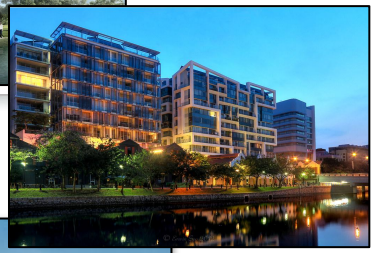


Predicting Resale Private Properties Prices in Singapore

(Non landed)

Data Science Project

Team 3| Team mates

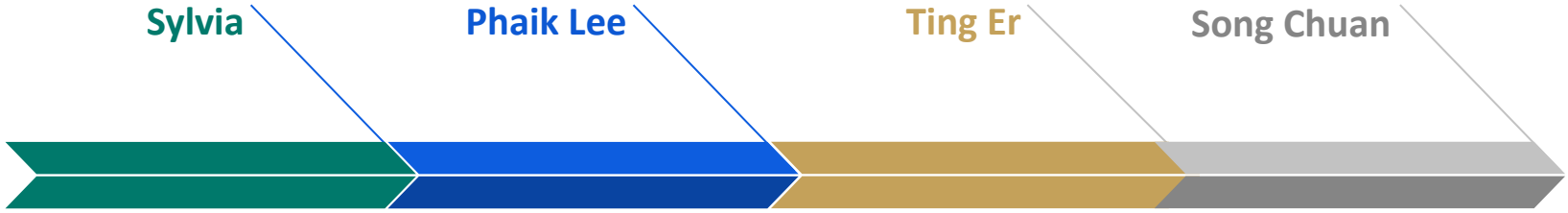
Song Chuan
(Team leader)

Ting Er

Phaik Lee

Sylvia

Agenda



- Business Understanding
- Data Exploration & Preparation
- Data Representation & Transformation
- Data Visualization & Presentation
- Train Data Models
- Validate Data Models
- Deploy Data Models
- Environmental feedback

Property news

EdgeProp

Home > Property News > News

Unit at The Nassim reaps $2.6 mil profit

SINGAPORE (EDGEPROP) - The seller of a unit at **The Nassim**, located on Nassim Hill, made the top gain of $2.6 million over the week of Oct 26 to Nov 2. The 3,122 sq ft unit on the third floor was bought for $10.4 million ($3,332 psf) in February 2018 and sold for $13 million ($4,165 psf) on Oct 26. The seller therefore made a 25% profit, or an annualised profit of 6% over almost four years.

EdgeProp Home Sale Rent New Launches Cast Tools Property News

Home > Singapore Condo & Apartment Directory > Orchard View

Orchard View

On the other hand, the most unprofitable deal of the week was the resale of a 2,530 sq ft unit at Orchard View on Oct 27. Having sold the property for $7.1 million ($2,807 psf), the seller suffered an 18% loss of $1.59 million. The unit was purchased in August 2010 for $8.69 million ($3,434 psf). Over a holding period of 11 years, this translates into an annualised loss of 2%.

Property news

$11.8 mil loss!

CNA
about a year ago

James Dyson and his wife bought the Wallich Residence apartment a year ago for a reported S$73.8 million.

CHANNELNEWSASIA.COM

British billionaire Dyson sells Singapore's priciest penthouse for S...

British billionaire James Dyson, the inventor of the bagless vacuum cleaner, and hi…

BBC | Sign in | Home | News | Sport | Reel | Wo

NEWS

Home | Coronavirus | Climate | Video | World | Asia | UK | Business | Tech | Science | Stories

Business | Market Data | New Economy | New Tech Economy | Companies | Entrepreneurship

Sir James Dyson to sell Singapore penthouse at a loss

19 October 2020

Sir James Dyson has agreed to sell his Singapore penthouse at a loss, just one year after buying it.

Singapore's Business Times reported that he accepted an offer for $62m Singapore dollars ($47m; £36m) from US-based billionaire Leo Koguan.

The sale price is lower than the reported purchase price of S$73.8m.

Sir James bought the luxurious flat - said to be Singapore's largest - last year after announcing he was moving Dyson's headquarters to the city state.

The company is best known for vacuum cleaners, air purifiers and hair dryers.

Dyson also considered building an electric car factory in Singapore, before pulling the plug over concerns about the car's commercial viability.

Business understanding

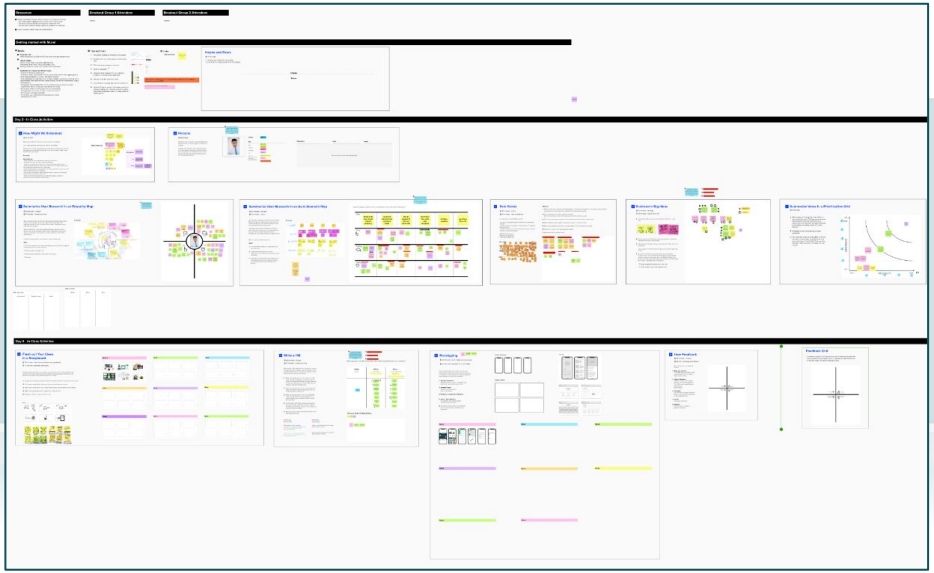
Singapore property market is dynamic, owning a private property touches a special place in many hearts. Ownership means security for self and families and properties are often seen as safe haven investments.

Purchase of a property is always fraught with emotions, fear seems to be inevitable. Is it the right time, right place and right price ? Fear of buyers' regret often kept people on a constant hunt for the 'perfect' property.

Rise and fall of property values are more than just numbers for homeowners, anxiety can be real. Will machine learning be able to predict property prices and the good buys?

Let's explore...

EDT



EDT | HMW, Persona and Empathy map

HMW

How might we	forecast prices	
for	resale private property buyers	
so that	they can make a better decision	and see value appreciation in the future?

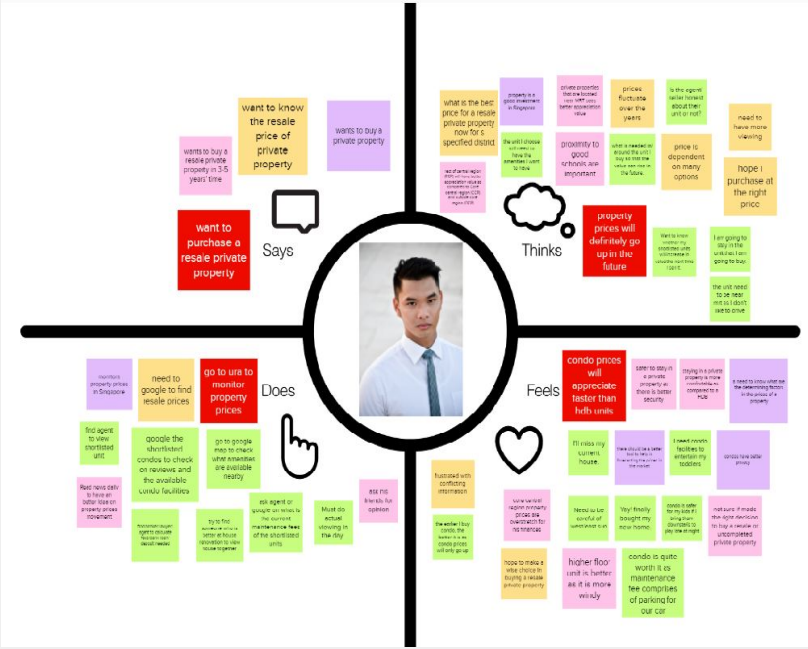
Persona

Name Justin

Bio

Age	37
Location	Singapore
Education	Tertiary
Job	Manager
Family	Married with 1 child
Work experience	12 years
Education	
Technical literacy	Banking and finance

Empathy map



EDT | Pain points & Big ideas

Pain points

★ Lack of information available

- need to google to find resale prices
- what is the best price for a resale private property now for s specified district
- monitors property prices in Singapore
- frustrated with conflicting information

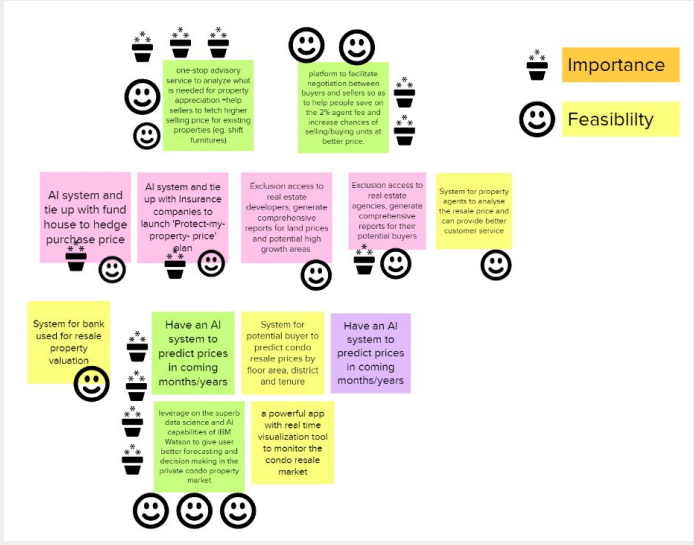
★ Worry about wrong choices made

- Want to know whether my shortlisted units will increase in value the next time i sell it.
- hope to make a wise choice in buying a resale private property
- not sure if made the right decision to buy a resale or uncompleted private property
- prices fluctuate over the years
- hope i purchase at the right price
- need to have more viewing

★ Is the price of the unit going to rise or not?

- what is needed in/around the unit I buy so that the value can rise in the future.
- private properties that are located near MRT sees better appreciation value
- property prices will definitely go up in the future
- rest of central region (RCR) will have better appreciation value as compared to Core central region (CCR) and outside core region (OCR)
- price is dependent on many options

Big ideas

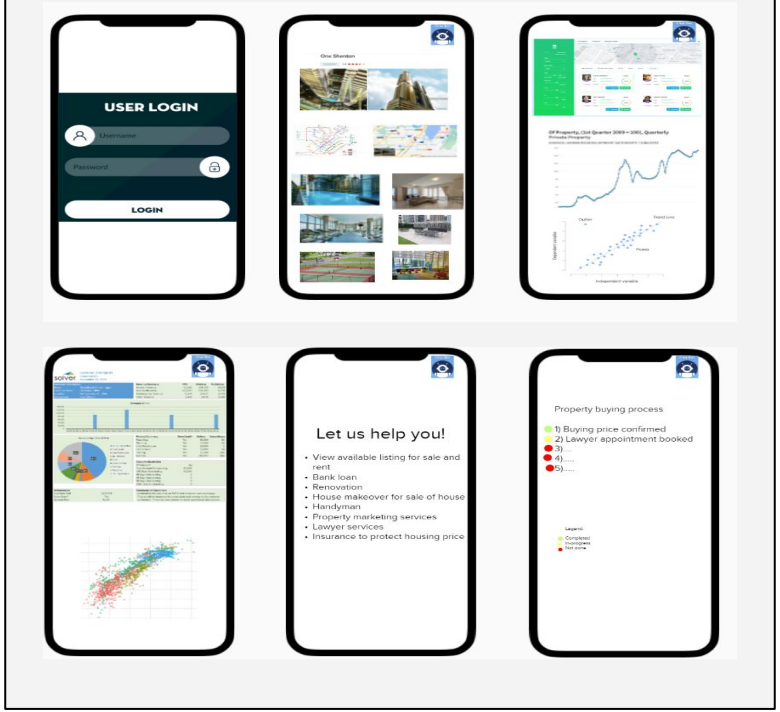


EDT | Hill statement & Prototype

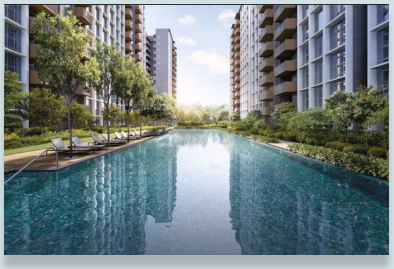
Hill Statement

Justin	is now able to see what factors affecting the price of his shortlisted units	and enabling him to make a more informed decision for his resale property purchase

Prototype



Business Opportunity



Using AI + machine learning tools, we seize the opportunity to help people to find the best value property and provide a one stop service solution needed for the property purchase journey.

