GovTech Presentation

Section 1

Tan Jun Yu (JY)

Question 1 (Prediction)

- Dataset: Annual Motor Vehicle Inspection - Passing Rate of Motor Vehicles on First Inspection
- Shape: 792 x 6

	year	type	age	number_reported	number_passed	passing_rate
0	2006	Cars	1	455	432	94.9
1	2006	Cars	2	1082	1026	94.8
2	2006	Cars	3	73558	68432	93.0
3	2006	Cars	4	627	560	89.3
4	2006	Cars	5	17963	16716	93.1

Task 1

What's the average passing rate on first inspection each year, taking into account motorcycles of all age groups?

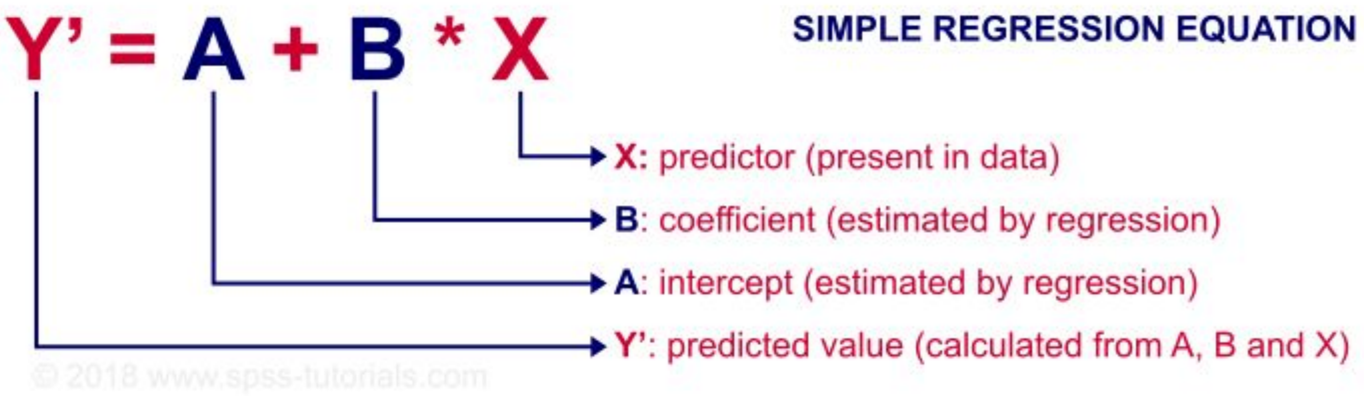
```
motor.groupby('year').mean()[['passing_rate']]
```

year	passing_rate
2006	92.990909
2007	93.686925
2008	93.870527
2009	93.873521
2010	93.981256
2011	94.045351
2012	92.791833
2013	89.050335
2014	83.116493
2015	86.459294
2016	87.506665
2017	87.604945

Task 2

For motorcycles of each age, estimate their passing rate next year.

- Ran a simple linear regression model on motorcycles of each age bracket (ages 1 - >10)



Task 2

$Y = A + Bx$

- Y: passing rate (target / predicted value)
- A: intercept of best-fit line
- B: gradient, or rate of growth/decrease of passing rate (coefficient)
- x: year (predictor)

Task 2

For motorcycles of each age, estimate their passing rate next year.

```
For motorcycles aged 1:
-----------------------
In year 2017, the passing rate was [0.].
A simple linear regression model forecasts a growth of -10.735, with an estimated [-10.735] passing rate in the next
year 2018.

For motorcycles aged 2:
-----------------------
In year 2017, the passing rate was [97.258].
A simple linear regression model forecasts a growth of 0.084, with an estimated [97.342] passing rate in the next yea
r 2018.

For motorcycles aged 3:
-----------------------
In year 2017, the passing rate was [96.96].
A simple linear regression model forecasts a growth of 0.307, with an estimated [97.267] passing rate in the next yea
r 2018.

For motorcycles aged 4:
-----------------------
In year 2017, the passing rate was [97.243].
A simple linear regression model forecasts a growth of 0.167, with an estimated [97.41] passing rate in the next year
2018.
```

Task 2

For motorcycles of each age, estimate their passing rate next year.

Motorcyle Age	2017 Passing Rate	Gradient	Predicted 2018 Passing Rate
1	0	-10.735	0
2	97.258	0.084	97.342
3	96.96	0.307	97.267
4	97.243	0.167	97.41
5	96.21	0.151	96.361
6	96.147	0.122	96.269
7	96.632	0.163	96.795
8	96.287	0.184	96.471
9	95.641	0.144	95.785
10	95.77	0.223	95.993
>10	95.506	0.091	95.597

Task 3

Assuming your estimated rates are true, can you suggest a sensible range of possible passing ranges for motorcycles in the 5-year age group next year, with at least 95% possibility of including the actual passing rate? If you can come up with multiple ranges that meet this criteria, use the one with the narrowest range. You may assume the number of motorcycles is the same as the number in the 4-year age group in the previous year.

