MINDEF Technical Assessment

Quantitative Strategy Case Study Interview

Gary Goh Shing Wee

25 October 2023

All source code and analysis can be found in this Github repository:
https://github.com/garygsw/mindef-assessment

Table of Contents

Section 1: Scenario 1

Section 1: Scenario 1

Problem statement:

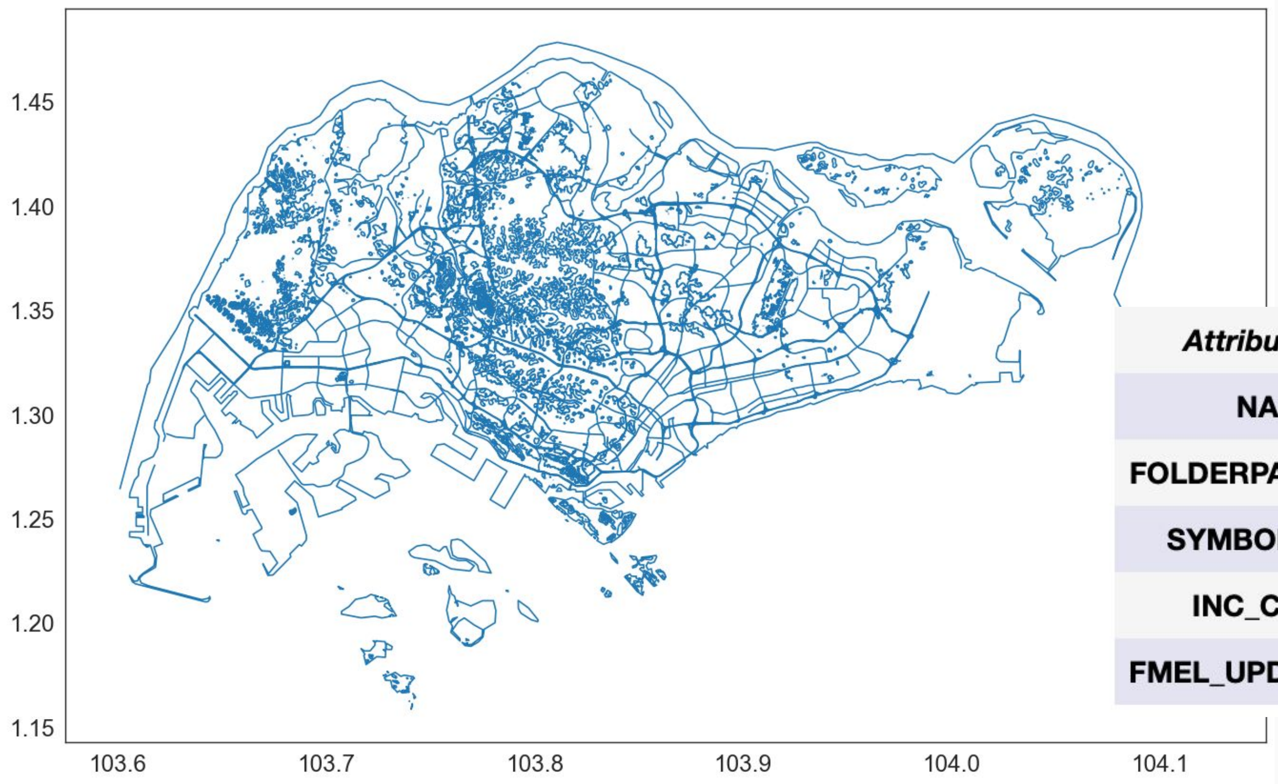
Analyse the relationship between HDB flat prices with proximity to expressways

Dataset(s) / source(s):

1. Resale Flat Prices from Jan 2017 onwards (link)
   - ➢ Filter by month from Oct 2021 onwards (2 years)
2. OneMap Geocode API (link)
3. National Map Line (link)
   - ➢ Filter expressways only
4. National Map Polygon (link)