Our analysis will be available to potential property buyers which was traditionally known only to the property agents. Our system can also guide buyers to complete property transactions on their own.

This in turn generate opportunities for our buyers, service partners and in turn our incentives from the successful collaborations.

Hypothesis & Assumptions

Hypothesis

(1) Resale private properties that are located near CBD area have higher price per square foot
(2) High level units can fetch higher prices than mid and low floor units
(3) Projects that are located nearer to MRT can fetch higher resale price

Assumptions

(1) Home loan interest rate is constant
(2) No adverse economic market conditions
(3) Persona does not have any financial constraint
(4) Persona is only looking at resale private condominiums or apartments

Watson Studio

Data Exploration & Preparation

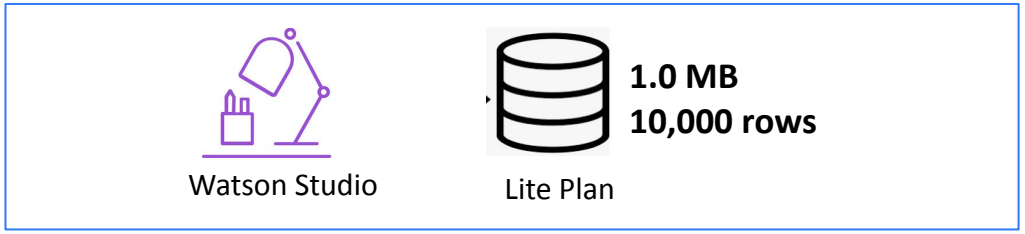
- Data of resale transaction price of condominiums in Singapore
- **URA e-Service** provides the following data:
  - Resale private property transactions within the last 60 months (from Nov 2016 till Nov 2021)
  - Data downloaded by district
  - District 1-28 , except district 24 (Lim Chu Kang area)

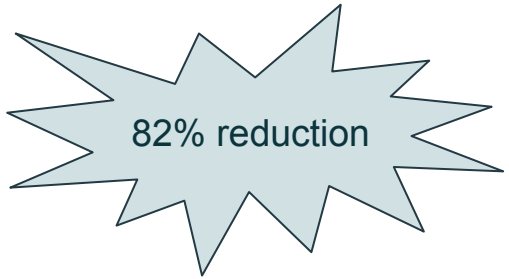
URBAN REDEVELOPMENT AUTHORITY

- **27 district files** are downloaded with the following features:

Project_Name	Street_Name	Type	Postal_District	Market_Segment	Tenure	Price	Area_Sqft	Floor_Level	Unit_Price_psf	Date_of_Sale

- **To combine all files**
  - pandas to concatenate 27 district data files



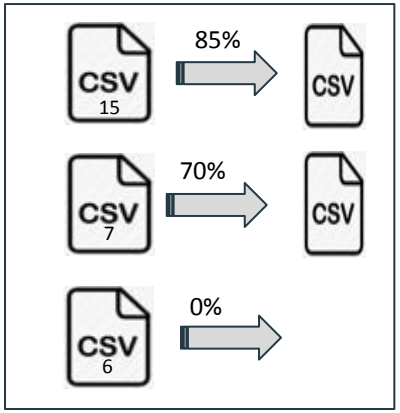



```
import pandas as pd

df1 = pd.read_csv('/Users/User1/Desktop/dataset/district/d1.csv')
df2 = pd.read_csv('/Users/User1/Desktop/dataset/district/d2.csv')
  :
  :
df28 = pd.read_csv('/Users/User1/Desktop/dataset/district/d28.csv')
df1 = pd.read_csv('/Users/User1/Desktop/dataset/district/d1.csv')
frames=[df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12,df13,df14,df15,df16,df17,df18
,df19,df20,df21,
        df22,df23,df25,df26,df27,df28]
df = pd.concat(frames)
df.to_csv('/Users/User1/Desktop/dataset/district/rawdata.csv',index=False)
```

- How to reduce the data file size to 1MB and 10,000 rows
  - Reduce the master file (5.6MB, 53,380 rows)
  - Reduce every district file (27 files) ✓

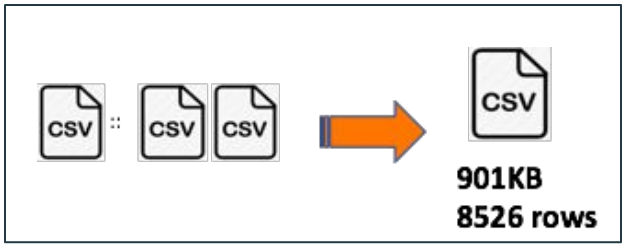
- Reduce 27 district files individually by different percentage
  - every district file has different data size
  - district 15 (512KB), district 11 (207KB), district 6(1KB)
  - reduce every district file by different percentage
  - to keep the data with smaller data size

- numpy to randomly reduce different file by the different percentage
- pandas to concatenate all the district files to become single file




```
import numpy as np

np.random.seed(10)

tdf1 = df1
rows = round(tdf1.shape[0] * .85)
remove_n = rows
drop_indices = np.random.choice(tdf1.index, remove_n, replace=False)
tdf1 = tdf1.drop(drop_indices)
:
tdf7 = df7
rows = round(tdf7.shape[0] * .70)
remove_n = rows
drop_indices = np.random.choice(tdf7.index, remove_n, replace=False)
tdf7 = tdf7.drop(drop_indices)
```

python



❑ Data Understanding

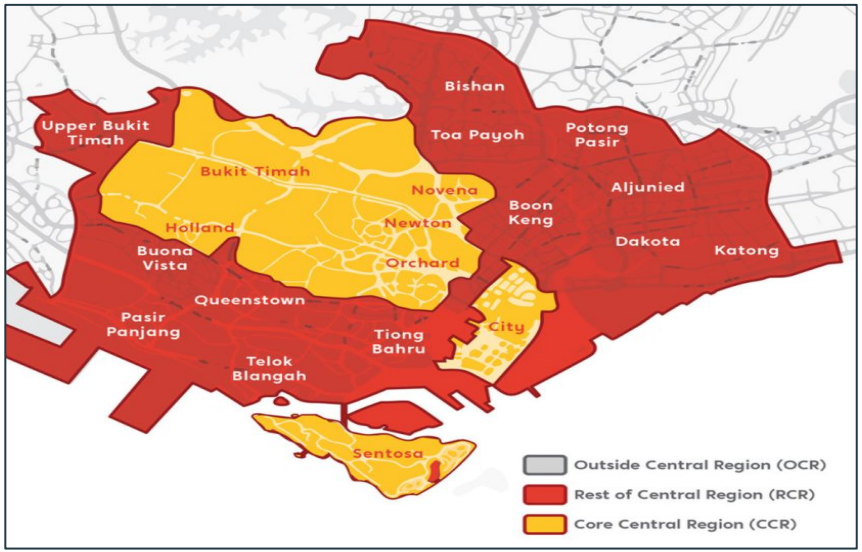
Project_Name	Street_Name	Type	Postal_District	Market_Segment	Tenure	Price	Area_Sqft	Floor_Level	Unit_Price_psf	Date_of_Sale
THE SAIL @ MARINA BAY	MARINA BOULE	Apartment	1	CCR	99 yrs lease commencing from 2002	5000000	2164	46 to 50	2311	1/11/21
MARINA ONE RESIDENCES	MARINA WAY	Apartment	1	CCR	99 yrs lease commencing from 2011	1575852	721	06 to 10	2185	1/11/21
MARINA ONE RESIDENCES	MARINA WAY	Apartment	1	CCR	99 yrs lease commencing from 2011	1660359	753	16 to 20	2204	1/11/21
V ON SHENTON	SHENTON WAY	Apartment	1	CCR	99 yrs lease commencing from 2011	3545100	1755	46 to 50	2021	1/11/21
MARINA ONE RESIDENCES	MARINA WAY	Apartment	1	CCR	99 yrs lease commencing from 2011	1579694	710	06 to 10	2224	1/11/21
THE SAIL @ MARINA BAY	MARINA BOULE	Apartment	1	CCR	99 yrs lease commencing from 2002	1530000	689	11 to 15	2221	1/10/21
MARINA ONE RESIDENCES	MARINA WAY	Apartment	1	CCR	99 yrs lease commencing from 2011	4421711	1593	21 to 25	2776	1/10/21

Project_Name	Condominium/Apartment/s name
Street_Name	Address
Type	Condominium, Apartment
Postal_District	District number in Singapore
Market_Segment	Market Segment
Tenure	Lease/Freehold
Price	Transaction price in $
Area_Sqft	Area in square feet
Floor_Level	Floor level from 01 onwards
Unit_Price_psf	Unit price per square foot
Date_of_Sale	Date of transcation

- Market Segment - 3 main regions
  - **CCR** (Core Central Region)
  - **OCR** (Outside Central Region)
  - **RCR** (Rest of Central Region)

- These regions contain all 28 districts

Market_Segment
RCR
OCR
OCR
RCR
OCR
OCR
RCR



Singapore Regions	Postal District It Covers
Core Central Region (CCR)	9, 10 and 11, and parts of 1, 2, 4, 6 and 7
Rest of Central Region (RCR)	3, 8 and 12, and parts of 1, 2, 4, 5, 6, 7, 13, 14, 15 and 20
Outside Central Region (OCR)	16 to 19, 21 to 28, and parts of 5, 14, 15 and 20

- Postal District
  - Postal code - First 2 digits of postal code, The Balmoral postal code **25**9802 is district 10
  - General location - Raffles Place is district 1 and Jurong is district 22

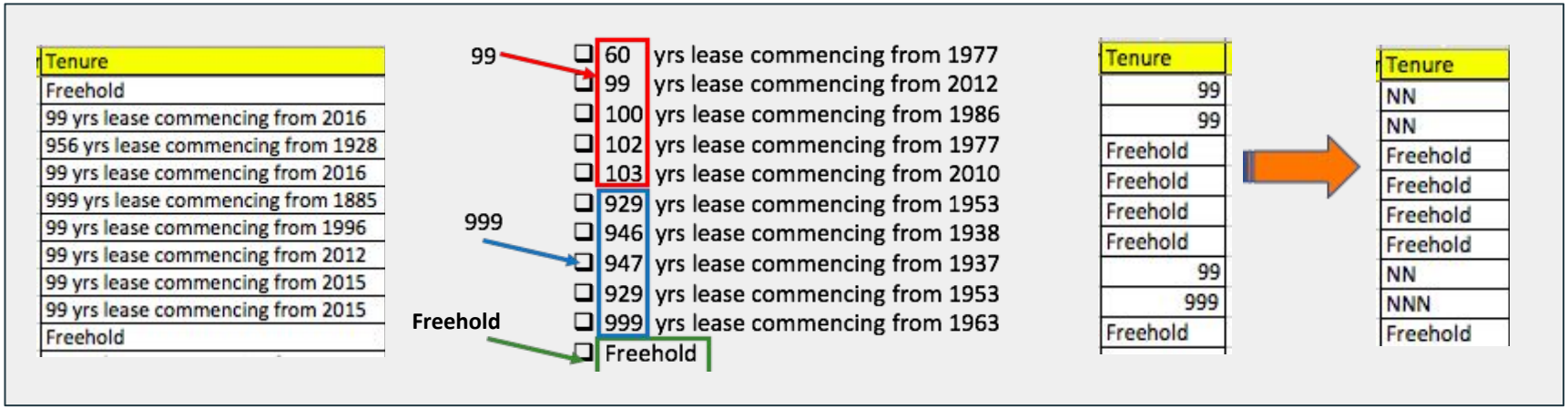
Postal_District
15
25
19
4
27
18
4

Postal District	Postal Sector (1st 2 digits of 6-digit postal codes)	General Location
01	01, 02, 03, 04, 05, 06	Raffles Place, Cecil, Marina, People's Park
02	07, 08	Anson, Tanjong Pagar
03	14,15, 16	Queenstown, Tiong Bahru
04	09, 10	Telok Blangah, Harbourfront
05	11, 12, 13	Pasir Panjang, Hong Leong Garden, Clementi New Town
06	17	High Street, Beach Road (part)
07	18, 19	Middle Road, Golden Mile
08	20, 21	Little India
09	22, 23	Orchard, Cairnhill, River Valley
10	24, 25, 26, 27	Ardmore, Bukit Timah, Holland Road, Tanglin
11	28, 29, 30	Watten Estate, Novena, Thomson
12	31, 32, 33	Balestier, Toa Payoh, Serangoon
13	34, 35, 36, 37	Macpherson, Braddell
14	38, 39, 40, 41	Geylang, Eunos
15	42, 43, 44, 45	Katong, Joo Chiat, Amber Road
16	46, 47, 48	Bedok, Upper East Coast, Eastwood, Kew Drive
17	49, 50, 81	Loyang, Changi
18	51, 52	Tampines, Pasir Ris
19	53, 54, 55, 82	Serangoon Garden, Hougang, Punggol
20	56, 57	Bishan, Ang Mo Kio
21	58, 59	Upper Bukit Timah, Clementi Park, Ulu Pandan
22	60, 61, 62, 63, 64	Jurong
23	65, 66, 67, 68	Hillview, Dairy Farm, Bukit Panjang, Choa Chu Kang
24	69, 70, 71	Lim Chu Kang, Tengah
25	72, 73	Kranji, Woodgrove
26	77, 78	Upper Thomson, Springleaf
27	75, 76	Yishun, Sembawang
28	79, 80	Seletar