Task 3

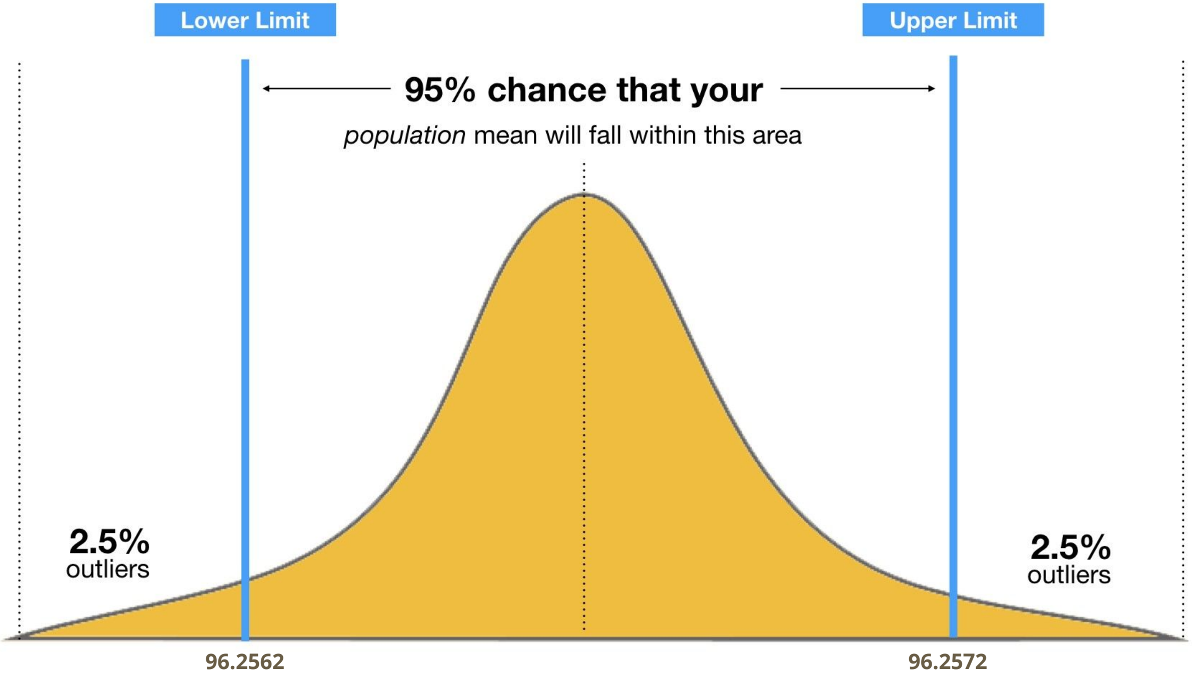
- Plot linear regression and obtain the **coefficient = 0.0465** from the model's summary statistics:

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.0465	0.000	147.873	0.000	0.046	0.047

- 2017 passing rate for motorcycles aged 5: **96.2102**
- We can say with 95% confidence that the true passing rate for 2018 lies between (96.2102 + 0.046) and (96.2102 + 0.047), OR

  [96.2562, 96.2572]

Task 3



Lower Limit
Upper Limit
95% chance that your
population mean will fall within this area
2.5%
outliers
2.5%
outliers
96.2562
96.2572

Question 2 (Association)

- Dataset: CEA Salespersons' Transaction Records (for HDB Resale)
- Shape: 101775 x 5

	complete_date_txt	town_txt	represented	salesperson_name	salesperson_reg_no
0	January 2017	JURONG WEST	Buyer	DERRICK YEO CHUN MENG	R018231E
1	January 2017	BUKIT MERAH	Buyer	LIM HOCK LEONG (LIN FULONG)	R027276D
2	January 2017	CENTRAL AREA	Buyer	LAWRENCE TAN CHOON KIAT (CHEN JUNJIE)	R006416I
3	January 2017	PUNGGOL	Buyer	LIM KIM HENG	R018637Z
4	January 2017	PASIR RIS	Buyer	ONG SHU LING	R024367E

Task 1

Based on the dataset, how many sales would you expect an agent to close each year? How much variation is there among agents?

Calculating average:

- Number of sales = 101775
- Number of agents = 13521
- Number of years = 3 *(there were 3 months from 2020 in the dataset, but for simplicity's sake, exclude 2020)*

Hence, expect each agent to close (101775 / 13521) / 3 = **2.51 sales per year**.

Task 1

Based on the dataset, how many sales would you expect an agent to close each year? How much **variation** is there among agents?

- First, groupby salesperson and count number of transactions, sorted by descending order:

	salesperson_reg_no	number_of_transactions
0	R043039D	719
1	R057585F	448
2	R007707D	389
3	R024302J	379
4	R026970D	262

Task 1

How much **variation** is there among agents?

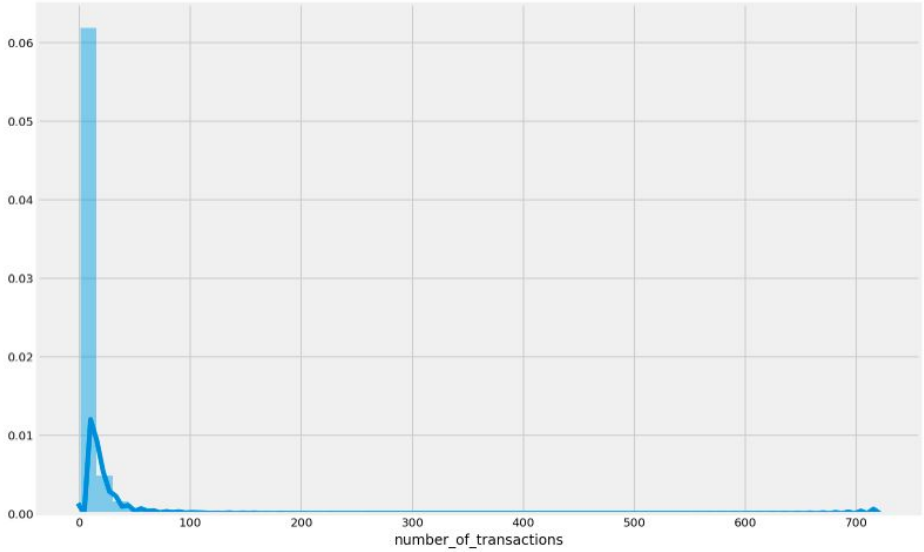
- Summary statistics:

	number_of_transactions
count	13521.000000
mean	7.527180
std	16.157642
min	1.000000
25%	1.000000
50%	3.000000
75%	8.000000
max	719.000000

Task 1

How much **variation** is there among agents?

- Distribution plot:



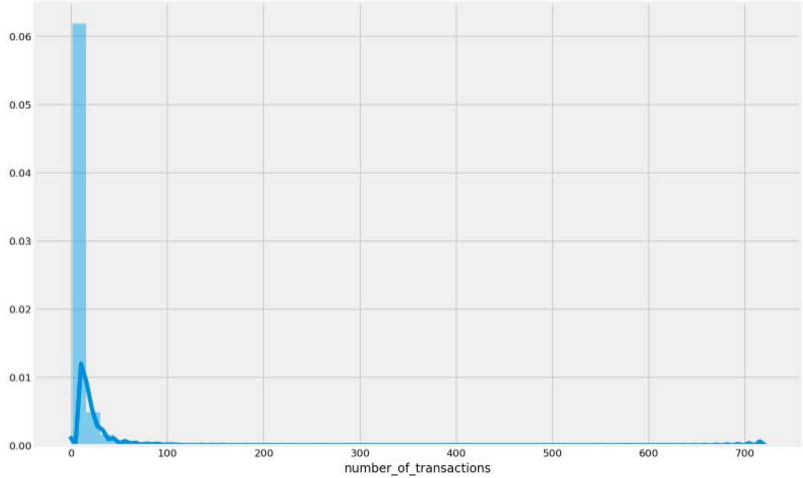
Task 1

How much **variation** is there among agents?

