes/coordinates. OneMap provides API calls https://docs.onemap.sg
- Statista for Singapore Inflation Rates https://www.statista.com/statistics/379423/inflation-rate-in-singapore/

2.2 Data Cleaning

Understanding the Data



Month	Town	Flat type	Block	Street name	Storey range	Floor area sqm (Sqm)	Flat model	Lease commence date	Remaining lease	Resale price ($)
2020-03	ANG MO KIO	3 ROOM	319	ANG MO KIO AVE 1	10 TO 12	73	New Generation	1977	56 years 03 months	340,000
2020-03	ANG MO KIO	3 ROOM	310C	ANG MO KIO AVE 1	04 TO 06	70	Model A	2012	91 years 06 months	460,000
2020-03	ANG MO KIO	3 ROOM	319	ANG MO KIO AVE 1	04 TO 06	73	New Generation	1977	56 years 02 months	330,000
2020-03	ANG MO KIO	3 ROOM	332	ANG MO KIO AVE 1	07 TO 09	68	New Generation	1981	59 years 10 months	278,000
2020-03	ANG MO KIO	3 ROOM	473	ANG MO KIO AVE 10	01 TO 03	81	New Generation	1984	63 years 04 months	310,000

Figure 1: Sample of Raw Data

Column Header	Remarks
Month	The Month and Year of transaction
Town	Singapore plans residential properties according to towns, (aka Boroughs in other countries)
Flat_Type	Public housing comes in various types – 2,3,4,5, Executive Flats
Block	Block Number where the transaction occured
Street_name	The street name of the block
Storey_Range	The range of storeys where the property sold is located
Floor_area_sqm	Area is measured in Square-Meter, according to the Metric system
Flat model	Over the years, several design iterations in flat designs have been created, classified under flat model. *Not to be confused with Flat-Type.*
Remaining_lease	The number of years left to the property. All public housing in Singapore have a 99-year tenure
Resale_price	Price of the resale

Data Wrangling

Changes to Raw Data

Raw Data Header	New Header	Remarks
Month	Year	Converted all resale dates to only account for year
Storey_range	Storey_range	Storey_range is a string i.e. '01 TO 05'. To measure how the the storey of the property affects its price, storey_range was converted to use the median and stored as an integer value.
Block, Street Name	Address	To facilitate the more precise distance measures, a new column "Address" was created using the "Block" and "Street name" columns. These addresses would be later used to map the exact distances between a facility to the specific address.
Resale_price	Resale_price	The resale price is adjusted for inflation based on the year of transaction, to the current year 2020

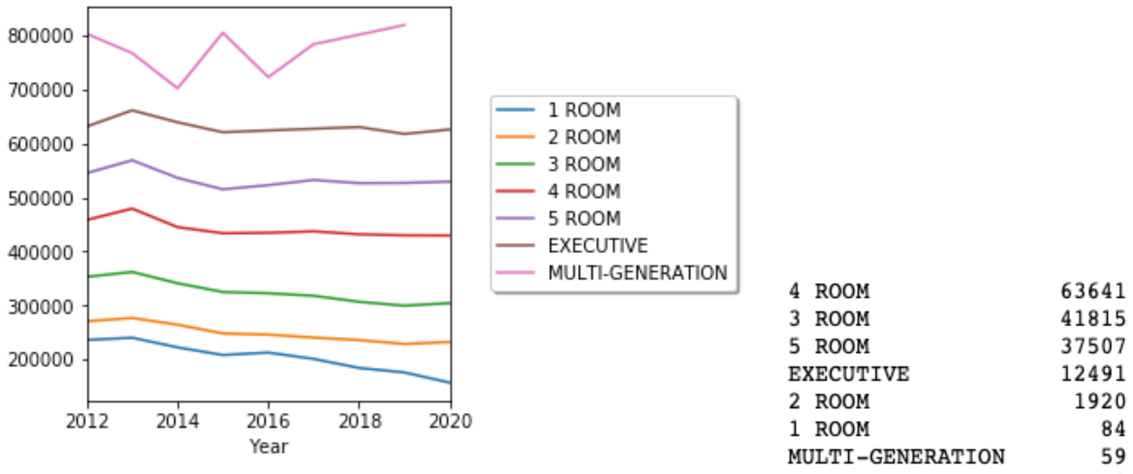
The merged dataset contains resale values from year 2012 – 2020 Q1. While the raw data provided by the Singapore government backdates further, for the purpose of relevance, the study was kept to within the last 8 years.