Section 1: Scenario 1

Raw data

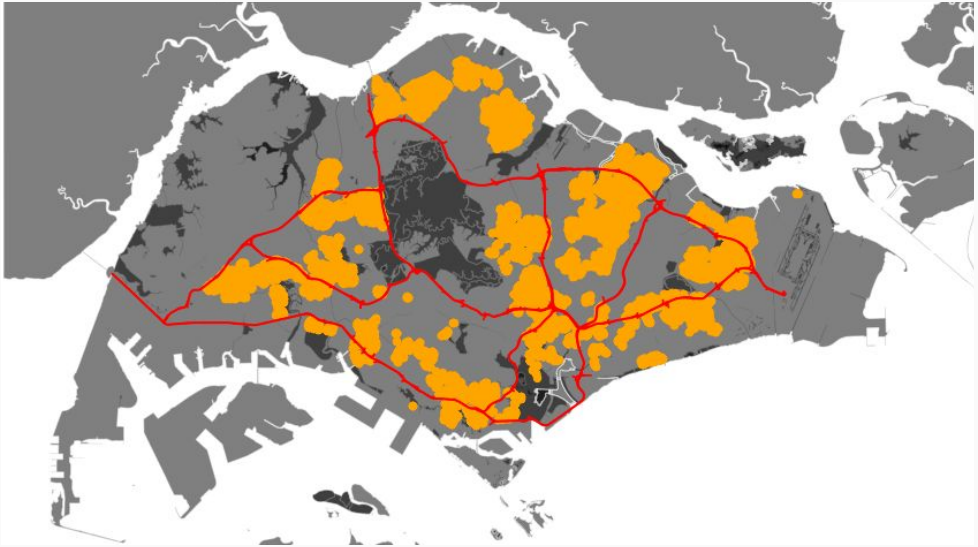


Filter: EXPRESSWAY only

Attributes	
NAME	CENTRAL EXPRESSWAY
FOLDERPATH	Layers/Expressway_Sliproad
SYMBOLID	2
INC_CRC	0C08DFFA475DDCCD
FMEL_UPD_D	20191008154530

Section 1: Scenario 1

Obtain location from OneMap Geocode API: search term= `block` + `street_name`



Next step: For each location, find the distance to the nearest expressway

Approach:

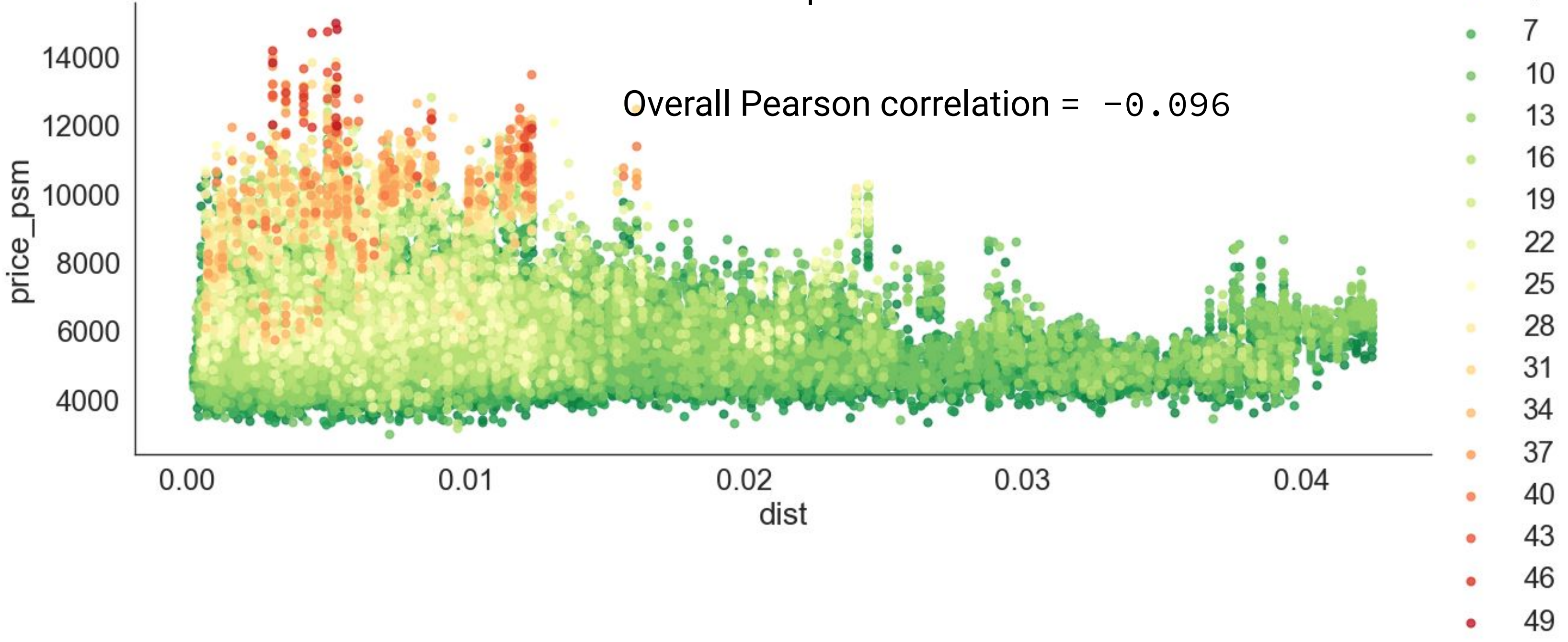
- ➢ Use spatial cKDTree to store all expressway lines for quick lookup for nearest points
- ➢ Use latitudes and longitudes to compute distances