- Extremely huge variation in the number of sales among agents.
- Around 75% of agents close less than 10 sales over 3 years (so on average, 3 cases a year), while the remaining 25% of agents close more than 10 cases over 3 years.
- There are extreme outliers (top performing agents) that close more than 300 cases to 719 cases (the maximum) over 3 years.

Task 2

Examine the distribution for number of sales closed by an agent in a year & suggest a probability distribution that may be suitable for modelling this set of values. What are some ways in which your suggested distribution is appropriate? What are some of its limitations?



Task 2

Power Law distribution

- Sales per agent per year observes a classic long tail distribution.
- Appropriate in the context of jobs related to sales, **or networks**.
    - The better the agent is (in terms of sales), the stronger the network effects (reputation increases, referrals increases), and his/her sales increases as well, perpetuating the cycle.
    - Not at all surprising to observe the large disparity between the top performers (with the top performer selling almost double of the second top performer) and the rest.

Task 2

Power Law distribution - Limitation

- Non-normal distribution: as we increase the number of samples we take, values will NOT converge to an average.
    - Central Limit Theorem doesn't hold!
- They will, in fact, diverge, with some exceptions. This explains why the above calculation of how much sales on average to expect of each agent (2.51 per year) is so different from the ground truth.

Task 3

Property agents tend to specialise in one or more specific geographical areas, rather than ply their trade equally island-wide. Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year? (*Note: you may wish to use association rules for this task.*)

Task 3

Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year?

- First, filter original DataFrame by agents who have closed sales in EITHER Sembawang or Yishun:

	complete_date_txt	town_txt	represented	salesperson_name	salesperson_reg_no
0	January 2017	JURONG WEST	Buyer	DERRICK YEO CHUN MENG	R018231E
1	January 2017	BUKIT MERAH	Buyer	LIM HOCK LEONG (LIN FULONG)	R027276D
2	January 2017	CENTRAL AREA	Buyer	LAWRENCE TAN CHOON KIAT (CHEN JUNJIE)	R006416I
3	January 2017	PUNGGOL	Buyer	LIM KIM HENG	R018637Z
4	January 2017	PASIR RIS	Buyer	ONG SHU LING	R024367E

	year	town_txt	salesperson_reg_no
9	2017	YISHUN	R043256G
18	2017	YISHUN	R047261E
21	2017	YISHUN	R046477I
31	2017	YISHUN	R050567Z
34	2017	YISHUN	R057473F

Task 3

Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year?

- Next, do a groupby salesperson and count of sales by town (‘town_txt’)
- Further filter this DataFrame by looking only at count of ‘town_txt’ > 1

	year	town_txt	salesperson_reg_no
9	2017	YISHUN	R043256G
18	2017	YISHUN	R047261E
21	2017	YISHUN	R046477I
31	2017	YISHUN	R050567Z
34	2017	YISHUN	R057473F

	salesperson_reg_no	town_txt
0	R047710B	52
1	R043232Z	26
2	R043256G	25
3	R027388D	22
4	R018918B	21

Task 3

Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year?

- From this resulting filtered DataFrame, I can get the list of salesperson who have closed sales in BOTH Sembawang AND Yishun (since the count of sales was set to >1)

	salesperson_reg_no	town_txt
0	R047710B	52
1	R043232Z	26
2	R043256G	25
3	R027388D	22
4	R018918B	21

Task 3

Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year?

- Going back to the original DataFrame, filter this once more by showing only sales made the list of salesperson who have closed sales in both towns

	complete_date_txt	town_txt	represented	salesperson_name	salesperson_reg_no
0	January 2017	JURONG WEST	Buyer	DERRICK YEO CHUN MENG	R018231E
1	January 2017	BUKIT MERAH	Buyer	LIM HOCK LEONG (LIN FULONG)	R027276D
2	January 2017	CENTRAL AREA	Buyer	LAWRENCE TAN CHOON KIAT (CHEN JUNJIE)	R006416I
3	January 2017	PUNGGOL	Buyer	LIM KIM HENG	R018637Z
4	January 2017	PASIR RIS	Buyer	ONG SHU LING	R024367E

	year	town_txt	salesperson_reg_no
9	2017	YISHUN	R043256G
12	2017	SENGKANG	R027089C
17	2017	HOUGANG	R044058F
20	2017	BEDOK	R045202I
26	2017	WOODLANDS	R044930C

Task 3

Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year?

- Now that I know this DataFrame contains ONLY agents that sold in both towns, I simply need to extract the **non-Sembawang/Yishun towns with the most count**

Task 3

Given a property agent who has closed sales in Sembawang and Yishun during a given year, which other town is he/she most likely to be active in that year?

SembYish_2017

town_txt	salesperson_reg_no
WOODLANDS	723
JURONG WEST	374
SENGKANG	345
PUNGGOL	310
TAMPINES	273

SembYish_2018

town_txt	salesperson_reg_no
WOODLANDS	936
SENGKANG	549
JURONG WEST	539
PUNGGOL	538
TAMPINES	449

SembYish_2019

town_txt	salesperson_reg_no
WOODLANDS	996
SENGKANG	682
JURONG WEST	531
PUNGGOL	421
TAMPINES	416

SembYish_2020

town_txt	salesperson_reg_no
WOODLANDS	101
SENGKANG	50
TAMPINES	43
PUNGGOL	42
PASIR RIS	37

Question 3 (Classification)