The entire dataset used in this investigation involved 3 CSV Files. Whilst they were largely similar, the oldest CSV file, containing data from Years 2012 – 2014 did not have the column “remaining_lease” but had a “lease_commence” date, which refers to the purchase year of the property. Therefore, as part of standardization, the remaining_lease can be drived from (99 years – lease_commence) as all public housing in Singapore have only a total lease of 99 years.

Further into the study, it was discovered that a particular street name “ST. GEORGE’S” could not be retrieved from the OneMap API given that the API only recognizes “SAINT GEORGE’S”. Therefore, these addresses were adjusted accordingly.

The dataset had no null values

2.3 Data Exploration



```
4 ROOM              63641
3 ROOM              41815
5 ROOM              37507
EXECUTIVE           12491
2 ROOM               1920
1 ROOM                 84
MULTI-GENERATION       59
```

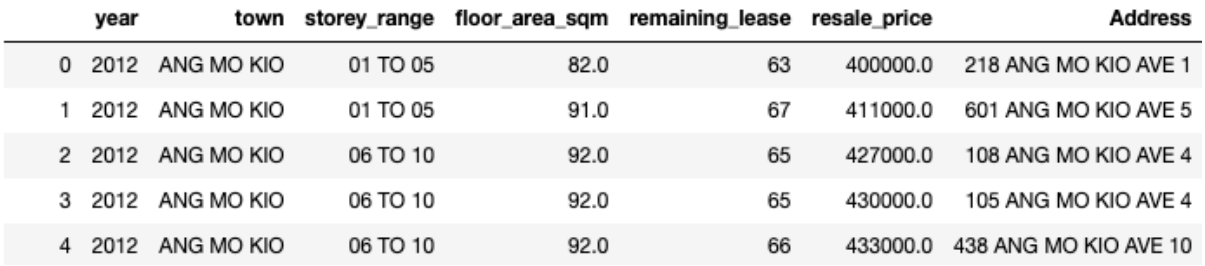
Figure 2: Resale Price trends by Flat Types

A quick visualization of the resale price trends over the years by house-type shows a rather similar trend across all types, except for Multi-Generation, which only has a sample size of 59 out of 157,000 samples in the dataset. For simpliclity and significance of study, we will focus on investiagting the resale price trends on 4-room flats and how the proxmity of a train station, shopping mall and distance to Singapore’s Central business District (CBD), Raffles Place, affects the resale price of the flat.

After cleaning the data, there were a total of 157,517 rows of data and 8 columns. After exploring and honing in particularly on 4-room flats, the data was reduced to 63,641 rows, which represents the number of 4 Room flats as seen in *Figure 2.*

Dropped Columns

Header Dropped	Reasons
Flat_Type	Since the sampled data focuses on 4-room flats, there is no need for flat_type moving forward
Flat_model	The study investigates resale price in general
Lease_commence_date	Column not needed as remaining_lease is already derived
Block & Street Name	Combined to form Address column



	year	town	storey_range	floor_area_sqm	remaining_lease	resale_price	Address
0	2012	ANG MO KIO	01 TO 05	82.0	63	400000.0	218 ANG MO KIO AVE 1
1	2012	ANG MO KIO	01 TO 05	91.0	67	411000.0	601 ANG MO KIO AVE 5
2	2012	ANG MO KIO	06 TO 10	92.0	65	427000.0	108 ANG MO KIO AVE 4
3	2012	ANG MO KIO	06 TO 10	92.0	65	430000.0	105 ANG MO KIO AVE 4
4	2012	ANG MO KIO	06 TO 10	92.0	66	433000.0	438 ANG MO KIO AVE 10

Figure 3: Cleaned and Wrangled Raw Data before Feature Engineering

The final cleaned data is as shown in *Figure 3*.

