A
HDB Resale Flat Price Prediction Model
(Data Science Final Project – IMVAI 1901)

Apr 2022

Group 1

Kho Liang Heng

Goh Qing Wen

Chew Jun Kang

Liew Kah Khong

Karthika Muruganandam

Koo Phey Shi

Agenda

	Topics	Member	Time Allocation
I	Introduction	Kho Liang Heng	9 min
II	Planning Analytics		
III	Descriptive Analytics	Goh Qing Wen	10 min
IV	Diagnostic Analytics	Liew Kah Khong	10 min
		Chew Jun Kang	10 min
V	Predictive Analytics	Pan Koo Phey Shi	5 min
		Karthika	10 min
VI	Model Test Case	Pan Koo Phey Shi	5 min
VII	Summary	Kho Liang Heng	1 min

I. Introduction

Background: Singapore Property Market Outlook 2022

...Property prices to continue rising for most of 2022.

With another estimate 31,000 HDB flats coming off their MOP in 2022...

...The "MOP effect' may only start driving in 2023 when the number of MOP flats will be around 15,700.

https://www.propertyguru.com.sg/property-guides/property-market-outlook-pmo-2022-58272

Conclusion

It has been about two years since COVID-19 shook the world, and in 2021, Singapore tried its best to pivot from managing the coronavirus as a pandemic to an endemic.

Although the country was not fully able to do so due to the Delta and Omicron variants, many Singapore residents have found a way to continue life as usual. We will probably still grapple with safe management measures in our daily lives in 2022. However, given how well the majority has adapted this year, it likely won't matter to the property market, which is expected to continue thriving.

> "Given the demand observed and the supply contraction, we believe that home prices will continue to be robust."
>
> **Dr. Tan Tee Khoon**
> Country Manager
> PropertyGuru Singapore

With the next wave of 31,000 HDB flats reaching MOP in 2022, the domestic buying frenzy will most likely continue. If foreign buyers are granted easier access to Singapore via VTLs, we expect their increased participation to fuel the market as well. Furthermore, mortgage rates are expected to remain in the buyers' favour for most of 2022.

Property sellers and developers are aligned as well – since the last quarter of the year, many developers have grown more confident and started raising their prices by 1% to 4%. Yet, properties continue to move quickly.

With these in mind, we look forward to a solid and resilient Singapore property market in 2022.

PropertyGuru

… developers have grown more confident and started raising their prices by 1% to 4%. Yet, properties continue to move quickly.

https://www.propertyguru.com.sg/property-guides/property-market-outlook-pmo-2022-58272

Who are we

We are a group of **freelance property agents** and we have a lot of potential customers on hands, sitting on the side bench and waiting for the right moment to plunge into the resale HDB property market.

We have built our credential over the years through our high quality customer service and “Trust” has been the most Critical Success Factor. It is our utmost duty to provide them with the most updated and relevant information for our customer to make his once in a life time decision making. In return, the happy customer would reference friends and relatives.

From the above report, the market outlook for year 2022 is positive and we are seeing great opportunity.

It is the right timing to embark on our first step to digitalisation with confidence and achieve a competitive edge in the property sector and enable growth.

Our Plan

Having recently graduated from IBM I am Vitalize – AI course, the team decided to put our new acquired knowledge into good use.

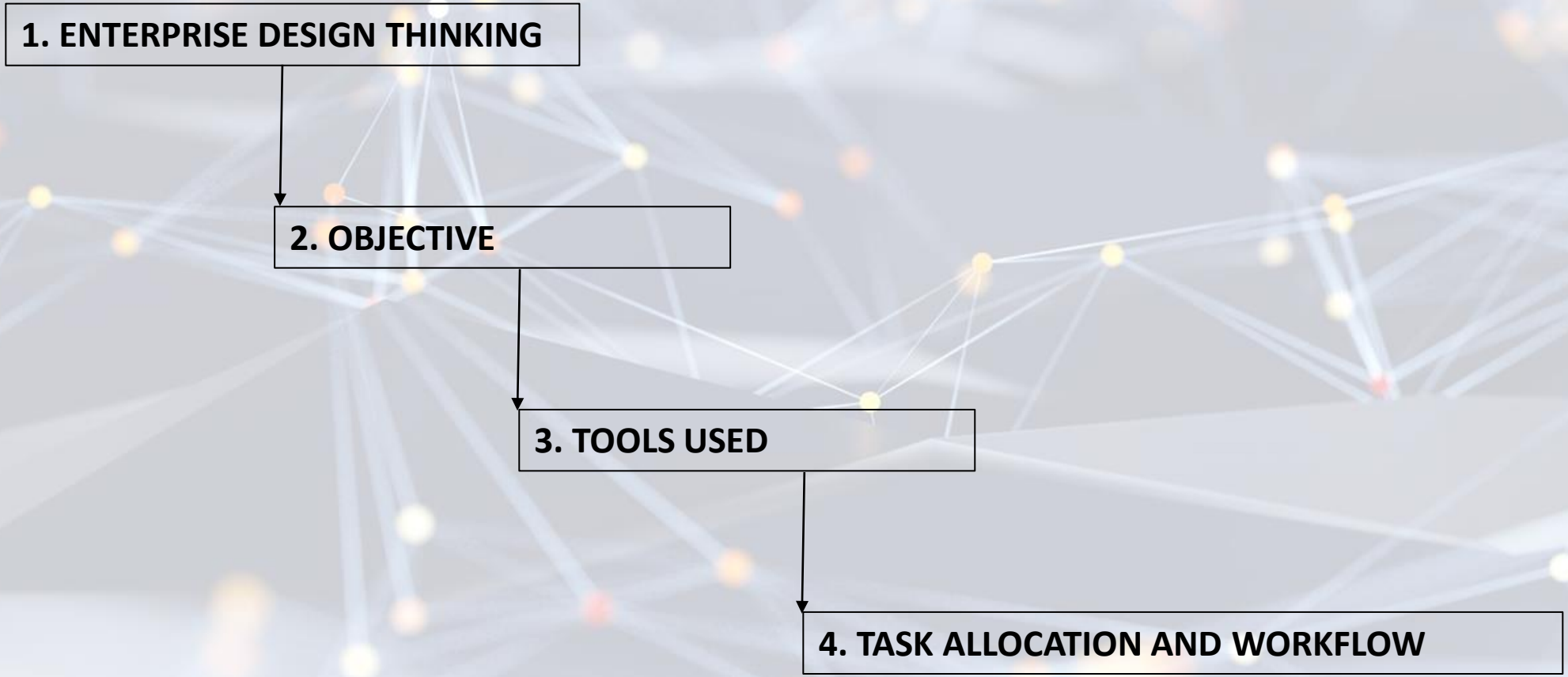
We need to instantly develop a reliable tool to collect buying pattern and predict the price of HDB Resale price

Our winning strategy is “Instant, Insight, Indispensable”

Instant : We provide any HDB Resale flat information instantly through our AI Model
Insight : We provide new invaluable insights of the property with our AI Model
Indispensable : The AI is an indispensable tool to our property agent.

II. Planning Analytics

Planning Analytics Method



1. Enterprise Design Thinking

1 How Might We Statement (Problem Statement)

How might we help HDB resale flat buyers and sellers price their resale flats optimally.

4 Persona

Name Harry De Beng (HDB)

Profile

Age	38
Location	Punggol 2 room flat
Education	Tertiary
Job	Engineer
Family Work experience Technical literacy	Single income family with two young children going to primary school in 3 years time, and another one is expecting early next year. Graduate with a Diploma in IT and has 15 years of experience in the engineering industry as Software Engineer. A family man whom besides work, spend most of his time with family.

4 Persona

Motivations

His wife can walk their children to and from school safely so that time is not wasted and expenses on travel is save up.

This will be his biggest long term investment

To be a proud father and husband whom provide a better home to his wife and children

To be near parent home so that the family could visit and take care of each other more conveniently

His wife can easily access to a nearby wet market for fresh product early in the morning

To set up a nice and cosy home office for himself.

Goals

To buy a 3 or 4 room resale flat near the children primary school and his parent house as well as other amenities within this year

To sell his unit at above market price

To buy his unit at below market price

Needs

Need to upgrade to a bigger unit with the increase in family size

6 Pain Points

Harry De Beng do not know what is the market price of his interested resale flat, and do not know if it is a right timing to make the purchase.

He also do not know what price to accept during price negotiation, and wonder if he should wait for the price to go up further.

7 Reframe the HMW Statement

Reframed HOW MIGHT WE STATEMENT:

How might we help Harry De Beng to identify and evaluate the factors that affects flat prices, so that they can price their flats accordingly.

11 Write a Hill

Harry De Beng is able to effectively price his existing flat to sell at an optimal price and also buy another resale HDB flat at a fair market price with the AI Model generated tool for HDB price predictions with up to 90% accuracy.

Hypothesis :

- Distance to the market / hawker center will have a higher impact on HDB flat resale prices than distance to the MRT.
- HDB prices are positively affected with nearer accessibility to a MRT station
- HDB resale prices are negatively correlated to the age of the property and proximity to a hospitals.

Master Prototype - Iteration 2

Reference : Resale@SG App

2. PROJECT OBJECTIVE

2. Project Objective

Harry De Beng is able to effectively price his existing flat to sell at an optimal price and also buy another resale HDB flat at a fair market price with the AI generated tool for HDB price predictions with up to 90% accuracy

There are 3 **hypotheses** :

Distance to the market / hawker centre will have a higher impact on HDB flat resale prices than distance to the MRT.

HDB prices are positively affected with nearer accessibility to a MRT station

HDB resale prices are negatively correlated to the age of the property and proximity to a hospitals.

3. TOOL USED

3. Tools used

- **Zoom, Google Meet** – for remote Meeting
- **Mural** – for collaboration and Enterprise Design Thinking process
- **Excel and Power Query** – For data collection and cleaning
- **Jupyter Notebook** (Anaconda, Kaggle)– for data refinery and data exploration
- **PowerBI** – for Data refinery, visualization
- **Tableau Prep Builder and Tableau** – for Data refinery, visualization
- **Orange** – for visualization , Machine Learning
- **IBM Watson Studio** – for Data Refinery, visualization, **SPSS**, **Auto AI**
- **OneMap API** – for Latitude & Longitude extraction

- Python Library & Modules
  - **Pandas** – For data manipulation and analysis
  - **Numpy** – For mathematical functions, multi-dimensional arrays and matrices
  - **Seaborn** – For data visualization
  - **Matplotlib** – For data visualization
  - **Scikit-learn** – For predictive data analysis
  - **Glob** – For pathnames specifications
  - **Json** – For working with JSON data.
  - **Request** – For HTTP request
  - **Geopy** – For Geographical coordinates locations and for measuring geographic distance

4. TASK ALLOCATION

4. Task Allocation & Workflow

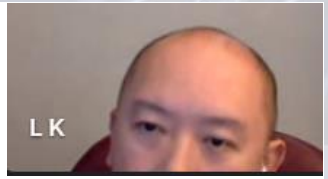
All
- Data Acquisition

Data Engineer
(Goh Qing Wen)
- Data cleansing
- Data Preparation

Deploy Data Models (Optional)
-

Business Owner
(Kho Liang Heng)
- Biz Understanding
- EDT

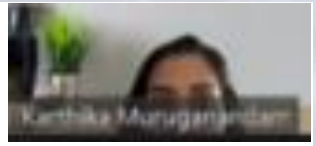
Data Analyst 1
(Liew Kah Khong)
- Data Exploration
- Data Visualisation



Data Scientist 2
(Pan Koo Phey Shi)
- Validate
- Train & Test Models

Data Analyst 2
(Chew Jun Kang)
- Data Exploration
- Data Visualisation

Data Scientist 1
(Karthika)
- Train Data Models



III. Descriptive Analytics

Data Acquisition

- Google search on:
  - HDB resale price data
  - HDB resale price prediction
    - Towards Data Science - Predict the Selling Price of HDB Resale Flats by Kok Jim Meng
    - GitHub
    - Kaggle - Drivers of HDB Resale Price and Prediction by Lau Teyang
- Data.gov.sg

*Sources URLs are at the end of the slides.

Data Acquisition

- Data sets used in the mentioned articles also used Data.gov.sg HDB data sets.
  - Total of 6 .csv files, data dating from 1990 to 2022
    - 11 columns* - Data from 1990 to 2014 only has 10 columns. (Missing remaining_lease column)
    - 870,092 rows of data in total.

	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	2015-01	ANG MO KIO	3 ROOM	174	ANG MO KIO AVE 4	07 TO 09	60	Improved	1986	70	255000
3	2015-01	ANG MO KIO	3 ROOM	541	ANG MO KIO AVE 10	01 TO 03	68	New Generation	1981	65	275000
4	2015-01	ANG MO KIO	3 ROOM	163	ANG MO KIO AVE 4	01 TO 03	69	New Generation	1980	64	285000
5	2015-01	ANG MO KIO	3 ROOM	446	ANG MO KIO AVE 10	01 TO 03	68	New Generation	1979	63	290000
6	2015-01	ANG MO KIO	3 ROOM	557	ANG MO KIO AVE 10	07 TO 09	68	New Generation	1980	64	290000

Data Acquisition & Cleaning

```
[1]: #Loading in the required environments
     import glob
     import pandas as pd
     import json
     import requests

[2]: price1999 = pd.read_csv('./data/resale-flat-prices-based-on-approval-date-1990-1999.csv')
     price2006 = pd.read_csv('./data/resale-flat-prices-based-on-approval-date-2000-2006.csv')
     price2012 = pd.read_csv('./data/resale-flat-prices-based-on-approval-date-2007-feb-2012.csv')
     price2014 = pd.read_csv('./data/resale-flat-prices-based-on-registration-date-from-mar-2012-to-dec-2014.csv')
     price2016 = pd.read_csv('./data/resale-flat-prices-based-on-registration-date-from-jan-2015-to-dec-2016.csv')
     price2017 = pd.read_csv('./data/resale-flat-prices-based-on-registration-date-from-jan-2017-onwards.csv')

[ ]: price1999.head()

[ ]: price2006.head()

[ ]: price2012.head()

[ ]: price2014.head()

[ ]: price2016.head()

[ ]: price2017.head()

[ ]: #Combing all datafiles together
     df = pd.concat([pd.read_csv(f) for f in glob.glob("./data/*.csv")], ignore_index=True)
     df.head()

[10]: df.isnull().sum()

[10]: month                   0
      town                    0
      flat_type               0
      block                   0
      street_name             0
      storey_range            0
      floor_area_sqm          0
      flat_model              0
      lease_commence_date     0
      resale_price            0
      remaining_lease        709050
      dtype: int64
```

```
[11]: df = df.dropna() # drop remaining_lease missing values
      df.isnull().sum()

[11]: month                 0
      town                  0
      flat_type             0
      block                 0
      street_name           0
      storey_range          0
      floor_area_sqm        0
      flat_model            0
      lease_commence_date   0
      resale_price          0
      remaining_lease       0
      dtype: int64
```

- Remaining rows of data – 161,042
- Our working data set size.
  - (2015 – 2022)

Data Acquisition & Cleaning

- Article from Kok Jim Meng on Towards Data Science, explored to find relationship between resale flat prices and distances to MRT and CBD areas.
  - Inspired us to look at more of such factors;
    - Distances to Market/ Hawker centres,
    - Schools (Primary, Secondary, Tertiary)
    - Parks,
    - Hospitals (Public / Govt),
    - Facilities (Stadium, Sports Centres, Swimming pools, Community Centres),

Data Acquisition & Cleaning

- Revised the list of MRT stations used in his Jupyter Notebook code.
  - Updated with the new Thomson-East Coast Line Stations.
  - 188 Stations, including LRT Stations.

- Gather info on the list of:
  - Market/ Hawker centres – 106
  - Schools (Primary, Secondary, Tertiary) – 359
  - Parks – 75
  - Hospitals (Public / Govt) – 51
  - Facilities (Stadium, Sports Centres, Swimming pools, Community Centres) - 116

*Sources URLs are at the end of the slides.

Data Acquisition & Cleaning

- Retrieving Latitude and Longitude of the Amenities. (MRT/ Market/ Schools…)

[12]:
```
list_of_mrt = [
    'Admiralty MRT Station',
'Aljunied MRT Station',
'Ang Mo Kio MRT Station',
'Bakau  LRT  Station',
'Bangkit  LRT  Station',
'Bartley MRT Station',
'Bayfront MRT Station',
'Bayshore MRT Station',
'Beauty World MRT Station',
'Bedok MRT Station',
'Bedok North MRT Station',
'Bedok Reservoir MRT Station',
'Tuas West Road MRT Station',
'Ubi MRT Station',
'Upper Changi MRT Station',
'Upper Thomson MRT Station',
'Woodlands MRT Station',
'Woodlands North MRT Station',
'Woodlands South MRT Station',
'Woodleigh MRT Station',
'Yew Tee MRT Station',
'Yio Chu Kang MRT Station',
'Yishun MRT Station'
]
print (len(list_of_mrt))
```

```
188
```

[13]:
```
mrt_lat = []
mrt_long = []

requests.packages.urllib3.util.connection.HAS_IPV6 = False

for i in range(0, len(list_of_mrt)):
    query_address = list_of_mrt[i]
    query_string = 'https://developers.onemap.sg/commonapi/search?searchVal='+str(query_address)+'&returnGeom=Y&getAddrDetails=Y'

    #print (query_string)
    headers = {"User-Agent": "Mozilla/5.0 (X11; CrOS x86_64 12871.102.0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.141 Safari/537.36"}
    resp = requests.get(query_string)
    data_mrt=json.loads(resp.content)

    if data_mrt['found'] != 0:
        mrt_lat.append(data_mrt["results"][0]["LATITUDE"])
        mrt_long.append(data_mrt["results"][0]["LONGITUDE"])

        print (str(query_address)+",Lat: "+data_mrt['results'][0]['LATITUDE'] +" Long: "+data_mrt['results'][0]['LONGITUDE'])

    else:
        mrt_lat.append('NotFound')
        mrt_lat.append('NotFound')
        print ("No Results")
```

```
Admiralty MRT Station,Lat: 1.44058856161847 Long: 103.800990519771
Aljunied MRT Station,Lat: 1.3164326118157 Long: 103.882906044385
Ang Mo Kio MRT Station,Lat: 1.36993284962264 Long: 103.849558091776
Bakau  LRT  Station,Lat: 1.38799431054768 Long: 103.905415300171
Bangkit  LRT  Station,Lat: 1.38002223010088 Long: 103.772647370452
Bartley MRT Station,Lat: 1.34250117805245 Long: 103.880177899184
Bayfront MRT Station,Lat: 1.28187378879209 Long: 103.859079764874
Bayshore MRT Station,Lat: 1.31312248299118 Long: 103.942307986645
Beauty World MRT Station,Lat: 1.34090149001429 Long: 103.775746717778
Bedok MRT Station,Lat: 1.32040148729112 Long: 103.957183823478
Bedok North MRT Station,Lat: 1.33474211664091 Long: 103.91797832995
Bedok Reservoir MRT Station,Lat: 1.33660782955099 Long: 103.9322346232
Bedok South MRT Station,Lat: 1.31668400658419 Long: 103.94931120983
Bencoolen MRT Station,Lat: 1.29891843369422 Long: 103.850353762717
```

[14]:
```
mrt_location = pd.DataFrame({
    'MRT': list_of_mrt,
    'latitude': mrt_lat,
    'longitude': mrt_long
})
```

Data Acquisition & Cleaning

- Retrieving Latitude and Longitude of the HDB Addresses.

```
[17]: #Get coordinates for address
latitude = []
longitude = []
blk_no = []
road_name = []
postal_code = []
address = []
count = 0

requests.packages.urllib3.util.connection.HAS_IPV6 = False

for row in range(len(address_list)):
    query_address = address_list[row]
    query_string='https://developers.onemap.sg/commonapi/search?searchVal='+s
    resp = requests.get(query_string)


    data_geo_location=json.loads(resp.content)
    if data_geo_location['found'] != 0:
        latitude.append(data_geo_location['results'][0]['LATITUDE'])
        longitude.append(data_geo_location['results'][0]['LONGITUDE'])
        blk_no.append(data_geo_location['results'][0]['BLK_NO'])
        road_name.append(data_geo_location['results'][0]['ROAD_NAME'])
        postal_code.append(data_geo_location['results'][0]['POSTAL'])
        address.append(query_address)
        print (str(query_address) + " ,Lat: " + data_geo_location['results'][0]['LATITUDE'] +
    else:
        print ("No Results")
```

```
174 ANG MO KIO AVE 4 ,Lat: 1.37509746867904 Long: 103.83761896123
541 ANG MO KIO AVE 10 ,Lat: 1.37392239168826 Long: 103.855621371068
163 ANG MO KIO AVE 4 ,Lat: 1.37354853919927 Long: 103.838176471398
446 ANG MO KIO AVE 10 ,Lat: 1.36776095130953 Long: 103.855357145908
557 ANG MO KIO AVE 10 ,Lat: 1.3716257020332 Long: 103.857736107527
603 ANG MO KIO AVE 5 ,Lat: 1.38020079047279 Long: 103.83575571651
709 ANG MO KIO AVE 8 ,Lat: 1.37113720765377 Long: 103.847662320064
333 ANG MO KIO AVE 1 ,Lat: 1.3613425564061 Long: 103.851698621454
109 ANG MO KIO AVE 4 ,Lat: 1.3700965375834 Long: 103.837687766047
564 ANG MO KIO AVE 3 ,Lat: 1.36984837555524 Long: 103.859404131956
218 ANG MO KIO AVE 1 ,Lat: 1.36511908595886 Long: 103.841742478489
556 ANG MO KIO AVE 10 ,Lat: 1.37203236308683 Long: 103.857625375797
156 ANG MO KIO AVE 4 ,Lat: 1.37549519574964 Long: 103.839947470774
471 ANG MO KIO AVE 10 ,Lat: 1.36346600647245 Long: 103.856702918462
434 ANG MO KIO AVE 10 ,Lat: 1.3678915069194 Long: 103.85345502653
560 ANG MO KIO AVE 10 ,Lat: 1.37081626777721 Long: 103.859192334256
332 ANG MO KIO AVE 1 ,Lat: 1.36167030761833 Long: 103.851955453764
421 ANG MO KIO AVE 10 ,Lat: 1.36538422642267 Long: 103.852966961527
506 ANG MO KIO AVE 8 ,Lat: 1.37440041920745 Long: 103.84893617353
```

```
[18]: df_coordinates = pd.DataFrame({
    'latitude': latitude,
    'longitude': longitude,
    'blk_no': blk_no,
    'road_name': road_name,
    'postal_code': postal_code,
    'address': address
})
len(df_coordinates)

[18]: 9263
```

Data Acquisition & Cleaning

- Measuring geographic distances of HDB addresses to CBD, nearest MRT, etc.

```
[19]: list_of_lat = df_coordinates['latitude'
      list_of_long = df_coordinates['longitud
      mrt_lat = mrt_location['latitude']
      mrt_long = mrt_location['longitude']

[20]: list_of_coordinates = []
      list_of_mrt_coordinates = []

      for lat, long in zip(list_of_lat, list_
          list_of_coordinates.append((lat,lon
      for lat, long in zip(mrt_lat, mrt_long)
          list_of_mrt_coordinates.append((lat
```

```
[21]: pip install geopy

Requirement already satisfied: geopy in c:\users\mytip\anaconda3\lib\site-packages (2.2.0)
Requirement already satisfied: geographiclib<2,>=1.49 in c:\users\mytip\anaconda3\lib\site-packages (from geopy) (1.52)
Note: you may need to restart the kernel to use updated packages.
```

```
[22]: # Distance to nearest MRT. Getting the distances via OneMap API:
      from geopy.distance import geodesic

      list_of_dist_mrt = []
      min_dist_mrt = []

      for origin in list_of_coordinates:
          i=i+1
          print("orgin" +str(i))
          for destination in range(0, len(list_of_mrt_coordinates)):
              list_of_dist_mrt.append(geodesic(origin,list_of_mrt_coordinates[dest
          shortest = (min(list_of_dist_mrt))
          min_dist_mrt.append(shortest)
          list_of_dist_mrt.clear()
```

```
          min_dist_mrt.append(shortest)
          list_of_dist_mrt.clear()

orgin188
orgin189
orgin190
orgin191
orgin192
orgin193
orgin194
orgin195
orgin196
orgin197
orgin198
orgin199
orgin200
```

```
[23]: # Distance from CDB
      cbd_dist = []

      for origin in list_of_coordinates:
          cbd_dist.append(geodesic(origin,(1.2830, 103.8513)).meters) #CBD coordinates

[24]: # Put MRT and CBD distance together
      df_coordinates['cbd_dist'] = cbd_dist
      df_coordinates['min_dist_mrt'] = min_dist_mrt
```

Data Acquisition & Cleaning

```
[25]: df_coordinates
```

[25]:

	latitude	longitude	blk_no	road_name	postal_code	address	cbd_dist	min_dist_mrt
0	1.37509746867904	103.83761896123	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVE 4	10296.855747	418.541691
1	1.37392239168826	103.855621371068	541	ANG MO KIO AVENUE 10	560541	541 ANG MO KIO AVE 10	10065.227934	806.176499
2	1.37354853919927	103.838176471398	163	ANG MO KIO AVENUE 4	560163	163 ANG MO KIO AVE 4	10118.355593	291.928107
3	1.36776095130953	103.855357145908	446	ANG MO KIO AVENUE 10	560446	446 ANG MO KIO AVE 10	9383.300907	688.600961
4	1.3716257020332	103.857736107527	557	ANG MO KIO AVENUE 10	560557	557 ANG MO KIO AVE 10	9825.917255	929.164159
...	...	...	...	...	...	...	...	...
9258	1.41604013194338	103.843489405235	513A	YISHUN STREET 51	761513	513A YISHUN ST 51	14736.556279	1178.950136
9259	1.36886409962902	103.83395420063	260B	ANG MO KIO STREET 21	562260	260B ANG MO KIO ST 21	9688.670875	408.893418
9260	1.31844042901657	103.899852177666	410	EUNOS ROAD 5	400410	410 EUNOS RD 5	6674.885834	403.758477
9261	1.3100634072092	103.784498087029	32	GHIM MOH LINK	271032	32 GHIM MOH LINK	8014.143708	668.855750
9262	1.41371070675887	103.839528173264	509A	YISHUN AVENUE 4	761509	509A YISHUN AVE 4	14512.574309	834.245910

9263 rows × 8 columns

```
[26]: df_coordinates.to_csv('df_coordinates.csv',index=False)

[27]: df_coordinates = pd.read_csv('df_coordinates.csv')

[28]: df_new = df_coordinates.merge(df, on="address", how='outer')
      df_new
```

Data Acquisition & Cleaning

	latitude	longitude	blk_no	road_name	postal_code	address	cbd_dist	min_dist_mrt	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price	remaining_lease
0	1.375097	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVE 4	10296.855747	418.541691	2015-01	ANG MO KIO	3 ROOM	174	ANG MO KIO AVE 4	07 TO 09	60.0	Improved	1986	255000.0	70
1	1.375097	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVE 4	10296.855747	418.541691	2015-12	ANG MO KIO	3 ROOM	174	ANG MO KIO AVE 4	10 TO 12	60.0	Improved	1986	275000.0	69
2	1.375097	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVE 4	10296.855747	418.541691	2016-05	ANG MO KIO	3 ROOM	174	ANG MO KIO AVE 4	04 TO 06	69.0	Improved	1986	310000.0	68
3	1.375097	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVE 4	10296.855747	418.541691	2016-06	ANG MO KIO	2 ROOM	174	ANG MO KIO AVE 4	07 TO 09	45.0	Improved	1986	253000.0	68
4	1.375097	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVE 4	10296.855747	418.541691	2016-11	ANG MO KIO	3 ROOM	174	ANG MO KIO AVE 4	04 TO 06	61.0	Improved	1986	290000.0	68
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
161037	1.318440	103.899852	410	EUNOS ROAD 5	400410	410 EUNOS RD 5	6674.885834	403.758477	2022-04	GEYLANG	3 ROOM	410	EUNOS RD 5	04 TO 06	74.0	Model A	1984	405000.0	61 years 07 months
161038	1.310063	103.784498	32	GHIM MOH LINK	271032	32 GHIM MOH LINK	8014.143708	668.855750	2022-04	QUEENSTOWN	3 ROOM	32	GHIM MOH LINK	31 TO 33	68.0	Model A	2018	695000.0	95 years 03 months
161039	1.413711	103.839528	509A	YISHUN AVENUE 4	761509	509A YISHUN AVE 4	14512.574309	834.245910	2022-04	YISHUN	4 ROOM	509A	YISHUN AVE 4	07 TO 09	93.0	Model A	2018	520000.0	95 years 01 month
161040	1.413711	103.839528	509A	YISHUN AVENUE 4	761509	509A YISHUN AVE 4	14512.574309	834.245910	2022-04	YISHUN	4 ROOM	509A	YISHUN AVE 4	07 TO 09	93.0	Model A	2018	525000.0	95 years 01 month
161041	1.413711	103.839528	509A	YISHUN AVENUE 4	761509	509A YISHUN AVE 4	14512.574309	834.245910	2022-04	YISHUN	5 ROOM	509A	YISHUN AVE 4	07 TO 09	113.0	Improved	2018	635000.0	95 years 01 month

161042 rows × 19 columns

Data Acquisition & Cleaning

Rinse and repeat steps for:

- Market / Hawker Centres
- Schools
- Hospitals
- Parks
- Facilities

Data Acquisition & Cleaning

- Issues encountered:
  - OneMap API unable to retrieve coordinates for certain locations.
- Resolved by teamwork!
  - Manually searched and verify coordinates for the missing / “No Results”.

```
In [5]: facil_lat = []
        facil_long = []

        requests.packages.urllib3.util.connection.HAS_IPV6 = False

        for i in range(0, len(list_of_facil)):
            query_address = list_of_facil[i]
            query_string = 'https://developers.onemap.sg/commonapi/search?searchVal='+str(query_address)+'&returnGeom=Y&getAddrDetails=Y'

            #print (query_string)
            headers = {"User-Agent": "Mozilla/5.0 (X11; CrOS x86_64 12871.102.0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.
            resp = requests.get(query_string)
            data_facil=json.loads(resp.content)

            if data_facil['found'] != 0:
                facil_lat.append(data_facil["results"][0]["LATITUDE"])
                facil_long.append(data_facil["results"][0]["LONGITUDE"])

                print (str(query_address)+",Lat: "+data_facil['results'][0]['LATITUDE'] +" Long: "+data_facil['results'][0]['LONGITUDE'])

            else:
                facil_lat.append('NotFound')
                facil_lat.append('NotFound')
                print ("No Results")
```

```
No Results
Farrer Park Tennis Centre,Lat: 1.31171149305997 Long: 103.849
No Results
Geylang East Swimming Complex,Lat: 1.32125731314702 Long: 103
Golazo,Lat: 1.35284351489551 Long: 103.707107979873
No Results
No Results
No Results
No Results
No Results
No Results
Hougang Sports Hall,Lat: 1.37057348859012 Long: 103.888071727
Hougang Stadium,Lat: 1.36950400713381 Long: 103.885987504955
Hougang Swimming Complex,Lat: 1.36991089483793 Long: 103.8870
No Results
No Results
Jalan Besar Sport Centre,Lat: 1.30952190174503 Long: 103.8602
Jalan Besar Stadium,Lat: 1.30965069033363 Long: 103.860086286
Jalan Besar Swimming Complex,Lat: 1.31085490828994 Long: 103.
No Results
```

	A	B
31	'Delta ActiveSG Gym',	No Results
32	'Delta Hockey Pitch',	No Results
33	'Delta Sport Centre',	Delta Sport Centre,Lat: 1.29052498417749 Long: 103.820588112164
34	'Delta Sports Hall',	Delta Sports Hall,Lat: 1.29053485917109 Long: 103.820494844747
35	'Delta Swimming Complex',	Delta Swimming Complex,Lat: 1.28937428705782 Long: 103.820016544794
36	'Enabling Village ActiveSG Gym',	No Results
37	'Farrer Park Tennis Centre',	Farrer Park Tennis Centre,Lat: 1.31171149360745 Long: 103.849500507986
38	'Futsal Arena @ Yishun',	No Results
39	'Geylang East Swimming Complex',	Geylang East Swimming Complex,Lat: 1.32125731314702 Long: 103.8880244762
40	'Golazo',	Golazo,Lat: 1.35284351489551 Long: 103.707107979873
41	'Bedok ActiveSG Gym',	No Results
42	'Bedok ActiveSG Sport Centre',	No Results
43	'Bedok ActiveSG Sports Hall',	No Results
44	'Bedok ActiveSG Swimming Complex',	No Results
45	'Bedok ActiveSG Tennis Centre',	No Results
46	'Hougang ActiveSG Gym',	No Results
47	'Hougang Sports Hall',	Hougang Sports Hall,Lat: 1.37057348859012 Long: 103.888071727225
48	'Hougang Stadium',	Hougang Stadium,Lat: 1.36950400713381 Long: 103.885987504955
49	'Hougang Swimming Complex',	Hougang Swimming Complex,Lat: 1.36991089483793 Long: 103.887094818221
50	'Hyfa',	No Results
51	'Jalan Besar ActiveSG Gym',	No Results
52	'Jalan Besar Sport Centre',	Jalan Besar Sport Centre,Lat: 1.30952190174503 Long: 103.860211336261
53	'Jalan Besar Stadium',	Jalan Besar Stadium,Lat: 1.30965069033363 Long: 103.86008628687
54	'Jalan Besar Swimming Complex',	Jalan Besar Swimming Complex,Lat: 1.31085490828994 Long: 103.859961709542
55	'Jurong East ActiveSG Gym',	No Results
56	'Jurong East Stadium',	Jurong East Stadium,Lat: 1.34672644192914 Long: 103.729456501587
57	'Jurong East Swimming Complex',	Jurong East Swimming Complex,Lat: 1.34662688498665 Long: 103.729184512581
58	Jurong Lake Gardens',	Jurong Lake Gardens,Lat: 1.33348319560371 Long: 103.726807806345
59	Jurong Lake Gardens',	Jurong Lake Gardens,Lat: 1.33348319560371 Long: 103.726807806345

Data Acquisition & Cleaning

- Prepping and Cleaning our data set.
  - Changing data types.
    - “resale price” – float
    - “floor area” – float
    - “lease commence date” – integer
  - Standardizing
    - “Flat Types”,
    - “Flat Model”

```
[29]: df_new['resale_price'] = df_new['resale_price'].astype('float')
      df_new['floor_area_sqm'] = df_new['floor_area_sqm'].astype('float')
      df_new['lease_commence_date'] = df_new['lease_commence_date'].astype('int64')
      df_new['lease_remain_years'] = 99 - (2021 - df_new['lease_commence_date'])

      df_new.dropna(inplace=True)
```

```
[30]: #Data cleaning of flat_type
      df_new['flat_type'].unique()

[30]: array(['3 ROOM', '2 ROOM', '4 ROOM', '5 ROOM', 'EXECUTIVE', '1 ROOM',
             'MULTI-GENERATION'], dtype=object)

[31]: df_new['flat_type'] = df_new['flat_type'].str.replace('MULTI-GENERATION', 'MULTI GENERATION')
      df_new['flat_type'].unique()

[31]: array(['3 ROOM', '2 ROOM', '4 ROOM', '5 ROOM', 'EXECUTIVE', '1 ROOM',
             'MULTI GENERATION'], dtype=object)

[32]: #Data cleaning of flat_model
      df_new['flat_model'].unique()

[32]: array(['Improved', 'New Generation', 'Adjoined flat', 'Model A',
             'Standard', 'Simplified', 'Premium Apartment', 'Maisonette',
             'Apartment', 'Model A2', 'Type S1', 'Type S2',
             'Premium Maisonette', 'Improved-Maisonette', 'Terrace', 'DBSS',
             'Model A-Maisonette', '2-room', 'Multi Generation',
             'Premium Apartment Loft'], dtype=object)
```