- Dataset: Wireless HotSpots (GeoJSON format)
- Lots of cleaning and data wrangling first
    - Explore the data and access the relevant keys/values
    - Clean data using BeautifulSoup

```
wifi.description[0]
```

```
'<center><table><tr><th colspan=\'2\' align=\'center\'><em>Attributes</em></th></tr><tr bgcolor="#E3E3F3"> <th>Y</th>
<td>30059.55365961</td> </tr><tr bgcolor=""> <th>X</th> <td>24230.13882604</td> </tr><tr bgcolor="#E3E3F3"> <th>LOCAT
ION_NAME</th> <td>IHIS-NUHS AH Campus - Ward3</td> </tr><tr bgcolor=""> <th>LOCATION_TYPE</th> <td>Healthcare</td> </
tr><tr bgcolor="#E3E3F3"> <th>POSTAL_CODE</th> <td>159964</td> </tr><tr bgcolor=""> <th>STREET_ADDRESS</th> <td>378 A
lexandra Road</td> </tr><tr bgcolor="#E3E3F3"> <th>OPERATOR_NAME</th> <td>M1</td> </tr><tr bgcolor=""> <th>INC_CRC</t
h> <td>7805E8A17671DB33</td> </tr><tr bgcolor="#E3E3F3"> <th>FMEL_UPD_D</th> <td>20190527093724</td> </tr></table></c
enter>'
```

```
['Attributes',
 'Y',
 'X',
 'LOCATION_NAME',
 'LOCATION_TYPE',
 'POSTAL_CODE',
 'STREET_ADDRESS',
 'OPERATOR_NAME',
 'INC_CRC',
 'FMEL_UPD_D']
```

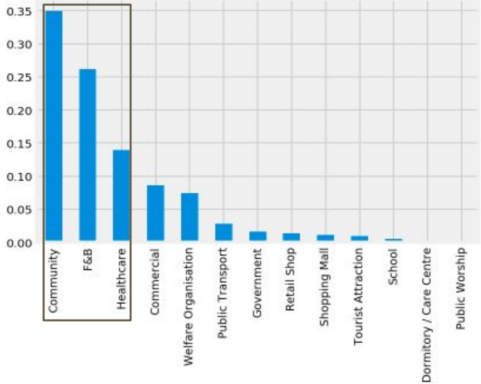
```
['30059.55365961',
 '24230.13882604',
 'IHIS-NUHS AH Campus - Ward3',
 'Healthcare',
 '159964',
 '378 Alexandra Road',
 'M1',
 '7805E8A17671DB33',
 '20190527093724']
```

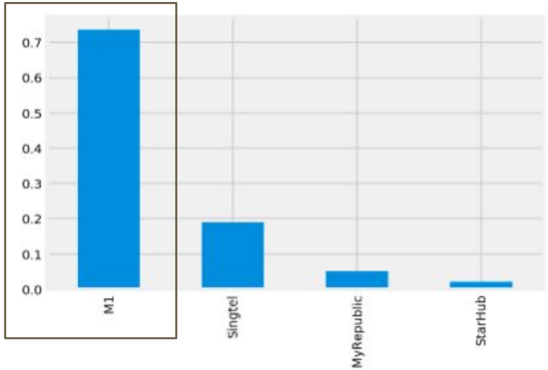
Task 1

From the table, what are some of the information you can deduce for each hotspot?

	lat	long	Y	X	LOCATION_NAME	LOCATION_TYPE	POSTAL_CODE	STREET_ADDRESS	OPERATOR_NAME	
0	103.799445	1.288122	30059.55365961	24230.13882604	IHIS-NUHS AH Campus - Ward3	Healthcare	159964	378 Alexandra Road	M1	7805E8A
1	103.799445	1.288122	30059.55365961	24230.13882604	IHIS-NUHS AH Campus - default location	Healthcare	159964	378 Alexandra Road	M1	7805E8
2	103.948071	1.340719	35875.76459983	40770.67589728	IHIS-Singhealth CGH Campus - IB - L1	Healthcare	529898	6 Simei Street 3	M1	2EB30FB
3	103.948071	1.340719	35875.76459983	40770.67589728	IHIS-Singhealth CGH Campus - IB - L2	Healthcare	529898	6 Simei Street 3	M1	2EB30FB
4	103.948071	1.340719	35875.76459983	40770.67589728	IHIS-Singhealth CGH Campus - IB - L3	Healthcare	529898	6 Simei Street 3	M1	2EB30FB

Task 1





	Y	X
OPERATOR_NAME		
M1	36569.122831	29327.493671
MyRepublic	34024.590095	28592.073393
Singtel	35107.800694	28181.615860
StarHub	32200.275675	28034.987791

1) 35% of hotspot areas are at Community areas, followed by F&B (26%) and Healthcare (14%) - totalling around 75% of Singapore's hotspots.
2) M1 provides the most hotspot coverage around the state at close to 74%, with Singtel lagging in second at 19% island-wide coverage.
3) M1 covers, on average, more of the northern and eastern parts of Singapore than other operators.

Task 2

Due to a system error, the location type column for the last 200 rows of the dataset has become garbled. Using all earlier rows as well as all other columns in the dataset, build a classification model to predict the location type for these hotspots. You may treat the three rarest location types as one category. *(Note: you may wish to create some additional features based on available ones.)*

Task 2

Due to a system error, the location type column for the last 200 rows of the dataset has become garbled. Using all earlier rows as well as all other columns in the dataset, build a classification model to predict the location type for these hotspots. You may treat the three rarest location types as one category.

(Note: you may wish to create some additional features based on available ones.)

```
School                      9
Dormitory / Care Centre     5
Public Worship              2
Name: LOCATION_TYPE, dtype: int64
```

```
Community                 0.349177
F&B                       0.262035
Healthcare                0.139549
Commercial                0.085923
Welfare Organisation      0.074345
Public Transport          0.028032
Government                0.016453
Retail Shop               0.014016
Shopping Mall             0.010969
Tourist Attraction        0.009750
Rare                      0.009750
Name: LOCATION_TYPE, dtype: float64
```

Task 2

Pre-modelling phase:

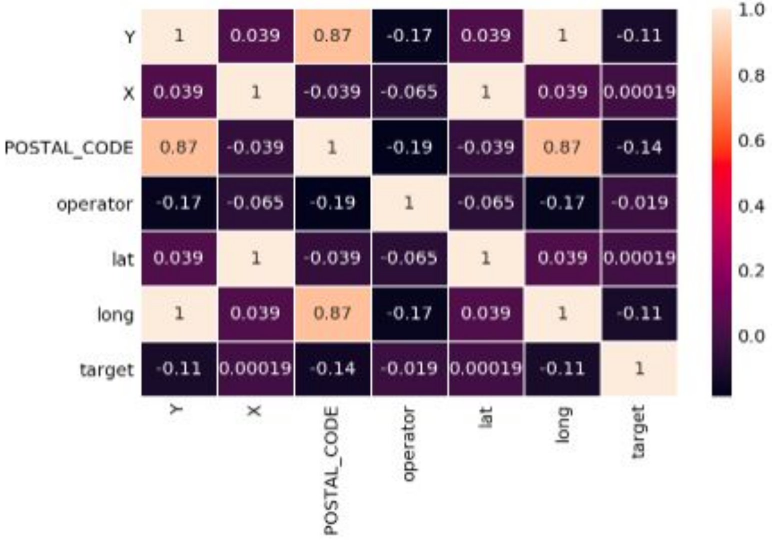
- Convert certain features and target to numerical type via Label Encoding
    - Feature: ‘operator’ (M1, MyRepublic, etc)
    - Target: ‘LOCATION_TYPE’ (Community, F&B, Healthcare, etc)
        - e.g.

BRIDGE-TYPE (TEXT)	BRIDGE-TYPE (NUMERICAL)
Arch	0
Beam	1
Truss	2
Cantilever	3
Tied Arch	4
Suspension	5
Cable	6

Task 2

Pre-modelling phase:

- Check correlation heatmap
  - ‘Y’ and ‘X’ are perfectly correlated with ‘long’ and ‘lat’ respectively (both are coordinates)
    - Hence, exclude ‘long’ and ‘lat’ from modelling



Task 2

Pre-modelling phase:

- Features: 'Y', 'X', 'POSTAL_CODE', 'operator'
- Target: 'target' ('LOCATION_TYPE' converted to integers for each location)
    - Total of 11 unique target values (or outcomes)
- Train-test-split with target being split in a stratified manner, due to slight imbalance in dataset

```
Community                0.349177
F&B                      0.262035
Healthcare               0.139549
Commercial               0.085923
Welfare Organisation     0.074345
Public Transport         0.028032
Government               0.016453
Retail Shop              0.014016
Shopping Mall            0.010969
Tourist Attraction       0.009750
Rare                     0.009750
Name: LOCATION_TYPE, dtype: float64
```

Task 2

Modelling phase:

- **Multi-class classification problem**
- Choose between 3 models
  - KNearestNeighbours (KNN) Classifier
  - Random Forests (RF)
  - Support Vector Machine (SVM)
- Evaluation metric: Accuracy score (classification rate)
  - Measures how well (how correctly) the model classifies the target (location type), and ranges from 0 to 1.
    - A score of 1 indicates perfect classification (all location types are predicted and classified correctly).

	family	model	classification_rate
0	KNN	KNN-1	0.459854
1	KNN	KNN-2	0.484185
2	KNN	KNN-3	0.445255
3	KNN	KNN-4	0.459854
4	KNN	KNN-5	0.445255
5	KNN	KNN-6	0.450122
6	KNN	KNN-7	0.440389
7	KNN	KNN-8	0.452555
8	KNN	KNN-9	0.459854
9	RF	RF-10	0.686131
10	RF	RF-100	0.669100
11	RF	RF-1000	0.656934
12	SVM	SVM-linear	0.201946
13	SVM	SVM-rbf	0.085158
14	SVM	SVM-sigmoid	0.262774

Task 3

The information has now been recovered from a backup copy of the file. Compared to the true location types, how good was your model? Be prepared to explain the metrics you use to evaluate your model.

- Chosen model: **RF-10,** or Random Forest Classifier with n_estimators = 10
- Upon predicting and scoring on completely unseen data (after training the model on whole training set):

```
accuracy_score(y_test, y_pred)
```

```
0.96
```

Task 3

```
accuracy_score(y_test, y_pred)
```

```
0.96
```

- Interpreting this score, it means that out of the 200 rows of unseen data, 96% of them were correctly predicted in terms of the location type.
    - 192 out of 200 predictions were correct, 8 were wrongly misclassified

Task 3

- 0 represents Commercial
- 1 represents Community
- 2 represents F&B
- 3 represents Government
- 4 represents Healthcare
- 5 represents Public Transport
- 6 represents Rare
    - School, Dormitory / Care Centre, Public Worship
- 7 represents Retail Shop
- 8 represents Shopping Mall
- 10 represents Welfare Organisation

Misclassifications

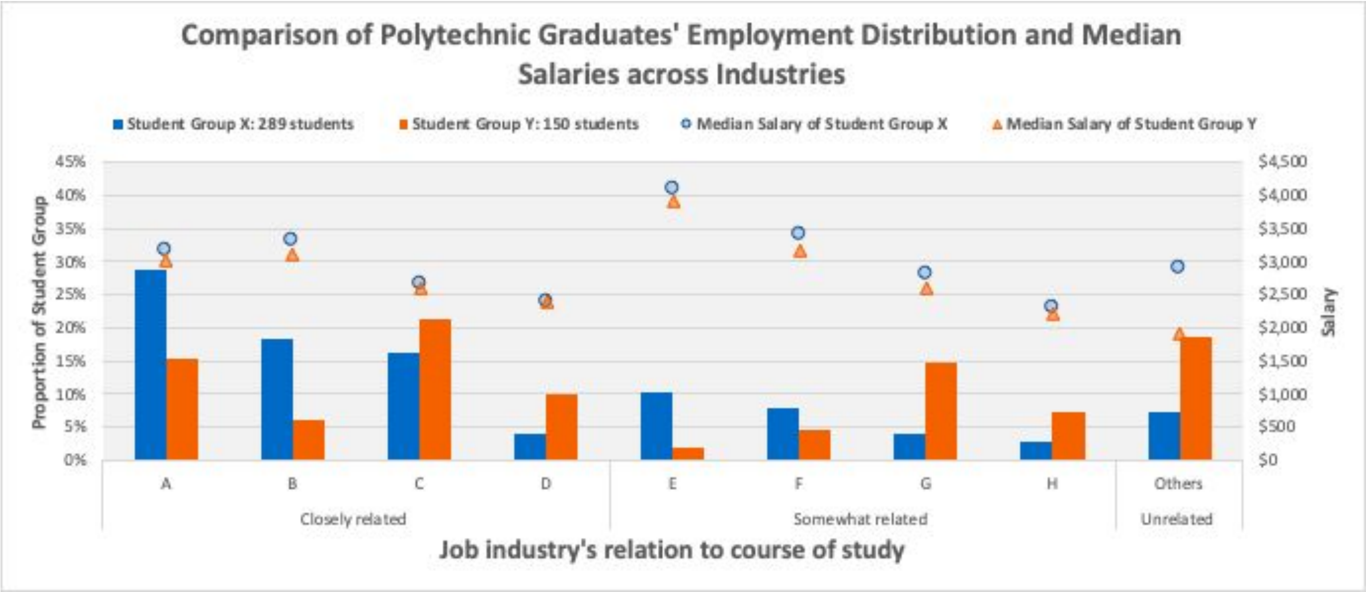
target	operator	predictions
2	2	5
5	2	2
0	2	2
1	0	0
0	3	8
8	3	3
6	3	0
10	0	1