Section 1: Scenario 1

Find correlation between resale price per square meter and distance

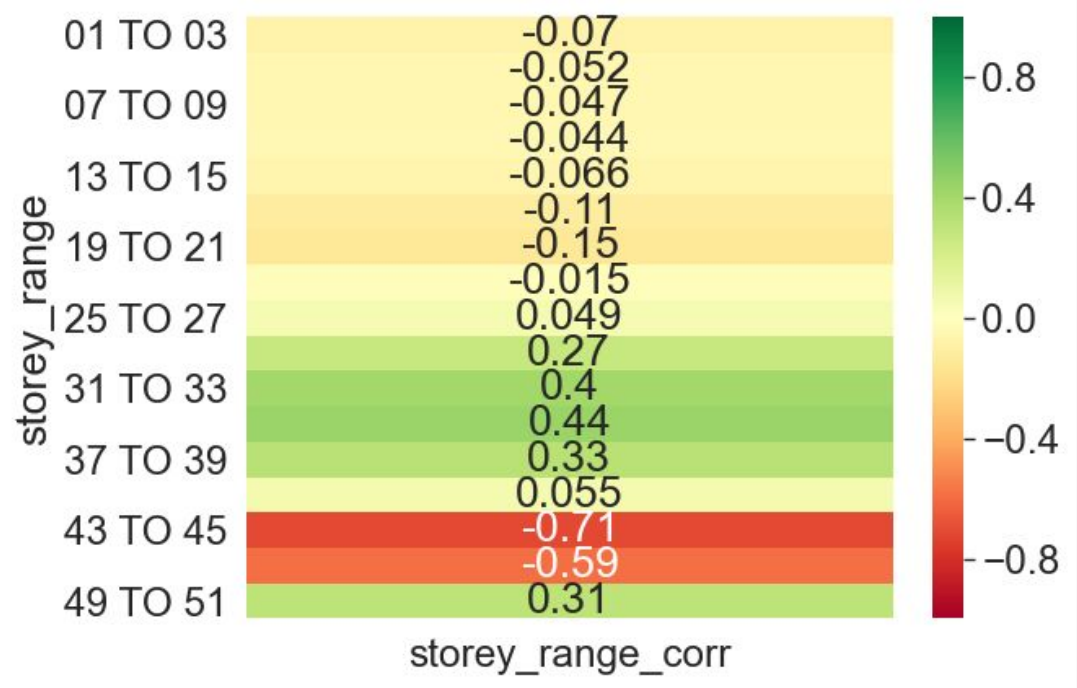
Observations:

- Most higher floors have higher prices given same distance
- There seems to be different relationships for different floors



Section 1: Scenario 1

storey_range	
01 TO 03	9496
04 TO 06	12542
07 TO 09	11700
10 TO 12	10405
13 TO 15	5319
16 TO 18	2574
19 TO 21	1079
22 TO 24	733
25 TO 27	507
28 TO 30	352
31 TO 33	202
34 TO 36	181
37 TO 39	144
40 TO 42	68
43 TO 45	25
46 TO 48	12
49 TO 51	6



Conclusions:

- Very weak negative correlation for low floors (27 and below)
- Medium positive correlation for mid floors (27 to 42)
- Somewhat strong negative correlation for high floors (>42 floors) – caution low sample size

Section 1: Scenario 1

Further potential improvements:

- ➢ Consider underground (less noisier?) or above ground expressways
- ➢ Consider traffic flows of expressways (different parts of expressway have different noise levels)
- ➢ Consider spatial distribution of different floors
- ➢ Set a distance threshold and convert to a binary value either “near” or “not near” and then analyze correlation

Section 2: Question 1

Section 2: Question 1 (Association)

Problem statement:

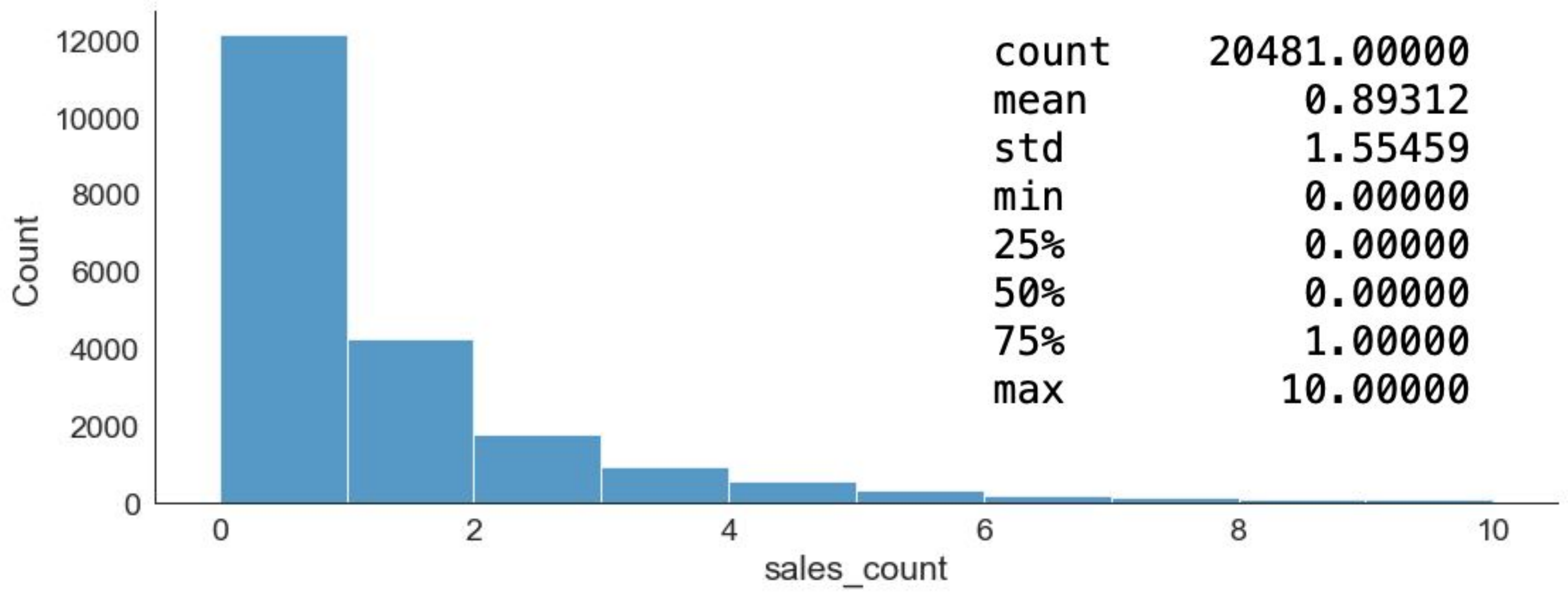
Find a suitable probability distribution to model the distribution for the number of HDB resale flat sales in a year closed by a property agent representing the seller.

Dataset(s) / source(s):

1. CEA Salesperson's Residential Property Transaction Records (link)
   - ➢ Filter by transaction dates from 2022-Oct onwards (1 year)

Section 2: Question 1 (Association)

Distribution (remove outliers by ignoring `counts > 10`)



Note: `salesperson_reg_num` is used as the unique identifier for the property agent

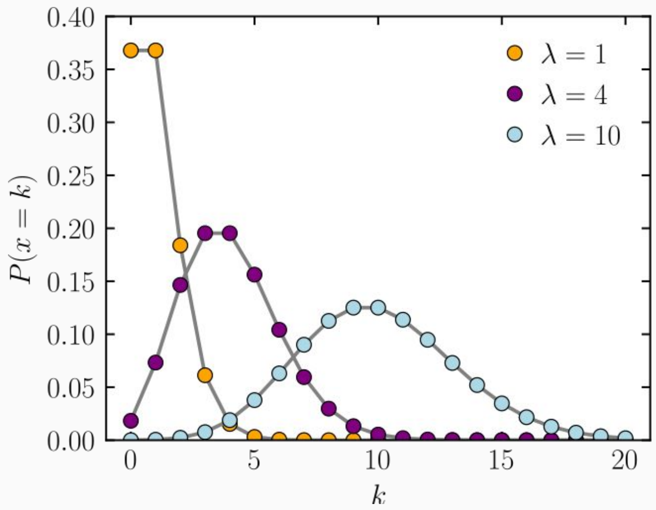
Section 2: Question 1 (Association)

Distribution to fit: Poisson distribution

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where

- k is the number of occurrences ($k = 0, 1, 2, \ldots$)
- e is Euler's number ($e = 2.71828\ldots$)
- $!$ is the factorial function.