Data Cleaning

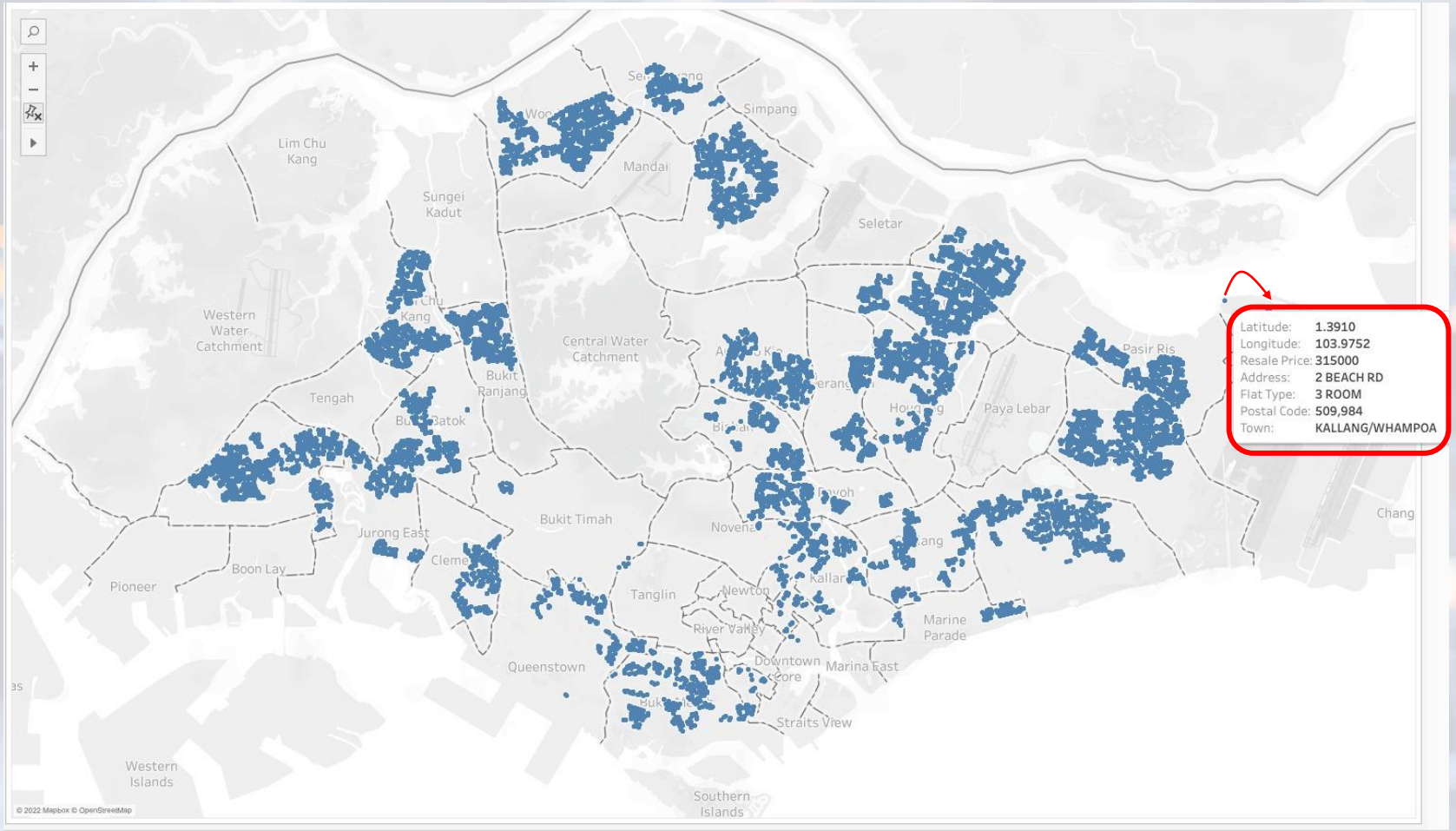
```
[34]: # Rename flat model duplicates
replace_values = {'Model A-Maisonette':'Maisonette', 'Improved-Maisonette':'Maisonette', 'Premium Maisonette':'Maisonette', 'Premium Apartment Loft':'Premium Apartment', 'Type S1':'Type S1S2', 'Type S2':'Type S1S2'}

df_new = df_new.replace({'flat_model': replace_values})

df_new['flat_model'].value_counts()
```

```
[33]: df_new['flat_model'].value_counts()

[33]: Model A                 51963
      Improved                40432
      New Generation          22273
      Premium Apartment       17803
      Simplified               6724
      Apartment                6363
      Maisonette               4754
      Standard                 4710
      DBSS                     2559
      Model A2                 2026
      Type S1                   363
      Model A-Maisonette        293
      Adjoined flat             275
      Type S2                   188
      Terrace                    98
      Premium Apartment Loft     86
      Multi Generation           69
      Improved-Maisonette        22
      2-room                     18
      Premium Maisonette         15
      Name: flat_model, dtype: int64
```

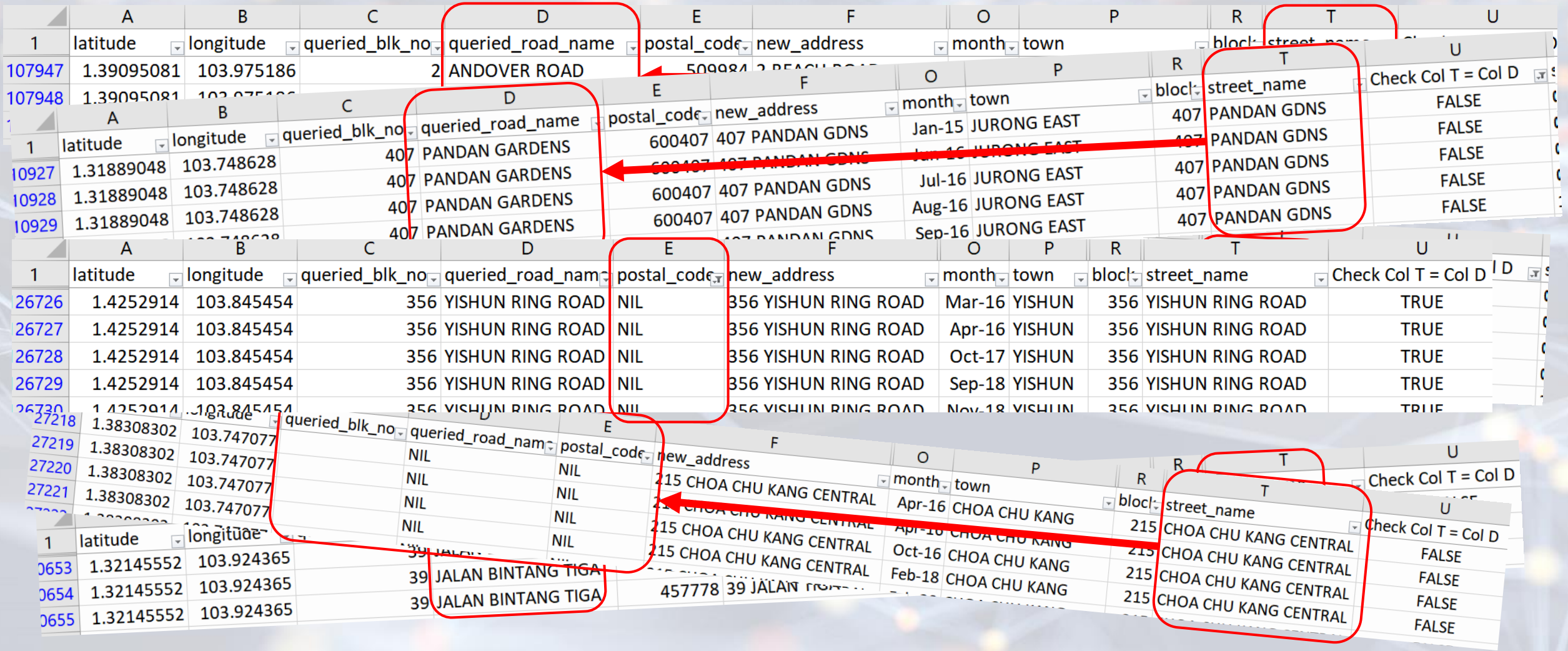
```
[34]: Model A                 51963
      Improved                40432
      New Generation          22273
      Premium Apartment       17889
      Simplified               6724
      Apartment                6363
      Maisonette               5084
      Standard                 4710
      DBSS                     2559
      Model A2                 2026
      Type S1S2                 551
      Adjoined flat             275
      Terrace                    98
      Multi Generation           69
      2-room                     18
      Name: flat_model, dtype: int64
```

Data Exploration

- Big red flags!
  - Some addresses queried on OneMap were wrong.
  - Lat & Long, postal code for these addresses are incorrect.



Data Exploration



	A	B	C	D	E	F	O	P	R	T	U
1	latitude	longitude	queried_blk_no	queried_road_name	postal_code	new_address	month	town	block	street_name	Check Col T = Col D
107947	1.39095081	103.975186	2	ANDOVER ROAD							
107948	1.39095081										
10927	1.31889048	103.748628	407	PANDAN GARDENS	600407	407 PANDAN GDNS	Jan-15	JURONG EAST	407	PANDAN GDNS	FALSE
10928	1.31889048	103.748628	407	PANDAN GARDENS	600407	407 PANDAN GDNS	Jul-16	JURONG EAST	407	PANDAN GDNS	FALSE
10929	1.31889048	103.748628	407	PANDAN GARDENS	600407	407 PANDAN GDNS	Aug-16	JURONG EAST	407	PANDAN GDNS	FALSE
					600407	407 PANDAN GDNS	Sep-16	JURONG EAST	407	PANDAN GDNS	FALSE
26726	1.4252914	103.845454	356	YISHUN RING ROAD	NIL	356 YISHUN RING ROAD	Mar-16	YISHUN	356	YISHUN RING ROAD	TRUE
26727	1.4252914	103.845454	356	YISHUN RING ROAD	NIL	356 YISHUN RING ROAD	Apr-16	YISHUN	356	YISHUN RING ROAD	TRUE
26728	1.4252914	103.845454	356	YISHUN RING ROAD	NIL	356 YISHUN RING ROAD	Oct-17	YISHUN	356	YISHUN RING ROAD	TRUE
26729	1.4252914	103.845454	356	YISHUN RING ROAD	NIL	356 YISHUN RING ROAD	Sep-18	YISHUN	356	YISHUN RING ROAD	TRUE
27218	1.38308302	103.747077		NIL	NIL	215 CHOA CHU KANG CENTRAL	Apr-16	CHOA CHU KANG	215	CHOA CHU KANG CENTRAL	FALSE
27219	1.38308302	103.747077		NIL	NIL	215 CHOA CHU KANG CENTRAL	Oct-16	CHOA CHU KANG	215	CHOA CHU KANG CENTRAL	FALSE
27220	1.38308302	103.747077		NIL	NIL	215 CHOA CHU KANG CENTRAL	Feb-18	CHOA CHU KANG	215	CHOA CHU KANG CENTRAL	FALSE
27221	1.38308302	103.747077		NIL	NIL						
10653	1.32145552	103.924365	39	JALAN BINTANG TIGA							
10654	1.32145552	103.924365	39	JALAN BINTANG TIGA	457778	39 JALAN					
10655	1.32145552	103.924365									

Data Cleaning & Collection (again…)

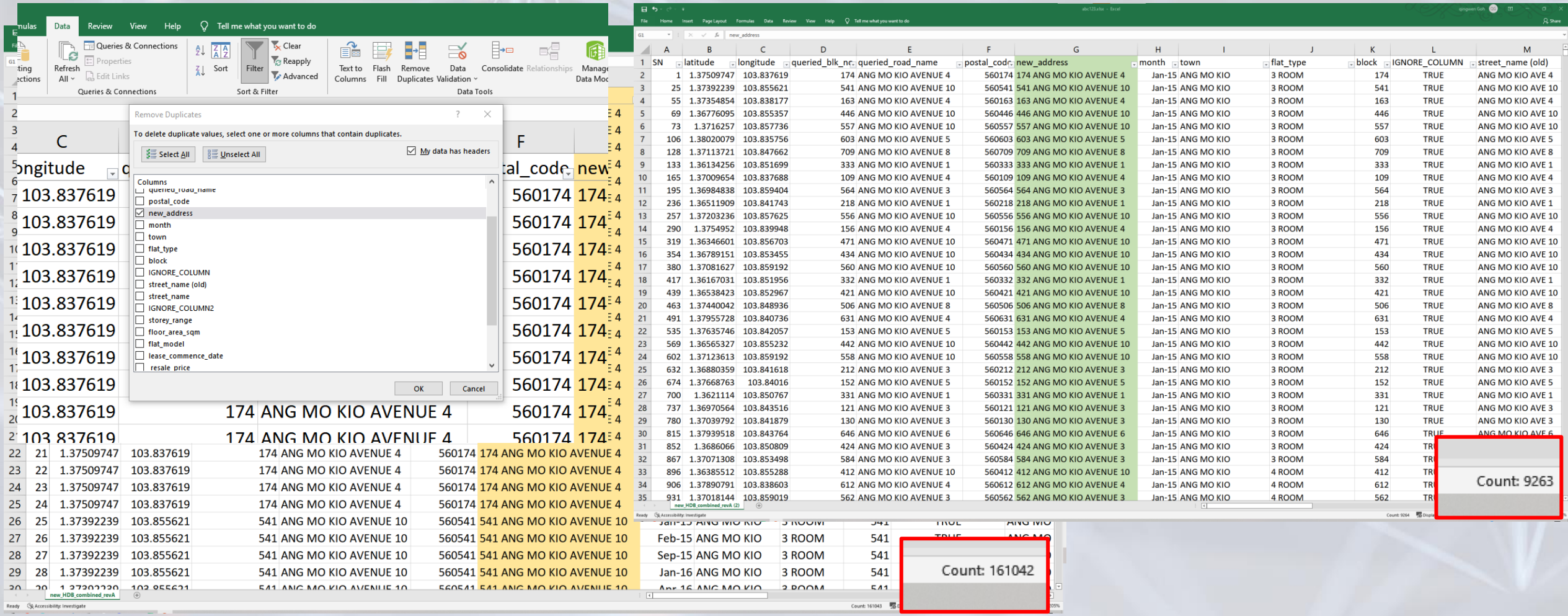
- Using Excel,
  - Converted abbreviations into long form.
  - Clean up the errors in the queried_road_name, and
  - Manually searched for the latitude, longitude and postal code using the new_address and Google Maps.

	A	B	C	D	E	F
1	street_name	STREETreet_name			ST.	SAINT
2	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	k		AVE	AVENUE
3	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	k		RD	ROAD
4	ANG MO KIO AVE 5	ANG MO KIO AVENUE 5	k		ST	STREET
5	ANG MO KIO AVE 8	ANG MO KIO AVENUE 8	k		STH	SOUTH
6	ANG MO KIO AVE 1	ANG MO KIO AVENUE 1	k		NTH	NORTH
7	ANG MO KIO AVE 3	ANG MO KIO AVENUE 3	k		CL	CLOSE
8	ANG MO KIO AVE 6	ANG MO KIO AVENUE 6	k		DR	DRIVE
9	ANG MO KIO ST 52	ANG MO KIO STREET 52	k		UPP	UPPER
10	ANG MO KIO ST 21	ANG MO KIO STREET 21	k		CTRL	CENTRAL
11	ANG MO KIO ST 31	ANG MO KIO STREET 31	k		BT	BUKIT
12	BEDOK RESERVOIR RD	BEDOK RESERVOIR ROAD	k		CRES	CRESCENT
13	BEDOK STH RD	BEDOK STREETH ROAD	k		LOR	LORONG
14	BEDOK NTH ST 3	BEDOK NORTH STREET 3	k		C'WEALTH	COMMONWEALTH
15	BEDOK NTH AVE 1	BEDOK NORTH AVENUE 1	k		TER	TERRACE
16	BEDOK NTH RD	BEDOK NORTH ROAD	k		PL	PLACE
17	NEW UPP CHANGI RD	NEW UPPER CHANGI ROAD	k		HTS	HEIGHTS
18	CHAI CHEE ST	CHAI CHEE STREET	k		TG	TANJONG
19	BEDOK NTH ST 1	BEDOK NORTH STREET 1	k		PK	PARK
20	BEDOK NTH AVE 4	BEDOK NORTH AVENUE 4	k		KG	KAMPONG
21	BEDOK NTH ST 2	BEDOK NORTH STREET 2	k			
22	CHAI CHEE AVE	CHAI CHEE AVENUE	k			
23	BEDOK NTH AVE 3	BEDOK NORTH AVENUE 3	k			
24	BEDOK STH AVE 1	BEDOK STREETH AVENUE 1	k			
25	BEDOK CTRL	BEDOK CENTRAL	k			
26	BEDOK NTH AVE 2	BEDOK NORTH AVENUE 2	k			
27	BEDOK STH AVE 2	BEDOK STREETH AVENUE 2	k			
28	BEDOK RESERVOIR VIEW	BEDOK RESERVOIR VIEW	k			

	T	U	V
	street_name (old)	street_name	new_address
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
4	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	174 ANG MO KIO AVENUE 4
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10
1	ANG MO KIO AVE 10	ANG MO KIO AVENUE 10	541 ANG MO KIO AVENUE 10

Data Cleaning & Collection (again…)

- As the Lat & Long for the HDB addresses were wrong, it also resulted in the wrong calculation of geographic distance to the amenities. Hence, we will need to recalculate again.
- Prior to that, used Excel drop duplicates function on new_address: 161,042 → 9263 rows



Data Cleaning & Collection (again…)

- Modify and re-run the Jupyter notebook code to re-calculate and re-retrieve the distances again.

```
[148]: address_coordinates
```

[148]:

	dis_to_cbd	min_dist_to_mrt	min_dist_to_school	min_dist_hospital	min_dist_to_park	min_dist_to_market	min_dist_to_facility
0	10296.855145	418.542931	234.381094	900.905352	601.981635	175.969647	984.548713
1	10065.228122	806.179213	438.152527	1107.495900	674.787991	167.324255	1297.742566
2	10118.355112	291.930042	388.251745	839.198973	535.485804	120.537648	802.404478
3	9383.300627	688.596184	383.450281	1298.879279	632.458098	122.571965	802.127936
4	9825.917191	929.163338	636.255854	1374.726144	814.413249	393.436638	1289.585608
...	...	...	...	...	...	...	...
9258	14736.556319	1178.949557	553.368035	1047.683571	940.822224	2396.833513	1350.405005
9259	9688.670929	408.893439	175.179969	1437.632112	1160.117151	630.569068	790.221014
9260	6674.887845	403.756166	828.183712	1078.776677	2671.487657	281.968968	1352.659450
9261	8014.142360	668.857162	391.317167	1732.093876	1043.308831	426.987500	2078.373506
9262	14512.574067	834.248496	185.974572	1126.413820	1030.395934	2325.748509	865.693471

9263 rows × 7 columns

```
[149]: address_coordinates.to_csv('address_coordinates.csv',index=False)
```

- Excel copy & paste to the table with unique new_address, then perform VLOOKUP to add all the data back to the base data set.

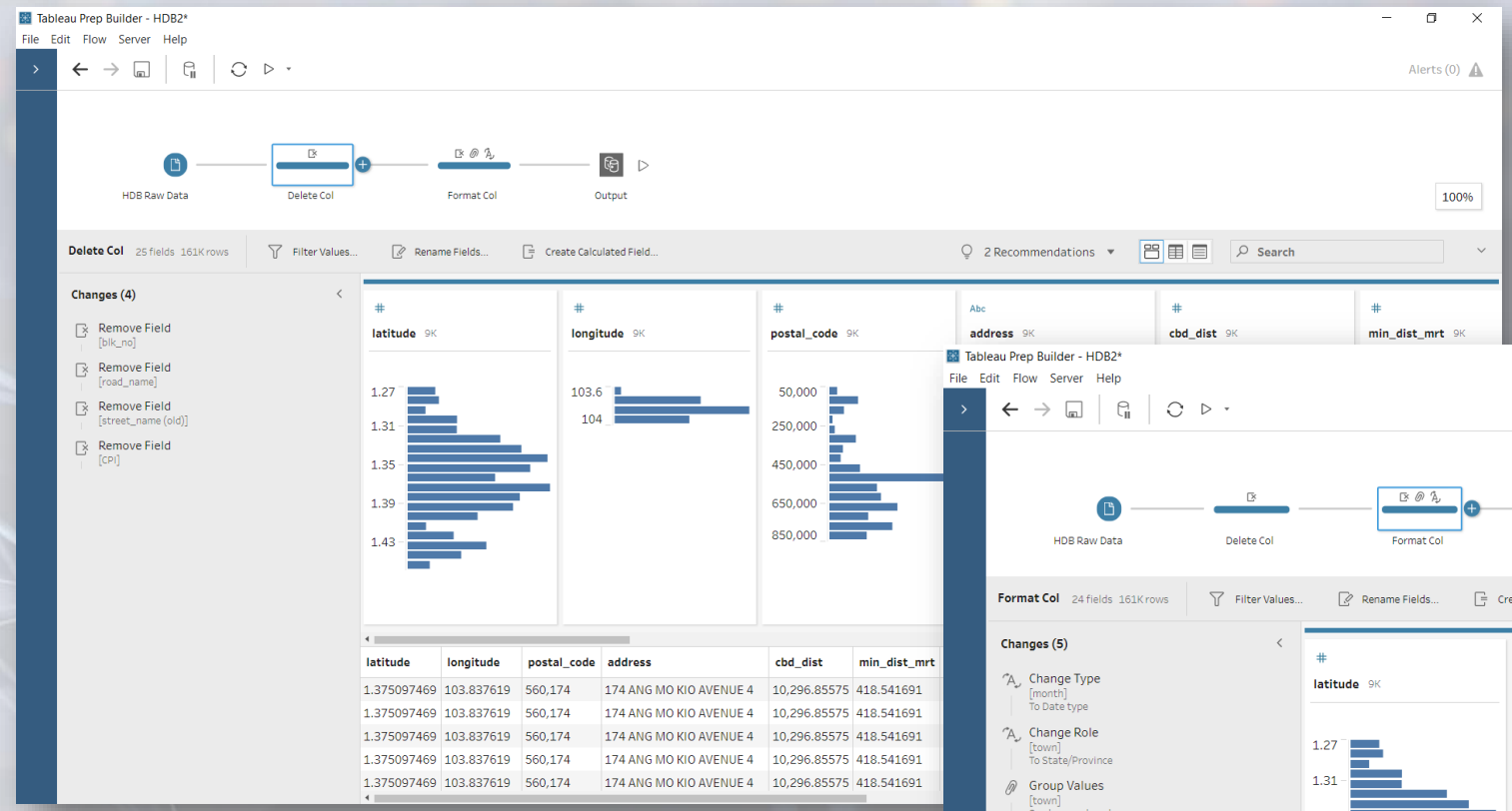
Finally…

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	SN	latitude	longitude	queried_blk_nc	queried_road_name	postal_code	new_address	month	town	flat_type	block	IGNORE_COLUMN	street_name (old)	street_name	IGNORE_COLUMN2	storey_range	floor_area_sqm	flat_mode
2	1	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	Jan-15	ANG MO KIO	3 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	07 TO 09	60	Improved
3	2	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	Dec-15	ANG MO KIO	3 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	10 TO 12	60	Improved
4	3	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	May-16	ANG MO KIO	3 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	04 TO 06	69	Improved
5	4	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	Jun-16	ANG MO KIO	2 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	07 TO 09	45	Improved
6	5	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	Nov-16	ANG MO KIO	3 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	04 TO 06	61	Improved
7	6	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	Apr-17	ANG MO KIO	2 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	04 TO 06	45	Improved
8	7	1.375097469	103.837619	174	ANG MO KIO AVENUE 4	560174	174 ANG MO KIO AVENUE 4	Jun-17	ANG MO KIO	2 ROOM	174	TRUE	ANG MO KIO AVE 4	ANG MO KIO AVENUE 4	TRUE	07 TO 09	45	Improved

S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
lease_commence_date	resale_price	remaining_leas	CPI	cpi_adjusted_resale_price	remaining_lease_in_months	remaining_lease_in_years	dis_to_c	min_dist_to_r	min_dist_to_sch	min_dist_hosp	min_dist_to_p	min_dist_to_mar	min_dist_to_faci
1986	$ 255,000.00	70	111.423	$228,858	840	70	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132
1986	$ 275,000.00	69	107.421	$256,002	828	69	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132
1986	$ 310,000.00	68	105.496	$293,850	816	68	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132
1986	$ 253,000.00	68	104.838	$241,325	816	68	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132
1986	$ 290,000.00	68	103.676	$279,718	816	68	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132
1986	$ 210,000.00	67 years 09 months	103.413	$203,069	813	67.8	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132
1986	$ 233,000.00	67 years 07 months	102.533	$227,244	811	67.6	10296.855	418.5429311	234.3810935	900.9053521	601.9816351	175.9696465	984.5487132

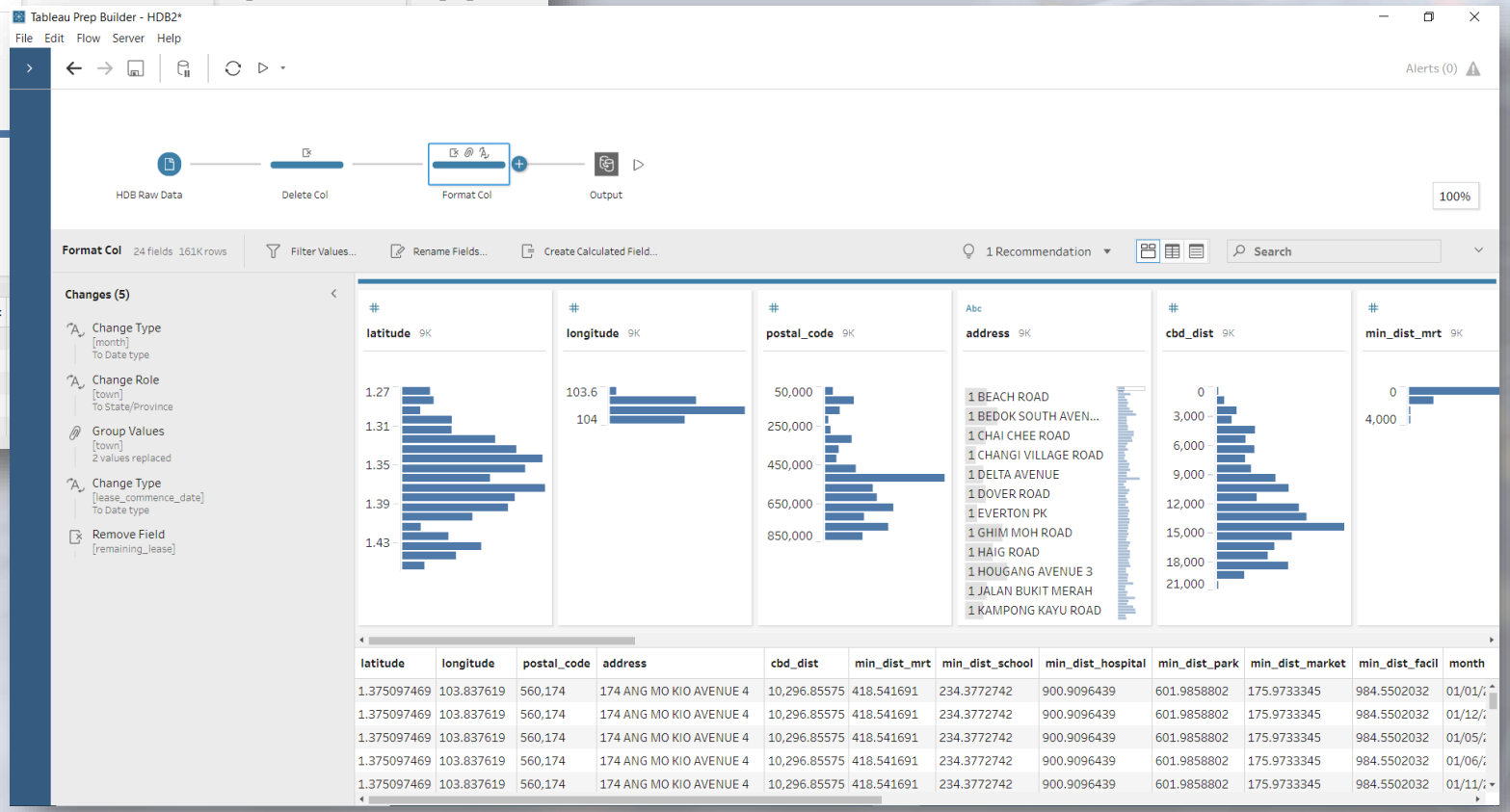
IV. Diagnostic Analytics

Data Refinery using Tableau Prep Builder



In order to start analyzing data, we will have to conduct data cleansing first. Using Tableau Prep Builder, we deleted unnecessary columns like extra Road Name columns, Block Numbers, etc.

We continued with formatting the data, ensuring correct data format such as: date, state/province, decimal, string, etc. File will be output has “hyper” type to be used in Tableau.



Data Visualization – Data overview & Tool

Overview of the Dataset from Data Engineer:

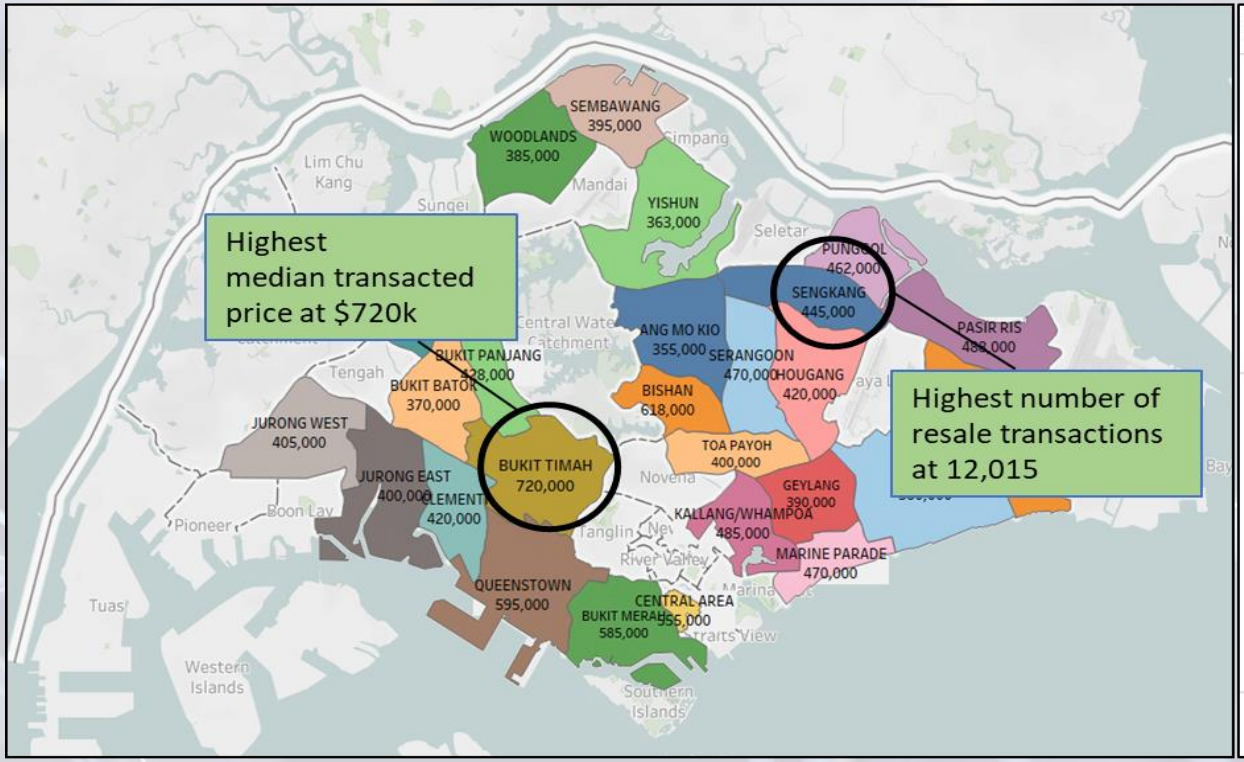
- 161,042 resale transactions recorded from 1 Jan 2015 to 1 April 2022
- Visualize on Tableau using 19 features consisting of:

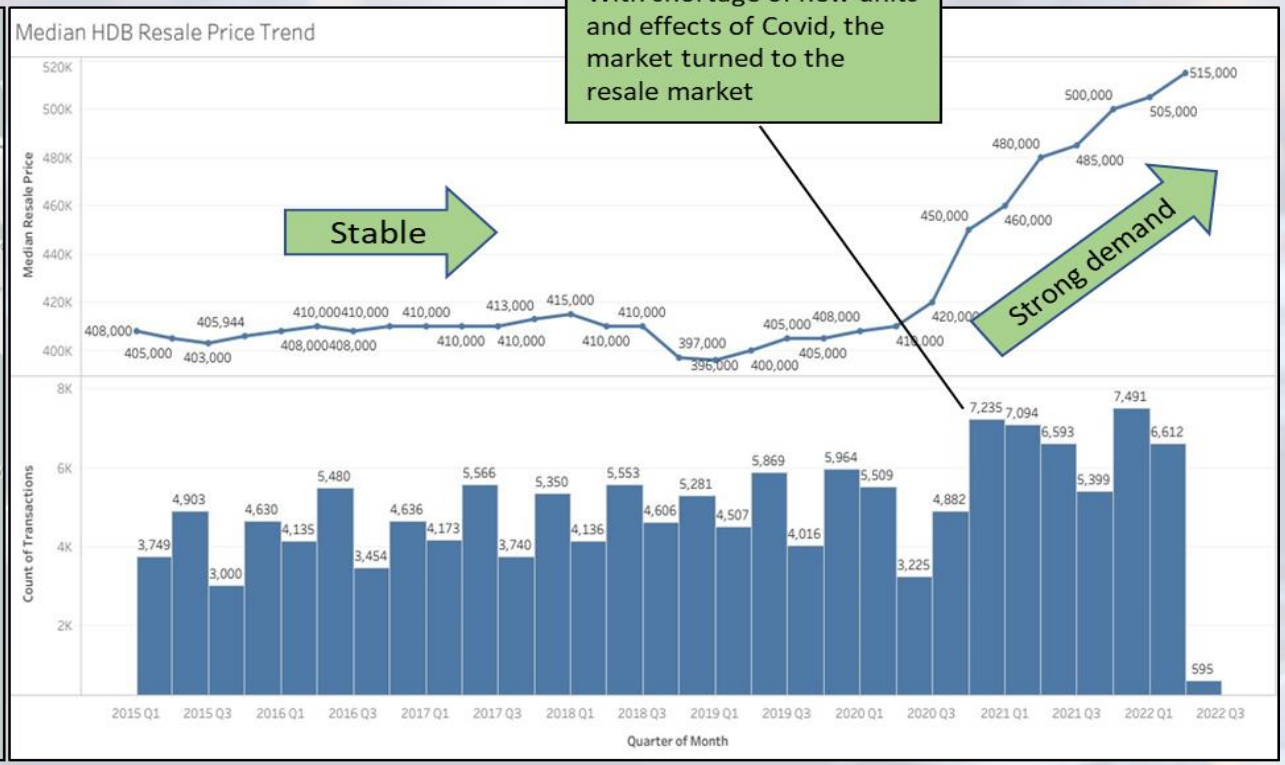
1. Flat Model	11. Distance to Nearest MRT Station
2. Flat Type	12. Distance to Nearest Park
3. Month	13. Distance to Nearest School
4. New Address	14. Distance to Nearest Hospital
5. Postal Code	15. Distance to Nearest Market
6. Remaining Lease (Years)	16. Distance to Nearest Sports Facilities
7. Storey Range	17. Resale Price (Target Variable)
8. Town	18. Latitude
9. Floor Area (sqm)	19. Longitude
10.Distance to Nearest CBD	

Fields

Type	Field Name	Phy...	Remote Field Name
🌐	Longitude	2104...	longitude
#	Postal Code	2104...	postal_code
Abc	New Address	2104...	new_address
📅	Month	2104...	month
🌐	Town	2104...	town
Abc	Flat Type	2104...	flat_type
Abc	Storey Range	2104...	storey_range
#	Floor Area Sqm	2104...	floor_area_sqm
Abc	Flat Model	2104...	flat_model
#	Lease Commence Date	2104...	lease_commence_date
#	Resale Price	2104...	resale_price
#	Remaining Lease In Years	2104...	remaining_lease_in_years
#	Dis To Cbd	2104...	dis_to_cbd
#	Min Dist To Mrt	2104...	min_dist_to_mrt
#	Min Dist To School	2104...	min_dist_to_school
#	Min Dist Hospital	2104...	min_dist_hospital
#	Min Dist To Park	2104...	min_dist_to_park
#	Min Dist To Market	2104...	min_dist_to_market

Data Visualization – Market Overview





- Median resale HDB price of our data set is at $410,000 with Bukit Timah township commanding the highest median price of $720,000
- Sengkang Town was the most active resale market with 12,015 transactions, much higher than the overall average of 5,668 transactions.