Question 4 (Data Visualisation)

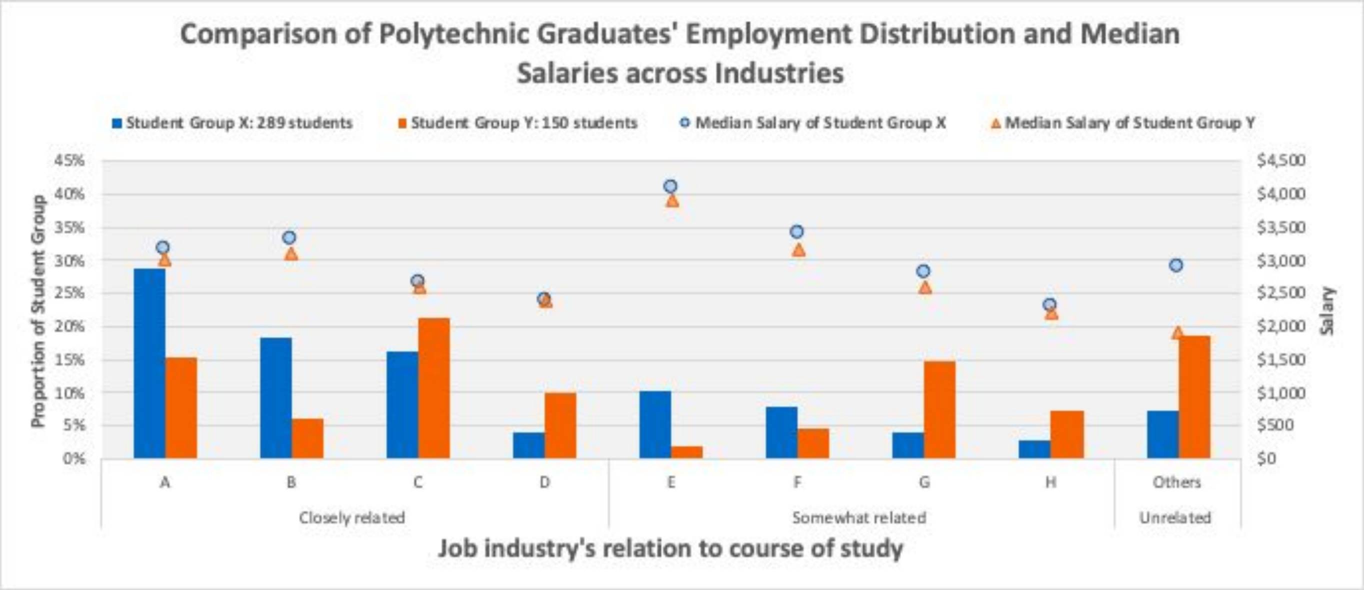
A colleague is working with a salary dataset based on recent poly graduates in a specific course of study highly subsidised by the government, to compare whether the career choices made by students from Group X are different from those from Group Y in any manner. She has already produced the following summary table and listed out the main insight she wishes to highlight, as well as pertinent observations on the dataset's characteristics, but is struggling to come up with a good way to communicate the insight to her audience in one visualisation while also accurately reflecting the dataset's characteristics.

Task

Help your colleague present the insight in an intuitive manner that is easily understood by a non-technical audience, and that reflects as many characteristics in the list as possible. Be prepared to justify any and every aspect of your visualisation (e.g. chart choice, colour palette, labels, orientation, etc.).



- Students from Group X > students from Group Y in this course of study.
- Proportionately more students from Group Y are in jobs unrelated to their course of study.
- The distribution of students among various industries is considerably different between the two student groups.
- Students from Group X tend to command higher salaries, for the same type of job & industry.
- Salary differential between the two student groups differs by job nature and industry.



Main insight: We should review the policy behind subsidising this course of study, as a considerable proportion of students from each group <u>do not</u> go on to work in industries closely related to it.

For Group X, this may be partially due to higher / comparable salaries offered by other industries.
For Group Y, non-salary factors may play a more prominent role.

GovTech Presentation

Section 2

Scenario 1

Some forum posters have complained that the value of their HDB flats suffer because they are near expressways, which are very noisy. Others say expressway proximity is good, due to the unblocked view (at least for higher floors).