Watson Studio

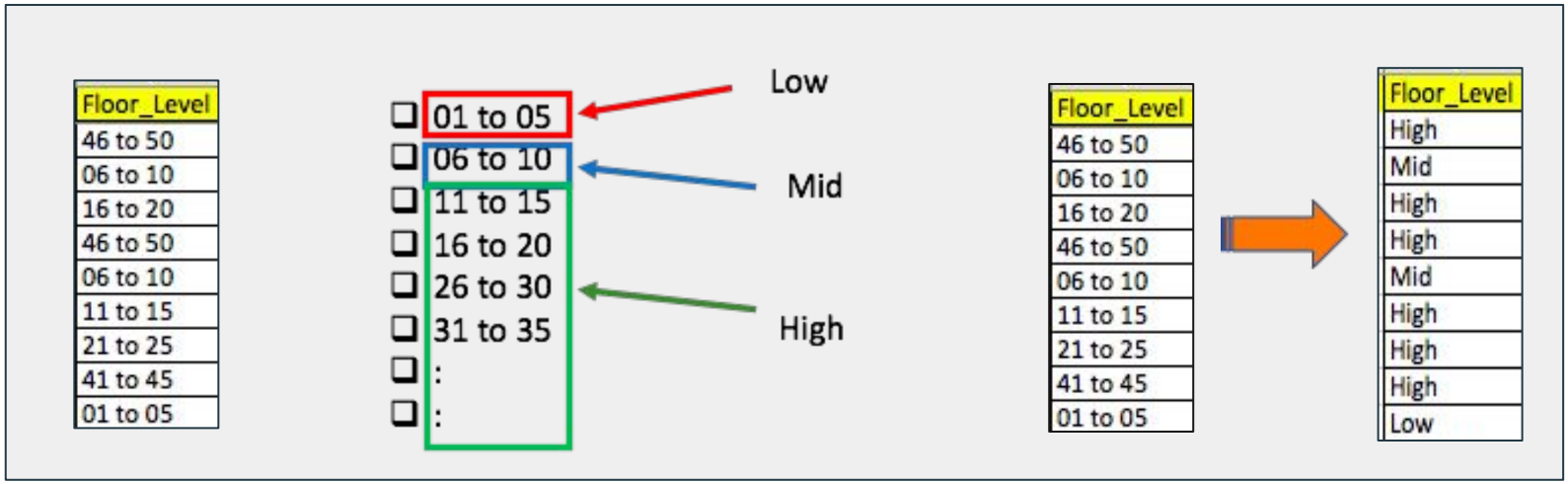
Data Representation & Transformation

- **Tenure**
  - Categorize tenure to 3 categories --> 99, 999 and freehold
  - numpy to do the categorization
  - Further categorize to become NN, NNN and Freehold




```
import numpy as np

col = "Tenure"
rdf[col] = rdf[col].str[:3]
conditions  = [ rdf[col] == "60 ",rdf[col] == "99 ", rdf[col] == "100",rdf[col] ==
"102",rdf[col] == "103", rdf[col] == "999",  rdf[col] == "946",rdf[col] ==
"947",rdf[col] == "956", rdf[col] == "929"]
choices     = [ "99","99", "99","99","99", "999","999", "999","999","999"]
rdf[col] = np.select(conditions, choices, default="Freehold")
```

- **Floor_Level**
  - Categorize floor level to “High”, “Mid” and “Low”
  - numpy to do the categorization




```
import numpy as np

col = 'Floor_Level'
conditions  = [ rdf[col] == "01 to 05", rdf[col] == "06 to 10"]
choices     = [ "Low", "Mid"]
rdf[col] = np.select(conditions, choices, default="High")
```

- **Postal_District**
  - Categorize postal district by adding “D” in front of the district number
  - Python to append “D” after converting the district number to String

Postal_District
15
25
19
4
27
18
4

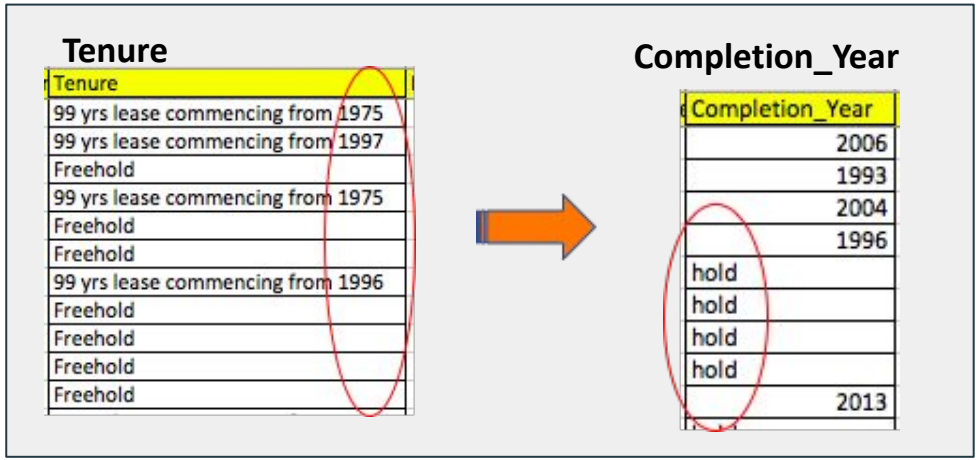
Postal_District
D15
D25
D19
D4
D27
D18
D4

python

```
xdf['Postal_District']= "D" + xdf['Postal_District'].astype(str)
```

- **Completion Year / TOP (Temporary Occupation Permit)** is a useful feature to analyse resale price for private property
  - No ready datasets of completion year for private property available
  - Useful information found in the raw data under Tenure feature
  - Python to extract the last 4 digits and new feature “Completion_Year” is created

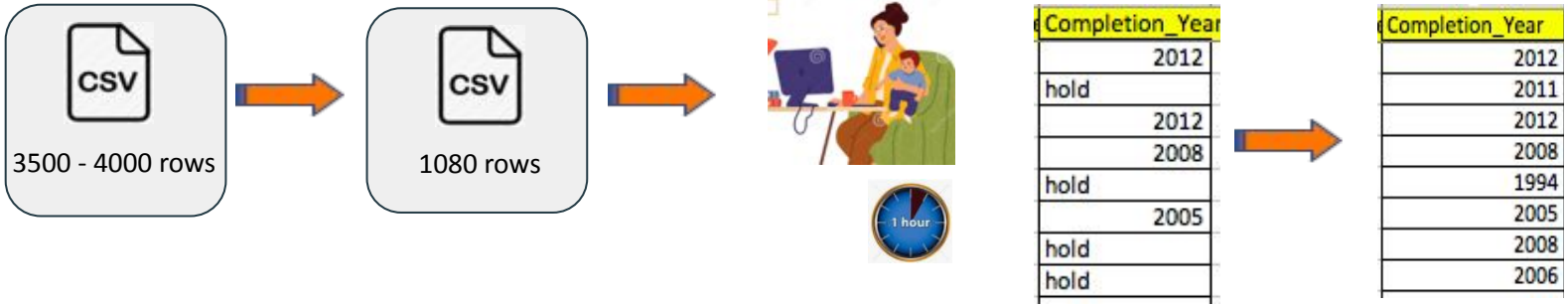
- Extracting completion year for 99 and 999 years




```
rdf["Completion_Year"] = rdf['Tenure'].str[-4:
```
python

❑ Completion year for freehold private property

Key Issues	Solution
➢ No ready data available ➢ To search on internet for each condominium ➢ 3500 to 4000 rows of data	➢ To remove rows with similar condominium's name ➢ Example The Balmoral has 10 rows → 1 row ➢ Python to drop duplicating row for condominium's name ➢ Generate a file with 1 row per 1 condo




```
freehold_df = rdf.loc[rdf['Tenure'] == 'Freehold']
freehold_df.to_csv('/Users/User1/Desktop/dataset/freehold.csv',index=False)
dup_df = freehold_df.drop_duplicates(subset=['Project_Name'])
dup_df.to_csv('/Users/kUser1/Desktop/dataset/freehold.csv',index=False)
```

python

- Age
  - number of year from Completion Year to 2021
  - Python to calculate the age

- Age = 2021 - Completion Year

- New features “Age_in_2021” is created

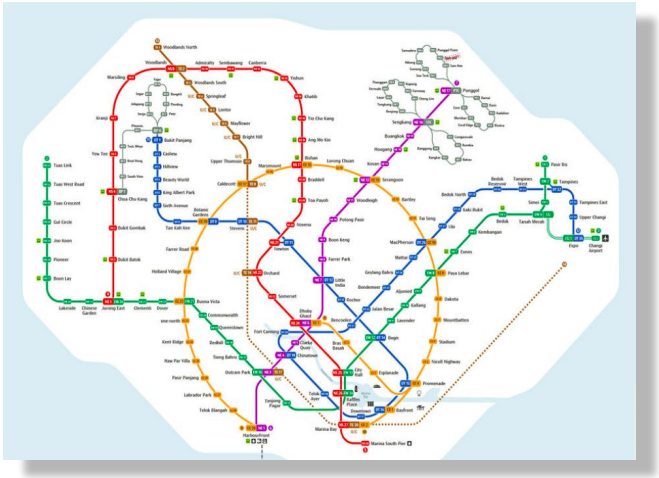
Completion_Year	Age_in _2021
2001	20
2011	10
2016	5
1999	22
2010	11
1992	29
2009	12
2006	15
1998	23

```
ydf["Age_in _2021"]=2021 - ydf["Completion_Year"]
```

python

- Distance between the condominium and MRT
  - the distance is a useful feature to analyse resale price for private property
- No ready datasets available for private property
- To find the distance between 2 locations
  - latitude and longitude for each condominium and the MRT stations
  - there are 8500 rows for condominium and 120 MRT stations

(Latitude, longitude) x 8526



(Latitude, longitude) x120

❑ Latitude and Longitude of a Condominium

➢ **OneMap REST API -** provides searching of a condominium's name and returns its latitude and longitude in JSON format

❑ **API endpoint**

"https://developers.onemap.sg/commonapi/search?searchVal=[+address+]&returnGeom=Y&getAddrDetails=Y&pageNum=1"

❑ **API response**

{"found":1,"totalNumPages":1,"pageNum":1,"results":[{"SEARCHVAL":"THE BALMORAL","BLK_NO":"14","ROAD_NAME":"BALMORAL PARK","BUILDING":"THE BALMORAL","ADDRESS":"14 BALMORAL PARK THE BALMORAL SINGAPORE 259846","POSTAL":"259846","X":"27394.4305731868","Y":"33086.3718265966","LATITUDE":"1.31549590683582","LONGITUDE":"103.827877231883","LONGTITUDE":"103.827877231883"}]}

API Response


```
{
  "found":1,
  "totalNumPages":1,
  "pageNum":1,
  "results":[
            {
            "SEARCHVAL":"THE BALMORAL",
            "BLK_NO":"14",
            "ROAD_NAME":"BALMORAL PARK",
            "BUILDING":"THE BALMORAL",
            "ADDRESS":"14 BALMORAL PARK THE
BALMORAL SINGAPORE 259846",
            "POSTAL":"259846",
            "X":"27394.4305731868",
            "Y":"33086.3718265966",
            "LATITUDE":"1.31549590683582",
            "LONGITUDE":"103.827877231883",
            "LONGTITUDE":"103.827877231883"
            }
 ]
}
```


- request to call to a simple http (API endpoint) and eval to get the response of latitude and longitude in JSON format
- the coordinates were filled up in another two new columns which are latitude and longitude