2.4 Feature Selection

The goal of the study is to investigate how the proximity of the following factors affect the pricing of the property

1. **Nearest Train Station**

Colloquially referred to as MRT Station, the nearest station was calculated based on the distance to the nearest MRT station.

2. **Distance to a shopping mall**

Shopping malls are also likely to contribute to the price of the property given how convenience is prioritized especially in Singapore given the relatively lower car ownership rates.

3. **Distance to Raffles Place, the central business district (CBD) in Singapore**

A strategic location would include minimizing the travel times to work. A lower public transport is also incurred the nearer one stays to central Singapore.

4. **Storey_range**

The higher a flat is located in a block, the higher the price. This follows as flats in Singapore are initially sold to the first owner at increments in price the higher the storey

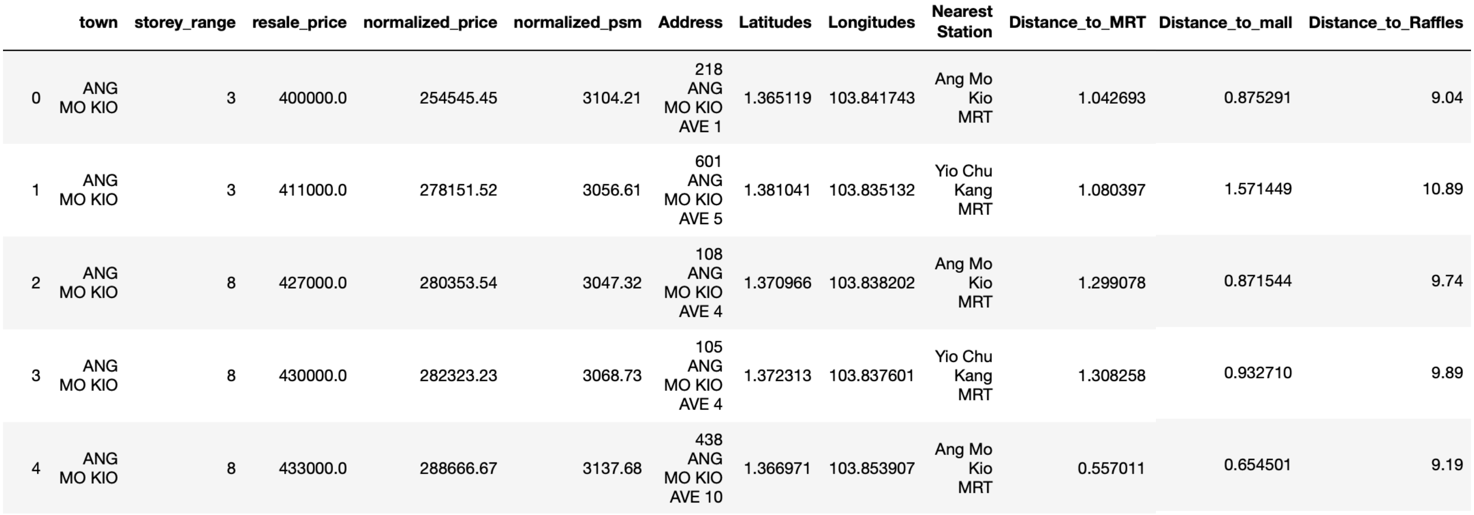
To achieve distance calculations, additional information had to be retrieved either through API or crawling the web.