- Median resale flat prices was generally stable from 2015 – 1Q 2020.
- Median resale flat prices started to increase from 3Q 2020 to 1Q 2022. It is believed that such "bullish resale market was (also) driven by delays in build-to-order (BTO) flats, which pushed homebuyers to look for more readily available alternatives." *(Business Times, HDB resale prices climb 12.7% in 2021, record growth since 2010, 28 Jan 2022)*.
- Increased in overall transaction volume from 3Q 2020

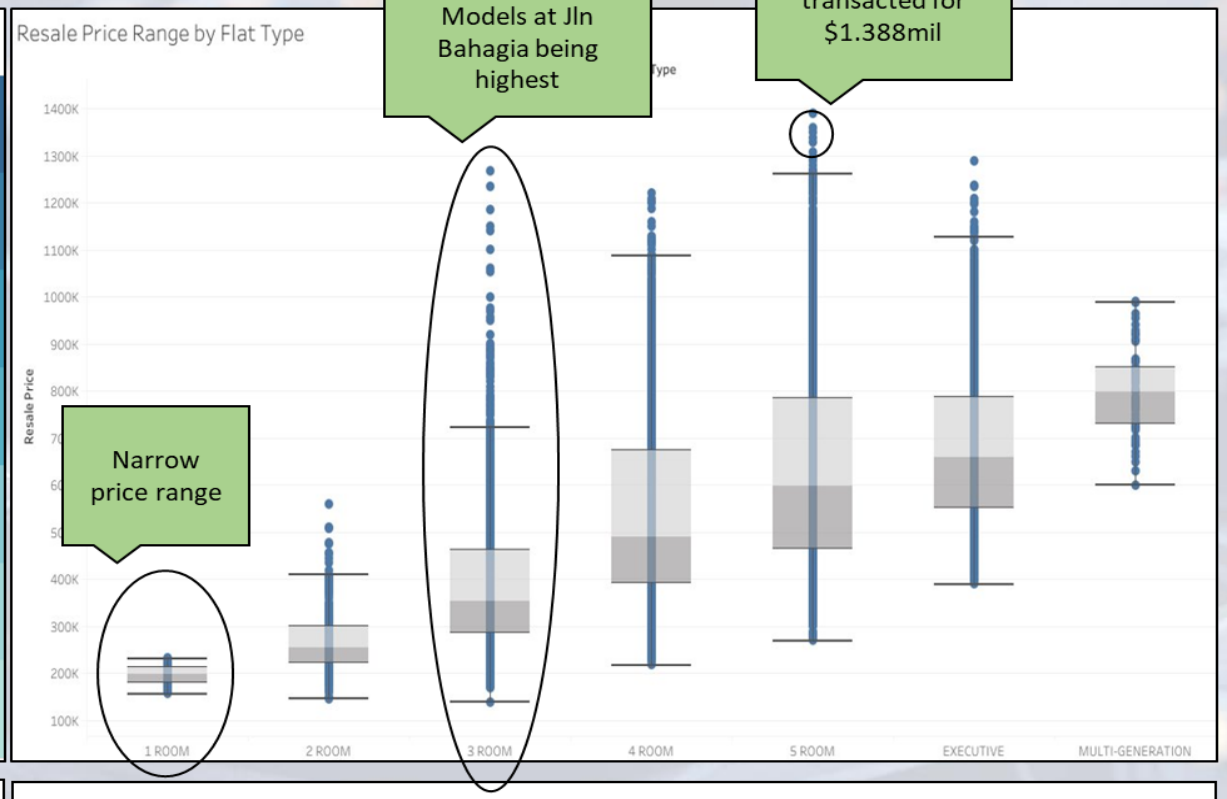
Data Visualization – Market Overview

Median Resale Price by Flat Type

Flat Type	2015	2016	2017	2018	2019	2020	2021	2022
MULTI-GENERATION	805,000	735,000	830,000	778,000	819,444	762,5	,000	842,000
EXECUTIVE	605,000	600,000	600,000	610,000	595,000	618,0	,000	730,000
5 ROOM	475,000	472,000	478,000	477,000	483,000	508,	,000	600,000
4 ROOM	404,000	405,000	408,000	400,000	400,000	420,0	,000	
3 ROOM	310,000	308,000	302,000	290,000	283,000	295,0	335,000	
2 ROOM	245,000	240,000	240,000	230,000	225,000	230,000	265,000	280,000
1 ROOM	208,000	213,000	203,000	180,000	176,634	174,000	198,000	215,000

11.40%
11.30%
12.20%
11.90%
13.60%
15.20%
13.80%

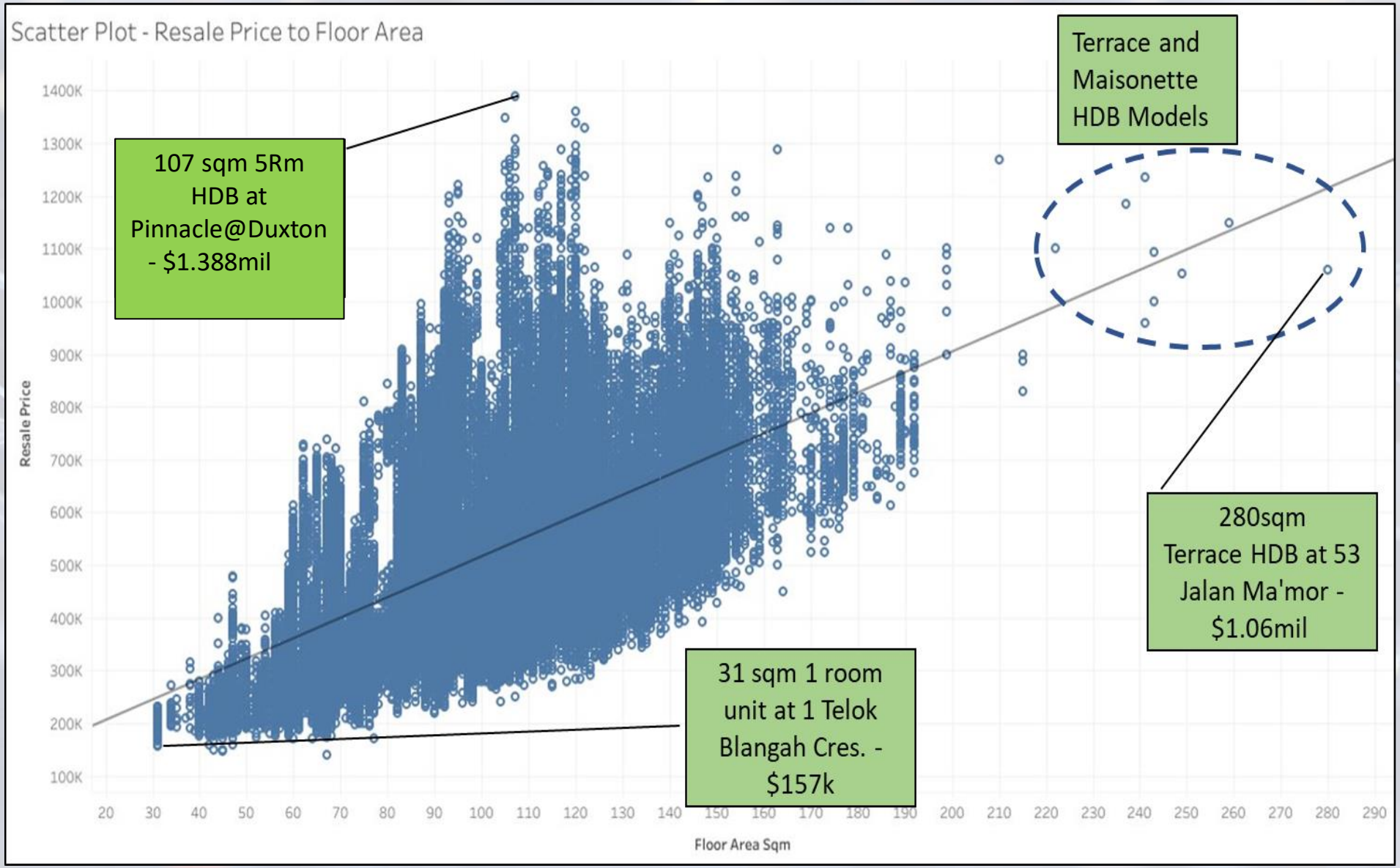
Smaller flat types experience higher % increase in prices



- Average resale HDB price of our data set was generally stable from 2015 – 2020. Prices start to increase from 2020 – 2022 likely due to covid, longer BTO waiting time and higher demands for flats.
- Observe that the increase in prices from 2021 was across all flat types where they experienced double digit % increase in resale prices.
- We could observe that 1-, 2- and 3-Room HDB experienced the highest % increase in prices

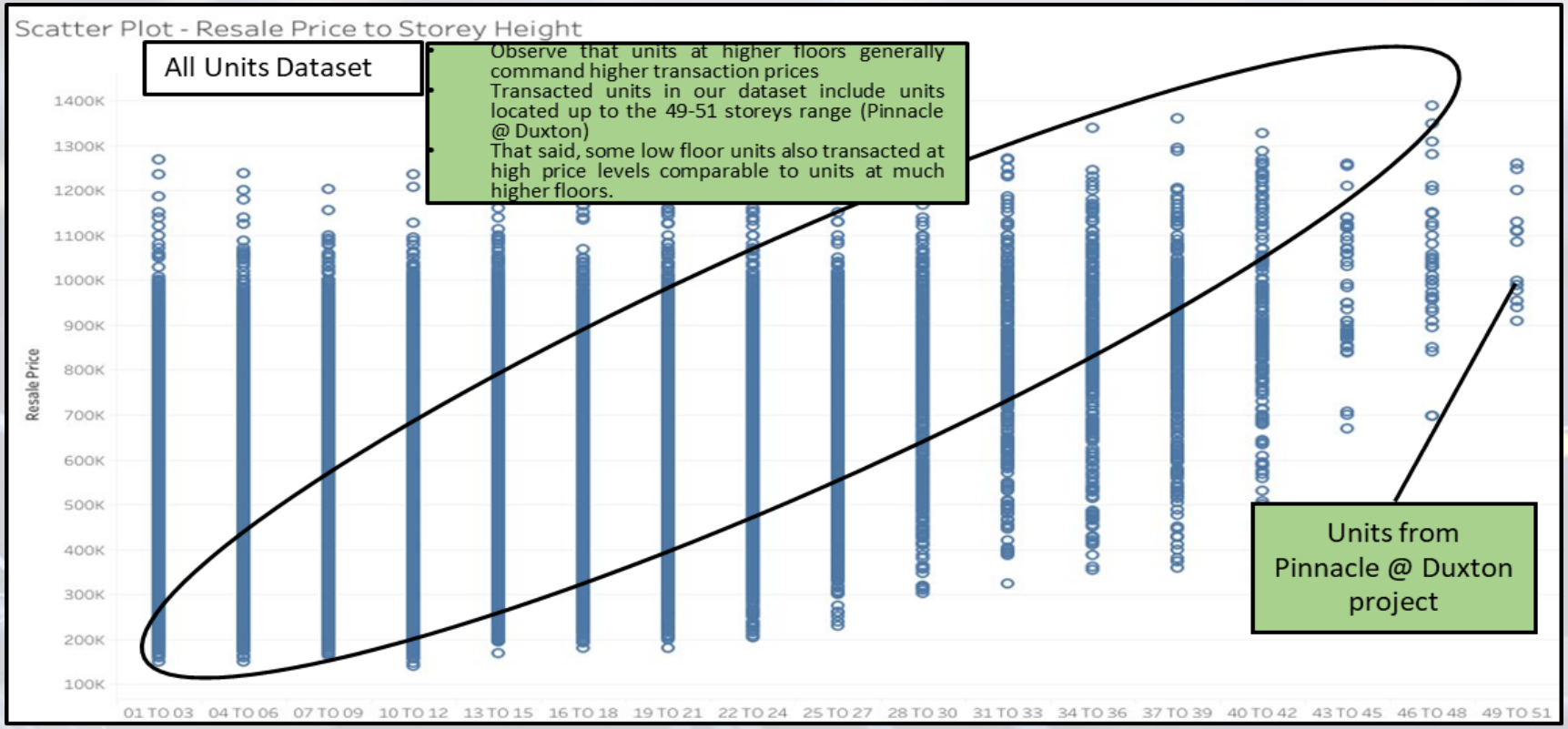
- Larger room type HDB units generally command higher transaction prices
- 1RM HDB units transact within the narrowest price range among the 7 flat types while 3R units transacted with the largest price range

Data Visualization – Resale Price to Flat Size



- Units with larger floor areas attract higher transaction prices
- Size of transacted HDB units range from 31sqm to 280 sqm with a median size of 95sqm
- Transacted prices of the units in the data range from $140k to $1.388mil with a median price of $425k.

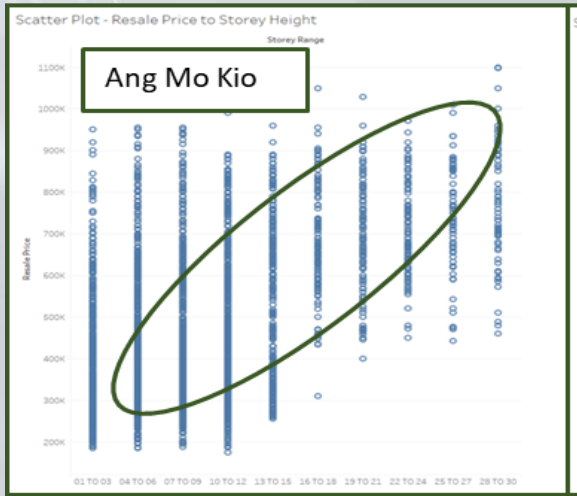
Data Visualization – Resale Price to Floor Height

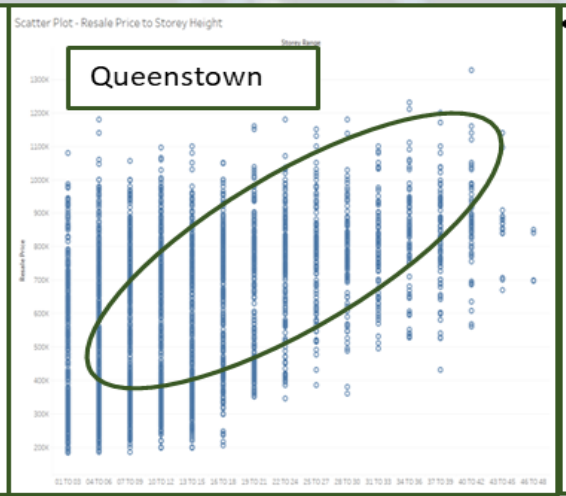


HDB terrace flats at Stirling Rd, Queenstown

HDB terrace flats in Whampoa

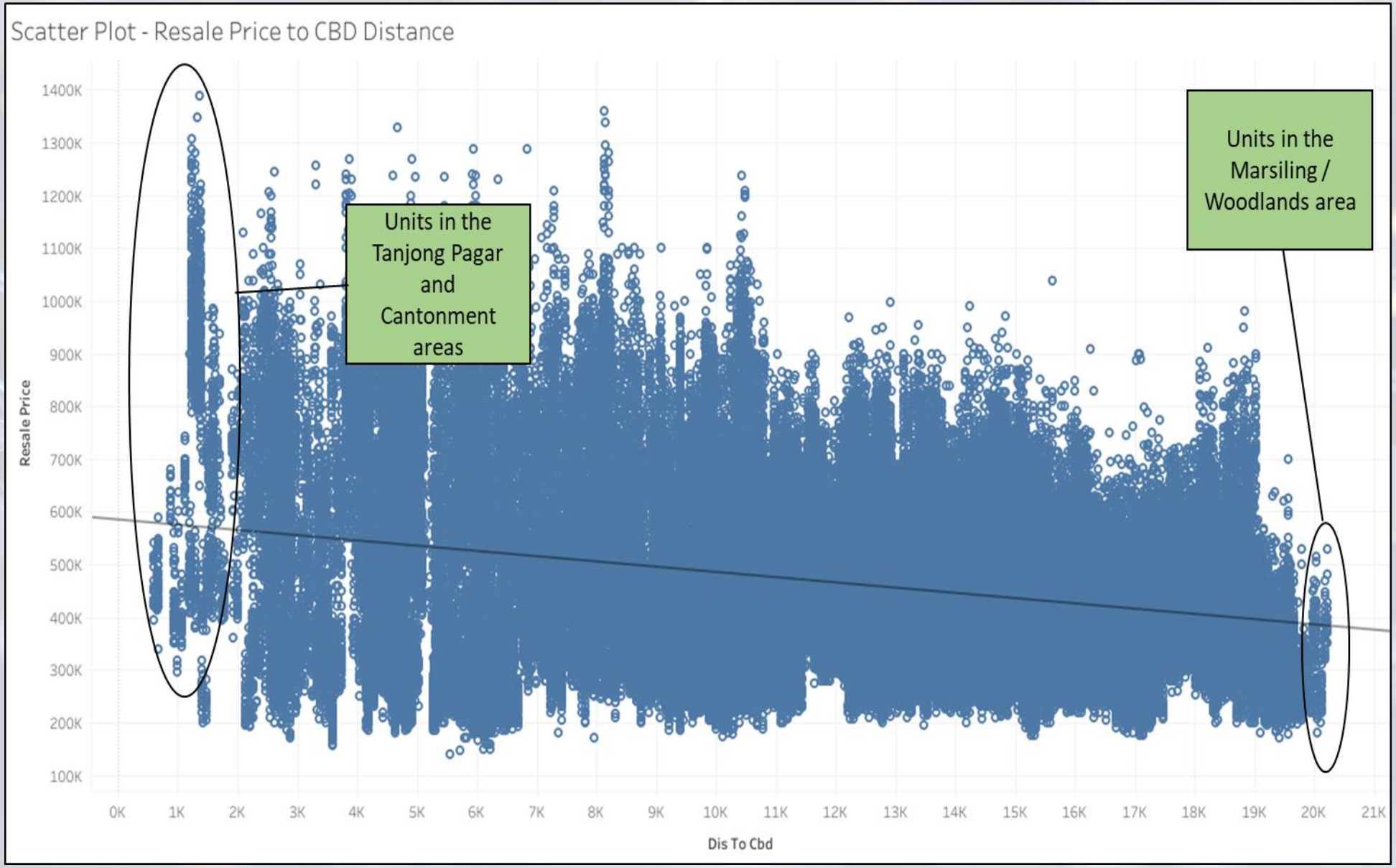
Source: https://www.99.co/singapore/insider/hdb-terrace-flats-can-find/





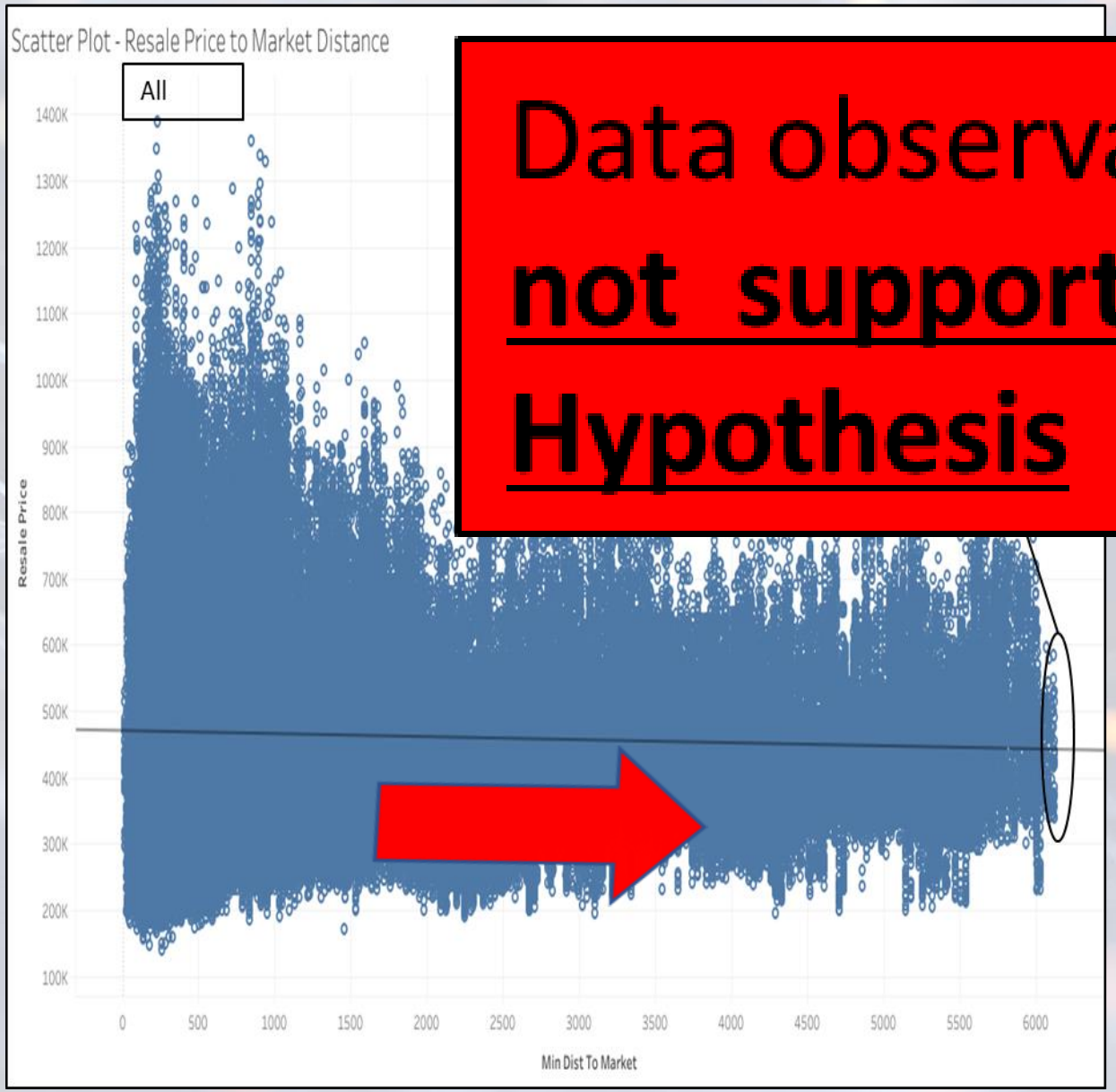
- The same pattern is also observed when we examine the price to storey height by Townships, where higher floor levels attract higher resale prices, such as the examples from Ang Mo Kio and Queenstown.

Data Visualization – Resale Price to Distance from CBD

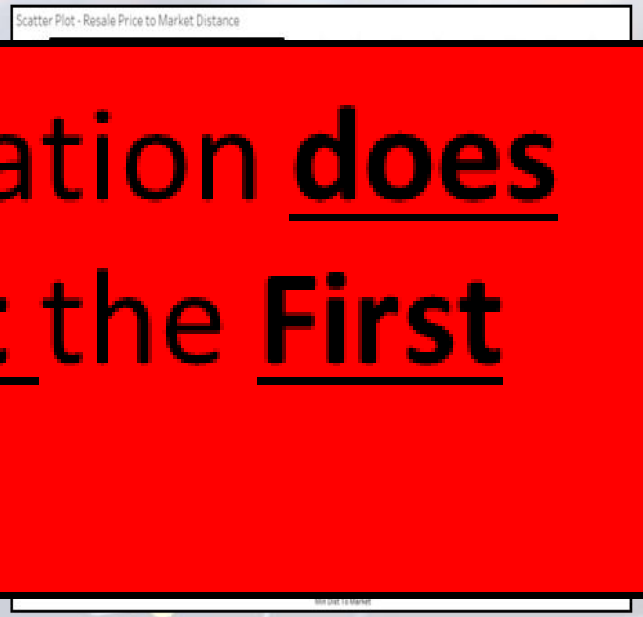


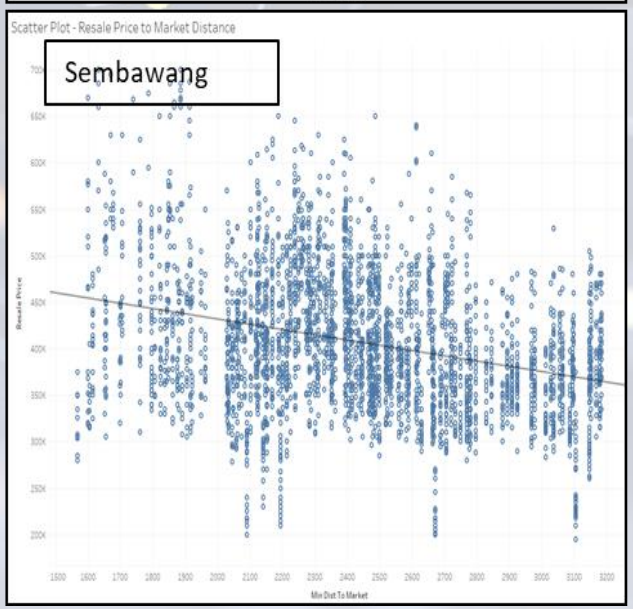
- Observations show that HDB resale units that are located near to the CBD attract higher prices than the ones further away. The observation seems to suggest that value given to its proximity to the core business district of Singapore
- Average distance to CBD of transacted units – 12.4km, with the furthest at 20.2km
- Half of the transacted units are located within 13.3km from the CBD

Data Visualization – Resale Price to Market Distance



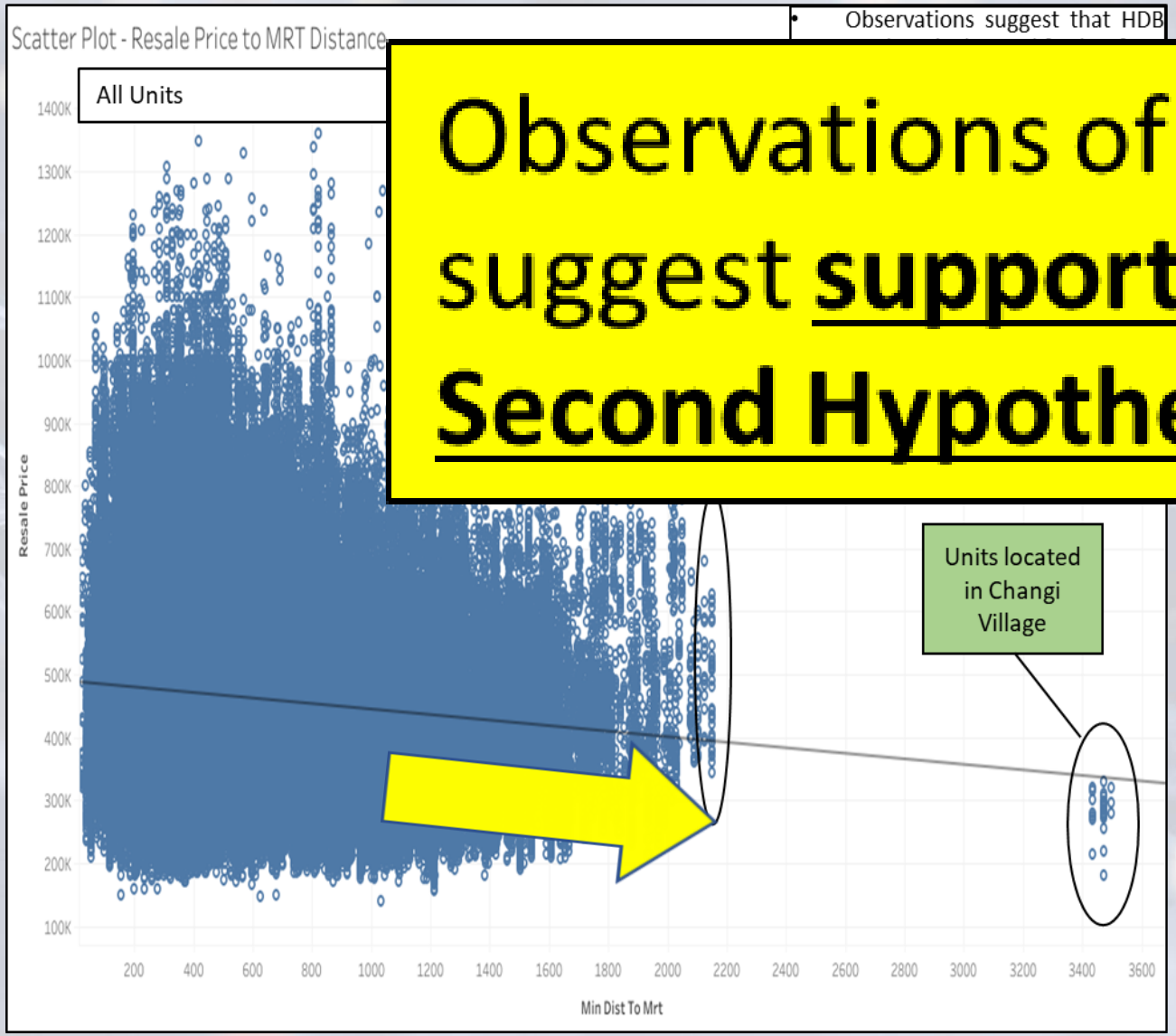
Data observation **does not support** the **First Hypothesis**





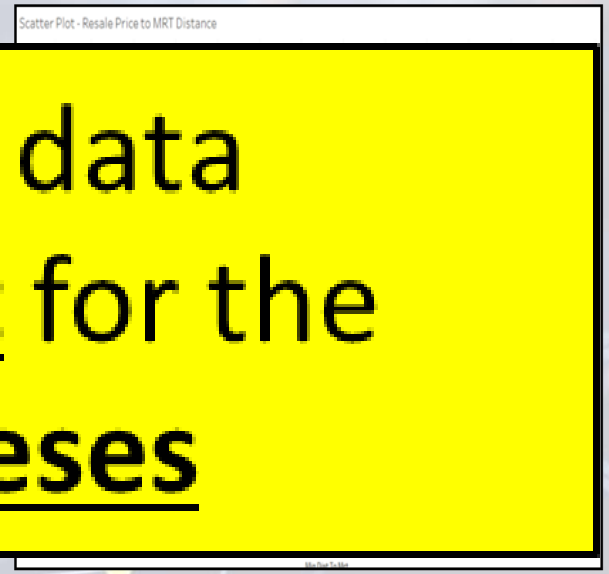
- Observation of the dataset slightly suggests that HDB resale units located nearer to markets transact at slightly higher prices than units located further away. Price influence by distance does not appear to be significant though.
- HDB units in the dataset are located ~1.9km from the nearest market on average with the furthest located 6.1km away from the nearest market (Choa Chu Kang St 62)
- 50% of the transacted units are within 1.42km distance to the nearest market location
- Zooming into data from the Townships of Bukit Batok and Sembawang shows similar observations where flats nearer to markets are priced higher than those located further away.