Project_Name	Latitude	Longitude
MARINA ONE RESIDENCES	1.276714985	103.8533269
MARINA ONE RESIDENCES	1.276714985	103.8533269
MARINA ONE RESIDENCES	1.276714985	103.8533269
MARINA ONE RESIDENCES	1.276714985	103.8533269
MARINA BAY RESIDENCES	1.279625799	103.8549872
V ON SHENTON	1.277083407	103.8491813
THE SAIL @ MARINA BAY	1.280769435	103.8526586
MARINA ONE RESIDENCES	1.276714985	103.8533269
THE SAIL @ MARINA BAY	1.280769435	103.8526586
THE SAIL @ MARINA BAY	1.280769435	103.8526586

```
import requests

def getcoordinates(address):
    req = requests.get('https://developers.onemap.sg/commonapi/search?searchVal='+address+'&returnGeom=Y&getAddrDetails=Y&pageNum=1')
    resultsdict = eval(req.text)
    if len(resultsdict['results'])>0:
        return resultsdict['results'][0]['LATITUDE'], resultsdict['results'][0]['LONGITUDE']
    else:
        return 0
```

- Latitude and Longitude of MRT Station
  - Similar method (OneMap REST API) is used to find the MRTs' latitude and longitude
  - 27 MRT coordinate files by district were generated

District 1

MRT_Station	Latitude	Longitude
MARINA BAY	1.27642735	103.854598
MARINA SOUTH P	1.27102704	103.862448
BAYFRONT	1.28187379	103.85908
DOWNTOWN	1.27944619	103.85284
NICOLL HIGHWAY	1.29976684	103.863637
PROMENADE	1.29289175	103.860892
TELOK AYER	1.28206895	103.848649
RAFFLES PLACE	1.28412561	103.851462
CHINATOWN	1.28422392	103.845144
CLARKE QUAY	1.28838602	103.846555
OUTRAM PARK	1.27973971	103.839514
TANJONG PAGAR	1.27656132	103.845725

District 14

MRT_Station	Latitude	Longitude
KALLANG	1.31148891	103.871387
ALJUNIED	1.31643261	103.882906
STADIUM	1.30281247	103.875338
DAKOTA	1.30854798	103.889065
MOUNTBATTEN	1.30620191	103.882528
UBI	1.32997463	103.899227
MACPHERSON	1.3260776	103.890391
PAYA LEBAR	1.31724739	103.892261
EUNOS	1.31978355	103.903226
KAKI BUKIT	1.3349673	103.90846
KEMBANGAN	1.32103825	103.912949

:::

District 28

MRT_Station	Latitude	Longitude
ANG MO KIO	1.36942856	103.849455
YIO CHU KANG	1.38175589	103.844947

- Shortest distance between a condominium and MRT
  - To compare one condominium with 120 MRT stations and get the shortest distance
  - To compare one condominium with MRT stations within the same district as the condominium ✓
- geopy to calculate distance from the latitude and longitude of 2 locations
- 2 new columns of feature Distance and nearest MRT station were created

D10

Project_Name	Latitude	Longitude
D'LEEDON	1.31521278	103.802002

D10

MRT_Station	Latitude	Longitude
STEVENS	1.32006556	103.826024
BOTANIC GARDENS	1.32248807	103.816014
TAN KAH KEE	1.32555624	103.80776
FARRER PARK	1.312679	103.854872
SIXTH AVENUE	1.33085765	103.796907
HOLLAND VILLAGE	1.3122403	103.796399

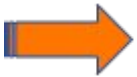
Project_Name	Distance	Mrt
D'LEEDON	0.70488254	HOLLAND VILLAGE
KENSINGTON SQUARE	0.4432076	BARTLEY
THE MINTON	0.88801498	SERANGOON
RIVER PLACE	0.93814885	OUTRAM PARK
ST THOMAS SUITES	0.4181853	SOMERSET
GLENTREES	1.36645925	HOLLAND VILLAGE
GUILIN VIEW	0.3280683	BUKIT GOMBAK
REFLECTIONS AT KEPPEL BAY	0.505334	TELOK BLANGAH
HILLVIEW GREEN	2.28139829	BEAUTY WORLD
SPOTTISWOODE RESIDENCES	0.99994916	TANJONG PAGAR
URBAN SUITES	0.37655642	SOMERSET
SKY@ELEVEN	0.71844685	CALDECOTT

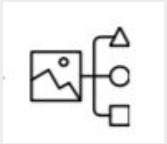
```
import geopy.distance
for i in range(0,gdf1.shape[0]):
    short_dist=100
    mrt="nil"
    coords_1=(gdf1.loc[i,"Latitude"],gdf1.loc[i,"Longitude"] )
    for j in range(0, gdf2.shape[0]):
        coords_2=(gdf2.loc[j,"Latitude"],gdf2.loc[j,"Longitude"] )
        cur_dist = geopy.distance.distance(coords_1, coords_2).km
        if cur_dist<short_dist:
            short_dist=cur_dist
            mrt=gdf2.loc[j,"MRT_Station"]
    gdf1.loc[i,"Distance"] = short_dist
    gdf1.loc[i,"Mrt"] = mrt
```

❑ **Final Dataset**

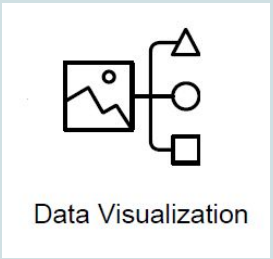
Project_Name	Street_Name	Type	Postal_District	Market_Segment	Tenure	Price	Area_Sqft	Floor_Level	Unit_Price_psf	Date_of_Sale	Completion_Year	Age_in_2021	Distance	Mrt