The Housing and Development Board has tasked your team to **analyse whether there is merit to either view, based on transaction prices for resale HDB flats in recent years.**

National Map Line Data Preparation

- Source: https://data.gov.sg/dataset/national-map-line
- GeoJSON format

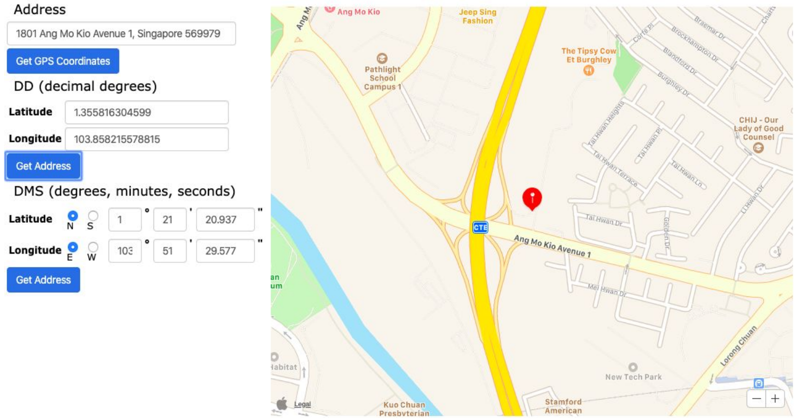
	NAME	FOLDERPATH	SYMBOLID	INC_CRC	FMEL_UPD_D
0	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	0C08DFFA475DDCCD	20191008154530
1	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	48A90A617CC124B8	20191008154530
2	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	051AA478B6209021	20191008154530
3	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	1C51FD53E1662A6B	20191008154530
4	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	44D0FFDF1EF47027	20191008154530

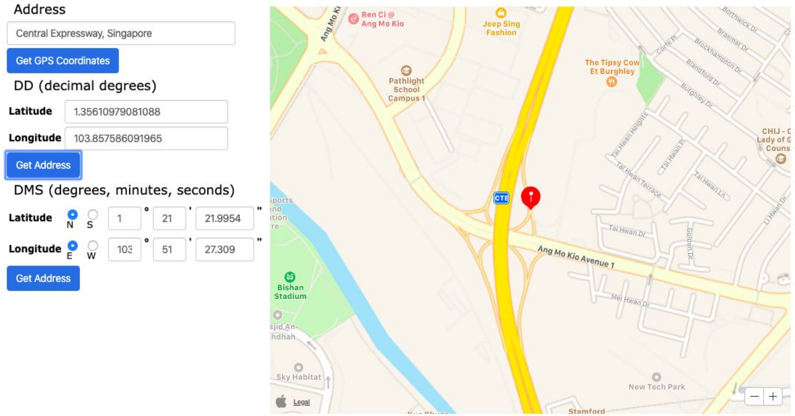
National Map Line Data Preparation

- Multiple coordinates in nested lists
- Plug in different values on actual map to check
    - Not much difference
    - Hence, simply take the first item in each nested list

```
NML.coords[0]
```

```
[[103.858333937416, 1.35595333317473, 0.0],
 [103.858215578815, 1.355816304599, 0.0],
 [103.858116866331, 1.35575566979974, 0.0],
 [103.857992826192, 1.35571405765487, 0.0],
 [103.85787572257, 1.35572105501446, 0.0],
 [103.85778107993, 1.35577301170758, 0.0],
 [103.857716551157, 1.35585094776557, 0.0],
 [103.857586091965, 1.35610979081088, 0.0]]
```





National Map Line Data Preparation

- Apply filter to get only expressways and expressway sliproads
- Final NML dataframe: 845 x 5

	lat	long	NAME	FOLDERPATH	SYMBOLID
0	103.858334	1.355953	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2
1	103.857586	1.356110	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2
2	103.860424	1.368165	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2
3	103.859780	1.372284	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2
4	103.859369	1.369135	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2

National Map Line Data Preparation

- Feature Engineer ‘Block’ and ‘Road’ from latitude and longitude input by querying OneMap Geocode API (https://docs.onemap.sg/#onemap-rest-apis)
- Understanding the data: search results from the inputted coordinates-pair are returned


```
[{'BUILDINGNAME': 'TANGLIN GROVE',
  'BLOCK': '32',
  'ROAD': 'TANGLIN HALT ROAD',
  'POSTALCODE': '142032',
  'XCOORD': '24223.6525182',
  'YCOORD': '31330.9784764',
  'LATITUDE': '1.2996205246199792',
  'LONGITUDE': '103.7993864042963',
  'LONGTITUDE': '103.7993864042963'},
 {'BUILDINGNAME': 'TANGLIN GROVE',
  'BLOCK': '31',
  'ROAD': 'TANGLIN HALT ROAD',
  'POSTALCODE': '141031',
  'XCOORD': '24233.1128068',
  'YCOORD': '31302.3827686',
  'LATITUDE': '1.2993619162962085',
  'LONGITUDE': '103.79947141266396',
  'LONGTITUDE': '103.79947141266396'},
 {'BUILDINGNAME': 'COMMONWEALTH VIEW',
  'BLOCK': '91',
  'ROAD': 'TANGLIN HALT ROAD',
  'POSTALCODE': '142091',
  'XCOORD': '24227.0909006',
  'YCOORD': '31453.4288456',
  'LATITUDE': '1.3007279228807094',
  'LONGITUDE': '103.79941728492155',
  'LONGTITUDE': '103.79941728492155'},
```

National Map Line Data Preparation

- Create function to automate API querying and data collection:

```python
counter = 0

for coords in coords_list:
    print(f'Fetching data for {coords}')
    block_list.append(get_blocks(coords))
    road_list.append(get_roads(coords))
    counter += 1
    print(counter)
```

```
Fetching data for 1.356,103.8583
1
Fetching data for 1.3561,103.8576
2
Fetching data for 1.3682,103.8604
3
Fetching data for 1.3723,103.8598
4
Fetching data for 1.3691,103.8594
5
Fetching data for 1.3693,103.8605
6
Fetching data for 1.3767,103.8589
7
Fetching data for 1.3769,103.8597
8
Fetching data for 1.3776,103.8587
9
```

- Save data into new features in NML DataFrame:

	long	lat	NAME	FOLDERPATH	SYMBOLID	nearest_block	nearest_road
0	103.858334	1.355953	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	None	None
1	103.857586	1.356110	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	None	None
2	103.860424	1.368165	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	459	ANG MO KIO AVENUE 10
3	103.859780	1.372284	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	558	ANG MO KIO AVENUE 10
4	103.859369	1.369135	CENTRAL EXPRESSWAY	Layers/Expressway_Sliproad	2	564	ANG MO KIO AVENUE 3

HDB Data Preparation

- Initially concatenated all 5 HDB datasets downloaded from https://data.gov.sg/dataset/resale-flat-prices `[hdb90to99, hdb00to12, hdb12to14, hdb15to16, hdb17onwards]`
- Resulted in a huge dataset of 812704 x 11, with sales from 1990 all the way till 2020

```
hdb_all.year.unique()
```

```
array([1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000,
       2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011,
       2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020])
```