Additional Information Sources

Proximity Factors	**Source**
1. Nearest Train Stations	Crawled Wikipedia for: 1. *List of MRT Stations in Singapore* 2. *MRT stations within each town (each town can have more* than one MRT station) OneMap API: - *Geocodes (Latitude and Longitude) for each MRT station*
2. Nearest Shopping Malls	Crawled Wikipedia for: - *List of Shopping Malls in Singapore* OneMap API: - *Geocodes (Latitude and Longitude) for each shopping mall*
3. Distance to Raffles Place	OneMap: - The geocodes for Raffles Place
4. Town premium	From the raw data, all resale prices were normalized according to remaining_lease and area_per_sqm linearly. Town Premium was calculated as the difference between the mean resale price within a town and the national mean resale price per sq meter.
5. Storey	The median of the storey_range from the raw data was being used to represent the storey of the property

<u>Columns Added</u>

In the final dataset set up for modelling, a total of 7 columns was added:



	town	storey_range	resale_price	normalized_price	normalized_psm	Address	Latitudes	Longitudes	Nearest Station	Distance_to_MRT	Distance_to_mall	Distance_to_Raffles
0	ANG MO KIO	3	400000.0	254545.45	3104.21	218 ANG MO KIO AVE 1	1.365119	103.841743	Ang Mo Kio MRT	1.042693	0.875291	9.04
1	ANG MO KIO	3	411000.0	278151.52	3056.61	601 ANG MO KIO AVE 5	1.381041	103.835132	Yio Chu Kang MRT	1.080397	1.571449	10.89
2	ANG MO KIO	8	427000.0	280353.54	3047.32	108 ANG MO KIO AVE 4	1.370966	103.838202	Ang Mo Kio MRT	1.299078	0.871544	9.74
3	ANG MO KIO	8	430000.0	282323.23	3068.73	105 ANG MO KIO AVE 4	1.372313	103.837601	Yio Chu Kang MRT	1.308258	0.932710	9.89
4	ANG MO KIO	8	433000.0	288666.67	3137.68	438 ANG MO KIO AVE 10	1.366971	103.853907	Ang Mo Kio MRT	0.557011	0.654501	9.19

Figure 4: The data after feature selection and crawling for necessary information

**All distances are in kilometres.*

3. Model Selection

3.1 Requirements

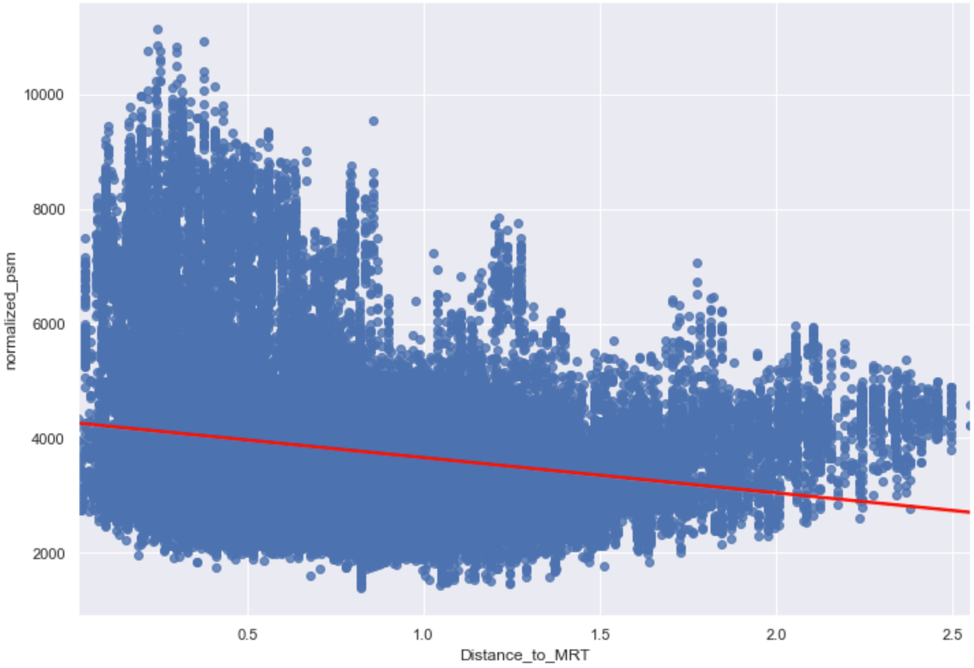
Because the goal is to discover how the various proximity factors affect the price, normalized_psm representing the normalized price of the properties is the dependant variable (Y)