962 KB
8475 rows



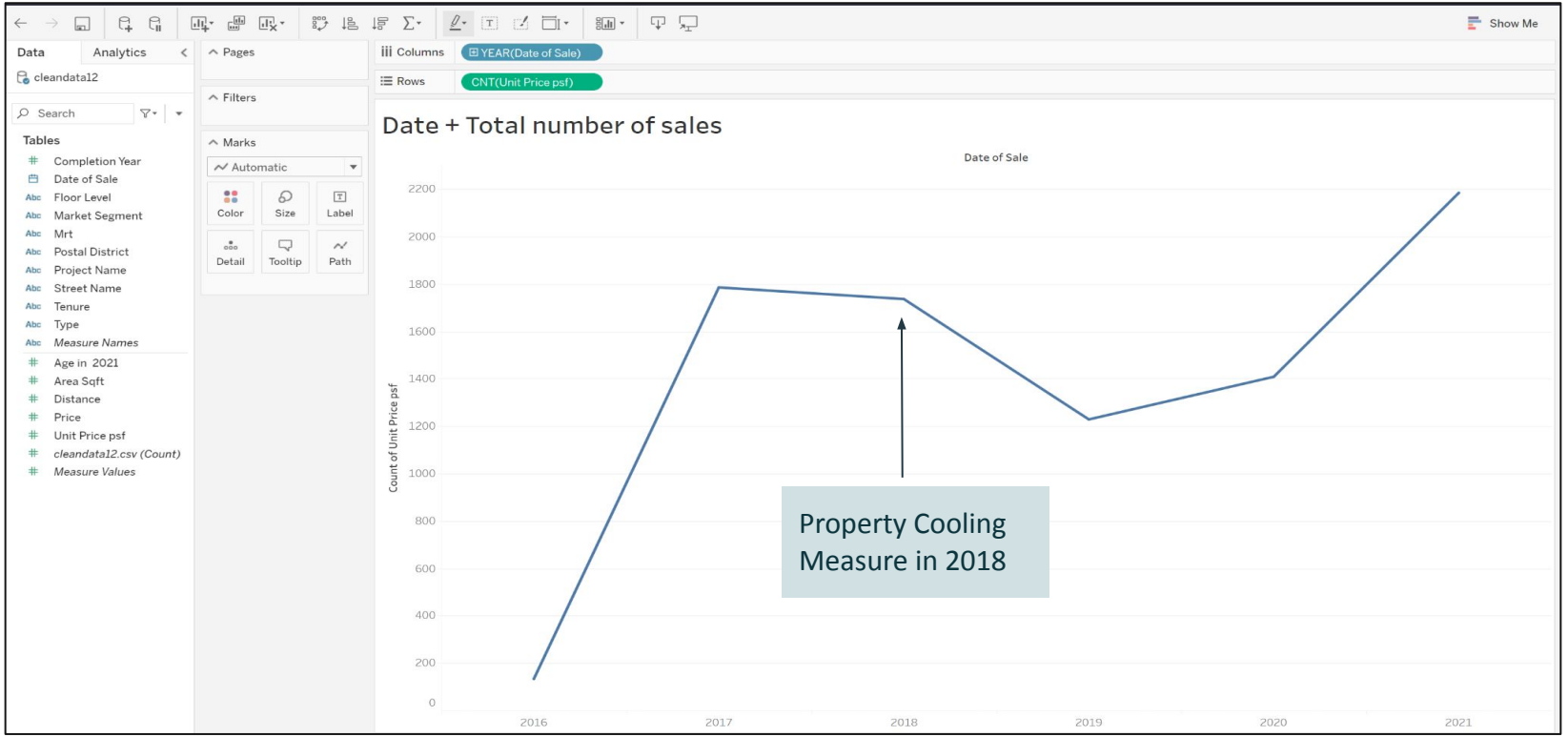


Data Visualization & Presentation

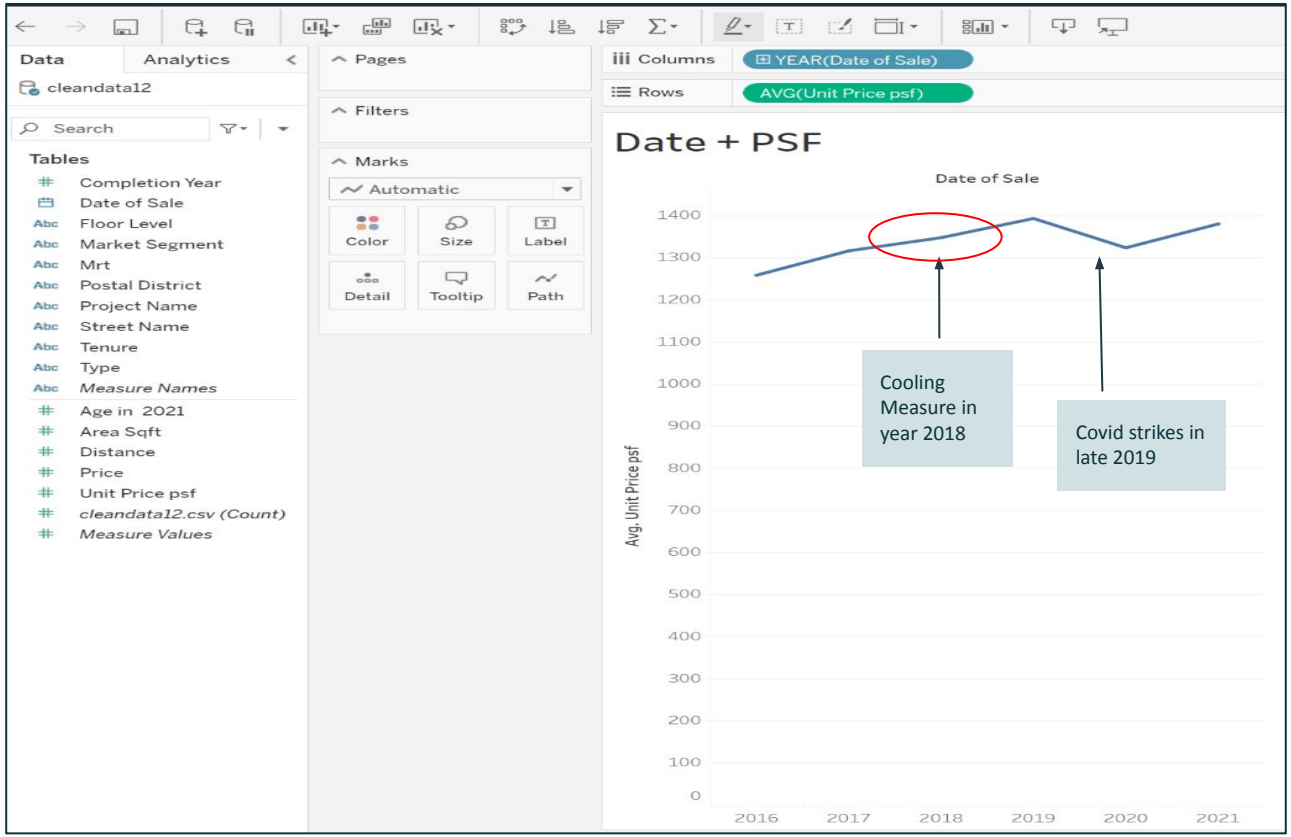


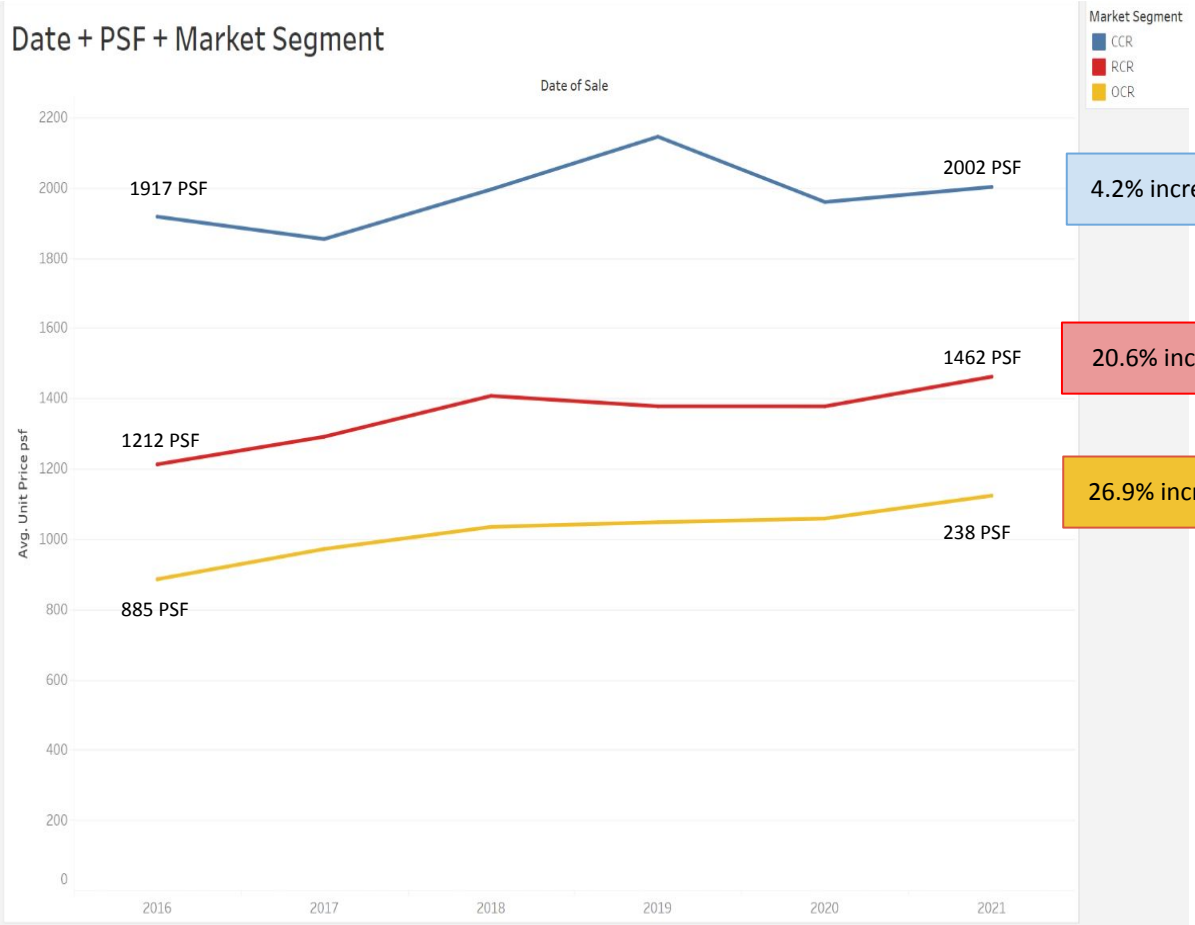
Data Visualization & Presentation

Market demand outlook

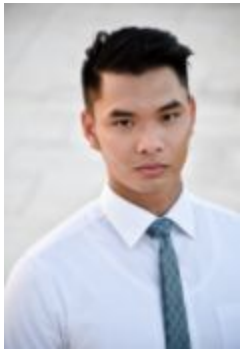


Market Demand outlook

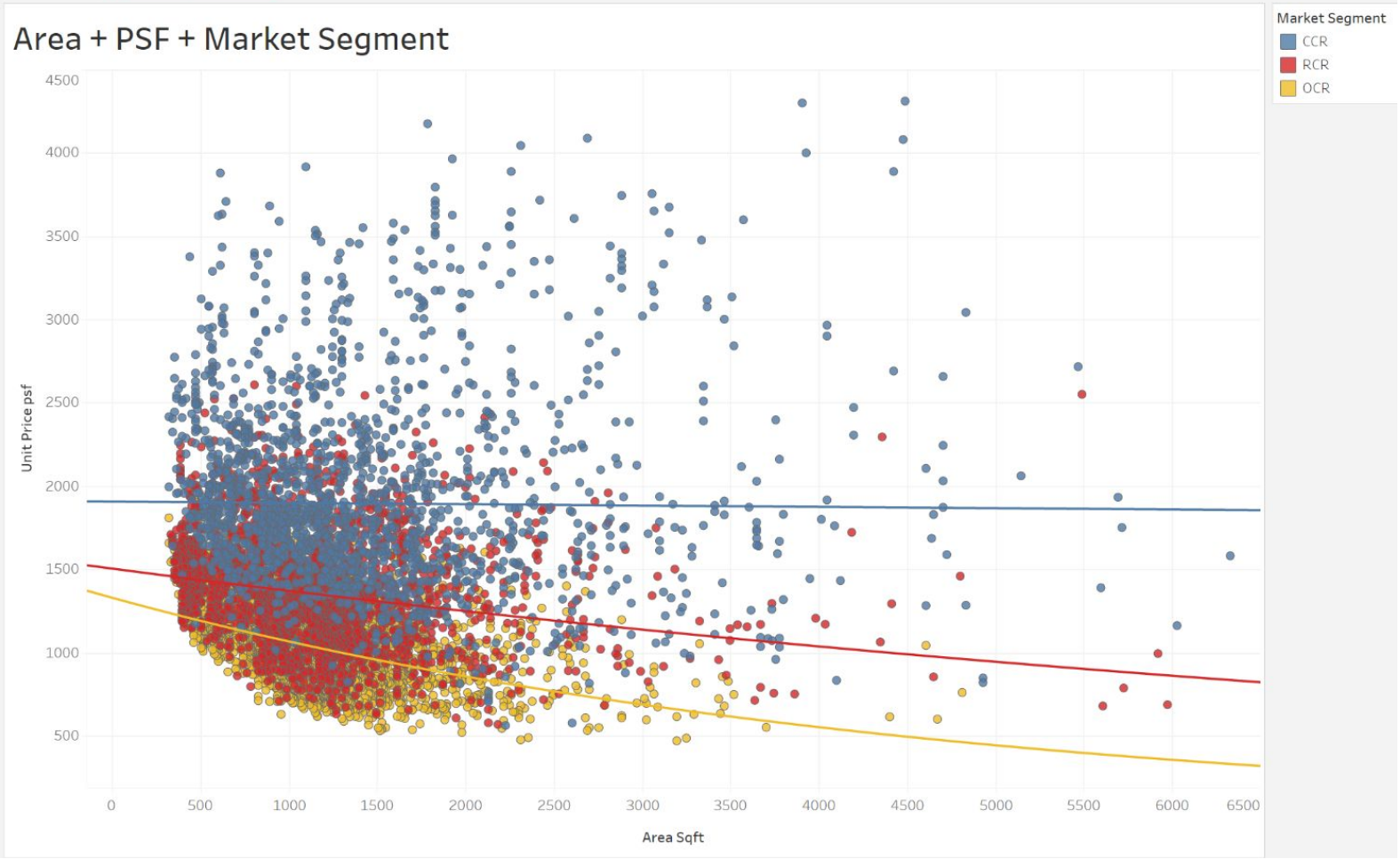


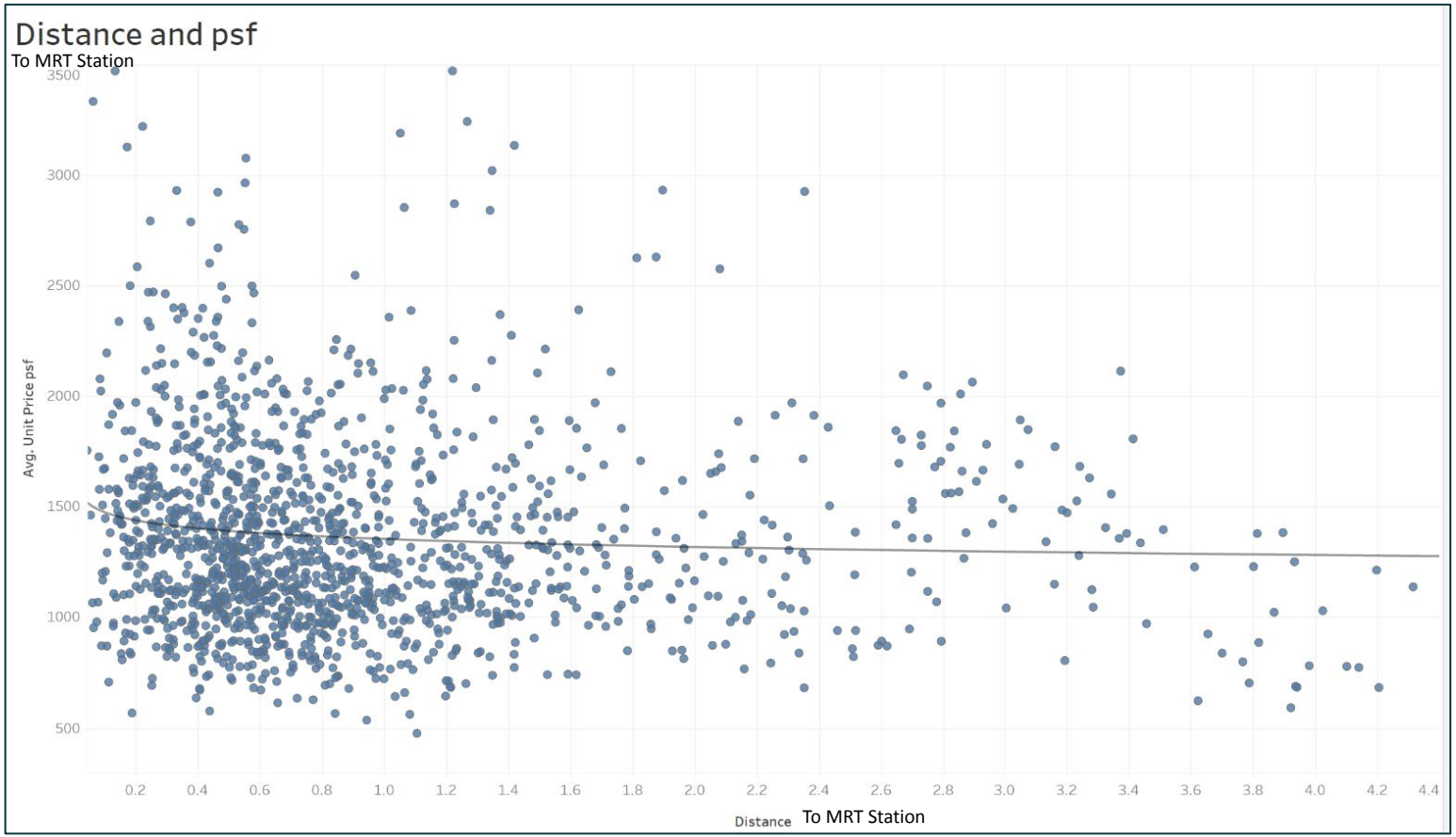


Does price of properties located nearer to CBD area rise faster ?