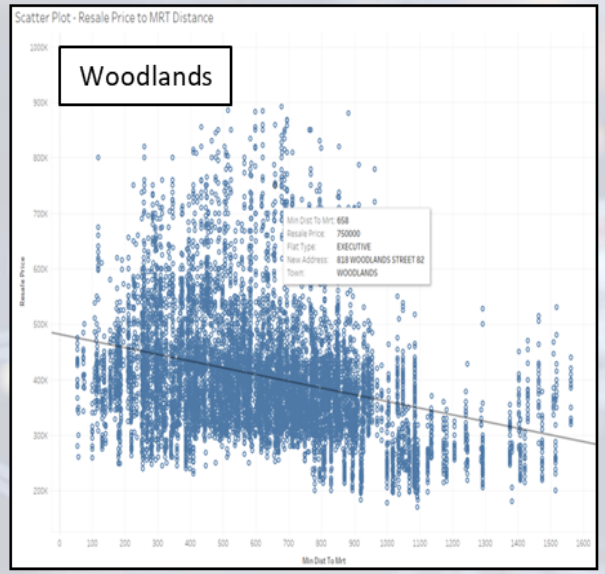
Data Visualization – Resale Price to Distance to MRT Station

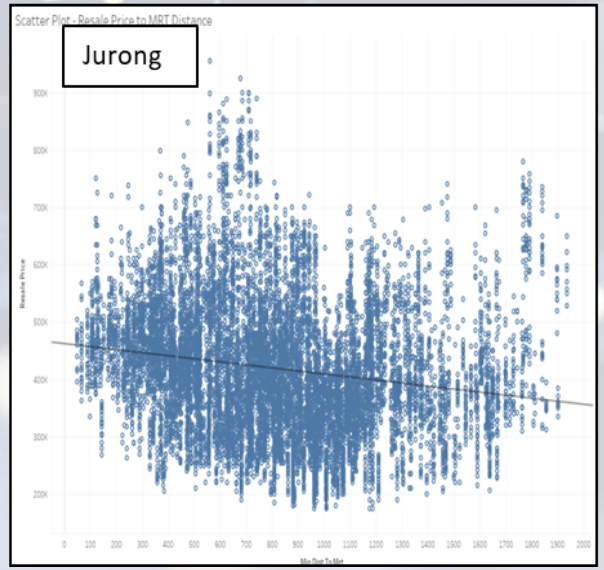


- Observations suggest that HDB

Observations of data suggest <u>support</u> for the <u>Second Hypotheses</u>

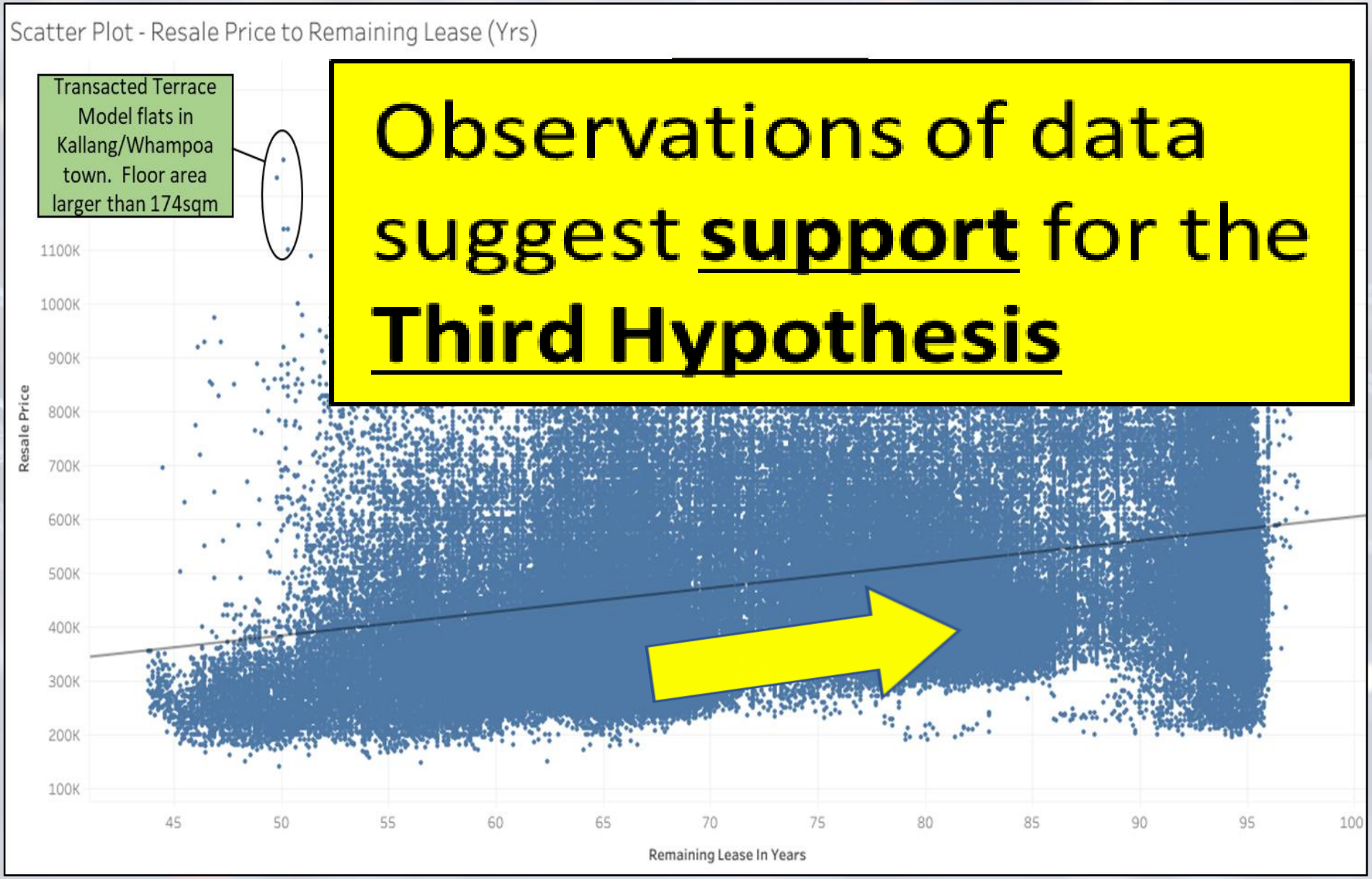






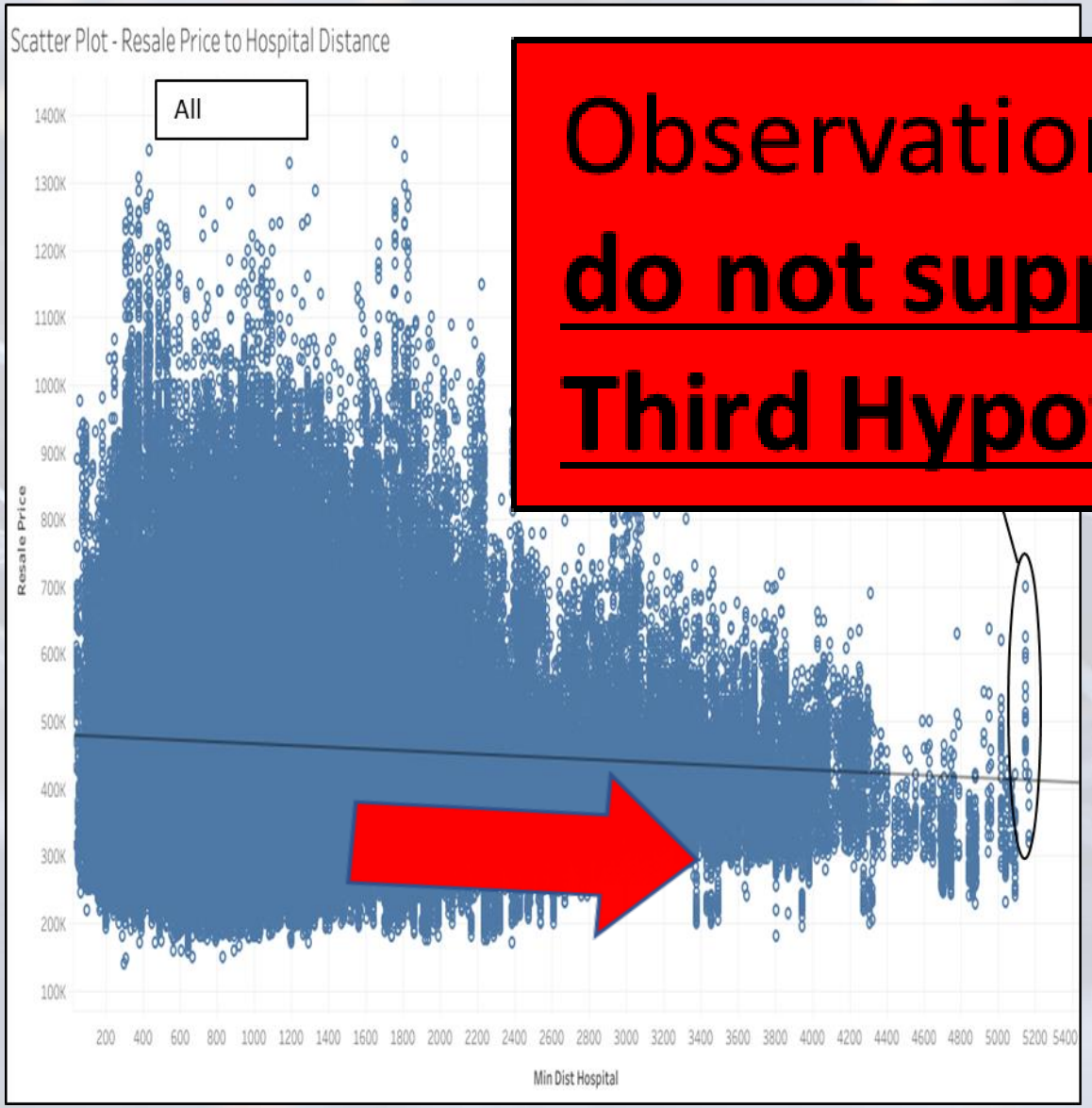
- Similar observations were made when we examine a sample of Townships where resale flats nearer to the MRT stations were priced higher that those located further away.
- Among the 3 Towns sampled, Woodlands Town's data seems to suggest the that distance to MRT station has the highest influence on resale flat prices.

Data Visualization – Resale Price to Remaining Lease

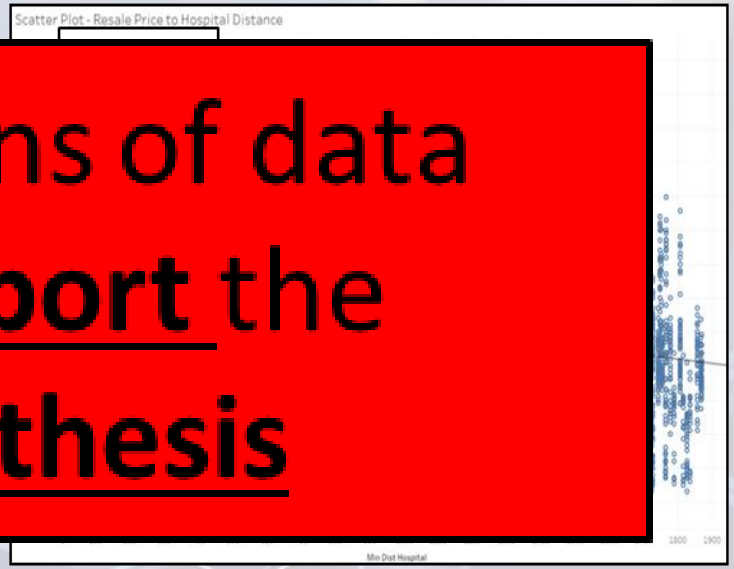


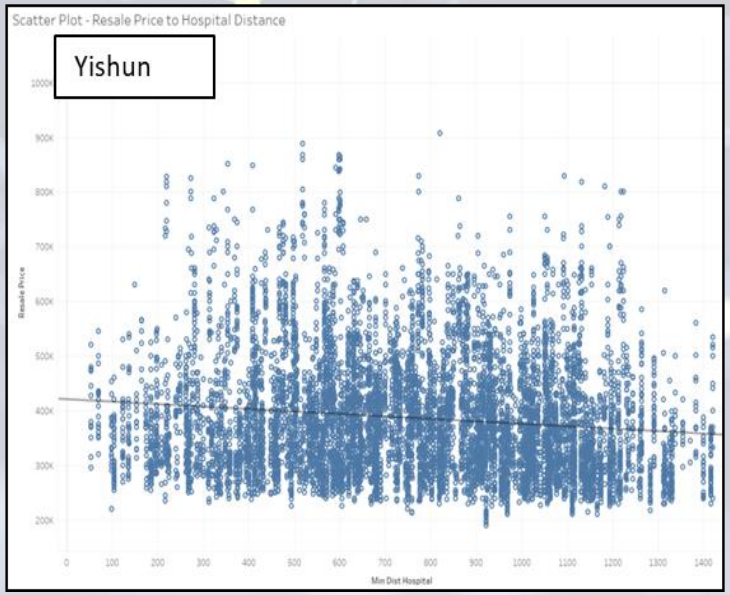
- Observe that units with more remaining lease years ("lower age" units) command higher prices
- Remaining lease years of HDB units in the dataset range from 43.8 years to 97.8 years, with a median of 74.5 years.
- Noted that there are several ~50-year-old units that transacted at above $1.1mil in the Kallang/Whampoa town.
- Some larger Maisonette type units in Bishan and Bukit Timah with remaining leas of ~65 years also transacted at above $1mil

Data Visualization – Resale Price to Healthcare Facility Distance



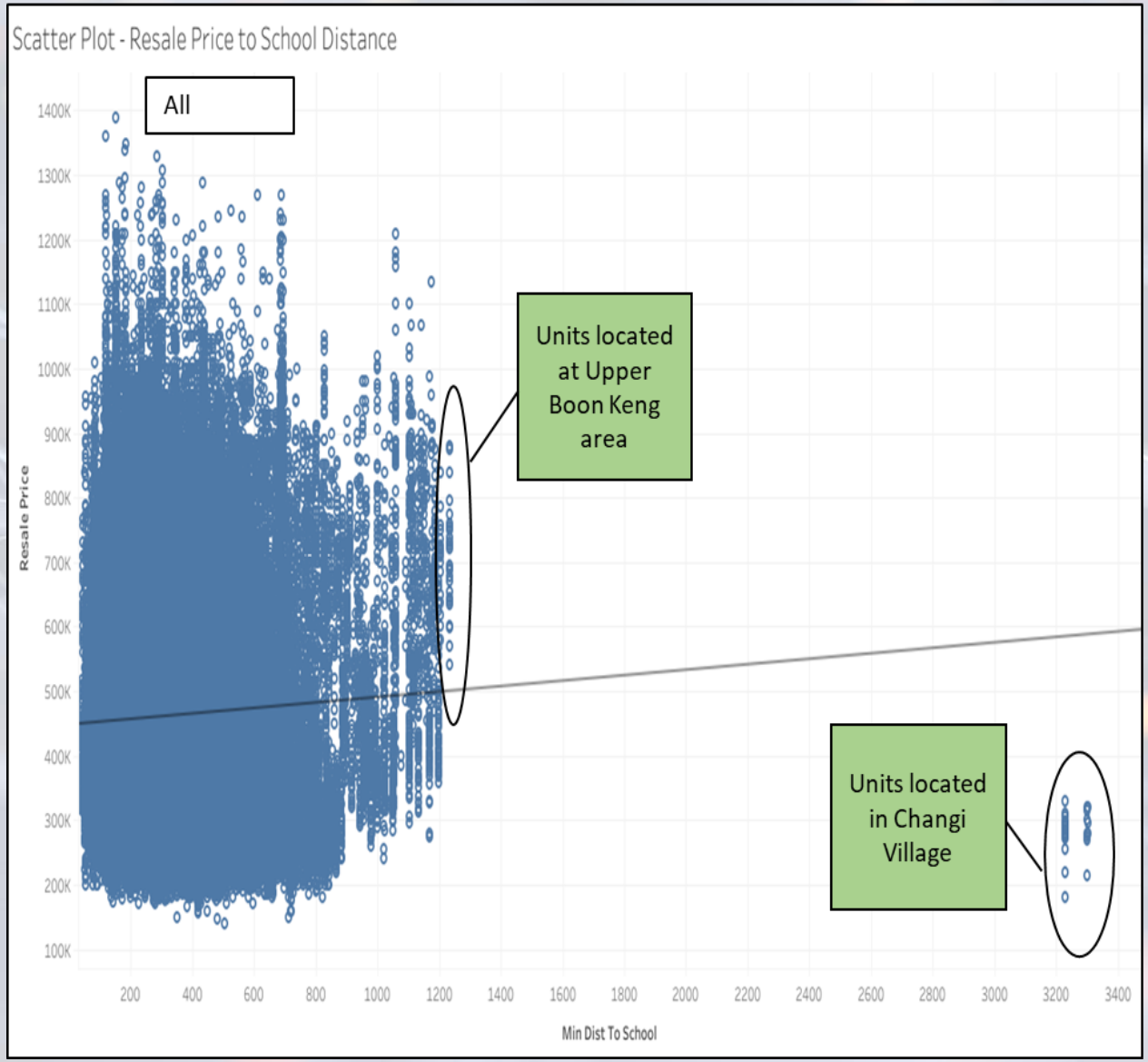
Observations of data **do not support** the **Third Hypothesis**

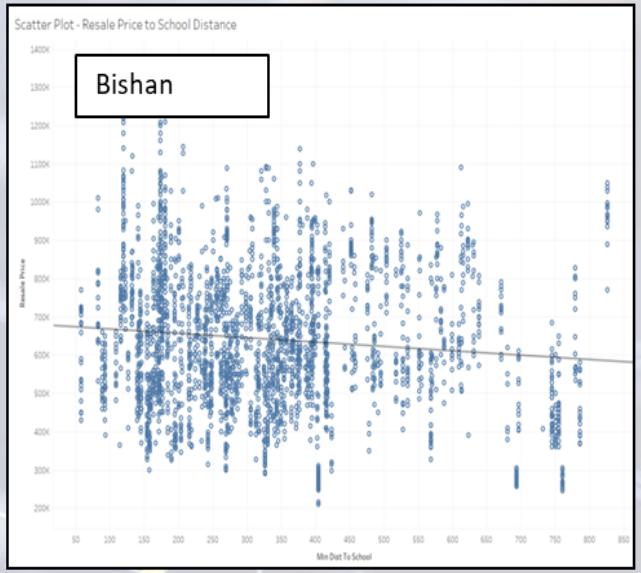


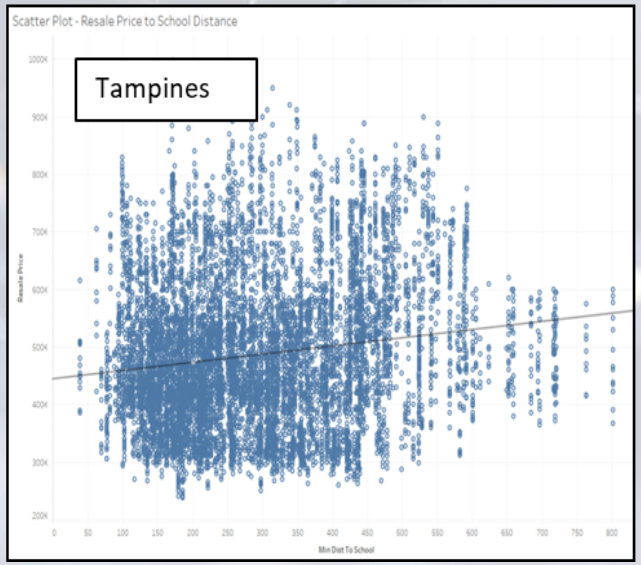


- Units located nearer to Healthcare Facilities are observed to transact at slightly higher prices than ones that further away but price influence by distance is does not appear to be significant.
- HDB units in the dataset are located ~1.3km from the nearest hospital on average with the furthest located ~5.1km away from the nearest market (Jurong West St 91)
- 50% of the transacted units are within ~1.1km distance to the nearest hospital
- Data from Sengkang and Yishun also show similar observations where units closer to Healthcare facilities are priced higher than units located further away

Data Visualization – Resale Price to School Distance

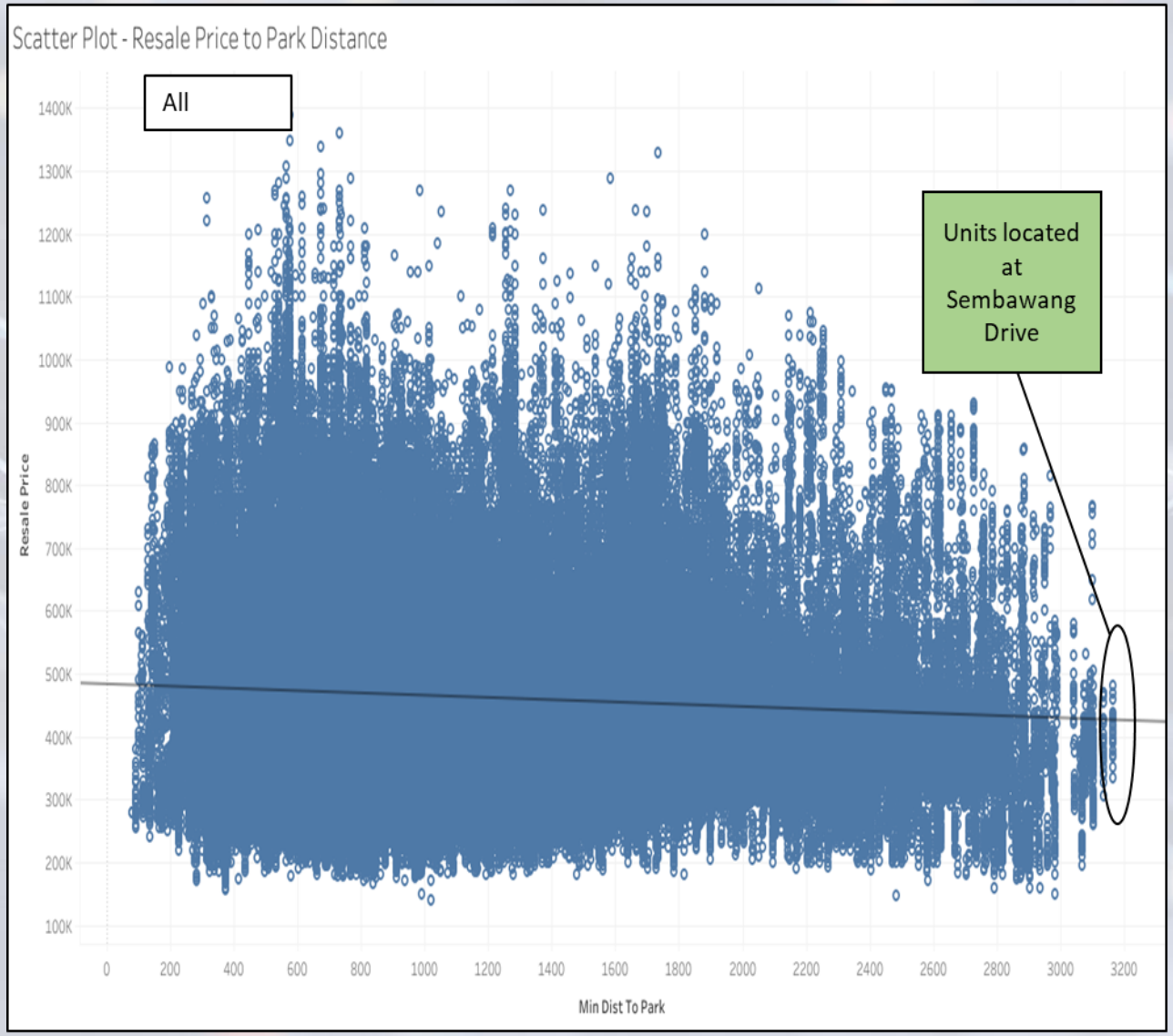


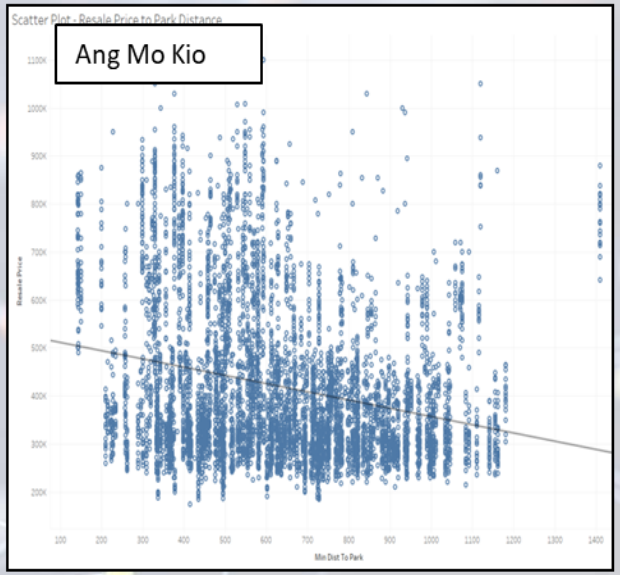


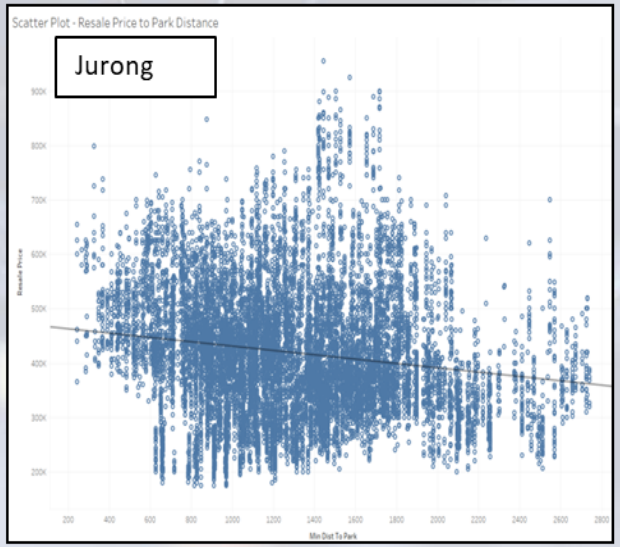


- Observation of the dataset shows that HDB resale units located nearer to schools transact at lower prices than units located further away.
- HDB units at Changi Village are located ~3.3km from the nearest school on average
- Excluding the Changi Village cluster, transacted units in the rest of the dataset are located 314m away from the nearest school on average with the furthest unit located 1.23km away (Upper Boon Keng area)
- Examining of Bishan and Tampines data shows differing observations. Units closer to schools in Bishan transact at higher prices than the ones located further away, while Tampines displays the opposite result

Data Visualization – Resale Price to Park Distance

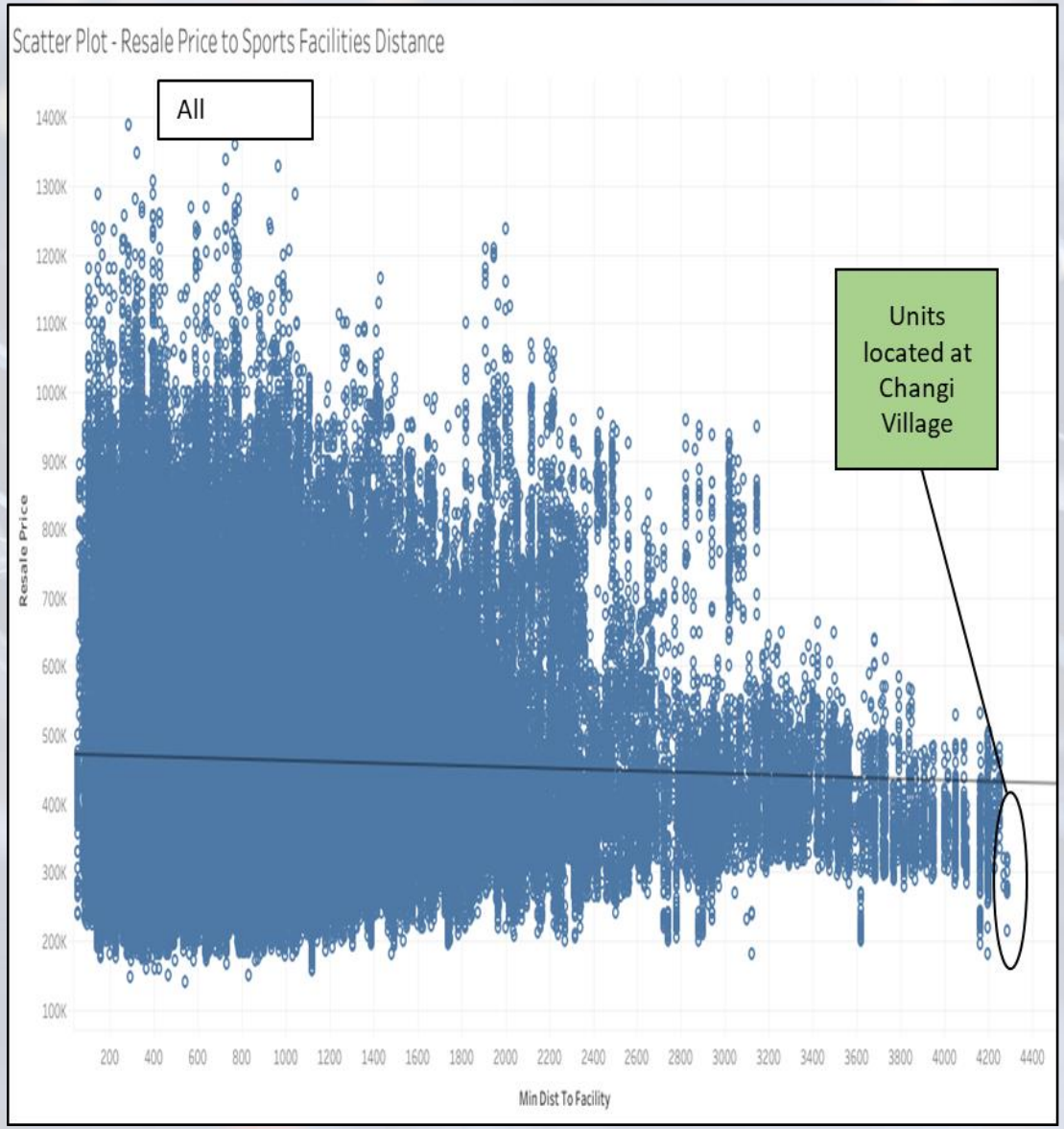


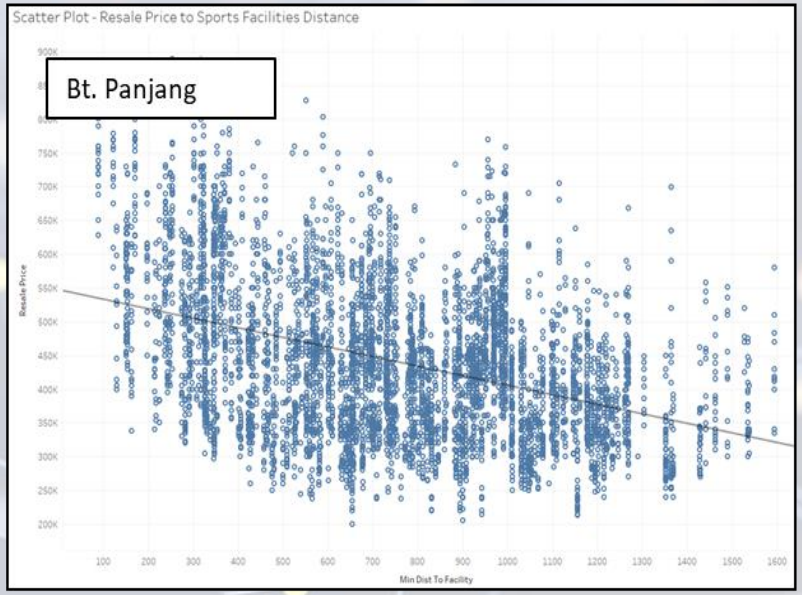


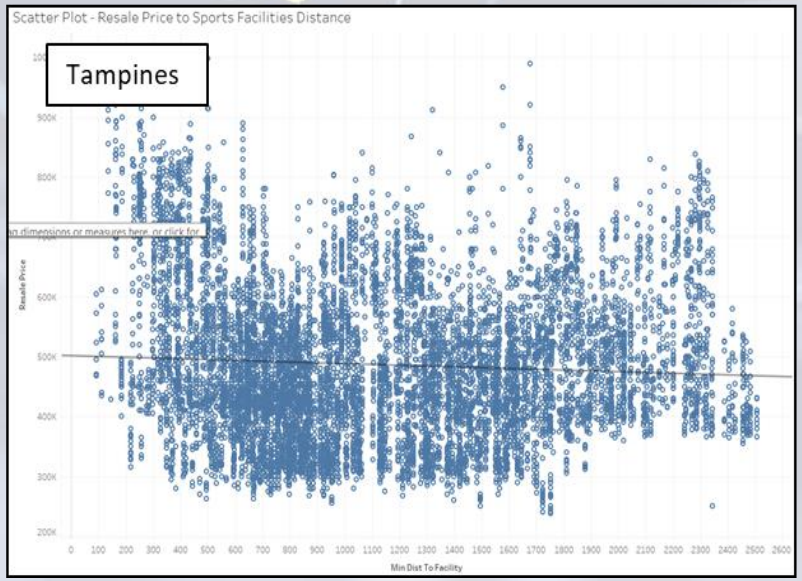


- Observation of the dataset suggests that HDB resale units located nearer to Parks attract higher prices than units located further away. Price influence by distance is does not appear to be significant.
- HDB units in the dataset are located ~1.2km from the nearest park on average with the furthest located ~3.2km away from the nearest market (Sembawang Drive area)
- 50% of the transacted units are within ~1.1km distance to the nearest park
- Further looking into data from Ang Mo Kio and Jurong also show similar observations

Data Visualization – Resale Price to Sports Facilities Distance





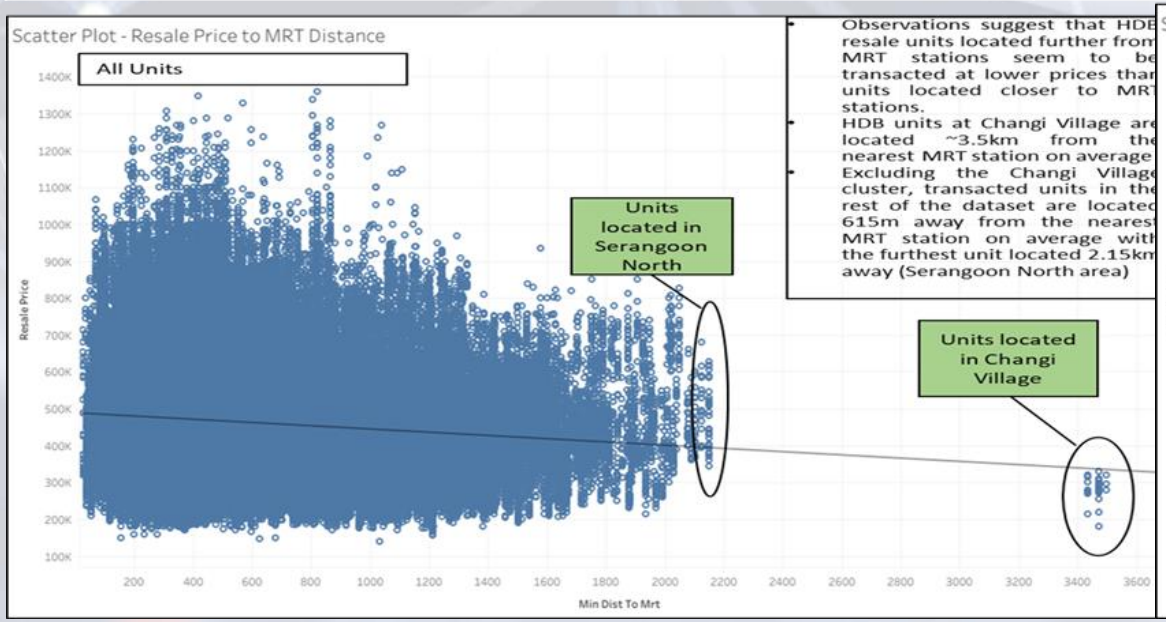


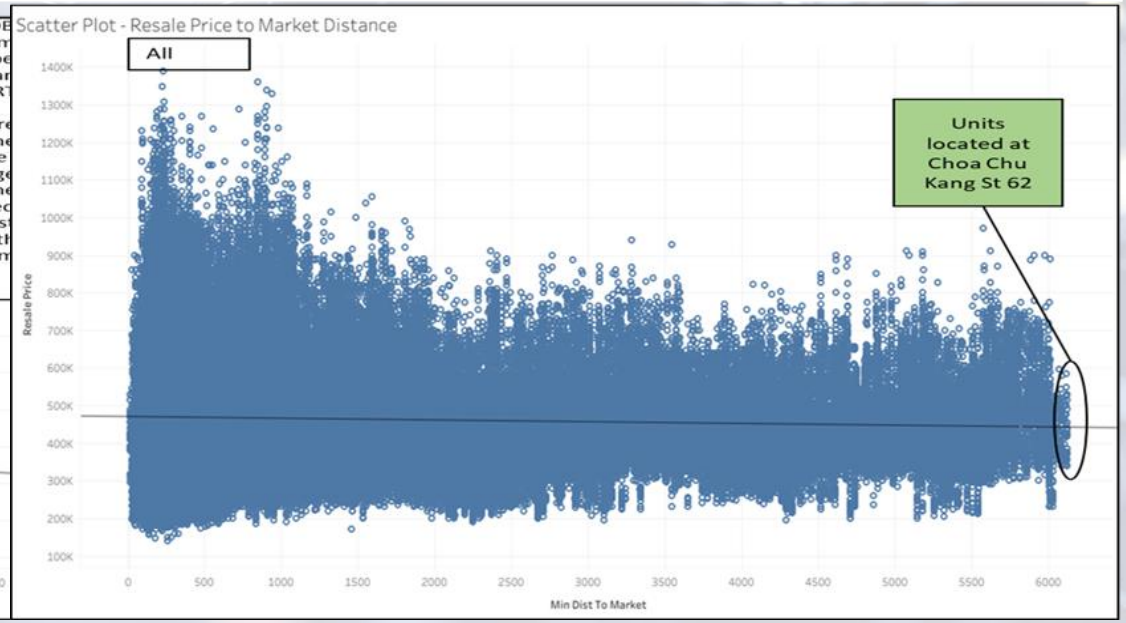
- Units located nearer to sports facilities are observed to attract slightly higher prices than ones that further away but price influence by distance is does not appear to be significant.
- HDB units in the dataset are located ~1.1km from the nearest sports facility on average with the furthest located ~4.2km away from the nearest sports facility (Changi Village area)
- 50% of the transacted units are within ~900m distance to the nearest sports facility
- Resale price and distance data from Bukit Panjang and Tampines also show similar observations where units nearer to sports facilities are priced higher than units further away.

Hypothesis to data visualization

Findings on our first hypothesis:

Hypothesis	Agree/ Disagree	Findings
Distance to the market / hawker center will have a higher impact on HDB flat resale prices than distance to the MRT.	Disagree	Based on our data, it appears that distance to MRT would affect the resale price more. As we can see bigger price fluctuation in the chart. As compared to distance to market/hawker, price fluctuation does not appear to be significant.