The Independent variables (X) are the 4 proximity factors - 'storey_range','Distance', 'Distance_to_mall', 'Distance_to_Raffles'

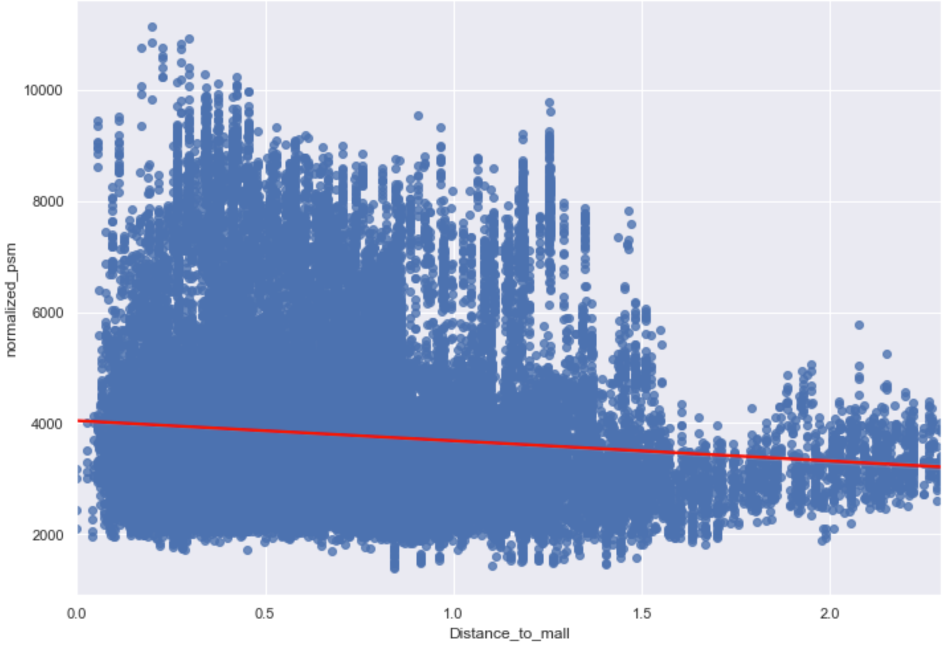
- All values are continuous.
- The model will involve multiple variables.
- Given the cleaned and well-labelled data, a supervised machine learning model will be utilized.

3.2 Understanding Data Points

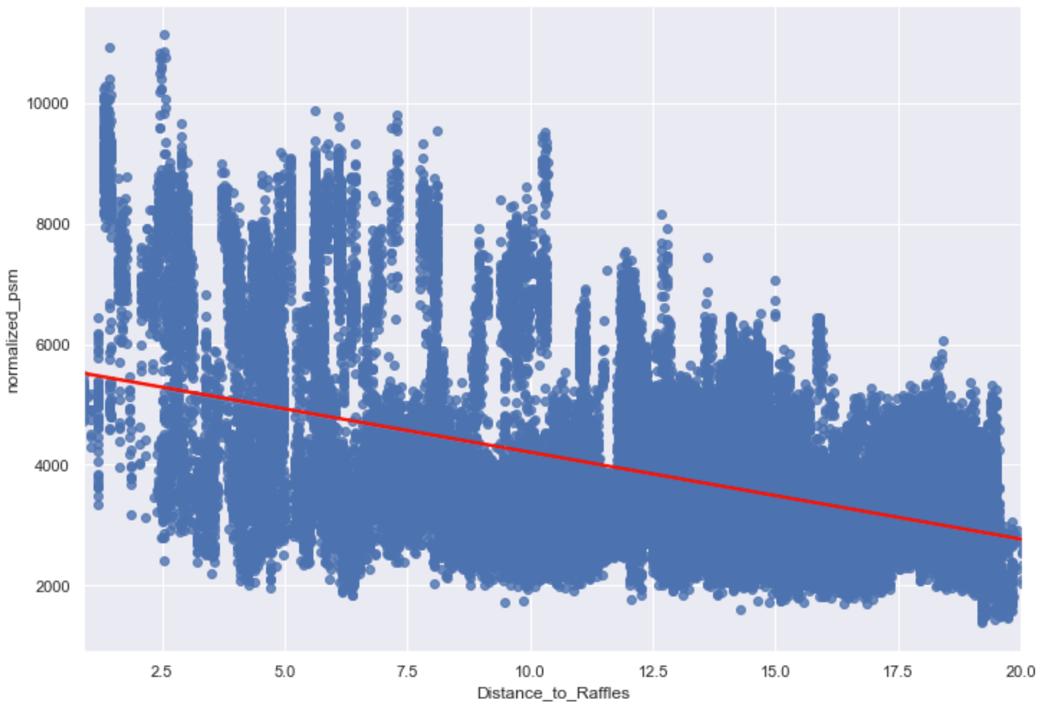
Below are several plots of the for each of the 3 proximity factors