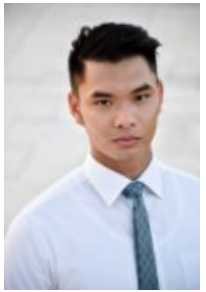


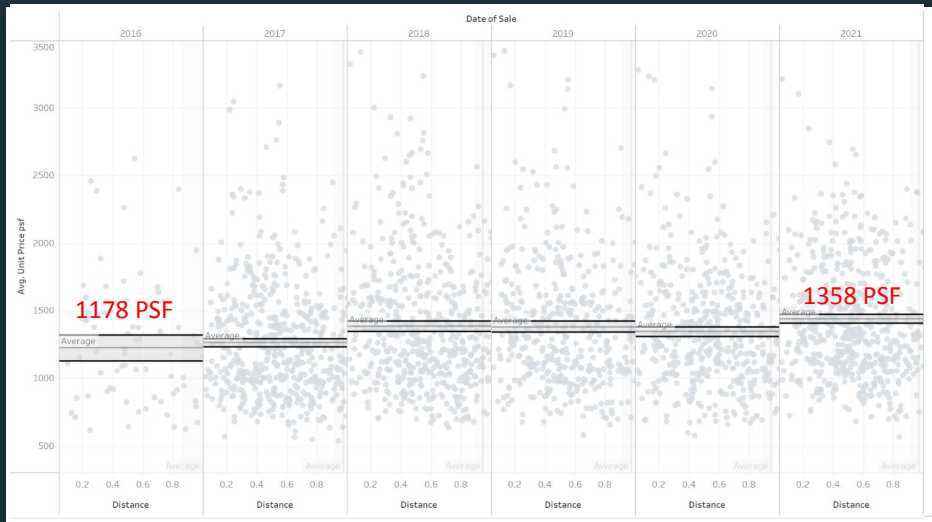
Further analysis on Market Segment





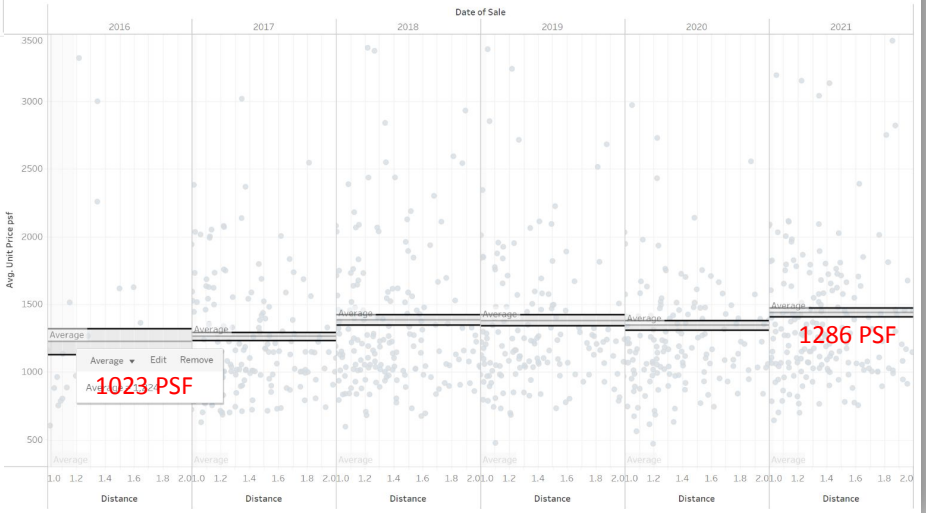
Should I choose a unit near MRT station?



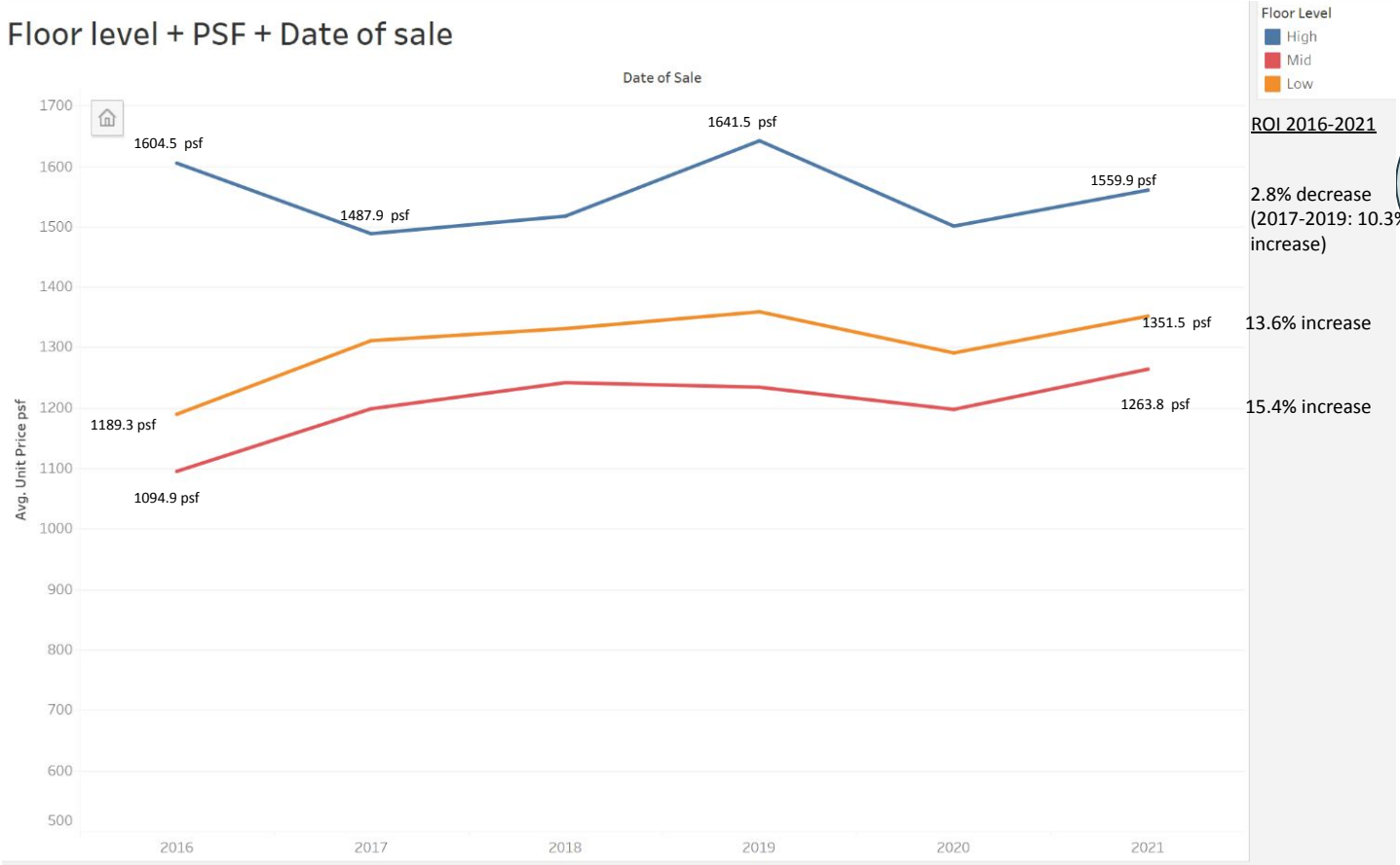


Distance within 1km from nearest MRT station	
From XXXX to 2021	Median PSF increase (%)
2016	15.3
2017	11.2
2018	3.4
2019	2.6
2020	3.0

Distance within 1km to 2km from nearest MRT station	
From XXXX to 2021	Median PSF increase (%)
2016	25.7
2017	13.8
2018	5.3
2019	6.4
2020	8.7



Floor level + PSF + Date of sale



ROI 2016-2021

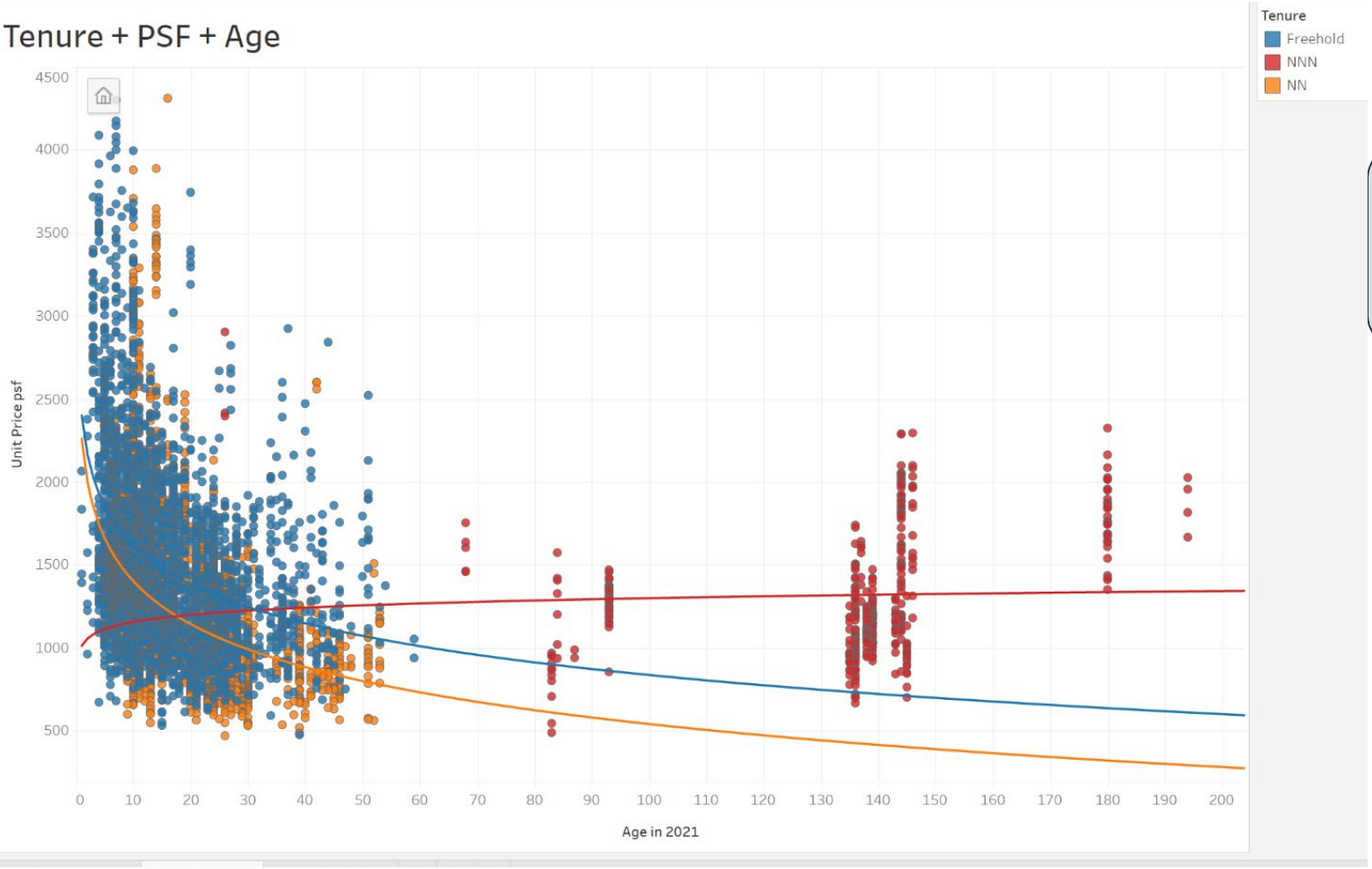
2.8% decrease (2017-2019: 10.3% increase)

13.6% increase

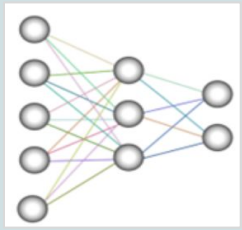
15.4% increase

Does higher floor level helps in property value appreciation?

Tenure + PSF + Age

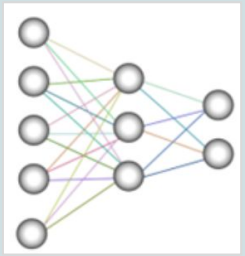


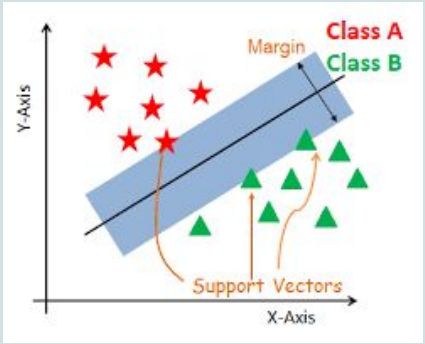
How fast does older units depreciate? Can freehold units hold its value?



Training Data Model

Modeling





- **Predictive modeling** is the process by which a model is created to predict an outcome.
  - If the outcome is **categorical** it is called **classification**
  - If the outcome is **numerical** it is called **regression**

- **Regression** is a data science task of predicting the value of **target** variable (**numerical**) by building a model based on one or more **feature** variables (**numerical** or **categorical**).

- We use **regression modelling** to predict the condominium resale price

Evaluation of Models for Regression

Regression models can be evaluated by many different criterias. Examples:

01	Root Mean Squared Error (RMSE)	• RMSE is a popular formula, but can only be compared between models whose errors are measured in the same units.
02	Relative Squared Error (RSE)	• Unlike RMSE, the relative squared error (RSE) can be compared between models whose errors are measured in the different units.
03	Mean Absolute Error (MAE)	• The mean absolute error (MAE) has the same unit as the original data, and it can only be compared between models whose errors are measured in the same units. It is usually similar in magnitude to RMSE, but slightly smaller.
04	Relative Absolute Error (RAE)	• Like RSE , the relative absolute error (RAE) can be compared between models whose errors are measured in the different units.
05	Coefficient of Determination (R2)	• The coefficient of determination (R2) summarizes the explanatory power of the regression model and is computed from the sums-of-squares terms.