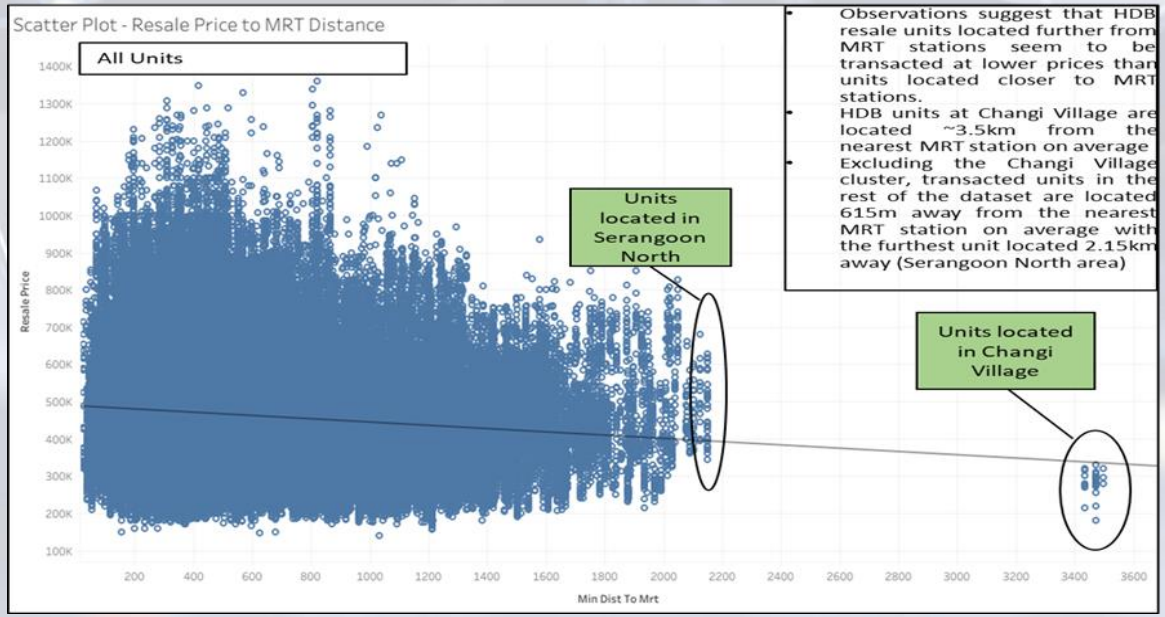


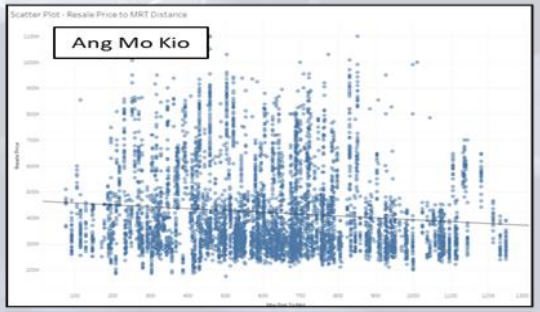


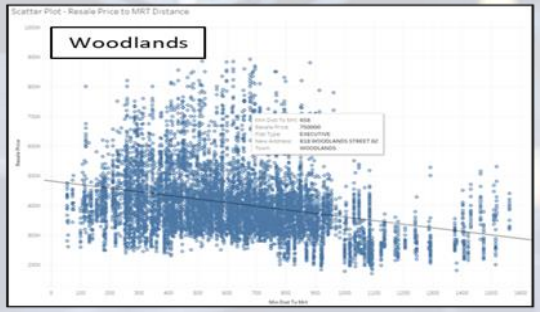
Hypothesis to data visualization

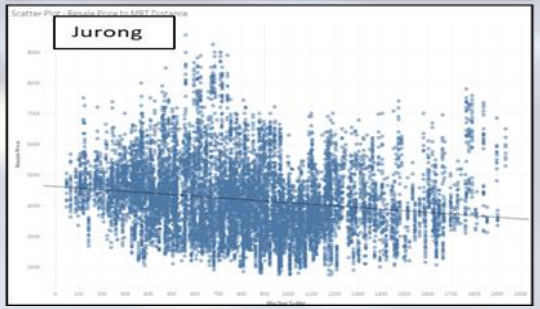
Findings on our Second hypothesis:

Hypothesis	Agree/ Disagree	Findings
HDB prices are positively affected with nearer accessibility to a MRT station	Agree	Based on our data, We could see that as distance to MRT got further, prices start to go lower. Thus, this could conclude that prices are positively affected with nearer accessibility to MRT station.







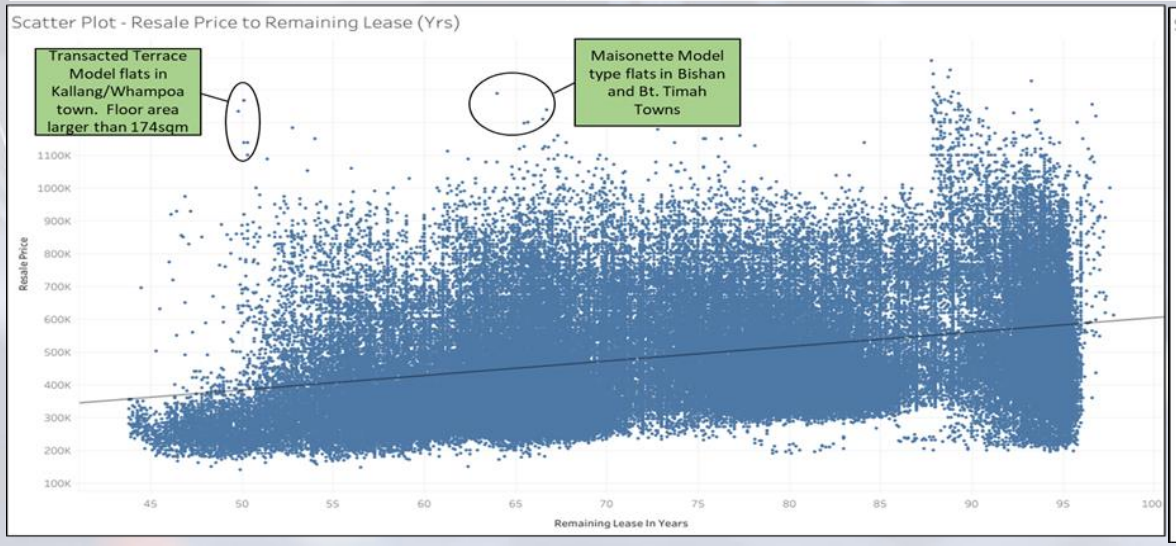


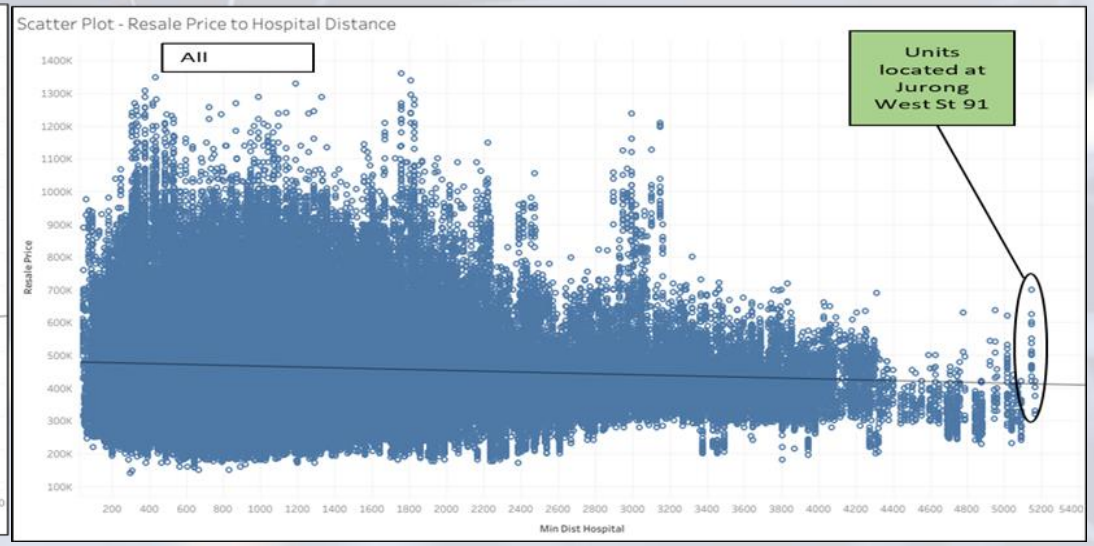
- Similar observations were made when we examine a sample of Townships where resale flats nearer to the MRT stations were priced higher that those located further away.
- Among the 3 Towns sampled, Woodlands Town's data seems to suggest the that distance to MRT station has the highest influence on resale flat prices.

Hypothesis to data visualization

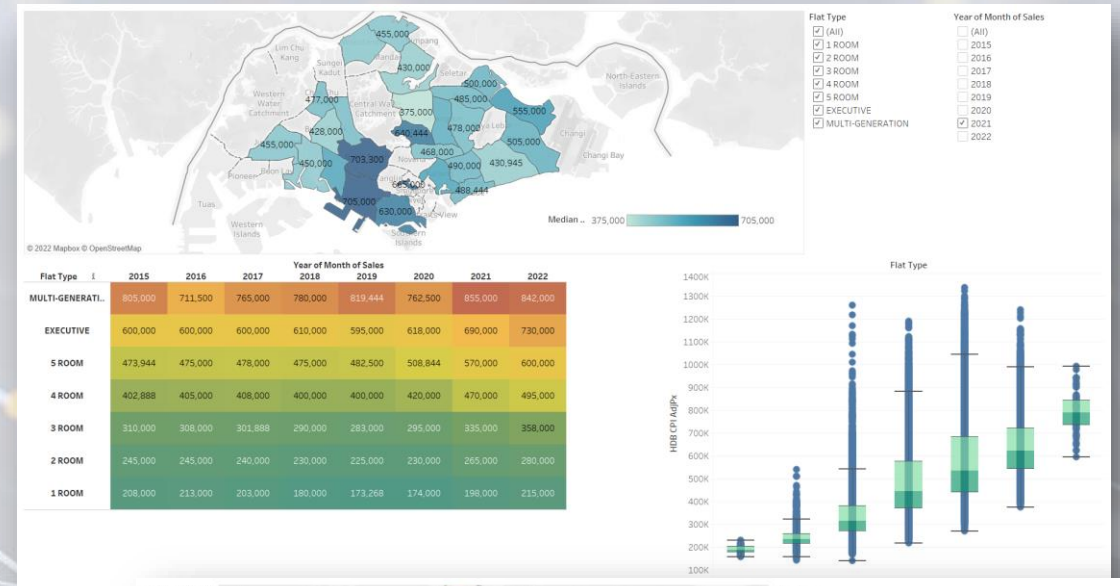
Findings on our Third hypothesis:

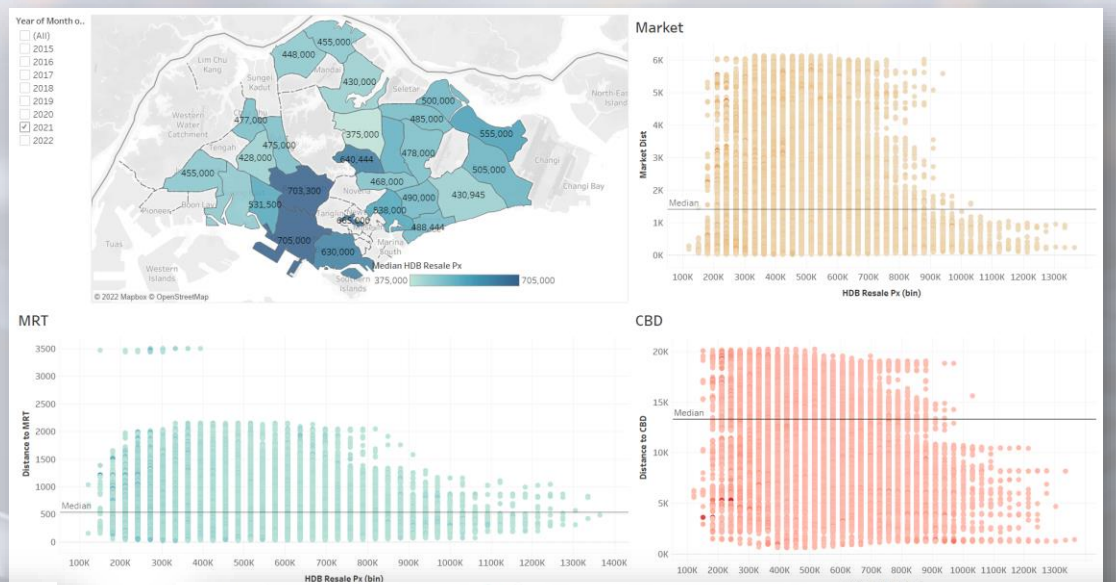
Hypothesis	Agree/ Disagree	Findings
HDB resale prices are negatively correlated to the age of the property and proximity to a hospitals.	Agree for remaining lease	Based on our data, we could see that it is negatively correlated because as remaining lease of HDB increases, the resale value increases as well.
	Disagree for Healthcare facilities	Based on our data, we could see that there's a positively correlation because there is a decrease in resale value as HDB is located further away from the healthcare facilities.

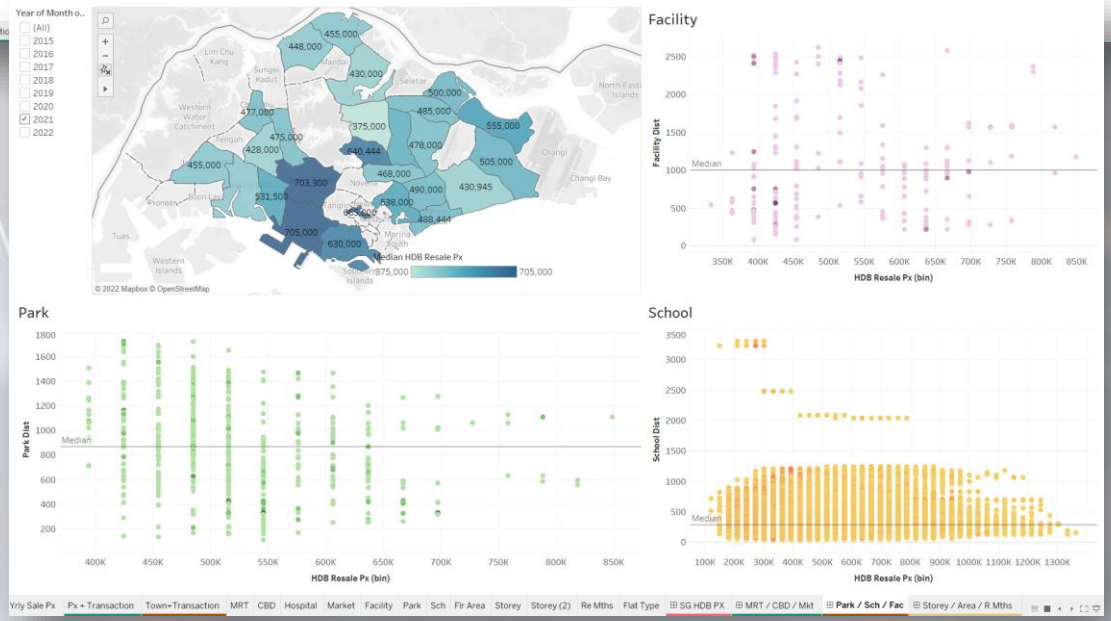


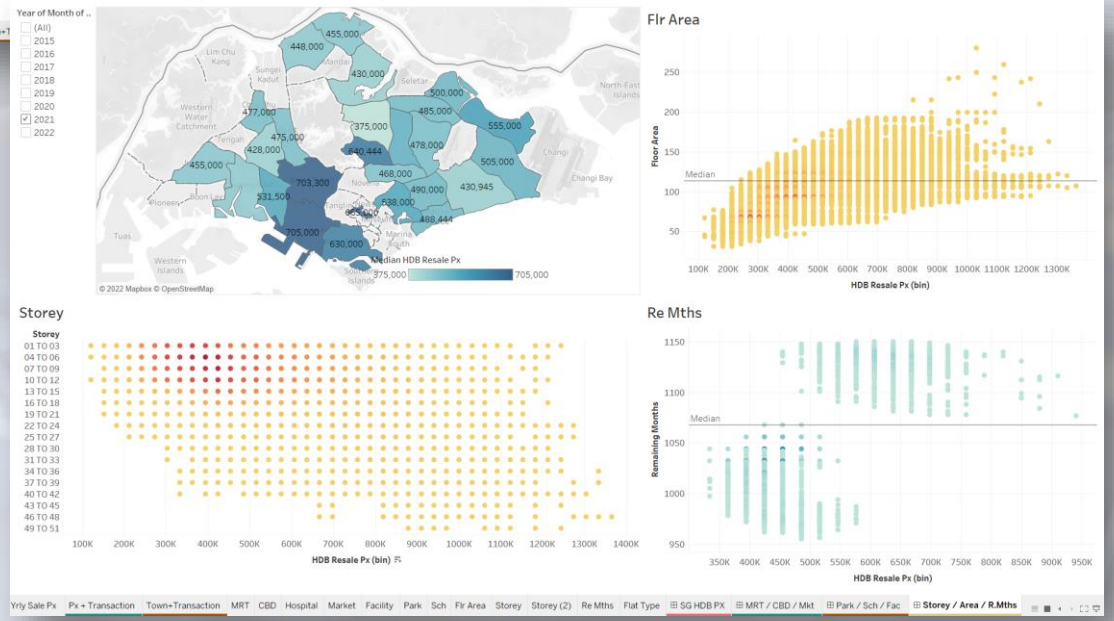


Dashboarding using Tableau









V. Predictive Analytics

Initial Issue (Snap boosting – 0.999)

Realised the prediction was far from expectation... …

No.	TEST DATA	ACTUAL	INITIAL AUTO AI
1	253 Ang Mo Kio Street 21, 560253, 10 to 12 floor, 138 sqm, 5 room	$750,000 (in Sept 2021)	$422,922 (-44.5%)
2	105 Ang Mo Kio Avenue 4, 560105, 04 to 06 floor, 92 sqm, -	$440,000 (in Jun 2021)	$421,615 (-43.6%)
3	310B Ang Mo Kio Avenue 1, 562310, 16 to 18 floor, 94 sqm, 4 room	$760,000 (in Mar 2022) $868,888 (in Feb 2022)	$421,615 (-4.2%)
	MSE ('1,000m)		**73.94**

Pipeline leaderboard

Rank	Name	Algorithm	RMSE (Optimized) Cross Validation	Enhancements	Build time
★ 1	Pipeline 4	Snap Boosting Machine Regressor	3573.191	HPO-1 FE HPO-2	00:04:01
2	Pipeline 3	Snap Boosting Machine Regressor	3596.407	HPO-1 FE	00:03:36

Measures	Holdout score	Cross validation score
Root mean squared error	3651.268	3573.191
R squared	0.999	0.999
Explained variance	0.999	0.999

Input list (3)

[Apr 2022, 253 Ang Mo Kio Street 21, 560253, null, 10 to 12, 138, 5 Room, null, null, null, null, null, null, null, null, null, null]

[Apr 2022, 105 Ang Mo Kio Avenue 4, 560105, null, 04 to 06, 92, null, null, null, null, null, null, null, null, null, null, null]

[apr 2022, 310B Ang Mo Kio Avenue 1, null, null, 16 to 18, 94, 4 Room, null, null, null, null, null, null, null, null, null, null]

redict (3)

Result


```
{
  "predictions": [
    {
      "fields": [
        "prediction"
      ],
      "values": [
        [
          422922.3587100832
        ],
        [
          421615.4006939126
        ],
        [
          421615.4006939126
        ]
      ]
    }
  ]
}
```


Feature name	Transformation	Feature importance
NewFeature_12 Most improved	divide(product(Park_Dist,HDB_CPI_AdjPx), Park_Dist)	100.00%
NewFeature_10	divide(product(Distance_to_MRT,HDB_CPI_AdjPx), Distance_to_MRT)	49.00%
NewFeature_7	divide(product(Floor_Area,HDB_CPI_AdjPx), Floor_Area)	46.00%
HDB_CPI_AdjPx	None	45.00%
NewFeature_14	divide(product(Remaining_Months,HDB_CPI_AdjPx), Remaining_Months)	13.00%
NewFeature_11	divide(product(School_Dist,HDB_CPI_AdjPx), School_Dist)	9.00%

Looking at the feature dependency, having high reliance on park distance is not an expected result. There could have been closer reliance on size of unit, age of unit.

Initial Issue – Signs of error (SPSS)

NIL Value

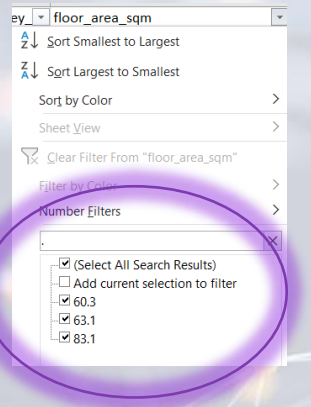
eried_blk_r	queried_road_name	postal_code
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL
356	YISHUN RING ROAD	NIL

Node:
HDB training

WDP Connector Error: CDICO9999E: Internal error occurred: SCAPI error: The following data for the postal_code column is invalid:NIL The data is invalid for the following reason: For input string: "NIL"

Execution was interrupted

Int + Dec in a Col



floor_area_sqm
60.3
60.3
60.3
60.3
60.3
60.3
83.1
63.1
63.1
63.1
83.1
63.1
63.1
63.1
63.1
63.1

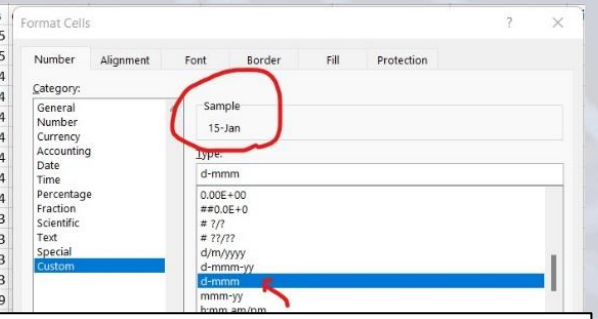
Node:
HDB training

WDP Connector Error: CDICO9999E: Internal error occurred: SCAPI error: The following data for the floor_area_sqm column is invalid:60.3 The data is invalid for the following reason: For input string: "60.3"

Execution was interrupted

Date Format

sold year	lease_commence_year	remaining_lease_in_years
15-Jan	2001	85
15-Aug	2001	85
15-Nov	2001	84
16-May	2001	84
16-May	2001	84
16-Jun	2001	84
16-Aug	2001	84
16-Sep	2001	84
16-Sep	2001	84
16-Oct	2001	83
16-Oct	2001	83
17-Jul	2001	83.3
17-Sep	2001	83
17-Oct	2001	82.9
17-Nov	2001	
18-Mar	2001	
18-May	2001	
18-May	2001	
18-Sep	2001	
18-Nov	2001	
18-Nov	2001	
19-Dec	2001	
20-Mar	2001	
20-Jun	2001	



Node:
Type

The storage class has changed for field: month (conversion applied)

Format Error

resale_price
$355,000.00
$440,000.00
$409,000.00
$356,000.00
$388,000.00
$360,000.00
$417,000.00
$365,000.00
$415,000.00
$393,000.00

resale_price
355000.00
440000.00
409000.00
356000.00
388000.00
360000.00
417000.00
365000.00
415000.00
393000.00

Node:
Type

The storage class has changed for field: resale_price (conversion applied)

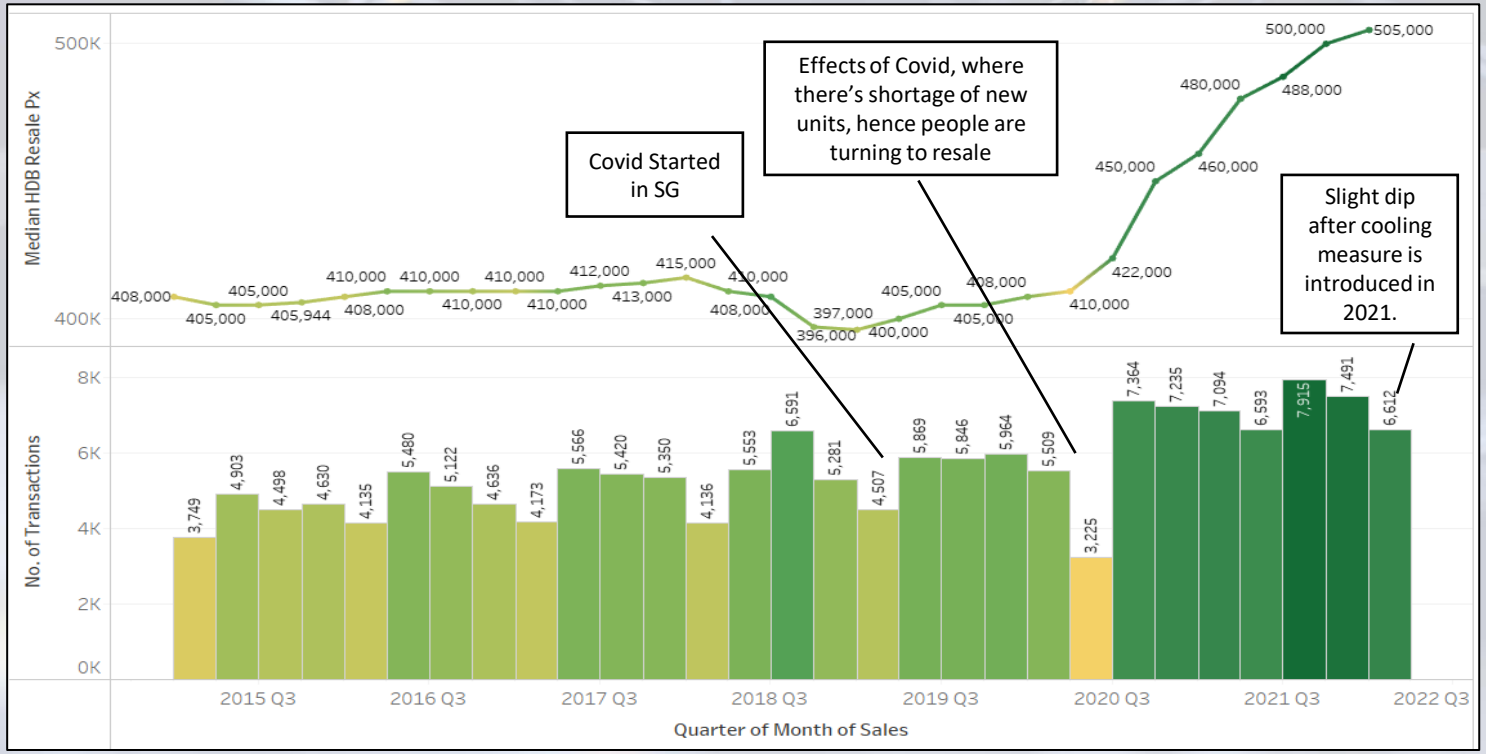
Node:
resale_price

No valid records found in the data source.

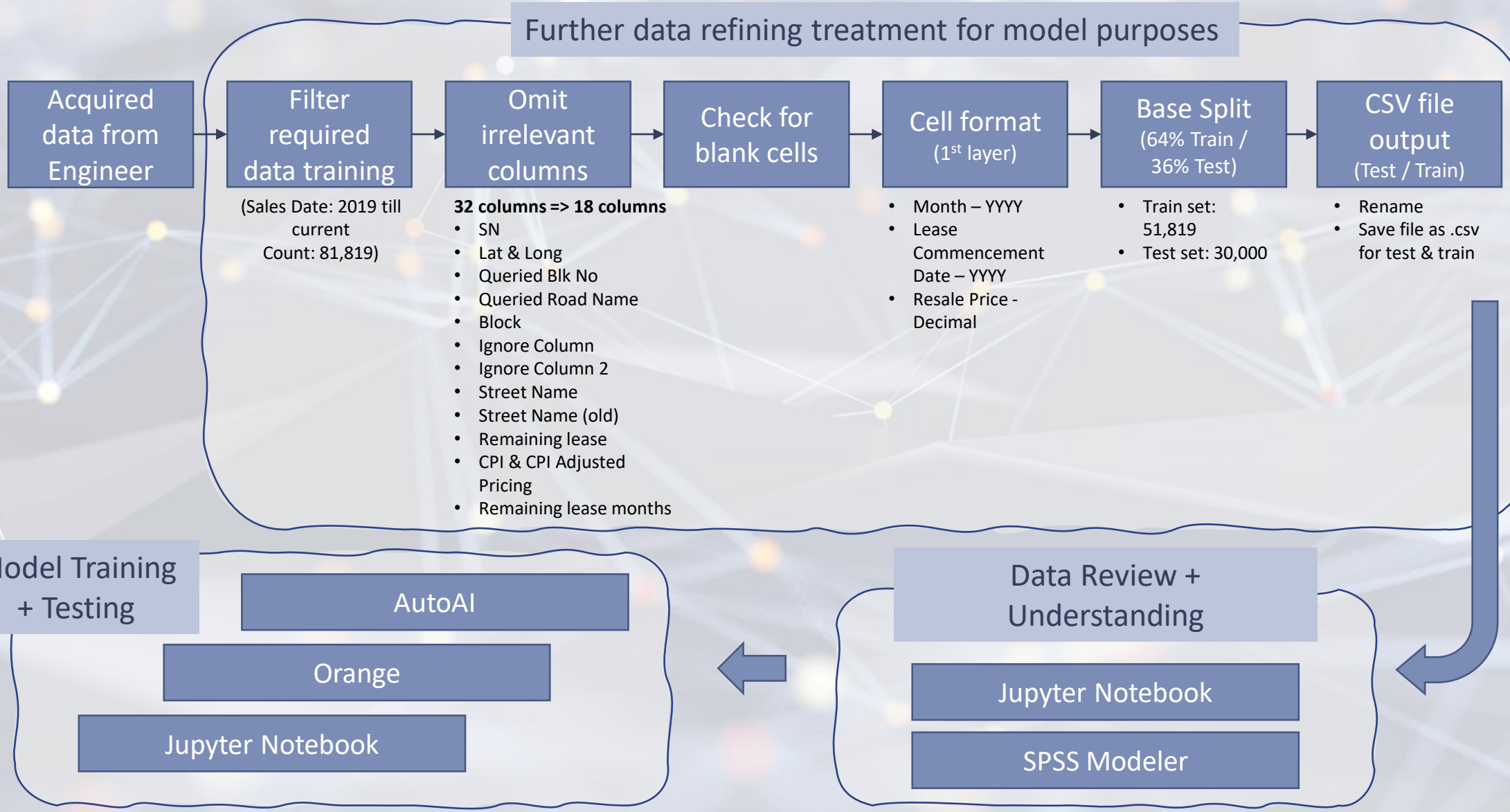
Execution was interrupted

Improvements

1. **For NIL values** – discovered 22 rows of data with “NIL” values for postal code, street name and or block number.
2. **Address** - Ensure fetched latitudes and longitudes are correct – discovered that the script used, replaced some street names with closest match. This resulted in the dataset having addresses of landed properties as well as inaccurate latitudes & longitudes. Which further resulted in providing the wrong distance for some address (e.g. Whampoa units were found in Changi area).
3. Sieving out only sales data from **2019 till Mar 2022** for training of machine.



Getting the Dataset Ready

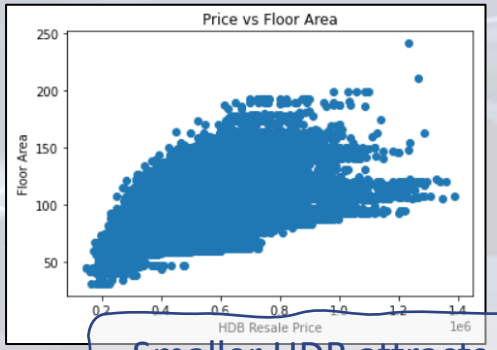


Further Understanding of Data Before ML

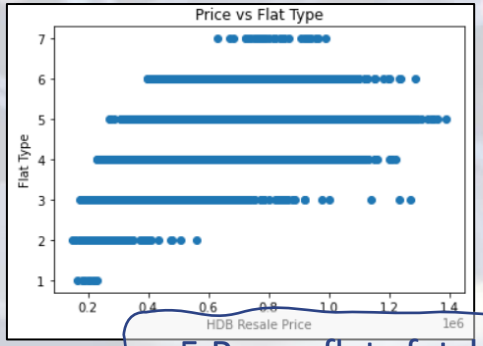
Before running more models using different tools, we did a quick co-relation analysis on dependencies which may affect the resale price of HDB, as well as to analyse some illustrations using Jupyter Notebook.

	Resale Price	CBD	MRT	School	Hospital	Park	Market	Facility	Storey_Median	Remaining Years	Floor Area
Resale Price	1	-0.26	-0.12	0.02	-0.07	-0.07	0.019	0.039	0.37	0.34	**0.62**
CBD	-0.26	1	0.087	-0.22	**0.43**	0.15	**0.5**	0.26	-0.17	0.32	0.24
MRT	-0.12	0.087	1	0.098	-0.04	0.054	-0.3	0.03	0.092	-0.12	0.035
School	0.02	-0.22	0.098	1	0.059	0.037	-0.24	0.022	0.026	-0.12	-0.12
Hospital	-0.07	**0.43**	-0.04	0.059	1	0.4	0.28	0.41	0.027	0.23	0.12
Park	-0.07	0.15	0.054	0.037	0.4	1	0.054	0.31	0.015	0.011	0.065
Market	0.019	**0.5**	-0.3	-0.24	0.28	0.054	1	0.2	0.018	**0.57**	0.18
Facility	0.039	0.26	0.03	0.022	0.41	0.31	0.2	1	0.047	0.16	0.081
Storey_Median	0.37	-0.17	0.092	0.026	0.027	0.015	0.018	0.047	1	0.27	-0.03
Remaining Years	0.34	0.32	-0.12	-0.12	0.23	0.011	**0.57**	0.16	0.27	1	0.1
Floor Area	**0.62**	0.24	0.035	-0.12	0.12	0.065	0.18	0.081	-0.03	0.1	1

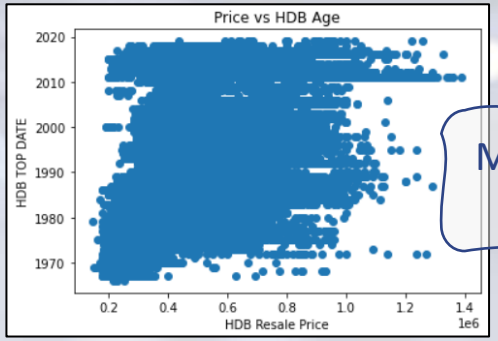
We may expect that floor area, remaining lease years, Storey and CDB could affect the resale values of HDB.



Smaller HDB attracts lesser price.