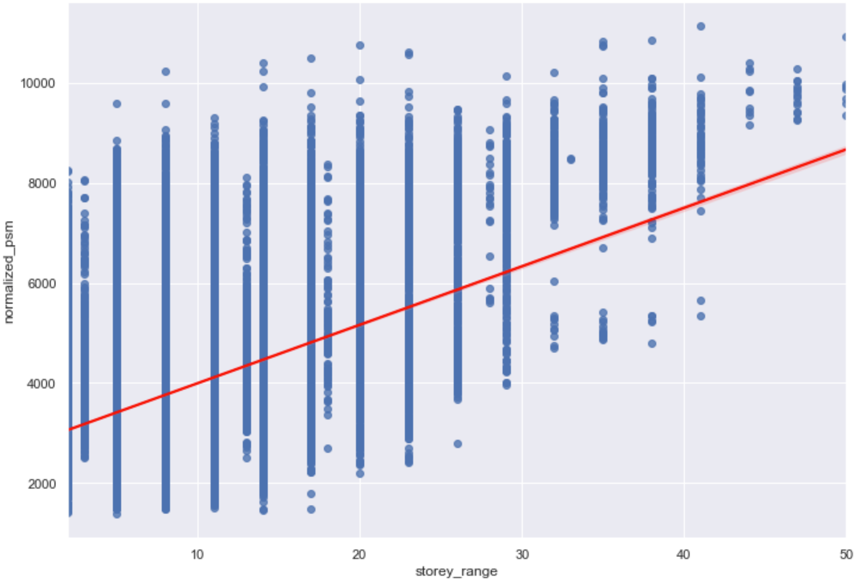
With a gradient of -600, we can see that the general trend is that the price of HDB flats declines as the distance to the nearest MRT Train station increases.



With a gradient of -355, there is a slight trend in the decrease in price of a flat as the distance to a nearby mall. One possibility to the slightly weaker trend would be that malls are not the only source of convenience, since Singapore has convenient stores, similar to 7-eleven, scattered around residential areas. These community areas also popularly include supermarkets and small businesses.