- **RMSE** and **R2** are chosen

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

a = actual target

p = predicted target

Coefficient of Determination → $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Sum of Squares Total → $SST = \sum(y - \bar{y})^2$

Sum of Squares Regression → $SSR = \sum(y' - \bar{y'})^2$

Sum of Squares Error → $SSE = \sum(y - y')^2$

- R2 is computed from the sums-of-squares terms
- If the SSE is zero, then R2 is 1, and the regression model is “perfect”. If SE is equal to SST, then R2 is zero, the regression model is a total failure

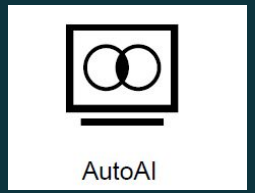
There are many regression algorithms available

Decision Tree
Multiple Linear regression (MLR)
K nearest neighbors
Artificial neural network (ANN)
Support Vector Machine (SVM)

Which one to use?
Watson's AutoAI is the solution.
It runs a several machine learning models and help the non-specialist choose the top performing one

AutoAI

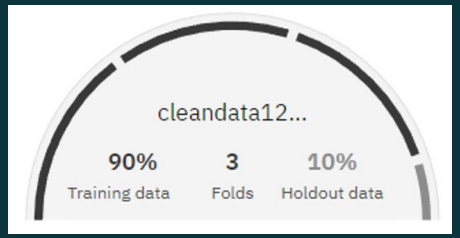
AutoAI is a variation of Automated Machine Learning (AutoML), while AutoML is the process of automating the manual tasks that data scientists must complete as they build and train machine learning models.



Prevent overfitting

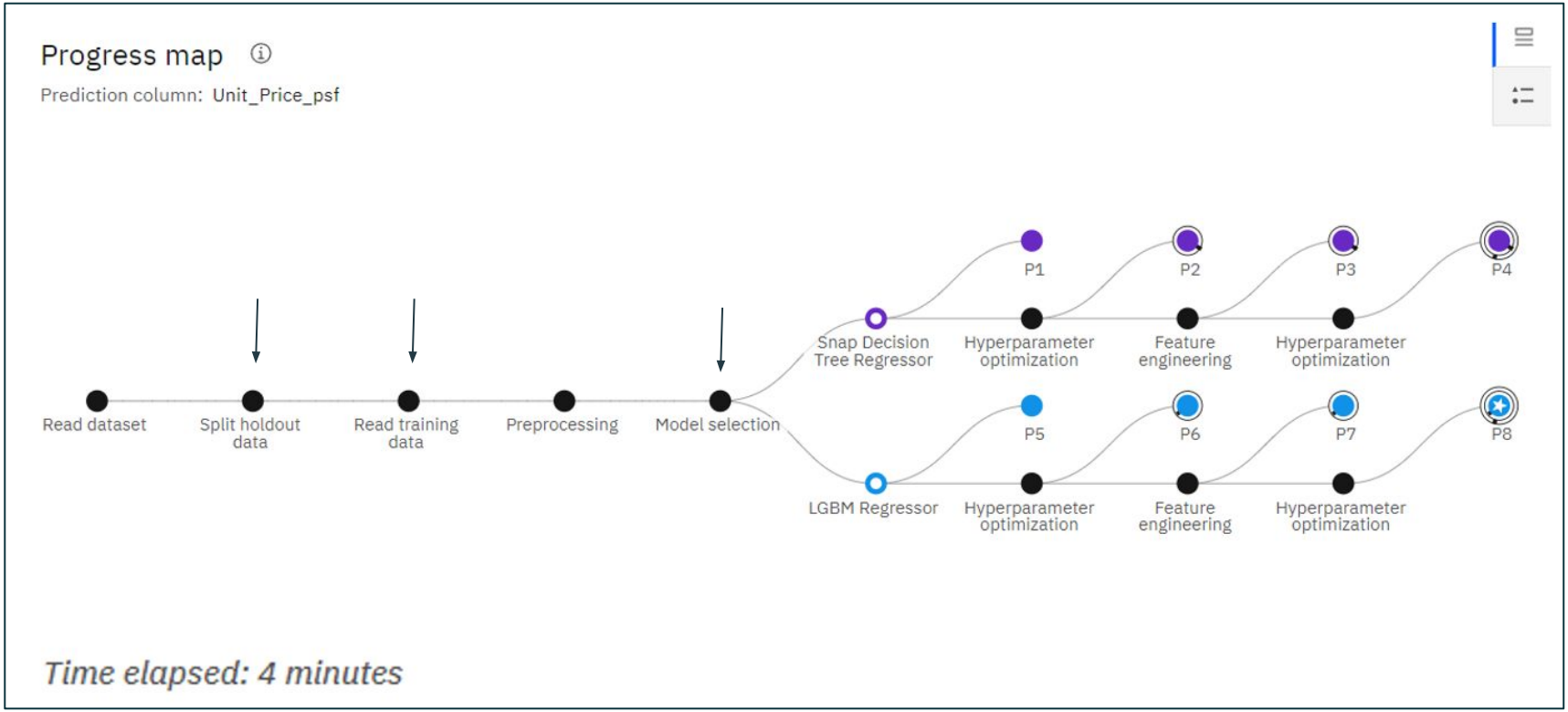
If the model works very well using training data, but badly on new data, that means overfitting has been committed.

To prevent the dangers of overfitting, Watson picks 90% of the data for training and 10% for holdout or testing



Specification of AutoAI Experiment

- Target variable: “**Unit_Price-psf**”
- Feature variables: Distance to MRT, Completion Year, Date of Sale, Postal District, Market Segment, Tenure, Floor level and Type (Condominium / Apartment)
- Prediction type: **Regression**
- Optimized metric: Root mean squared error (RMSE)


Progress map
Prediction column: Unit_Price_psf
Read dataset
Split holdout data
Read training data
Preprocessing
Model selection
Snap Decision Tree Regressor
Hyperparameter optimization
Feature engineering
Hyperparameter optimization
P1
P2
P3
P4
LGBM Regressor
Hyperparameter optimization
Feature engineering
Hyperparameter optimization
P5
P6
P7
P8
Time elapsed: 4 minutes

Pipeline leaderboard

	Rank	Name	Algorithm	RMSE (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 8	LGBM Regressor	215.771	HPO-1 FE HPO-2	00:00:31
	2	Pipeline 7	LGBM Regressor	224.648	HPO-1 FE	00:00:33
	3	Pipeline 5	LGBM Regressor	233.014	*None*	00:00:01
	4	Pipeline 6	LGBM Regressor	233.014	HPO-1	00:00:07

- **"LGBM Regressor"** model is chosen by Watson as the best-performing algorithm. It is used in the top 4 ranking pipelines
- **Light Gradient Boosting Machine**, is a free and open source distributed gradient boosting framework for machine learning originally developed by Microsoft
- **Gradient boosting** is a machine learning technique used in regression and classification tasks to gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Pipeline details

Pipeline 5

Rank	RMSE (Optimized)	Algorithm	Enhancements
3	239.473 (Holdout)	LGBM Regressor	*None*

Save as

Model viewer

Model information

Feature summary

Evaluation

Model evaluation

Feature summary

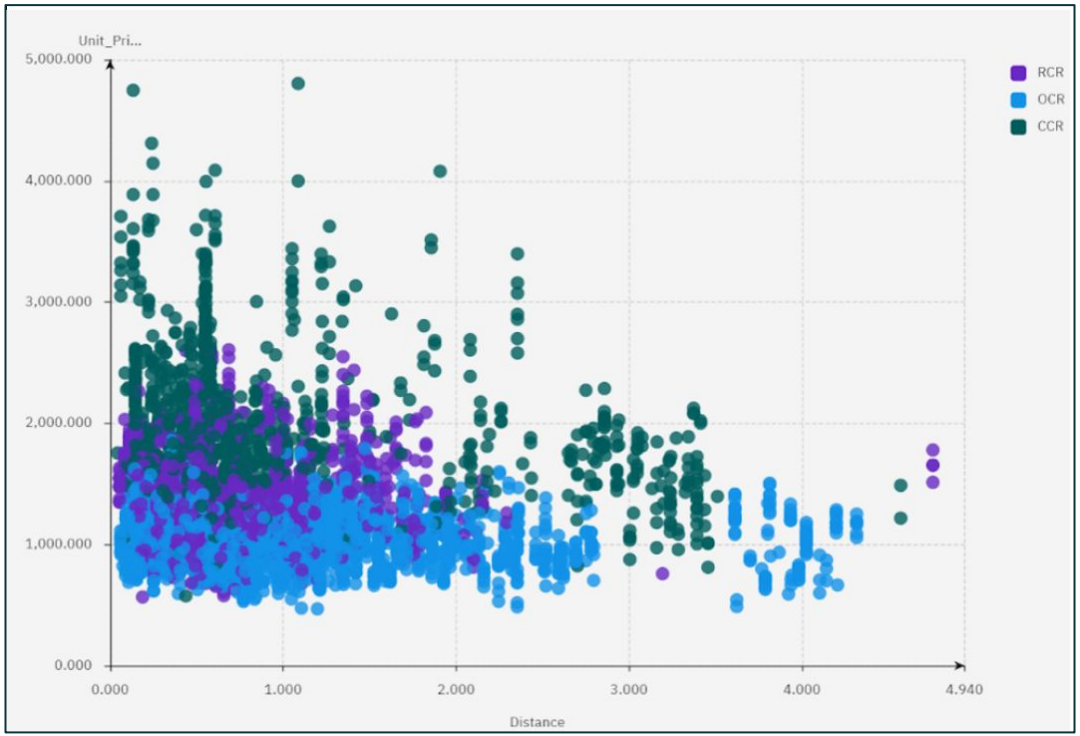
High correlation

All features

Search feature or transformer names

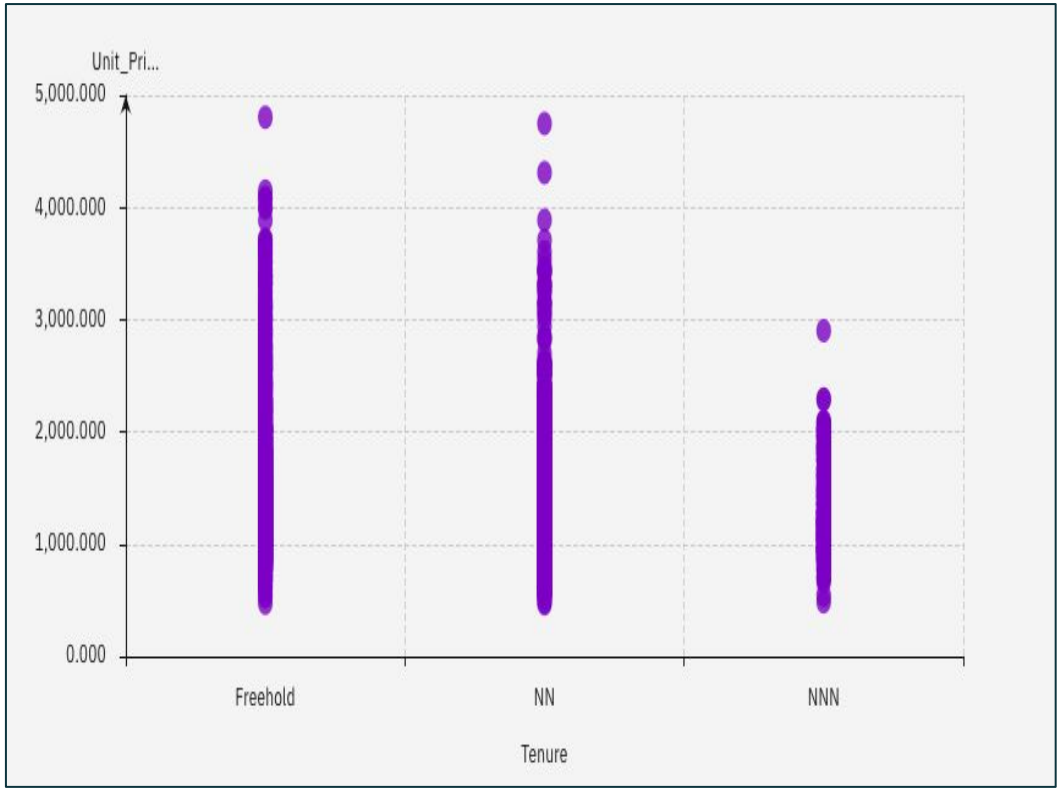
Feature name	Transformation	Feature importance
Distance	None	100.00%
Completion_Year	None	62.00%
Date_of_Sale	None	53.00%
Postal_District	None	41.00%
Market_Segment	None	6.00%
Tenure	None	3.00%
Floor_Level	None	3.00%
Type	None	0.00%