5 Room flats fetches the highest price



Majority of the older flats may struggle to fetch high prices

Quick Linear Regression Test (Jupyter)

Comparing previous test results vs scrubbed dataset

No.	TEST DATA	ACTUAL	INITIAL AUTO AI	QUICK TEST JUPYTER (LR-S)
1	253 Ang Mo Kio Street 21, 560253, 10 to 12 floor, 138 sqm, 5 room	$750,000 (in Sept 2021)	$422,922 (-44.5%)	$718,389 (-4.2%)
2	105 Ang Mo Kio Avenue 4, 560105, 04 to 06 floor, 92 sqm, -	$440,000 (in Jun 2021)	$421,615 (-43.6%)	$413,120 (-6.1%)
3	310B Ang Mo Kio Avenue 1, 562310, 16 to 18 floor, 94 sqm, 4 room	$760,000 (in Mar 2022)	$421,615 (-4.2%)	$631,431 (-16.9%)
	MSE ('1,000m)		**73.94**	**6.08**

```
In [16]: #Jupyter - Linear Short Set
         #Specify which columns for input variables
         X = dataset.iloc[:, :7]

         #Specify which column is the target or predicted column
         y = dataset.iloc[:, -1]

         #Training model using linear regression
         from sklearn.linear_model import LinearRegression
         regressor = LinearRegression()
         regressor.fit(X, y)

         #1 = 1 ROOM
         #2 = 2 ROOM
         #3 = 3 ROOM
         #4 = 4 ROOM
         #5 = 5 ROOM
         #6 = EXECUTIVE
         #7 = MULTI-GENERATIO
         #Sales Year, postal code, Room type, Floor from, Floor to, square area, TOP year

         #253 ANG MO KIO STREET 21, $750000 in 2021
         print(regressor.predict([[2022,560253,5,10,12,138,1996]]))

         [718389.54264478]

In [17]: #105 Ang Mo Kio Ave 4, $440k in Jun 2021
         print(regressor.predict([[2022,560105,4,4,6,92,1978]]))

         [413120.83800921]

In [18]: #310B ANG MO KIO AVENUE 1, $760000 in 2022
         print(regressor.predict([[2022,562310,4,16,18,94,2012]]))

         [631431.08309187]
```

- The results yields a better prediction as compared to the 1st attempt.
- ***With this, we are confident to proceed with more machine learning with different tools.***

Linear Regression Model (Jupyter)

```
In [36]: X = dataset.drop(['resale_price'], axis = 1)

In [37]: X

Out[37]:
```

	month	postal_code	flat_type	storey_from	storey_to	floor_area_sqm	lease_commence_date	remaining_lease_in_years	dis_to_cbd	min_dist_to_mrt
0	2019	530122	3	1						
1	2019	530122	5	1						
2	2019	530951	5	13						
3	2019	600021	4	4						
4	2019	600020	4	1						
...	...	...	...	...						
51814	2021	560116	2	4						
51815	2021	560225	3	10						
51816	2021	560219	3	4						
51817	2021	560219	3	10						
51818	2021	560330	3	10						

51819 rows × 15 columns

```
In [38]: Y = dataset['resale_price']

In [39]: Y

Out[39]: 0        260000.0
         1        710000.0
         2        510000.0
         3        382000.0
         4        380000.0
                    ...
         51814    230000.0
         51815    365000.0
         51816    305000.0
         51817    292000.0
         51818    320000.0
         Name: resale_price, Length: 51819, dtype: floa

In [40]: X.shape

Out[40]: (51819, 15)
```

```
In [42]: from sklearn.preprocessing import MinMaxScaler

         scaler = MinMaxScaler()
         X_scaled = scaler.fit_transform(X)

In [43]: X_scaled

Out[43]: array([[0.        , 0.61935446, 0.33333333, ..., 0.02726842, 0.06961268,
                 0.30280808],
                [0.        , 0.61935446, 0.66666667, ..., 0.02726842, 0.06961268,
                 0.30280808],
                [0.        , 0.62042387, 0.66666667, ..., 0.34461695, 0.29462381,
                 0.25081614],
                ...,
                [0.66666667, 0.65817973, 0.33333333, ..., 0.27651617, 0.0214236 ,
                 0.01008896],
                [0.66666667, 0.65817973, 0.33333333, ..., 0.27651617, 0.0214236 ,
                 0.01008896],
                [0.66666667, 0.65832292, 0.33333333, ..., 0.10012631, 0.05765642,
                 0.0247388 ]])

In [44]: scaler.data_max_

Out[44]: array([2.02200000e+03, 8.25195000e+05, 7.00000000e+00, 4.90000000e+01,
                5.10000000e+01, 2.41000000e+02, 2.01900000e+03, 9.78000000e+01,
                2.02251040e+04, 3.49640281e+03, 3.30250938e+03, 5.16411062e+03,
                3.16670325e+03, 6.12704905e+03, 4.28397253e+03])

In [45]: scaler.data_min_

Out[45]: array([2.01900000e+03, 5.00040000e+04, 1.00000000e+00, 1.00000000e+00,
                3.00000000e+00, 3.10000000e+01, 1.96600000e+03, 4.38000000e+01,
                5.92122414e+02, 2.26058811e+01, 3.87326699e+01, 3.85527644e+01,
                8.95880029e+01, 1.23958663e+01, 4.45731670e+01])

In [46]: print(X_scaled[:,0])

         [0.         0.         0.         ... 0.66666667 0.66666667 0.66666667]

In [47]: y = y.values.reshape(-1,1)

In [48]: y.shape

Out[48]: (51819, 1)

In [49]: y_scaled = scaler.fit_transform(y)
```

```
In [51]: X_train, X_test , y_train , y_test = train_test_split(X , y , test_size = 0.1)

In [52]: X_train.shape

Out[52]: (46637, 15)

In [53]: X_test.shape

Out[53]: (5182, 15)

In [54]: model = LinearRegression()
         model.fit(X_train , y_train)

Out[54]: LinearRegression()

In [57]: y_pred = model.predict(X_test)
         y_pred

Out[57]: array([[489374.41364466],
                [336287.3946487 ],
                [179540.76113561],
                ...,
                [832846.04201829],
                [672800.72999503],
                [475171.2649025 ]])

In [58]: y_test

Out[58]: array([[500000.],
                [345000.],
                [248000.],
                ...,
                [935000.],
                [650000.],
                [500000.]])

In [61]: int_array = y_pred. astype(float)
         y_pred

Out[61]: array([[489374.41364466],
                [336287.3946487 ],
                [179540.76113561],
                ...,
                [832846.04201829],
                [672800.72999503],
                [475171.2649025 ]])
```

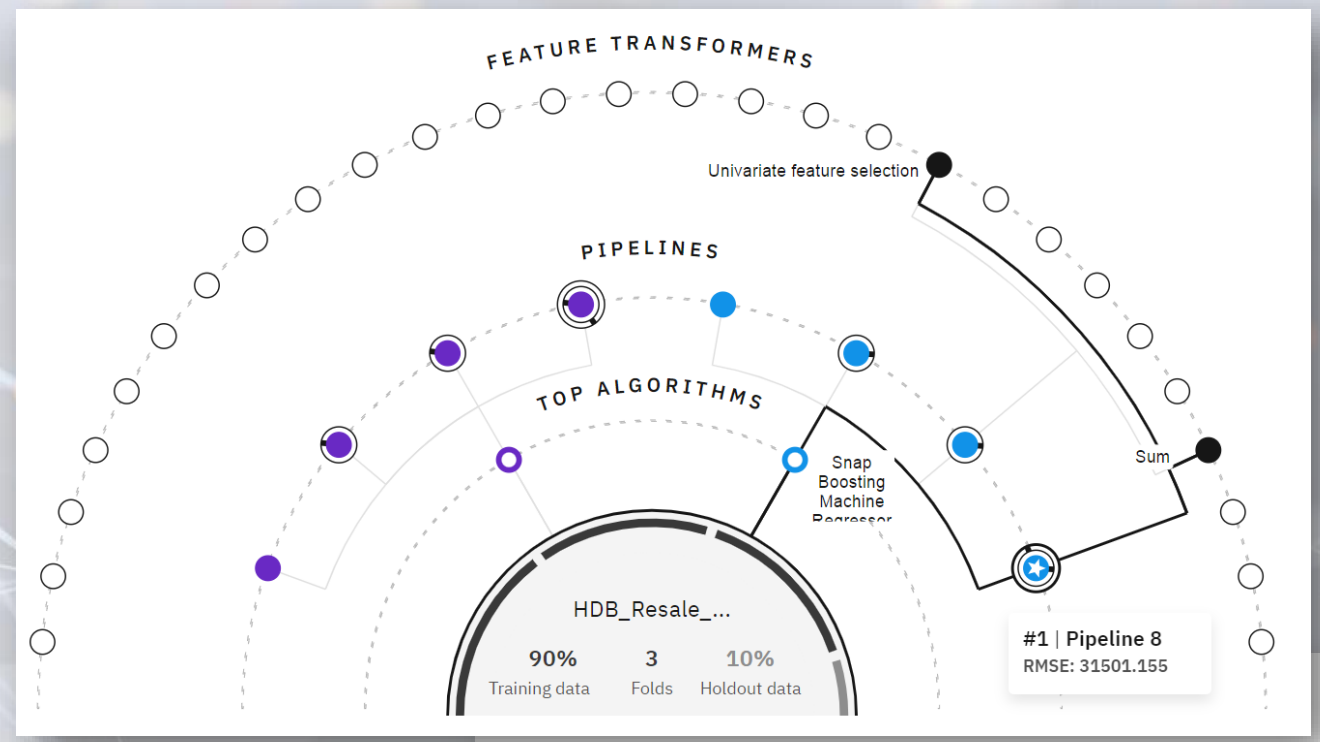
```
In [62]: int_array = y_test. astype(float)
         y_test

Out[62]: array([[500000.],
                [345000.],
                [248000.],
                ...,
                [935000.],
                [650000.],
                [500000.]])

In [63]: metrics.r2_score(y_test , y_pred)

Out[63]: 0.8411109808520748
```

Test on 90%; Holdout at 10%

Using Linear Regression Model

R2 score = 0.84

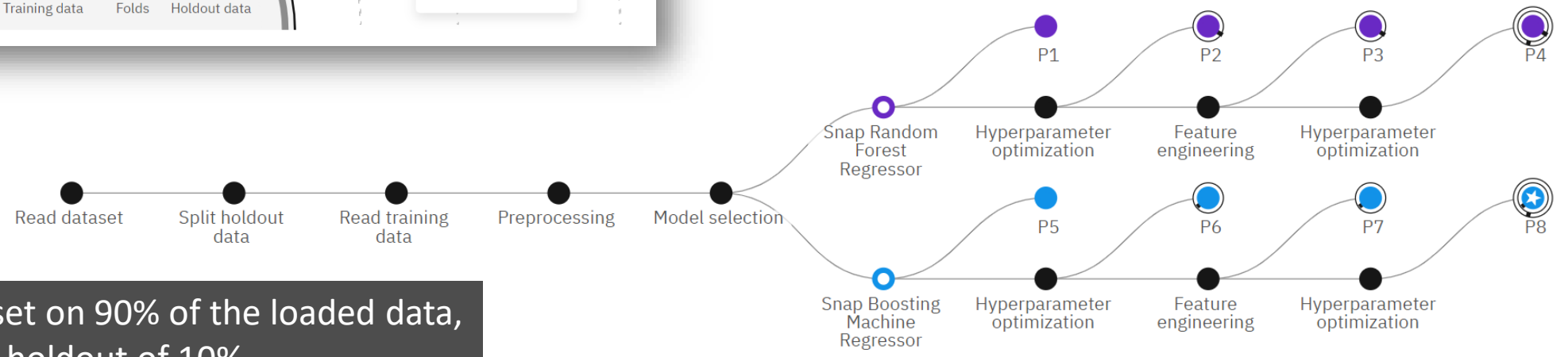
AutoAI – Model Training



Model evaluation

Model evaluation measure

Measures	Holdout score	Cross validation score
Root mean squared error	30782.419	31501.155
R squared	0.963	0.963
Explained variance	0.963	0.963
Mean squared error	947557325.599	992512333.297
Mean squared log error	0.004	0.004
Mean absolute error	22374.977	23003.625
Median absolute error	17001.123	17394.252
Root mean squared log error	0.063	0.063



Set to have training set on 90% of the loaded data, system holdout of 10%.

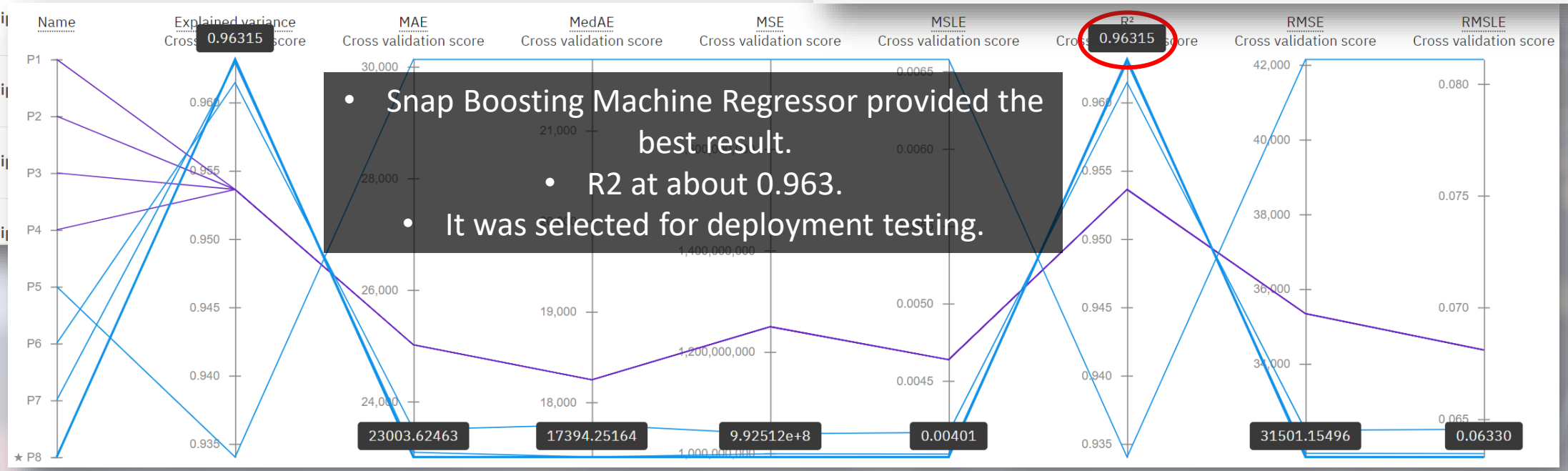
AutoAI – Model Training

	Rank	Name	Algorithm	RMSE (Optimized) Cross Validation	Enhancements
★	1	Pipeline 8	Snap Boosting Machine Regressor	31501.155	HPO-1 FE HPO-2
	2	Pipeline 7	Snap Boosting Machine Regressor	31603.529	HPO-1 FE
	3	Pipeline 6	Snap Boosting Machine Regressor	32219.134	HPO-1
	4	Pipeline 1	Snap Random Forest Regressor	35339.630	*None*
	5	Pi			
	6	Pi			
	7	Pi			
	8	Pi			

Model evaluation

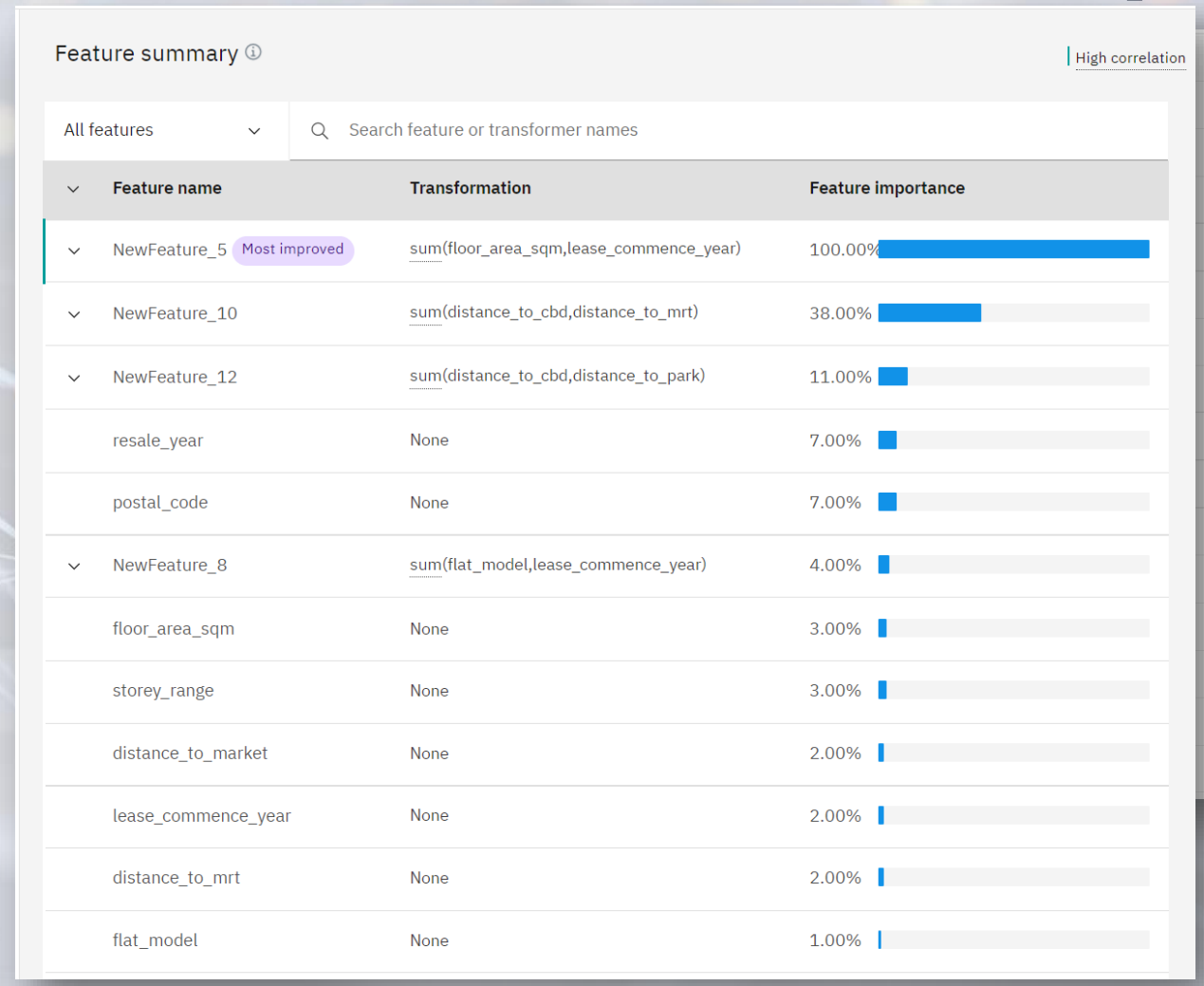
Model evaluation measure

Measures	Holdout score	Cross validation score
Root mean squared error	30782.419	31501.155
R squared	0.963	0.963
Explained variance	0.963	0.963
Mean squared error	947557325.599	992512333.297
Mean squared log error	0.004	0.004
Mean absolute error	22374.977	23003.625
Median absolute error	17001.123	17394.252
Root mean squared log error	0.063	0.063



- Snap Boosting Machine Regressor provided the best result.
- R2 at about 0.963.
- It was selected for deployment testing.

AutoAI – Model Dependencies



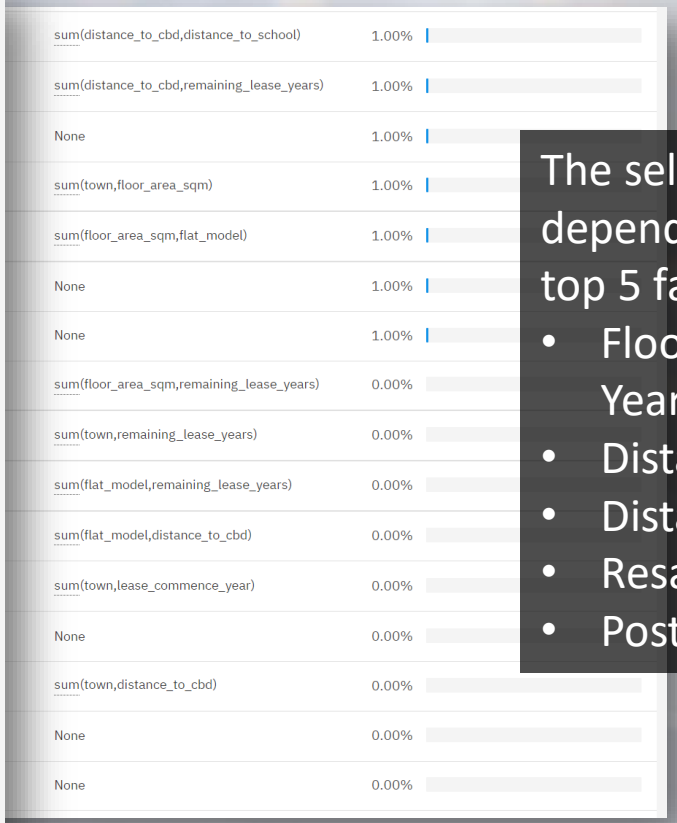
Feature summary

High correlation

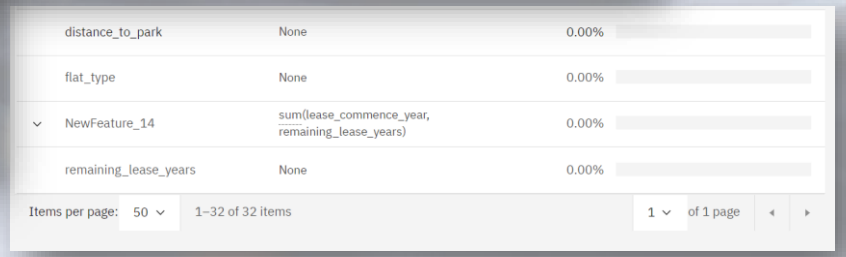
All features

Search feature or transformer names

Feature name	Transformation	Feature importance
NewFeature_5 Most improved	sum(floor_area_sqm,lease_commence_year)	100.00%
NewFeature_10	sum(distance_to_cbd,distance_to_mrt)	38.00%
NewFeature_12	sum(distance_to_cbd,distance_to_park)	11.00%
resale_year	None	7.00%
postal_code	None	7.00%
NewFeature_8	sum(flat_model,lease_commence_year)	4.00%
floor_area_sqm	None	3.00%
storey_range	None	3.00%
distance_to_market	None	2.00%
lease_commence_year	None	2.00%
distance_to_mrt	None	2.00%
flat_model	None	1.00%



sum(distance_to_cbd,distance_to_school)		1.00%
sum(distance_to_cbd,remaining_lease_years)		1.00%
None		1.00%
sum(town,floor_area_sqm)		1.00%
sum(floor_area_sqm,flat_model)		1.00%
None		1.00%
None		1.00%
sum(floor_area_sqm,remaining_lease_years)		0.00%
sum(town,remaining_lease_years)		0.00%
sum(flat_model,remaining_lease_years)		0.00%
sum(flat_model,distance_to_cbd)		0.00%
sum(town,lease_commence_year)		0.00%
None		0.00%
sum(town,distance_to_cbd)		0.00%
None		0.00%
None		0.00%



distance_to_park	None	0.00%
flat_type	None	0.00%
NewFeature_14	sum(lease_commence_year, remaining_lease_years)	0.00%
remaining_lease_years	None	0.00%

Items per page: 50 1–32 of 32 items 1 of 1 page

The selected model has high dependencies on the following top 5 factors for price prediction:

- Floor area, Lease Commence Year (100%)
- Distance to CBD + MRT (38%)
- Distance to CBD + Park (11%)
- Resale Year (7%)
- Postal Code (7%)

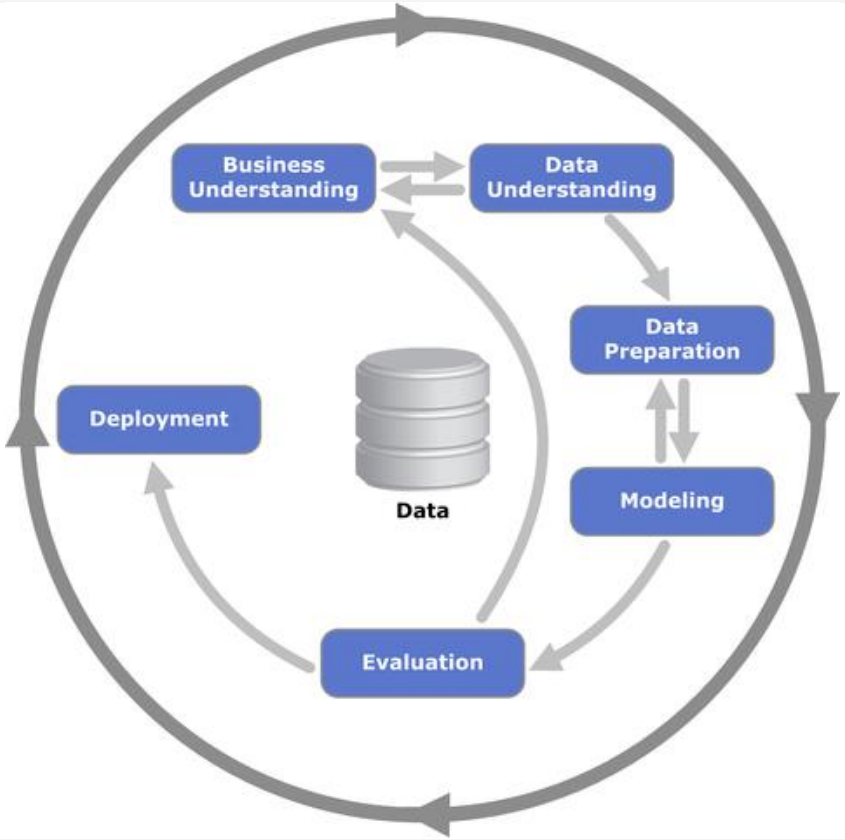
Machine Learning Builders

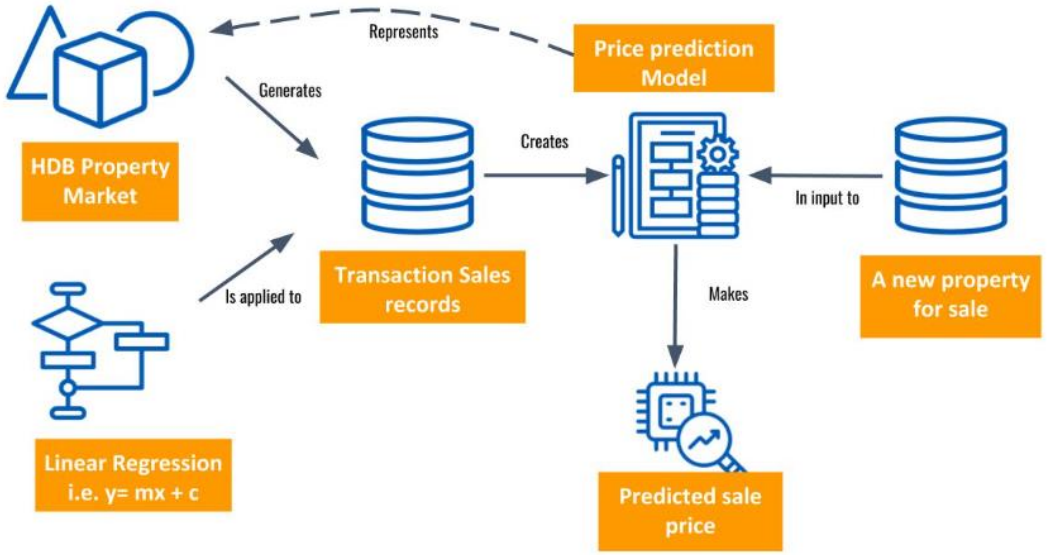
Predict the HDB flat resale price based on its characteristics

- **Data Scientists** use a training set - historical data in which the outcome of interest is known - to develop predictive models using the analytic approach. The modelling process is highly iterative.

Machine Learning Builders

Build a model that can calculate the HDB flat price based on historical data

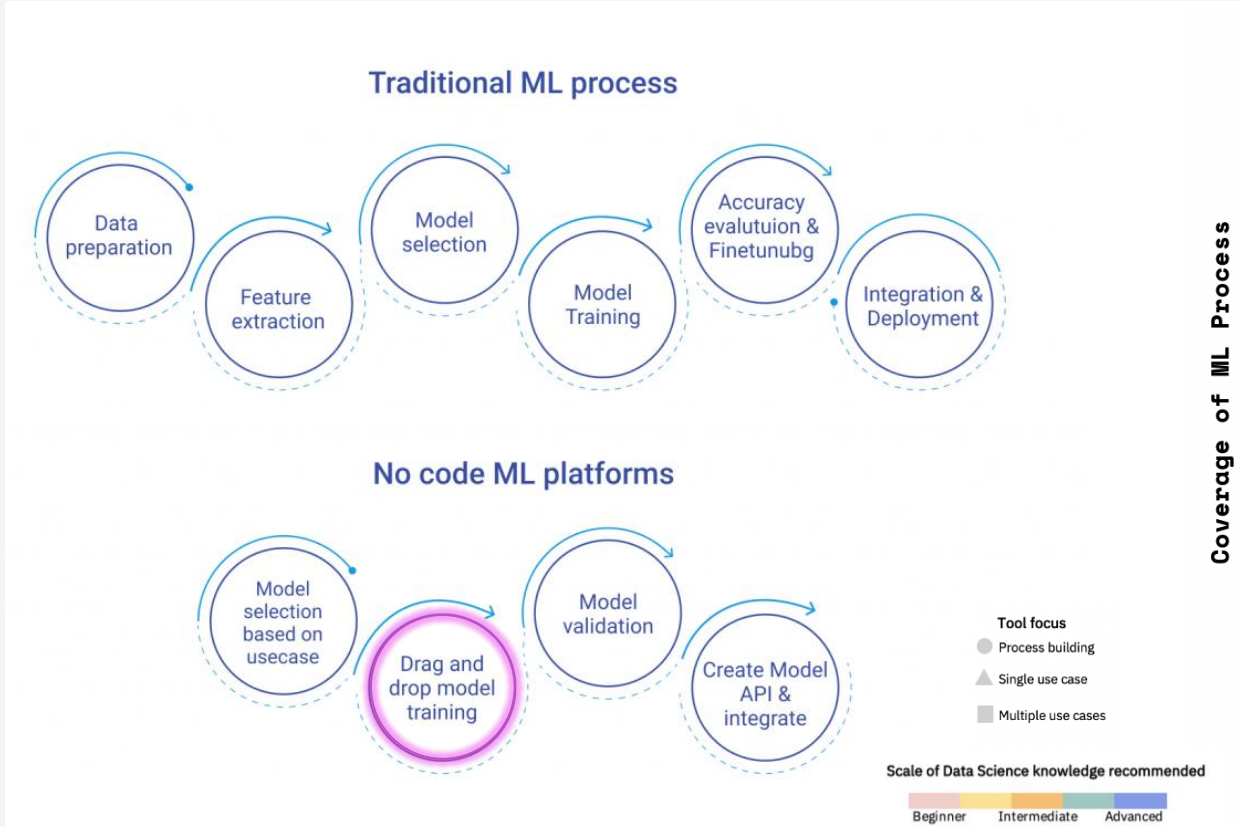




Any organization that uses predictive analysis will have a portfolio of models that needs to be governed over time. As data and business contexts change, **models will deteriorate and need to be re-calibrated**.

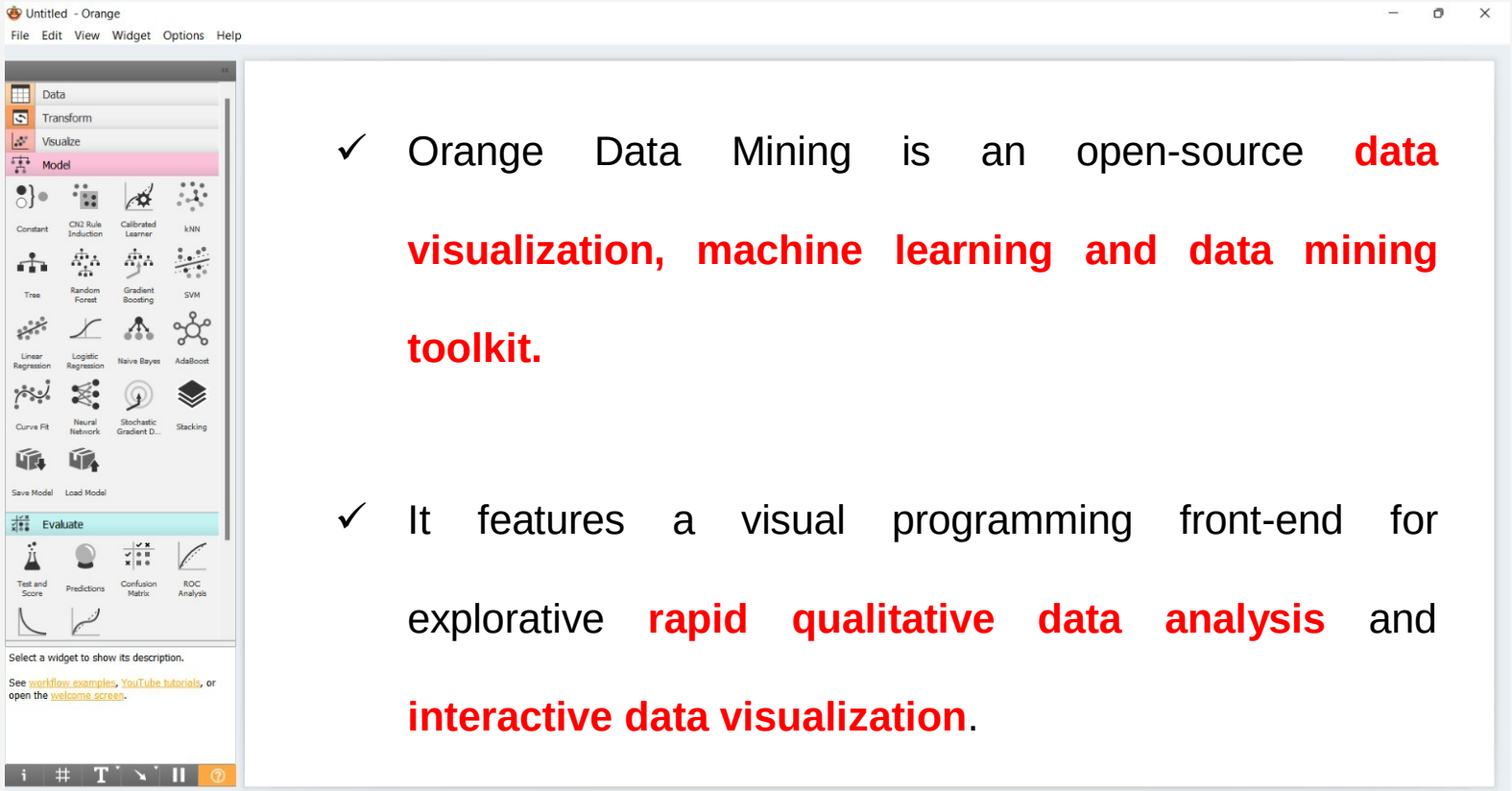
Machine Learning Builders

Traditional Vs “No Code” ML platforms

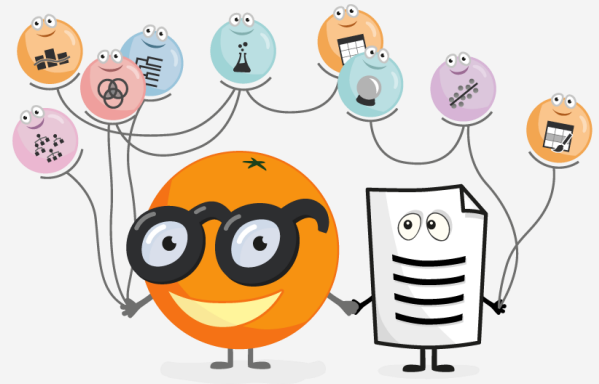


Machine Learning Builders

Orange Data Mining (Graphical ML Builder)



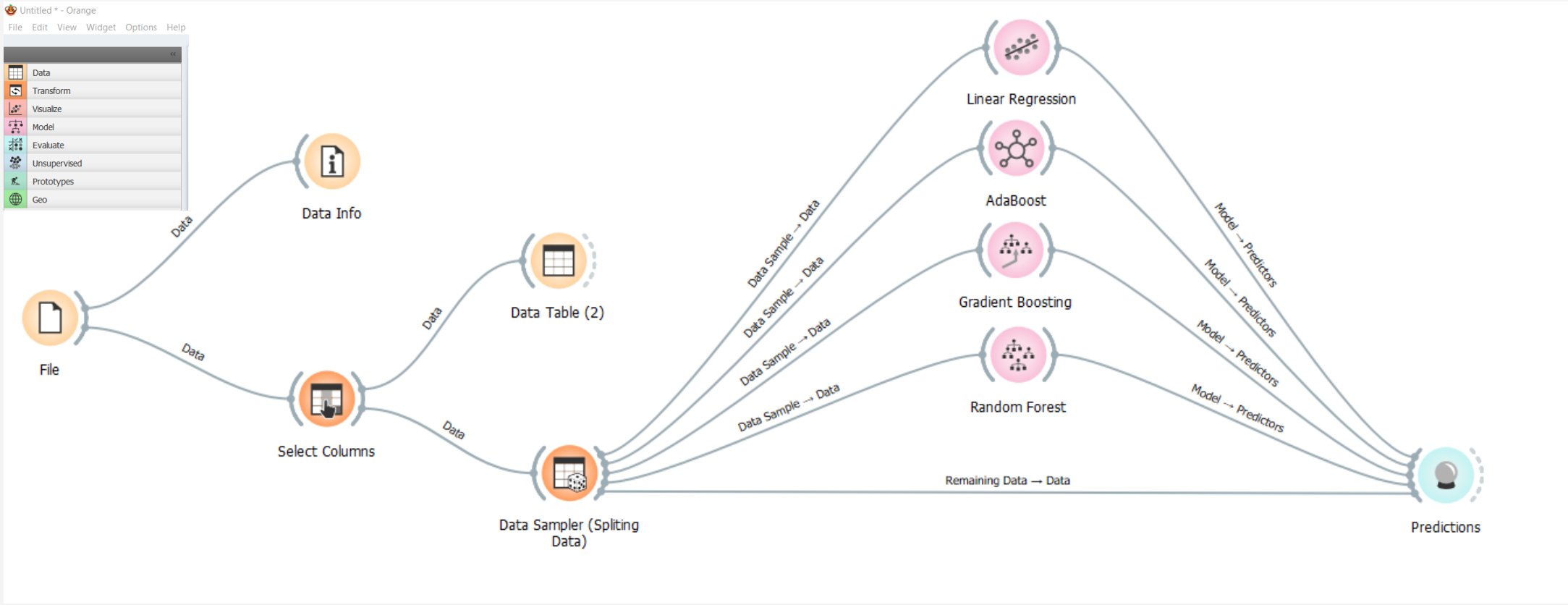
- ✓ Orange Data Mining is an open-source **data visualization, machine learning and data mining toolkit.**
- ✓ It features a visual programming front-end for explorative **rapid qualitative data analysis** and **interactive data visualization**.