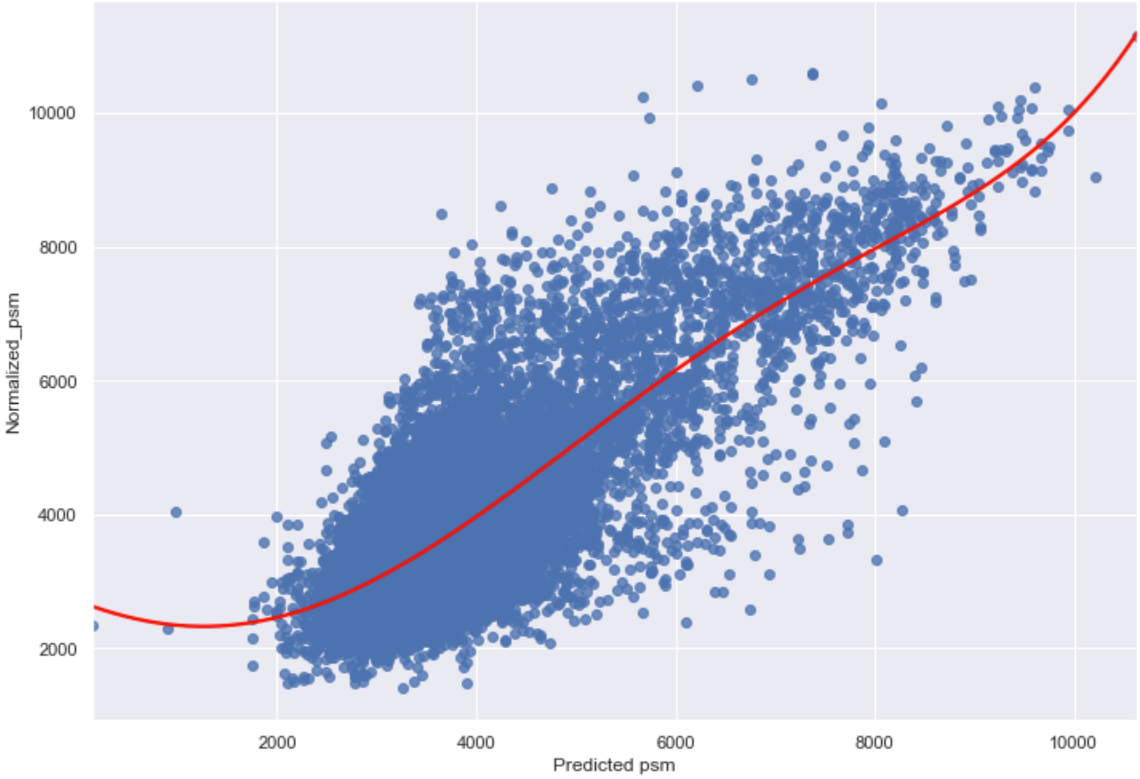
With a gradient of -142, the distance to Raffles place (central Singapore) has the lowest effect on buying decisions based on proximity factors. One consideration would be that majority of the population works outside of the Central Business District, specifically Raffles Place. *(note the scale of this graph is different to the previous graphs)*



With a gradient of 115, this suggests that an increase in floor of a house, there will be roughly a $115 increase in price per square meter on average, working out to be about $10,000 in total.

3.3 Better predictors

To comprehensively understand how the 3 proximity factors, I performed a multi-linear regression. The accuracy score came out to be 0.44. I then proceeded to perform a Multi-variable Polynomial Regression. For brevity, I tested to all n-degrees from 1 up to 10, with degree 5 returning the best R2 score of 0.608.



Summary

In summary, resale prices of HDB flats in Singapore are influenced by a multitude of factors, and about 60% of the price is possibly affected by its proximity to a MRT station, a nearby mall or even Raffles Place. It supports the hypotheses and general buyers’ sentiments in Singapore but definitely highlights how the buyers’ population do not place that great of an emphasis on these factors as commonly heard.

Considerations

The three proximity factors are only valid for individuals who purchase a property on the same premises. I acknowledge that buyers’ habits can vary greatly and that the study using these data points does not communicate the resale buying’s ground truth precisely.

1. Distance to MRT — Whilst a shorter walking distance is preferable, the desirability can be affected by personal considerations such as track noise, mode of transport and daily commute route.

2. Distance to a Mall — A shopping mall brings about great convenience given the diversity of products and services offered. However, patronage is largely dependant on the relevance of the retailers to the buyers. Some considerations include the lack or limited choice of grocery chains and costlier food options.

3. Distance to Raffles Place — It is a common trend for properties closer to the the center of a city to cost more, anywhere in the world. In Singapore, it is no different. However, the importance of staying close to Raffles Place largely depends on the lifestyle and daily commute of buyers.

Discussions

Other considerations to further improve the accuracy of the model can include modelling for the influence of inflation on housing prices and including the demographics of residents in each town. Furthermore, additional proximity factors could be added such as nearby educational facilities (in Singapore, children are enrolled into primary schools based on their residential address), Supermarkets and Hawker Centres given how these are facilities more frequented as opposed to shopping malls.