Validation: Scatter Plot of Unit Price vs Distance feature



The nearer to the MRT, the higher the unit price, especially the CCR segment

Validation: Scatter Plot of Unit Price vs Tenure feature

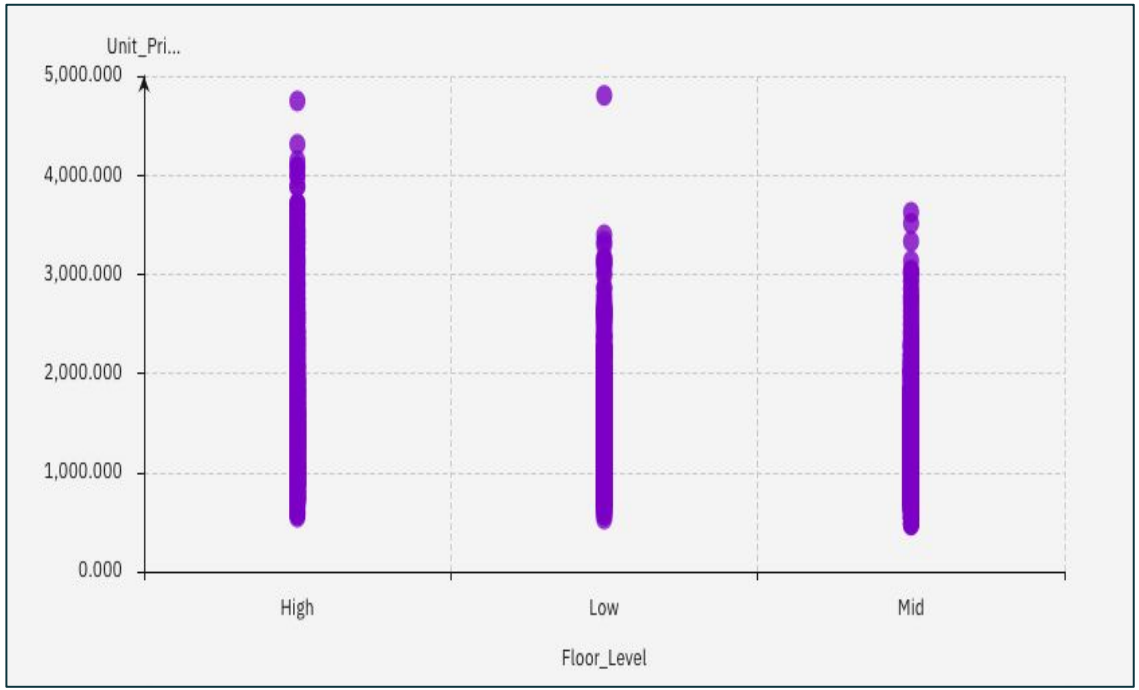


Unit price is the highest for freehold unit, as compared to the other two.

By looking at the average points on the graphs of the other two, the NN tenure is higher than the NNN tenure.

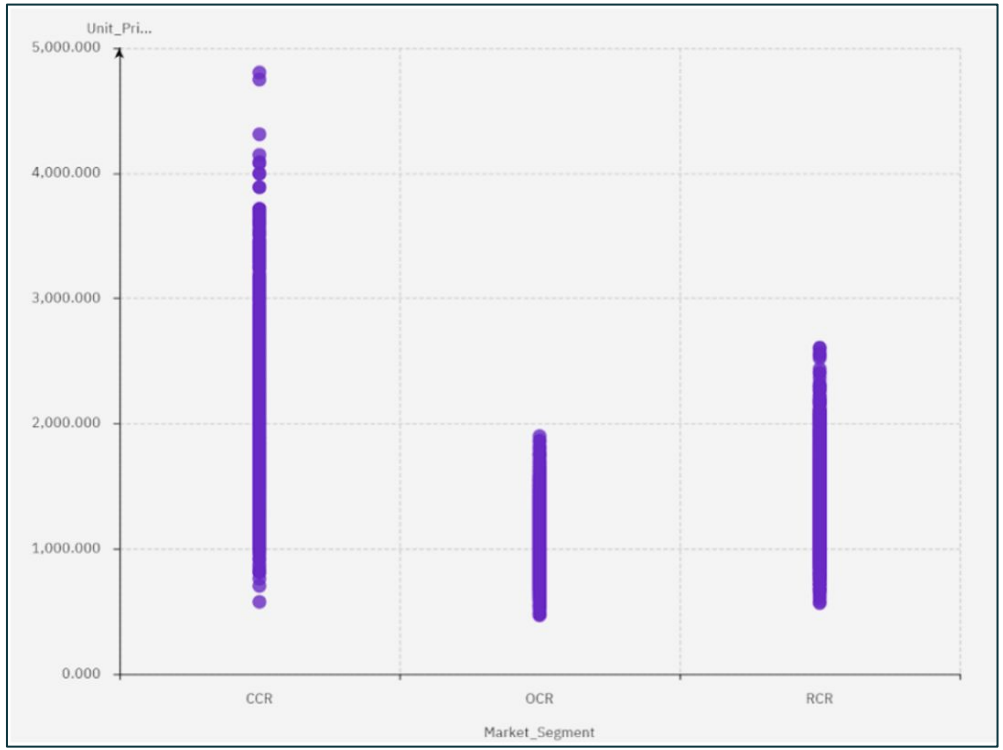
This is probably because the units in the NN are all very new in the data we collected

Validation: Scatter Plot of Unit Price vs Floor level feature



High floor generally fetch higher unit price when compared to low and middle floor levels

Validation: Scatter Plot of Unit Price vs Market segment feature



Unit price is highest in the CCR and lowest in the OCR

Cross Platform Cloud Applications

- The model can be deployed to be used across multiple-platform
- IBM Watson Studio provides direct link to the trained model in the API endpoint
- We manually define and pass the arrays of input value, set the API key, use programming language eg Python to request and get the response in terms of scoring in JSON format

Direct Link

https://us-south.ml.cloud.ibm.com/ml/v4/deployments/71bef014-2b1d-49c1-b4b9-6b0ec1c91ff3/predictions?version=2021-11-29

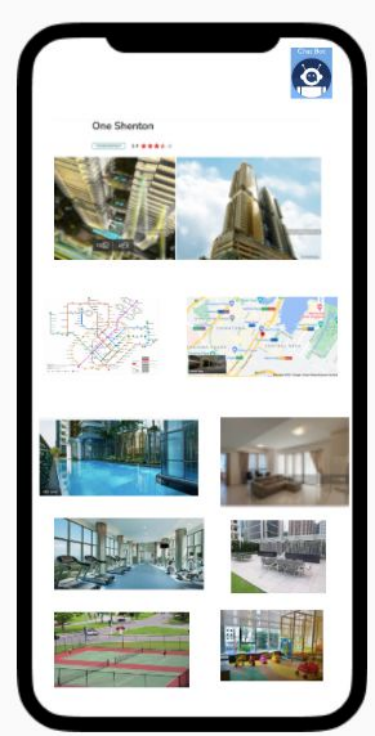
```
import requests

# NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud account.
API_KEY = "<your API key>"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey": API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

# NOTE: manually define and pass the array(s) of values to be scored in the next line
payload_scoring = {"input_data": [{"fields": [array_of_input_fields], "values": [array_of_values_to_be_scored, another_array_of_values_to_be_scored]}]}

response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/71bef014-2b1d-49c1-b4b9-6b0ec1c91ff3/predictions?version=2021-11-29',
json=payload_scoring, headers={'Authorization': 'Bearer ' + mltoken})
print("Scoring response")
print(response_scoring.json())
```

input:

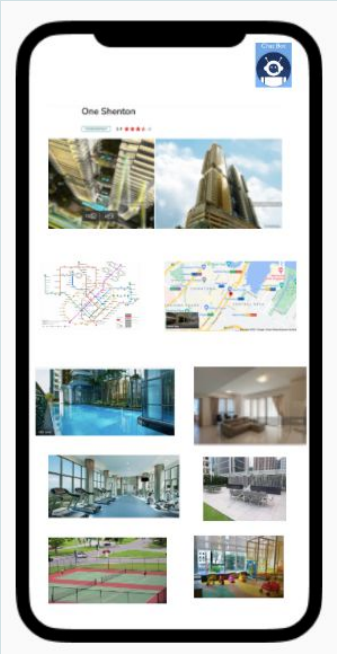
[null, null, Apartment, D15, RCR, Freehold, null, 1001, null, Mid, 2001, 20, null, null]

output:


```
{
    "predictions": [
        {
            "fields": [
                "prediction"
            ],
            "values": [
                [
                    1242.5463846886444
                ]
            ]
        }
    ]
}
```

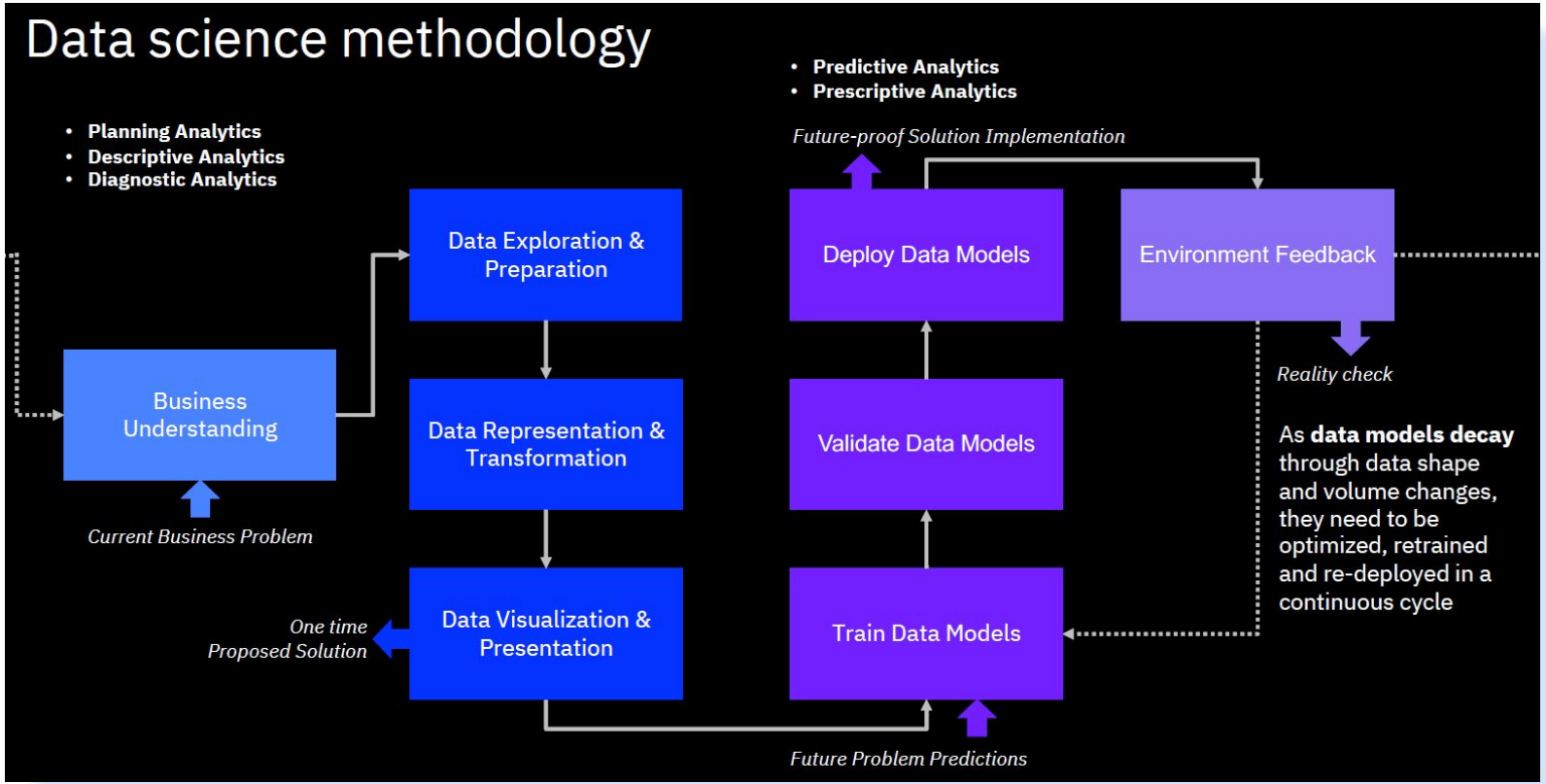



Environmental Feedback



- The API gateway can be used to measure usage statistics. In-app and social media feedbacks can also be obtained
- By collecting results from the implemented model, feedback on the model's performance and observation on how it affects its deployment environment can be obtained
- Based on the analysis of this feedback, the model can be refined, increasing its accuracy and thus its usefulness

Data science methodology




Thank you!