HDB Data Preparation

- Feature engineer latitude and longitude for each HDB resale flat by querying OneMap Geocode API (https://docs.onemap.sg/#onemap-rest-apis) and understanding the data

[{'SEARCHVAL': 'DBS REVENUE HOUSE', 'BLK_NO': '55', 'ROAD_NAME': 'NEWTON ROAD', 'BUILDING': 'DBS REVENUE HOUSE', 'ADDRESS': '55 NEWTON ROAD DBS REVENUE HOUSE SINGAPORE 307987', 'POSTAL': '307987', 'X': '28963.4088901328', 'Y': '33527.9738090914', 'LATITUDE': '1.3194895900724', 'LONGITUDE': '103.841975308494', 'LONGTITUDE': '103.841975308494'}, {'SEARCHVAL': 'REVENUE HOUSE', 'BLK_NO': '55', 'ROAD_NAME': 'NEWTON ROAD', 'BUILDING': 'REVENUE HOUSE', 'ADDRESS': '55 NEWTON ROAD REVENUE HOUSE SINGAPORE 307987', 'POSTAL': '307987', 'X': '28977.8507137401', 'Y': '33547.571269167594', 'LATITUDE': '1.3196668221166499', 'LONGITUDE': '103.84210507640101', 'LONGTITUDE': '103.84210507640101'}, {'SEARCHVAL': 'INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS)', 'BLK_NO': '55', 'ROAD_NAME': 'NEWTON ROAD', 'BUILDING': 'INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS)', 'ADDRESS': '55 NEWTON ROAD INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS) SINGAPORE 307987', 'POSTAL': '307987', 'X': '28983.753727264702', 'Y': '33554.4361084122', 'LATITUDE': '1.3197289051072298', 'LONGITUDE': '103.84215811826701', 'LONGTITUDE': '103.84215811826701'}]

HDB Data Preparation

- Created function to automate scraping of relevant information (latitude and longitude)...

```
for address in addresses:
    print(f'Fetching data for {address}')
    lat_list.append(get_lats(address))
    long_list.append(get_longs(address))

print('All done!')
print(lat_list[:10])
print(long_list[:10])
```

```
Fetching data for 309 ANG MO KIO AVE 1
Fetching data for 309 ANG MO KIO AVE 1
Fetching data for 216 ANG MO KIO AVE 1
Fetching data for 211 ANG MO KIO AVE 3
Fetching data for 202 ANG MO KIO AVE 3
Fetching data for 235 ANG MO KIO AVE 3
Fetching data for 235 ANG MO KIO AVE 3
Fetching data for 232 ANG MO KIO AVE 3
Fetching data for 232 ANG MO KIO AVE 3
Fetching data for 308 ANG MO KIO AVE 1
Fetching data for 308 ANG MO KIO AVE 1
Fetching data for 220 ANG MO KIO AVE 1
Fetching data for 219 ANG MO KIO AVE 1
Fetching data for 247 ANG MO KIO AVE 3
Fetching data for 320 ANG MO KIO AVE 1
Fetching data for 252 ANG MO KIO AVE 4
Fetching data for 223 ANG MO KIO AVE 1
Fetching data for 223 ANG MO KIO AVE 1
Fetching data for 230 ANG MO KIO AVE 3
Fetching data for 329 ANG MO KIO AVE 3
```

..BUT it’s taking too long! At least half a day probably.

HDB Data Preparation

- Hence, made the call to stop running the code to fetch data for all 812704 entries.
  - Taking too long!
  - hdb90to99 and hdb00to12 had minimum resale prices of 5k and 28k respectively. There is nothing wrong with values, but these sales were transacted at a time when inflation and cost of living weren't so high yet. Hence, including these long-ago datasets might not be very informative, and may even skew the insights.

	floor_area_sqm	lease_commence_date	resale_price
count	287200.000000	287200.000000	287200.000000
mean	93.351439	1983.206741	219541.850313
std	27.361839	6.085734	128144.384286
min	28.000000	1967.000000	5000.000000
25%	68.000000	1979.000000	127000.000000
50%	91.000000	1984.000000	195000.000000
75%	113.000000	1987.000000	298000.000000
max	307.000000	1997.000000	900000.000000

hdb90to99

	floor_area_sqm	lease_commence_date	resale_price
count	369651.000000	369651.000000	369651.000000
mean	96.586204	1987.984659	281271.860617
std	25.598886	9.122421	112118.967206
min	28.000000	1966.000000	28000.000000
25%	73.000000	1981.000000	195000.000000
50%	100.000000	1987.000000	263000.000000
75%	115.000000	1997.000000	350000.000000
max	297.000000	2012.000000	903000.000000

hdb00to12

HDB Data Preparation

- Focus on only HDB resale transactions from 2019 onwards.
  - Shape: 24113 x 15
- Perform API querying and collection once more... done:

storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price	remaining_lease	year	address	lat	long
01 TO 03	68.0	new_generation	1981	270000.0	61.0	2019	330 ANG MO KIO AVE 1	1.3624318640247899	103.851030689651
04 TO 06	73.0	new_generation	1976	295000.0	56.0	2019	215 ANG MO KIO AVE 1	1.36655830166124	103.841624082978
07 TO 09	67.0	new_generation	1978	270000.0	58.0	2019	225 ANG MO KIO AVE 1	1.3673961277686297	103.83815000746401
01 TO 03	67.0	new_generation	1978	230000.0	58.0	2019	225 ANG MO KIO AVE 1	1.3673961277686297	103.83815000746401
01 TO 03	68.0	new_generation	1981	262500.0	61.0	2019	333 ANG MO KIO AVE 1	1.3613425564061299	103.85169862145399

Reflections / Room For Improvement

- I faltered at close to 6am..
- Main obstacle: **unable to figure out a programmatic way to group each flat's coordinates with the appropriate expressway coordinates and perform distance calculation.**
- It gets even trickier when considering expressways such as PIE or CTE span a huge distance, and have many exits/sliproads.
- If use one single coordinate-pair for every expressway and perform the distance calculation, oversimplification and will skew results.
- Conclusion: need to spend more time and effort researching on how to overcome this problem.

Thank you!

Questions?