No code approach

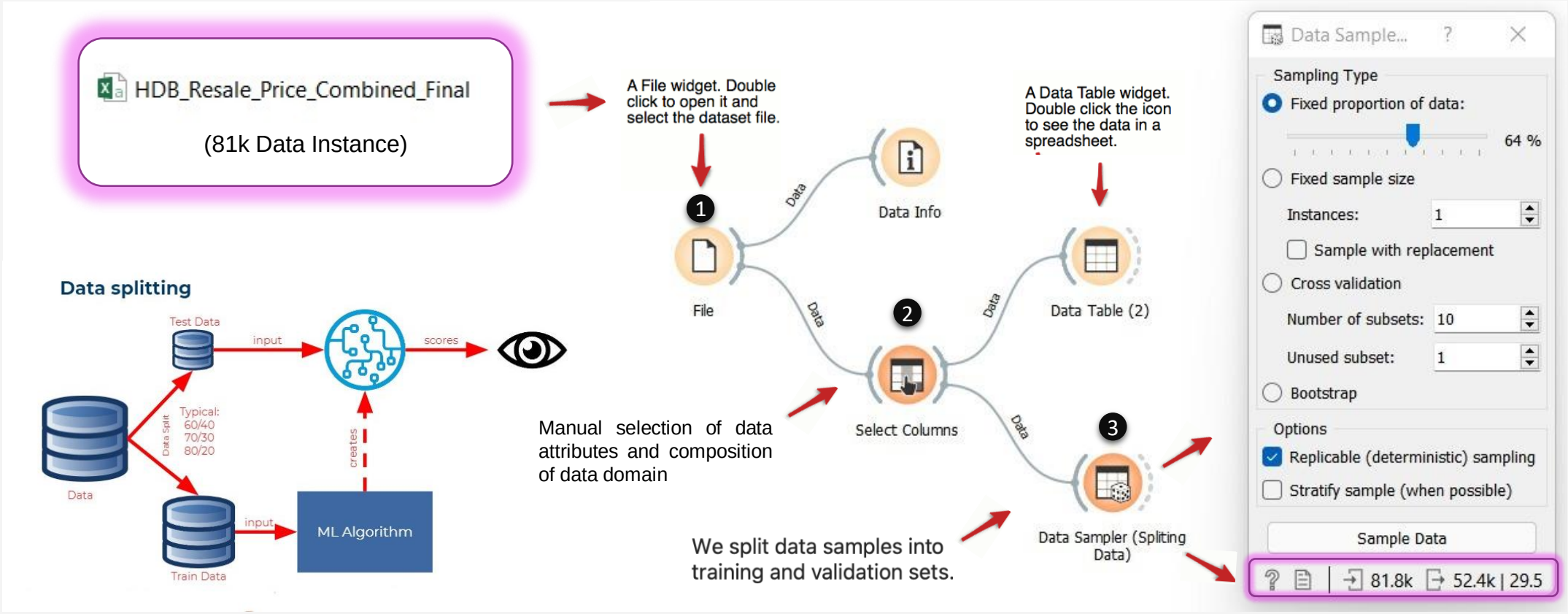
Orange Data Mining

Architecture



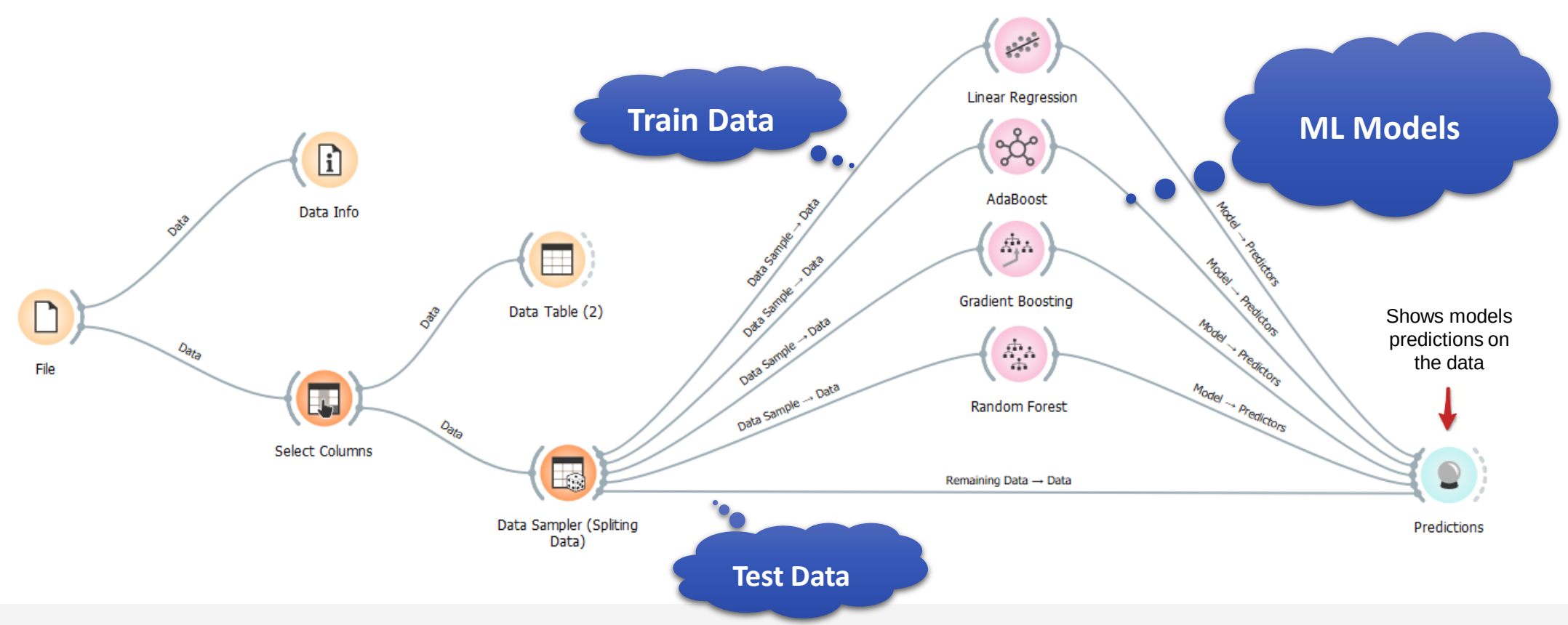
Orange Data Mining

Architecture



Orange Data Mining

Architecture



Validate Data Models

Predicated Models

- ✓ Target : Resale Price
- ✓ Refer the predicated models below,

Model	MSE	RMSE	MAE	R2
Linear Regression	2684411008.143	51811.302	39684.122	0.900
Gradient Boosting	2039605377.732	45161.990	32700.242	0.924
Random Forest	901399521.545	30023.316	21054.601	0.966
AdaBoost	919247651.299	30319.097	22207.646	0.966

- ✓ **Random Forest** and **AdaBoost** scored highest

  R^2 value **0.966**

Predicated resale price

Actual resale price

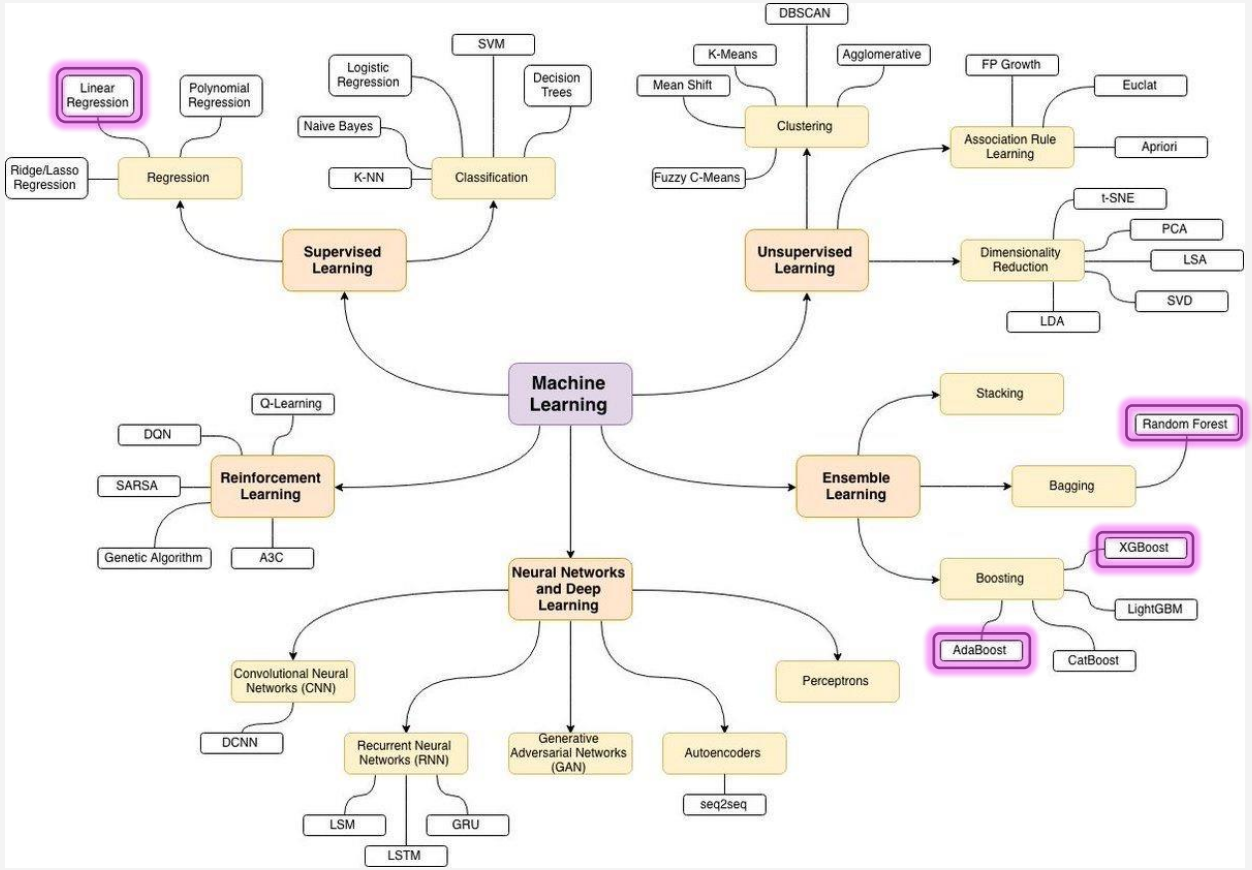
	Random Forest	Linear Regression	Gradient Boosting	AdaBoost	resale_price
1	168758.10	147981.77	204201.71	208000.00	165000.00
2	171341.67	241513.79	204116.50	210000.00	188000.00
3	177191.67	266906.91	210299.10	218000.00	190000.00
4	180908.33	260561.58	207651.74	220000.00	183888.00
5	181505.83	193476.29	212520.94	215000.00	180000.00
6	181948.10	140854.80	202638.02	208000.00	180000.00
7	183014.76	150742.46	214961.57	215000.00	165000.00
8	183014.76	146150.20	214961.57	215000.00	172000.00
9	183090.56	139315.23	210291.35	205000.00	188000.00
10	183323.21	152128.03	218048.60	210000.00	175000.00
11	185914.76	133926.74	202638.02	208000.00	180000.00
12	186494.52	190338.72	212084.63	215000.00	195000.00
13	186798.10	164209.76	221144.16	208000.00	187000.00
14	186811.51	126335.81	202282.18	205000.00	181000.00
15	187800.00	270935.05	238167.39	220000.00	200000.00
16	187956.55	194233.15	224231.20	210000.00	160000.00
17	188132.95	104542.89	227185.03	225000.00	208000.00
18	188403.17	148149.04	212364.90	208000.00	188000.00
19	189098.10	161873.96	221144.16	208000.00	173268.00
20	189098.10	159458.83	221144.16	208000.00	172000.00

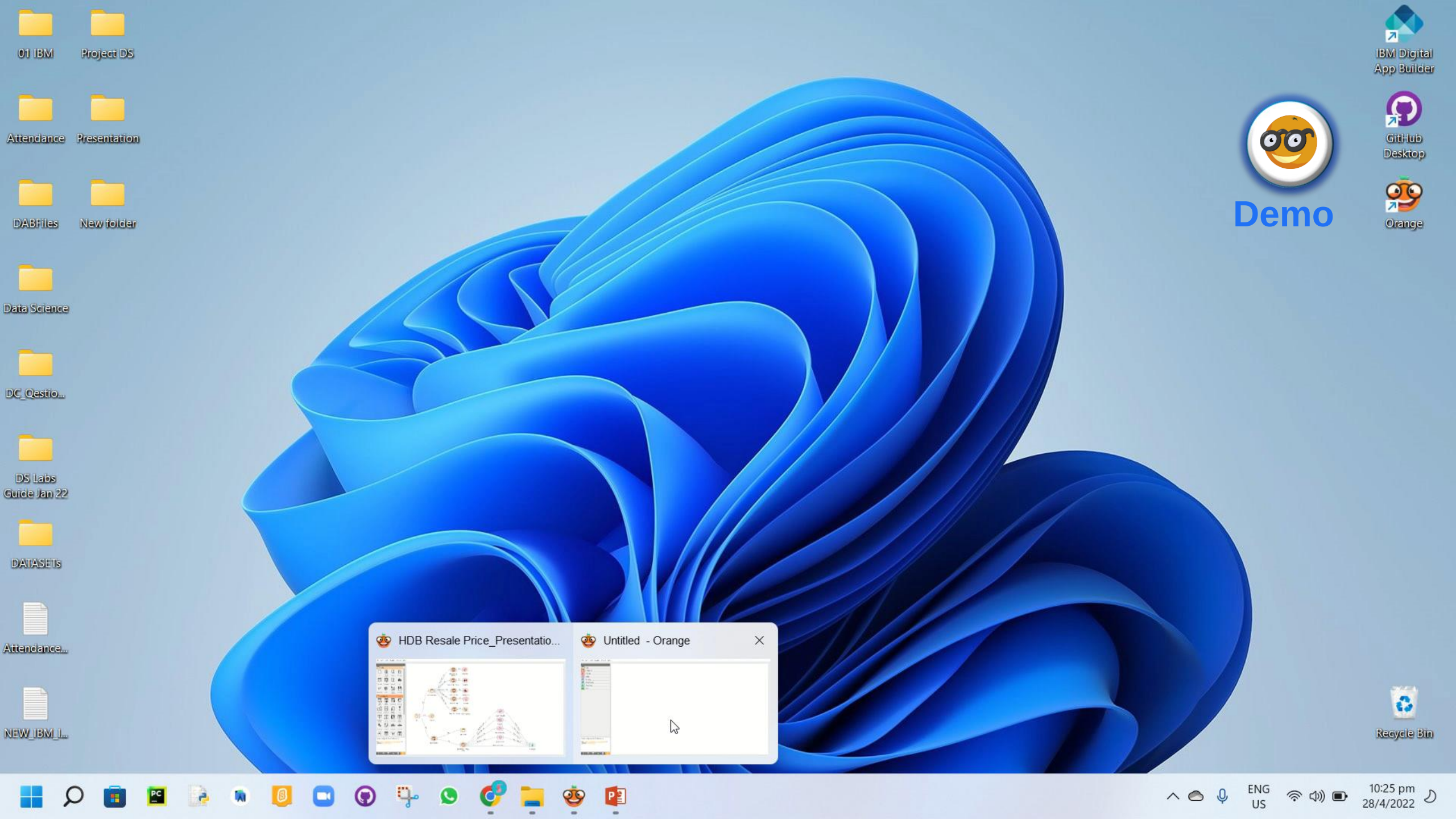
Validate Data Models

ML Models

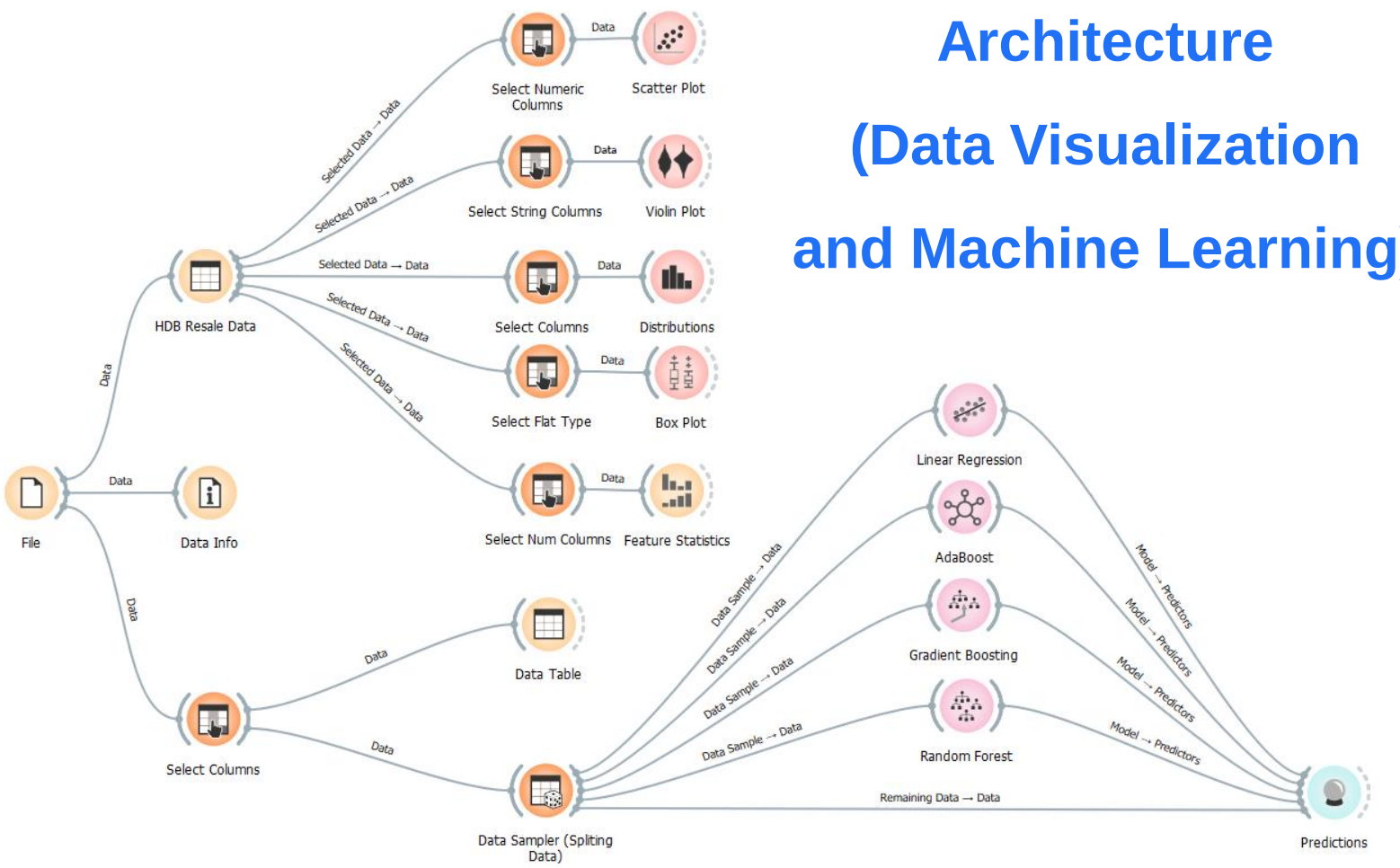
- ✓ **Random forest** is an ensemble Machine Learning Algorithm that is used widely in Classification and **Regression problems**

- ✓ **AdaBoost** is an ensemble learning method uses an iterative approach to learn from mistakes of weak classifiers and turn them into strong ones.




01 IBM
Project DS
Attendance
Presentation
DABFiles
New folder
Data Science
DC_Qestio...
DS Labs Guide Jan 22
DATASETs
Attendance...
NEW_IBM_I...
IBM Digital App Builder
GitHub Desktop
Orange
Demo
Recycle Bin
HDB Resale Price_Presentatio...
Untitled - Orange
ENG US
10:25 pm
28/4/2022

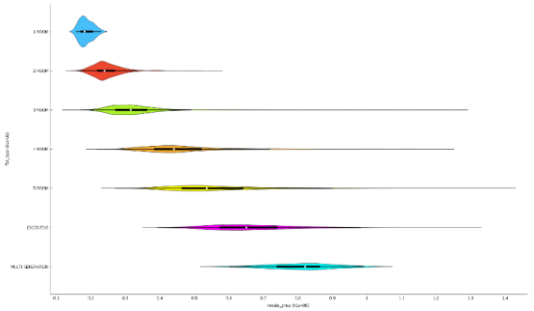
Orange Data Mining

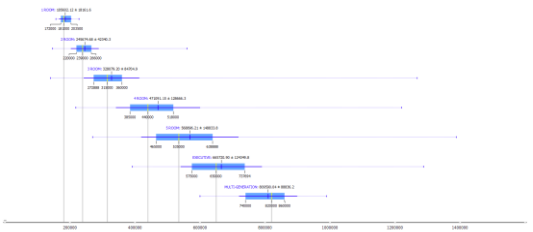


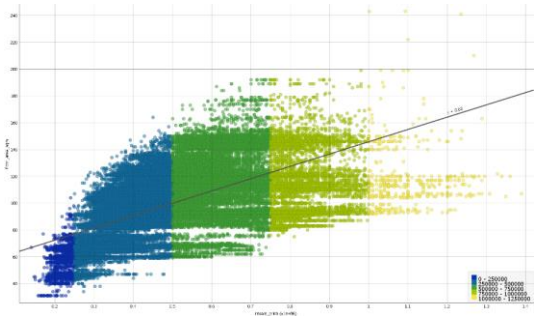
Architecture

(Data Visualization

and Machine Learning)







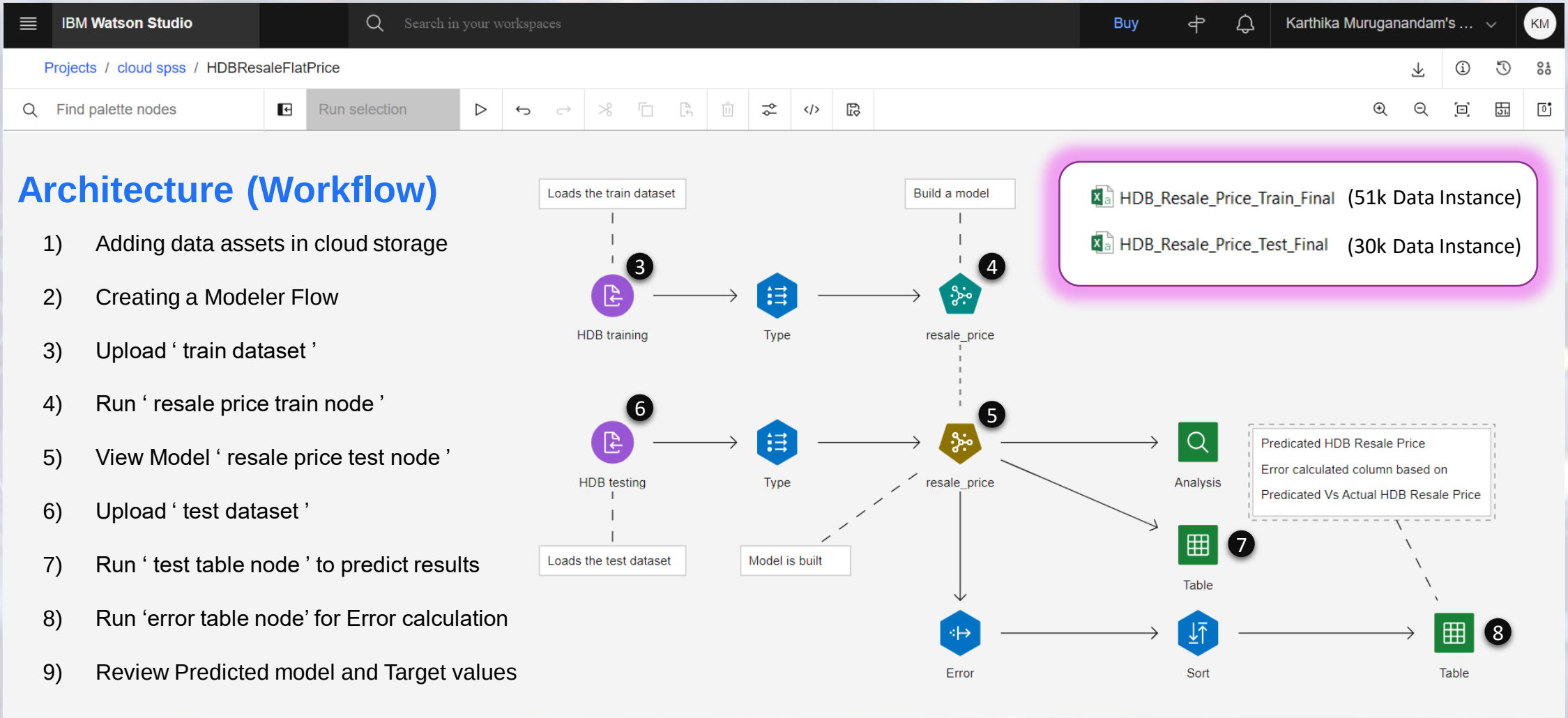
Machine Learning (ML) Builders

Graphical ML Builders

- ✓ Error Finding
- ✓ Data Updating (excel)
- ✓ Model Development

- ✓ Model Training
- ✓ Model Validation
- ✓ Resale Price Prediction

IBM SPSS Modeler



Architecture (Workflow)

1) Adding data assets in cloud storage
2) Creating a Modeler Flow
3) Upload ‘ train dataset ’
4) Run ‘ resale price train node ’
5) View Model ‘ resale price test node ’
6) Upload ‘ test dataset ’
7) Run ‘ test table node ’ to predict results
8) Run ‘error table node’ for Error calculation
9) Review Predicted model and Target values

IBM SPSS Modeler

Resale Price Format (Data Error)

- ✓ Amended the format to Numeric using excel

Data wrangling, is very important,

to achieve good accuracy, ML model

Node:
Type
The storage class has changed for field: resale_price (conversion applied)

Node:
resale_price
No valid records found in the data source.

Execution was interrupted

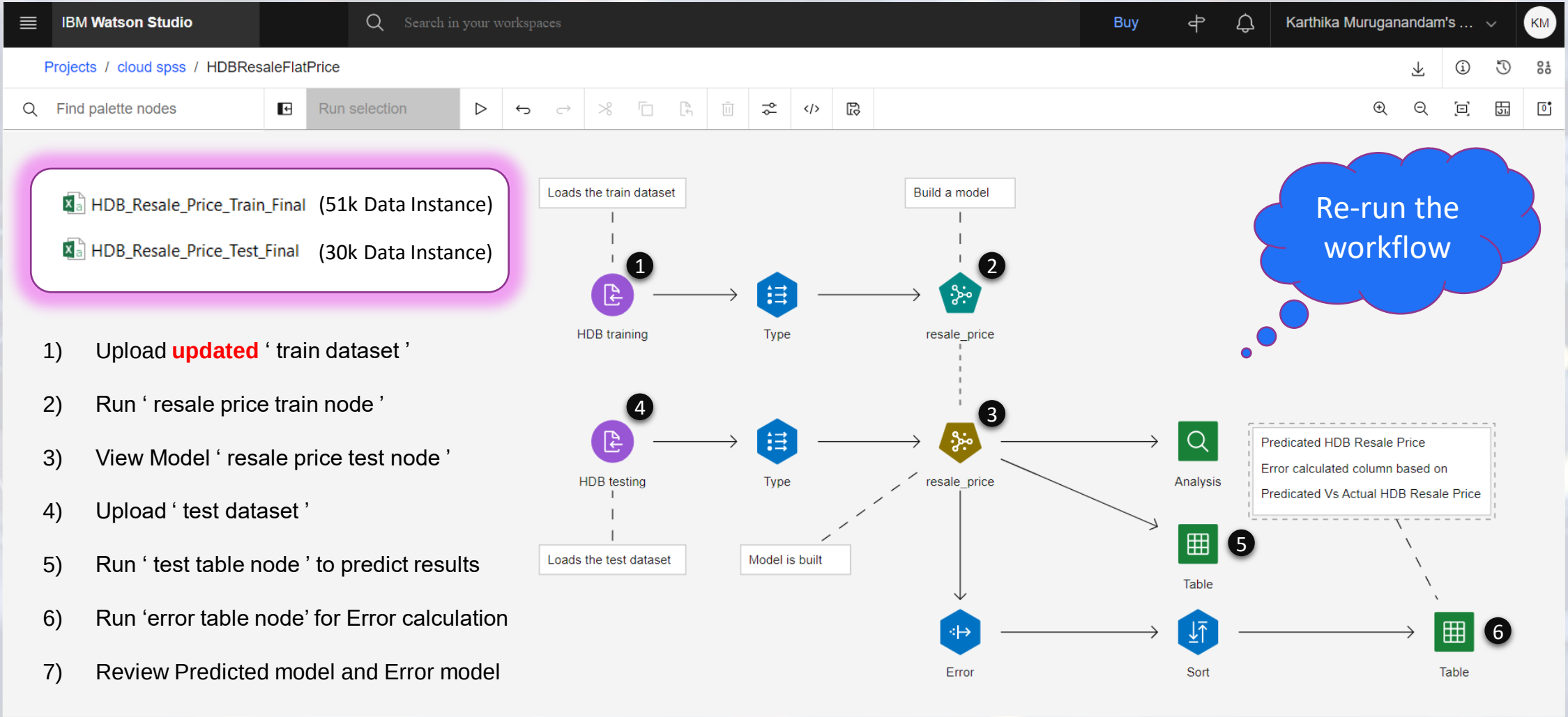
resale_price
$355,000.00
$440,000.00
$409,000.00
$356,000.00
$388,000.00
$360,000.00
$417,000.00
$365,000.00
$415,000.00
$393,000.00

Error

resale_price
355000.00
440000.00
409000.00
356000.00
388000.00
360000.00
417000.00
365000.00
415000.00
393000.00

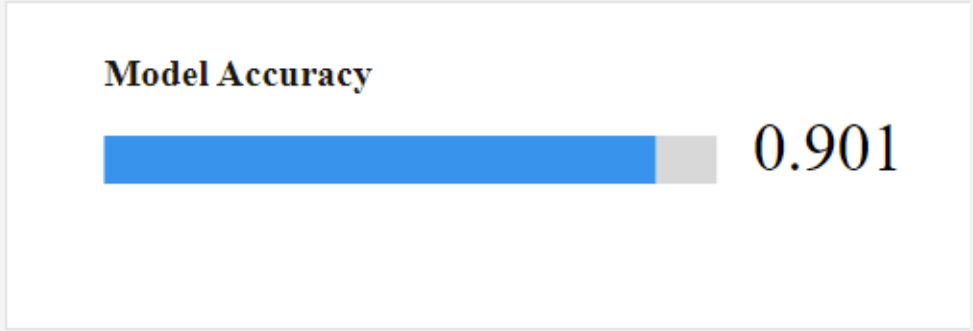
Amended

IBM SPSS Modeler



1) Upload **updated** ‘ train dataset ’
2) Run ‘ resale price train node ’
3) View Model ‘ resale price test node ’
4) Upload ‘ test dataset ’
5) Run ‘ test table node ’ to predict results
6) Run ‘error table node’ for Error calculation
7) Review Predicted model and Error model

IBM SPSS Modeler (View Model)



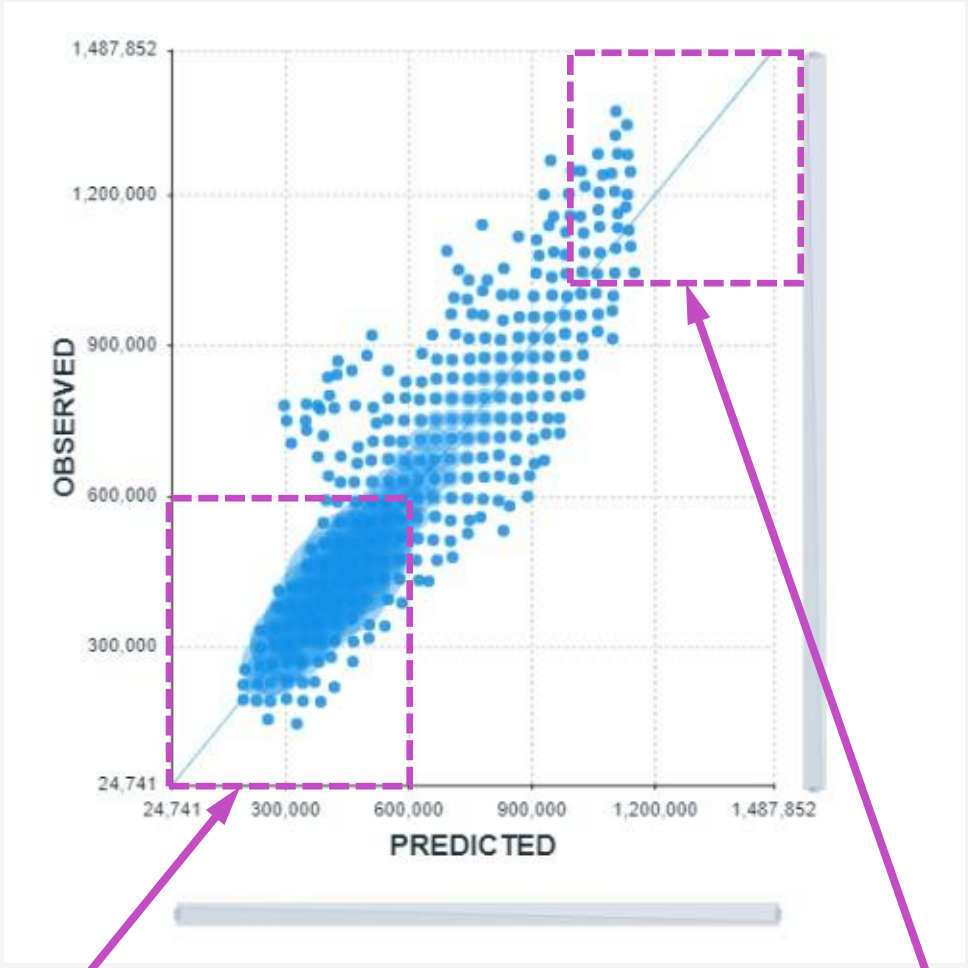
Model Evaluation Measures

R^2	0.901
Mean Squared Error (MSE)	2,653,162,395.003
Root Mean Squared Error (RMSE)	51,508.857

Model Evaluation

- ✓ Target : Resale Price
- ✓ R^2 value is **0.901**. It is pronounced "R squared", which is a measure of the goodness of fit of a model.
- ✓ Mean Squared Error (MSE) value is 2.6 Billion. It measures the average of the squares of the errors.
- ✓ Root Mean Square Error (RMSE) value is **$51,508**. It is a used measure of the differences between values predicted by a model and the values observed.

IBM SPSS Modeler (View Model)



Observed by Predicted

- ✓ Target : Resale Price
- ✓ The x-axis shows the model's predicted values, while the y-axis shows the dataset's actual values. The estimated regression line is the diagonal line in the center of the plot.
- ✓ Each data point is quite close to the projected regression line, we may conclude that the **regression model fits the data reasonably well**.

Area with more data have tighter tolerances

As compared to sparse data. This is because majority of HDB sales occur in the below SGD 600k range

IBM SPSS Modeler (View Model)

Target Field	resale_price
Model Building Method	Multilayer Perceptron (MLP)
Model Type	Regression
Number of Predictors	16
Number of Hidden Layers	1
Number of Neurons in Hidden Layer	12
Activation function in Hidden Layer	Hyperbolic Tangent (tanh)
Number of Neurons in Output Layer	1
Activation function in Output Layer	Identity

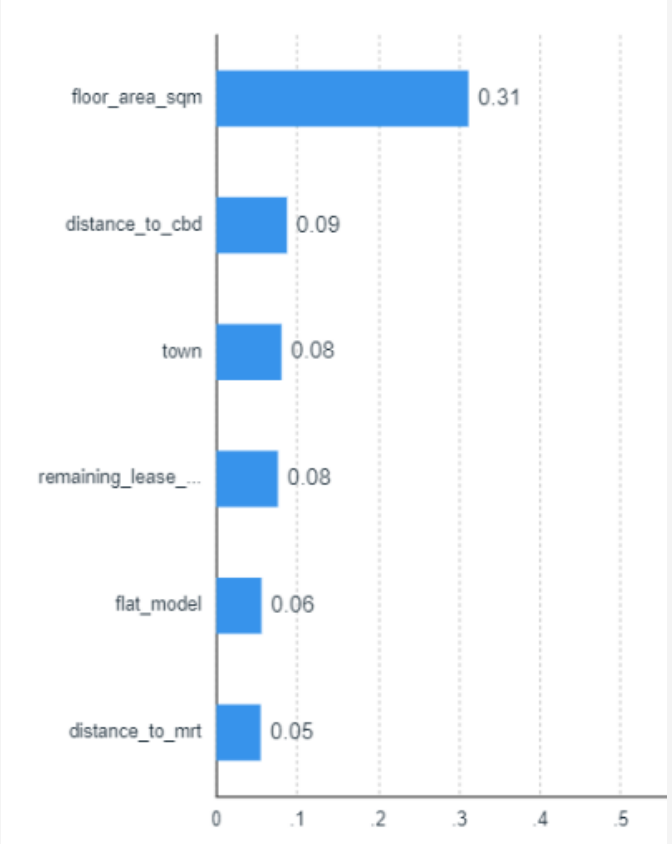
Model Information

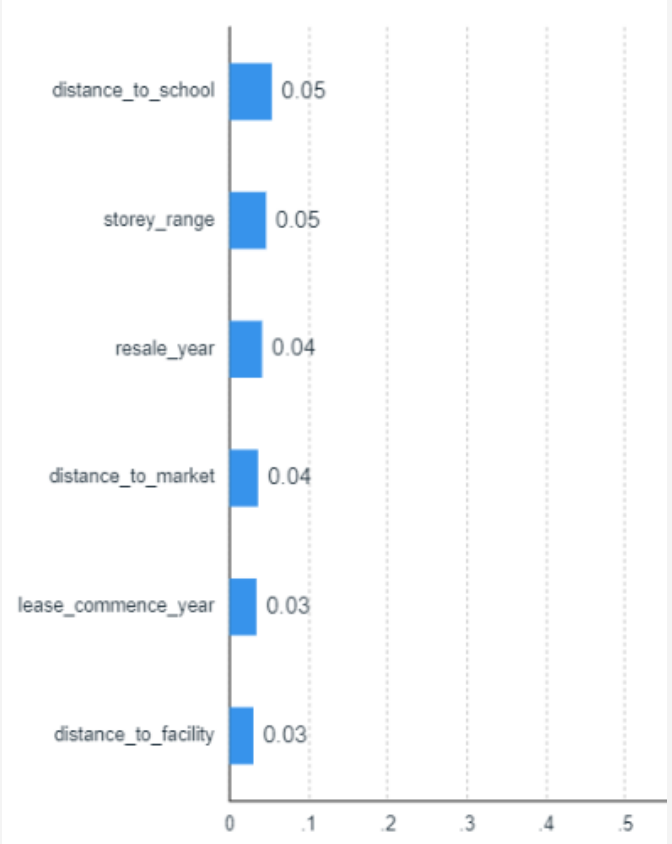
- ✓ Target : Resale Price
- ✓ Model building method is **Multilayer perceptron (MLP)**
- ✓ Model type is **Regression**
- ✓ Sixteen (16) Number of predictors, as follows

1) Town
2) Flat type
3) Flat model
4) Postal code
5) Resale year
6) Storey range
7) Floor area sqm
8) Distance to CBD
9) Distance to MRT
10) Distance to park
11) Distance hospital
12) Distance to facility
13) Distance to school
14) Distance to market
15) Lease commence year
16) Remaining lease years

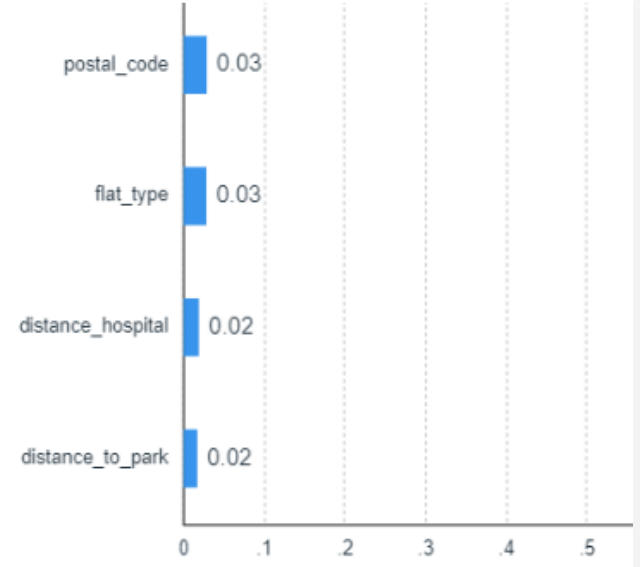
IBM SPSS Modeler (View Model)

Feature Importance (Target : Resale Price)





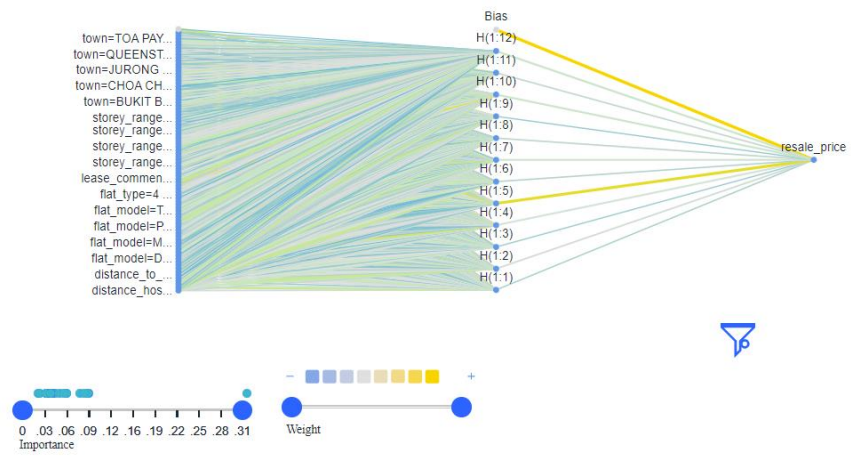
The Feature Importance chart shows the **relative importance of each predictor** in estimating the model



IBM SPSS Modeler (View Model)

Network Diagram , Build Settings, Training Summary

- ✓ IBM SPSS Neural Networks module offers you the ability to discover more **complex relationships in data** and generate better performing predictive model with **accuracy of 90%**



Algorithm	Neural net
Model type	Approximation
Date built	Sun Apr 24 07:26:37 UTC 2022
Elapsed time for model build	0 hours, 0 mins, 15 secs

Task type	Build new model
Overfit prevention set(%)	30.0
Replicate results	true
Random seed	229176228
Missing values in predictors:	Delete listwise
User-missing values	false
Calculate predictor importance	true
Neural network model	Multilayer Perceptron (MLP)
Hidden Layers	Automatically compute number of units
Hidden layer 1	1
Hidden layer 2	0
Use minimum accuracy	true
Accuracy (%)	90.0
Use maximum training time (per component model)	true
Minutes	15
Customize number of maximum training cycles	Compute automatically
Maximum number of cycles	250

Validate Data Models

Actual, Predicated Result & Error (Target : Resale Price)

View Output: Table (20 fields, 30,000 records) #2

Actual resale price

Predicted resale price

stance_hospital	distance_to_park	distance_to_market	distance_to_facility	lease_commence_year	resale_year	remaining_lease_years	resale_price	$N-resale_price	Error
656.770	804.719	1923.865	858.970	1988	2020	67.400	410000	410000.543	0.543
781.125	1569.748	270.152	658.060	1977	2021	55.300	339888	339889.982	1.982
1602.737	1410.977	4233.855	946.915	2017	2021	95.300	538000	538003.968	3.968
377.406	1491.863	419.155	907.870	1985	2020	64.500	520000	519995.947	4.053
1605.758	997.845	383.749	1162.601	1985	2019	64.600	250000	249991.522	8.478
760.038	815.577	1926.859	969.183	2015	2020	94.800	515000	515012.666	12.666
1430.754	1733.352	2201.456	2286.189	1996	2021	74.000	428000	427985.537	14.463
1792.426	772.751	3414.015	772.643	1993	2020	72.600	670000	669983.997	16.003
1760.044	2540.324	782.074	1111.751	1985	2021	63.100	375000	374982.756	17.244
3462.564	2580.482	2662.990	3684.272	2001	2020	80.100	405000	404981.074	18.926
2372.143	1129.760	3700.455	1349.642	1995	2021	73.300	545000	545020.176	20.176
2287.023	820.847	5826.795	1657.375	2016	2021	94.600	680000	679977.250	22.750
1773.803	1208.084	285.613	472.450	1976	2021	54.300	245000	244975.722	24.278
931.773	464.896	2210.095	978.139	2015	2019	94.800	488888	488863.615	24.385

VI. Model Test Case

Jupyter – Model Training (0.841)

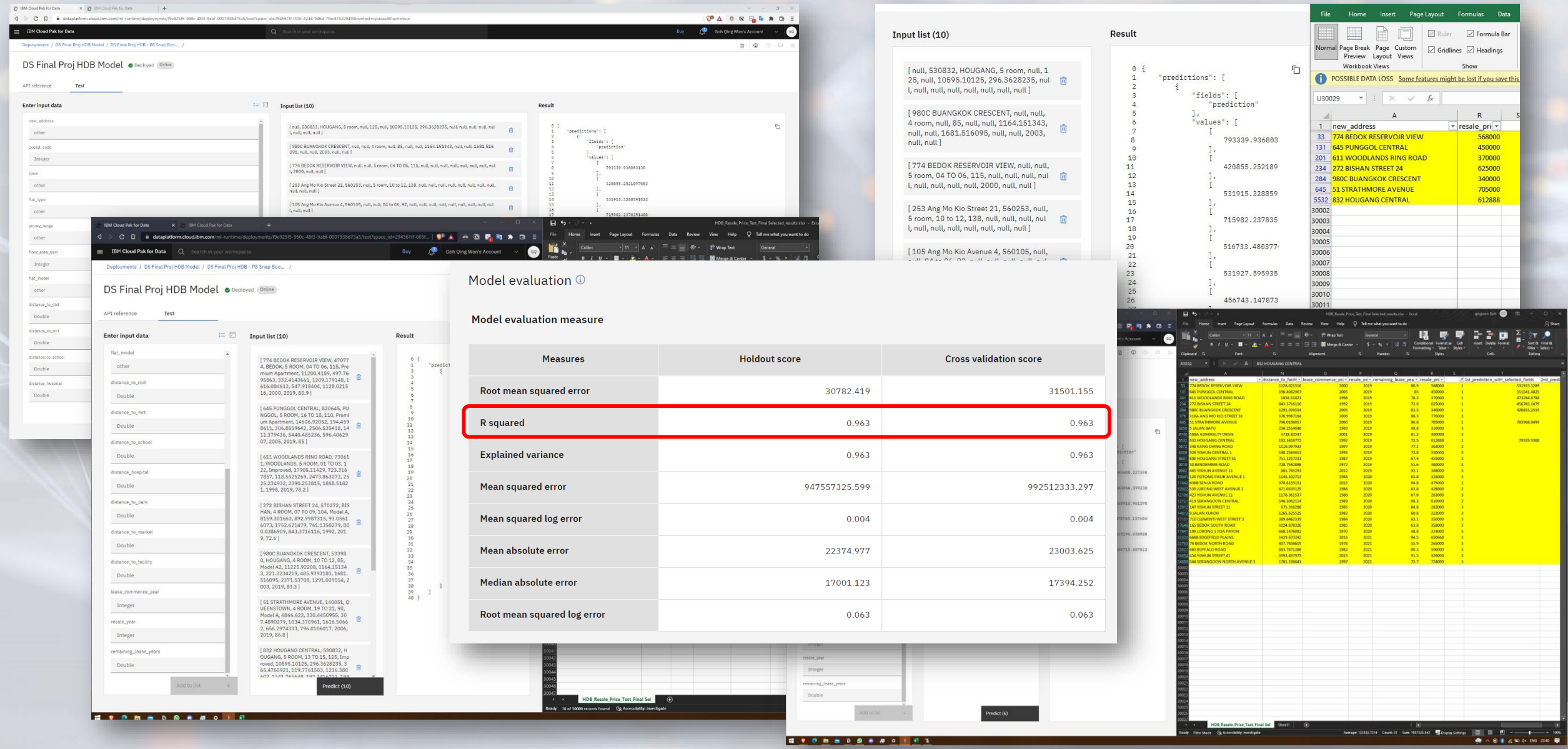
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		month	postal_co	flat_type	storey_fro	storey_to	floor_area	lease_com	remaining_	dis_to_cbc	min_dist_t	min_dist_t	min_dist_t	min_dist_t	min_dist_t	min_dist_t	Actual	Predicted (y_pred)
2	45158	2022	461220	3	10	12	69	2010	87.4	10403.51	1235.373	396.8694	307.7279	1669.3	209.223	247.222	540,000.00	510,000.00
3	49776	2022	762424	2	1	3	47	2015	92.8	15510.67	1654.999	467.4207	1166.217	688.139	2419.275	1183.799	290,000.00	239,000.00
4	26220	2021	520253	4	7	9	101	1996	74.4	13014.05	175.7851	162.1828	380.9692	2129.466	992.1189	1758.41	560,000.00	529,000.00
5	32653	2021	560179	4	4	6	91	1981	58.7	10518.23	602.8944	106.29	911.6245	807.8514	422.2705	1217.304	460,000.00	432,000.00
6	13005	2020	514526	4	7	9	93	2015	94	14276.4	533.0036	281.6197	1265.913	550.1783	2547.781	784.0463	520,000.00	525,000.00
7	26813	2021	682815	5	4	6	113	2017	94.9	15538.52	269.5412	521.0228	715.8087	1151.902	3834.899	1522.037	542,000.00	606,000.00
8	25617	2021	542190	4	7	9	95	2000	78.1	13541.79	212.8918	263.5639	910.1351	630.2413	4280.673	1601.74	385,000.00	465,000.00
9	10768	2019	560259	5	10	12	135	1982	61.3	9714.87	604.4257	327.0989	1694.851	1410.152	508.2068	1041.627	718,000.00	618,000.00
10	10971	2019	400349	4	1	3	84	1986	65.1	7246.194	482.5798	231.8055	1507.81	2601.302	736.4712	1486.025	326,000.00	386,000.00
11	33581	2021	650290	6	16	18	142	1997	75.3	12455.33	1016.898	169.7169	1368.782	1019.846	2151.703	1037.662	735,000.00	739,000.00
12	5934	2019	142050	5	22	24	117	2015	94.4	6353.941	195.406	348.2174	523.5727	738.032	91.6075	756.6424	1,030,000.00	881,000.00
13	10363	2019	460534	3	4	6	60	1986	65.2	9876.468	783.5608	470.1457	1065.017	312.8194	140.3923	1052.948	230,000.00	271,000.00

```
In [54]: model = LinearRegression()
         model.fit(X_train , y_train)

Out[54]: LinearRegression()

In [57]: y_pred = model.predict(X_test)
         y_pred

Out[57]: array([[489374.41364466],
                [336287.3946487 ],
                [179540.76113561],
                ...,
                [832846.04201829],
                [672800.72999503],
                [475171.2649025 ]])

In [58]: y_test

Out[58]: array([[500000.],
                [345000.],
```

```
In [62]: int_array = y_test. astype(float)
         y_test

Out[62]: array([[500000.],
                [345000.],
                [248000.],
                ...,
                [935000.],
                [650000.],
                [500000.]])

In [63]: metrics.r2_score(y_test , y_pred)

Out[63]: 0.8411109808520748
```

- Using Jupyter from Anaconda, we have trained our model using Linear Regression.
- R2 accuracy of 0.84 was achieved.

AutoAI – Model Testing (0.963)



Orange – Model Testing (0.996)

Predictions - Orange

	AdaBoost	Random Forest	Gradient Boosting	SVM	Linear Regression	Neural Network	resale_price	new_address	storey_range
1	630000....	628218.72	580211.79	655998...	562171.36	436558.50	630000.00	138C YUAN CHI...	13 TO 15
2	365000....	366953.57	348227.24	655996...	328825.64	369484.82	345000.00	472B FERNVALE...	07 TO 09
3	750000....	759288.91	787052.85	655996...	690827.23	696032.04	750000.00	18C CIRCUIT RO...	07 TO 09
4	345000....	334847.50	378293.70	655996...	384680.24	377584.15	315000.00	347 UBI AVENU...	01 TO 03
5	390000....	392584.64	382567.35	655996...	420796.44	327897.25	390000.00	6 SAINT GEORG...	07 TO 09
6	640000....	632127.78	607107.57	655997...	651900.48	706638.48	640000.00	37 LORONG 5 T...	19 TO 21
7	370000....	390620.46	394309.20	655997...	427505.29	429420.34	370000.00	350 CLEMENTI ...	01 TO 03
8	528000....	547976.57	556572.02	655995...	554099.77	530860.21	528000.00	864A TAMPINE...	04 TO 06
9	537000....	532967.14	495131.71	655998...	573503.41	544008.30	537000.00	817C KEAT HO...	19 TO 21
10	230000....	221310.71	237862.03	655996...	200304.77	269431.10	220000.00	633A SENJA RO...	01 TO 03
11	300000....	303091.19	297685.96	655994...	252525.28	291087.05	295000.00	101 TAMPINES ...	01 TO 03
12	380000....	367408.45	399709.59	655996...	406841.52	406920.78	368000.00	453D FERNVAL...	01 TO 03
13	335000....	337088.33	375642.10	655995...	383204.39	395178.08	333000.00	336A YISHUN S...	07 TO 09
14	550000....	549908.61	565923.16	655997...	576483.94	567010.04	563000.00	443B FAJAR RO...	07 TO 09
15	660000....	662872.21	609246.83	655998...	690199.36	668025.53	660000.00	470 JURONG W...	13 TO 15
16	950000....	951185.71	926488.51	656000...	881298.42	894690.40	950000.00	2 TOH YI DRIVE	01 TO 03
17	480000....	485735.48	450231.69	655995...	454340.57	431470.93	495000.00	200D SENGKAN...	07 TO 09
18	350000....	349936.07	372397.01	655995...	313892.72	275400.54	350000.00	203B COMPASS...	01 TO 03
19	585000....	565160.38	571686.57	655997...	625452.39	658399.93	585000.00	141 PASIR RIS S...	10 TO 12
20	340000....	344819.89	337961.36	655997...	316811.29	305872.84	340000.00	105 SERANGOO...	04 TO 06
21	450000....	448497.62	467196.39	655997...	417315.83	446890.21	450000.00	684A CHOA CH...	07 TO 09
22	625000....	614564.40	529283.60	655997...	567884.40	603679.46	625000.00	183 BEDOK NO...	16 TO 18
23	270000....	265166.47	270542.12	655996...	233690.30	224444.04	246000.00	445A FERNVALE...	04 TO 06
24	520000....	528311.57	516653.53	655996...	533728.53	533956.84	520000.00	144 PASIR RIS S...	01 TO 03
25	482000....	475810.30	500480.58	655997...	519124.49	542549.60	482000.00	336C ANCHORV...	16 TO 18

Show perfomance scores

Model	MSE	RMSE	MAE	R2
AdaBoost	96589893.436	9828.016	4768.758	0.996
Random Forest	244104604.426	15623.847	10747.406	0.991
Gradient Boosting	2175217062.442	46639.222	34046.498	0.918
SVM	59195964307.403	243302.208	214795.523	-1.238
Linear Regression	2713927624.443	52095.370	40032.799	0.897
Neural Network	2690022082.549	51865.423	38782.553	0.898

- A few models were tested using Orange:
  - AdaBoost
  - Random Forest
  - Gradient Boosting
  - SWM
  - Linear Regression
  - Neural Network
- Of the models, AdaBoost yields a model accuracy of 0.996.
- Per the circled results in the picture, there are many occurrences of accurate prediction of prices between AdaBoost model vs the actual retail price.

Model Testing Results Compile

No.	TEST DATA	ACTUAL	INITIAL AUTO AI	QUICK TEST JUPYTER (LR-S)	MODEL JUPYTER (LR-L)	MODEL AUTO AI (Snap Boost)	MODEL ORANGE (AdaBoost)	MODEL ORANGE (LR)
1	253 Ang Mo Kio Street 21, 560253, 10 to 12 floor, 138 sqm, 5 room	$750,000 (in Sept 2021)	$422,922 (-44.5%)	$718,389 (-4.2%)	$750,499 (0.1%)	$715,982 (-4.5%)	$750,000 (0.0%)	$715,121 (-4.7%)
2	105 Ang Mo Kio Avenue 4, 560105, 04 to 06 floor, 92 sqm, -	$440,000 (in Jun 2021)	$421,615 (-43.6%)	$413,120 (-6.1%)	$470,090 (6.8%%)	$516,773 (17.5%)	$440,000 (0.0%)	$464,497 (5.6%)
3	310B Ang Mo Kio Avenue 1, 562310, 16 to 18 floor, 94 sqm, 4 room	$760,000 (in Mar 2022) $868,888 (in Feb 2022)	$421,615 (-4.2%)	$631,431 (-16.9%)	$709,299 (-6.7%)	$531,927 (-30.0%)	$760,000 (0.0%)	$679,317 (-10.6%)
	MSE ('1,000m)		**73.94**	**6.08**	**1.16** 2	**19.69** 4	**0.00** 1	**2.78** 3

- Full address
- Unit details
- HDB Age
- Sales year

- Full address
- Unit details
- HDB Age
- Sales year
- All distances

Note : without full address data

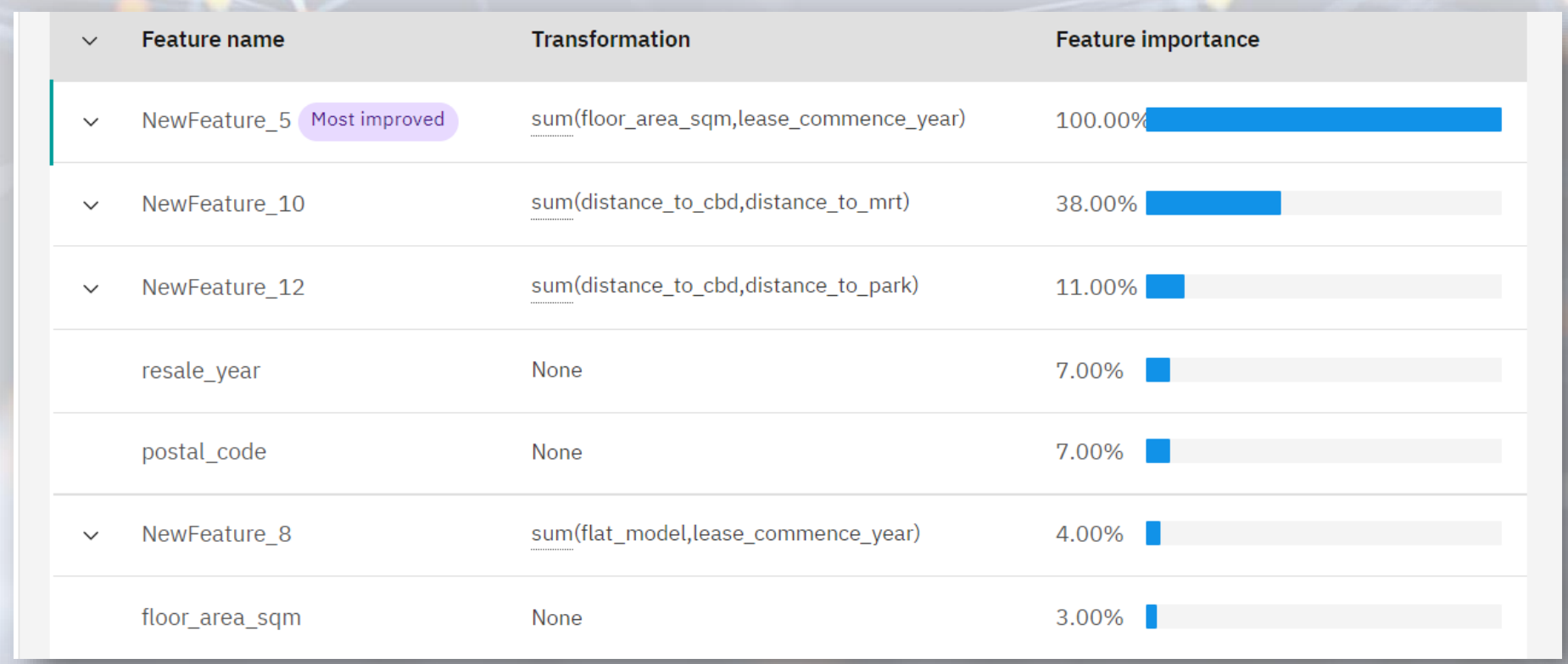
More Testing & Comparison

No.	TEST DATA	ACTUAL	MODEL JUPYTER (LR-L)	MODEL AUTO AI (Snap Boost)	MODEL ORANGE (AdaBoost)	MODEL ORANGE (LR)
1	832 HOUGANG CENTRAL, 530832, 5 RM, 13 TO 15, 125 SQM	$612,888	$625,407 (2.0%)	$715,926 (16.8%)	$612,888 (0.0%)	$580,395 (-5.3%)
2	774 BEDOK RESERVOIR VIEW, 470774, 5 RM, 04 TO 06, 115 SQM	$568,000	$579,867 (2.1%)	$523,586 (-7.82%)	$568,000 (0.0%)	$598,863 (5.4%)
3	645 PUNGGOL CENTRAL, 820645, 5 RM, 16 TO 18, 110 SQM	$450,000	$491,065 (9.1%)	$463,037 (2.9%)	$450,000 (0.0%)	$532,094 (18.2%)
4	611 WOODLANDS RING ROAD, 730611, 5 RM, 01 TO 03, 122 SQM	$370,000	$432,586 (14.5%)	$374,344 (1.17%)	$380,000 (2.7%)	$403,636 (9.1%)
5	272 BISHAN STREET 24, 570272, 4 RM, 07 TO 09, 104 SQM	$625,000	$502,520 (-19.6%)	$558,096 (-10.7%)	$625,000 (0.0%)	$574,603 (-8.1%)
6	980C BUANGKOK CRESCENT, 533980, 4 RM, 10 TO 12, 85 SQM	$340,000	$403,291 (18.6%)	$369,887 (8.8%)	$360,000 (5.9%)	$388,841 (14.4%)
7	51 STRATHMORE AVENUE, 140051, 4 RM, 19 TO 21, 90 SQM	$705,000	$666,971 (-5.4%)	$748,408 (6.2%)	$735,000 (4.3%)	$715,497 (1.5%)
	MSE ('1,000m)		**25.31** 4	**20.03** 3	**1.4** 1	**14.91** 2

Note : with full address data

Further Understanding of AutoAI

- Testing only with << 832 HOUGANG CENTRAL, 530832, 5 RM, 13 TO 15, 125 SQM >>, sold at $612,888.
- Based on Feature Summary, the following information is deemed ***more*** important:
  - Floor area of unit
  - Lease commencement year of HDB unit
  - Distance to CBD, Park, MRT
  - Resale year
  - Postal code



Feature name	Transformation	Feature importance
NewFeature_5 Most improved	sum(floor_area_sqm,lease_commence_year)	100.00%
NewFeature_10	sum(distance_to_cbd,distance_to_mrt)	38.00%
NewFeature_12	sum(distance_to_cbd,distance_to_park)	11.00%
resale_year	None	7.00%
postal_code	None	7.00%
NewFeature_8	sum(flat_model,lease_commence_year)	4.00%
floor_area_sqm	None	3.00%

Further Understanding of AutoAI

- Testing only with << 832 HOUGANG CENTRAL, 530832, 5 RM, 13 TO 15, 125 SQM >>, sold at $612,888.
- Various combinations were tested with AutoAI:

Line	Sales Month	Address	Postal	Town	Floor Area	T.O.P	Flat Model	Flat Type	Storey	Remain. Years	CBD	MRT	Park	Sch	Hos	Mkt	Facil	Pred Px	Var%	No. Items
1		X			X	X												$549,408	-10.4%	3
2		X									X	X						$663,066	8.2%	3
3		X									X		X					$66,516	-89.2%	3
4	X	X	X	X	X	X												$470,768	-23.2%	6
5	X	X	X	X							X	X						$619,196	1.03%	6
6	X	X	X	X							X		X					$606,715	-1.0%	6

- Comparing 3 items vs 6 items -> Keying more info, generate a closer prediction to actual.
- Observing line 2, 5, 6 -> Address and CBD is crucial in price prediction. This is however not aligned with the feature importance table by Auto AI.
- Observing line 2, 3, 5, 6 -> information on MRT or Park do not influence the price that much, if postal, town and sales month information are provided.
- Observing line 2, 3 -> if Sales month, postal and town information is unavailable, distance to MRT is crucial

VII. Summary & Reflection

Summary

Singapore Resale Flats price continue to raise in Year 2022 and the we, a team of freelance property agent decided to grab this opportunity to improve revenue growth and support government initiative for digital transformation. We decided to invested on the development of a HDB Resale Price predicting Model, which will add value to our customer service, where we can instantly provide insight of the HDB resale market and help them make a wise decision. This also help us to keep abreast with the competition.

With data acquired and prepared by our Data Engineer, our Data Analyst is able to shows that the general trend of HDB resale market has been stable from 2015 until 1Q2020, than it start to climb steeply from 3Q 2020 to 1Q 2022. This was driven by delays in build-to-order flats during the Covid 19 pandemic period , which pushed home buyers to look in resale market.

Our Data Analyst also put up 3 test hypothesis, which provide insight into the HDB Resale price factors:
H1: No --> **Distance to MRT** has a greater impact on the resale price as compared to **distance to market/hawker.**
H2: Yes -->Resale price increase as the **distance to nearest MRT station** decrease.
H3: Yes --> Resale price decrease as the **remaining lease of HDB** increases.
No --> Resale price increase as the **distance to nearest Healthcare facility** decrease.

In addition, the following buying patterns was observed

1. Flats located nearer to **Central Business District** (CBD) fetched higher prices
2. Flat unit at the higher **flat level** has the higher value
3. **Bukit Timah** Town command the highest median resale price
4. **Seng Kang** Town has the most active resale transactions.
5. The smaller flats type (**1,2,3-Room**) experience the highest % increase in price.

Our Data Scientist then proceed to developed Machine Learning model for predicting the HDB resale price. We explored the various tools:

- Jupyter Notebook,
- Orange
- IBM Watson Studio's Auto AI
- IBM Watson Studio's SPSS Modeler.

All tool give good result but Orange 'Adaboost' stand out from the rest with a accuracy of 0.996.

We are satisfied with the result of our first learning phase, and we will continue the deployment phase in our next project.

Thank you !

VIII. Appendix

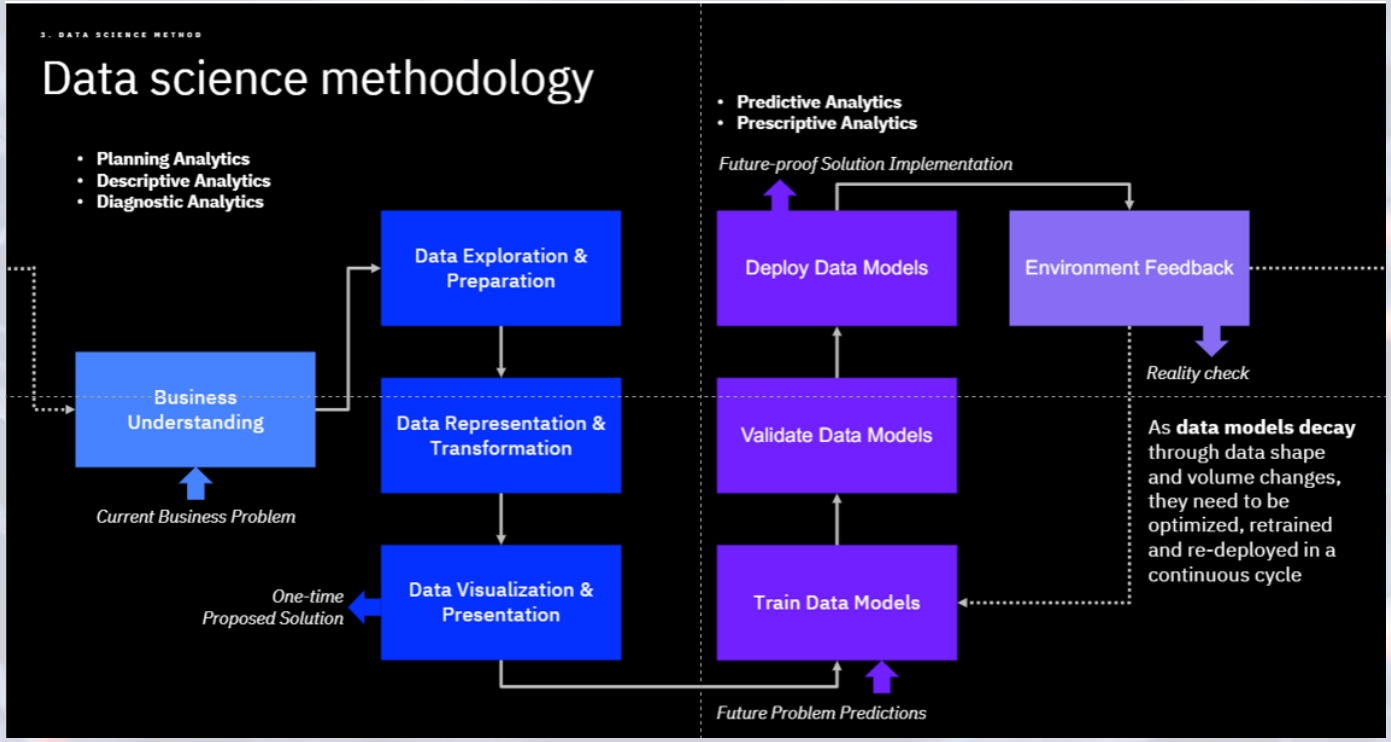
Sources

- HDB resale price prediction
  - https://towardsdatascience.com/predict-the-selling-price-of-hdb-resale-flats-50530391a845
  - https://github.com/jimmeng-kok-2017/hdb_resale_lr_predict
  - https://www.kaggle.com/code/teyang/drivers-of-hdb-resale-price-and-prediction
  - https://github.com/teyang-lau/HDB_Resale_Prices
- Data.gov.sg
  - https://data.gov.sg/dataset/resale-flat-prices
- School directory
  - https://data.gov.sg/dataset/school-directory-and-information

Sources

- List of Facilities
  - https://www.sportsingapore.gov.sg/about-us/facilities
- List of Hospitals
  - http://www.hospitals.sg/hospitals
- List of Parks
  - https://en.wikipedia.org/wiki/List_of_parks_in_Singapore
- List of Market & Hawker Centre
  - https://data.gov.sg/dataset/list-of-government-markets-hawker-centres

Reference Methodology



[Data Science Practitioners Course: Data Science Learning Module: Lecture 2 - Data Science Methodology Presentation (ibm.com)]

Initial Issue

Lesson learnt … …

- Looking at the feature dependency, having high reliance on park distance is not an expected result. There could have been closer reliance on size of unit, age of unit.

- We decided to relook into areas for fine-tuning… …

Feature name	Transformation	Feature importance
NewFeature_12 Most improved	divide(product(Park_Dist,HDB_CPI_AdjPx), Park_Dist)	100.00%
NewFeature_10	divide(product(Distance_to_MRT,HDB_CPI_AdjPx), Distance_to_MRT)	49.00%
NewFeature_7	divide(product(Floor_Area,HDB_CPI_AdjPx), Floor_Area)	46.00%
HDB_CPI_AdjPx	None	45.00%
NewFeature_14	divide(product(Remaining_Months,HDB_CPI_AdjPx), Remaining_Months)	13.00%
NewFeature_11	divide(product(School_Dist,HDB_CPI_AdjPx), School_Dist)	9.00%