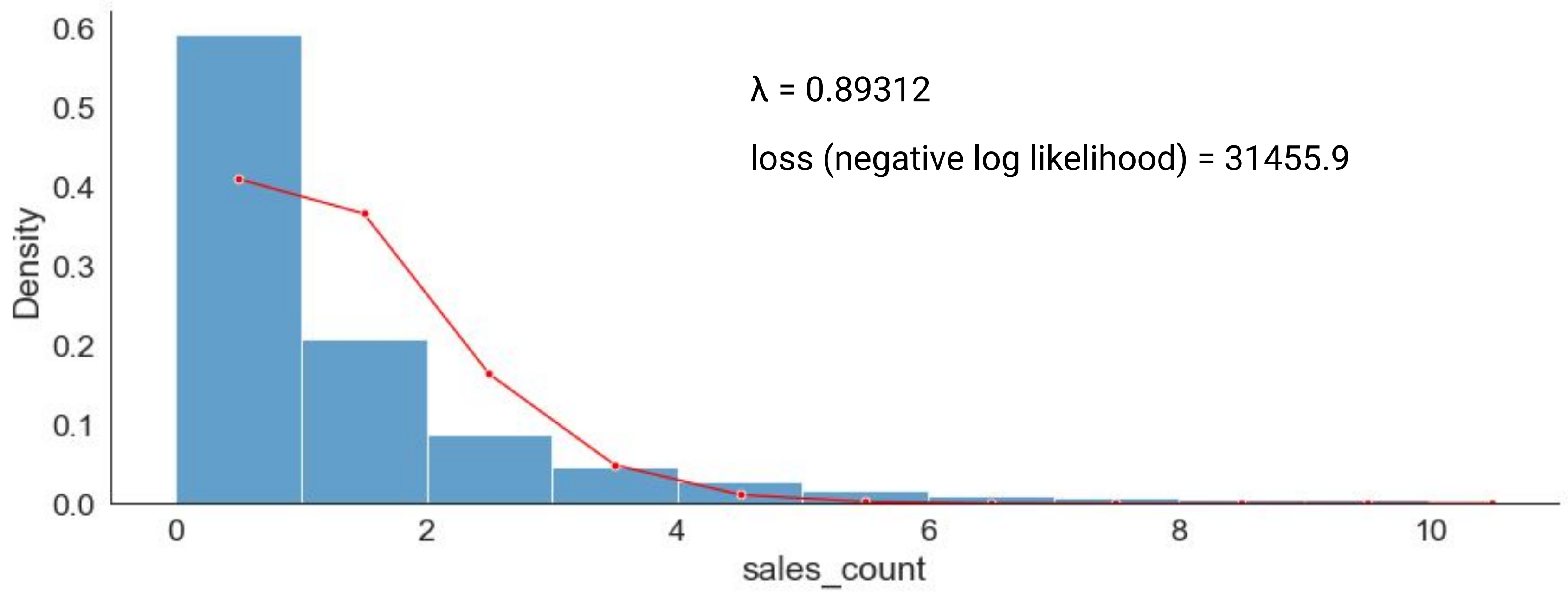


Characteristics:

- ➢ Discrete
- ➢ Expresses the probability of number of events occurring in a fixed time interval
- ➢ Counts the number of successes in independent Bernoulli trials

Section 2: Question 1 (Association)

Fitted using Maximum Likelihood Estimation



Section 2: Question 1 (Association)

Assumptions:

- ➢ Events occur at a fixed rate with a constant mean and variance
- ➢ Events occur independently from each other
- ➢ Probability of success in each trial is a constant

Further potential improvements:

- ➢ Choose to fit an empirical distribution using Kolmogorov-Smirnov test
- ➢ Use supervised statistical modelling to predict the random variable based on multiple factors e.g. agent's track record, agent's marketing expenses, agent's team size

Section 2: Question 2

Section 2: Question 2 (Classification)

Problem statement:

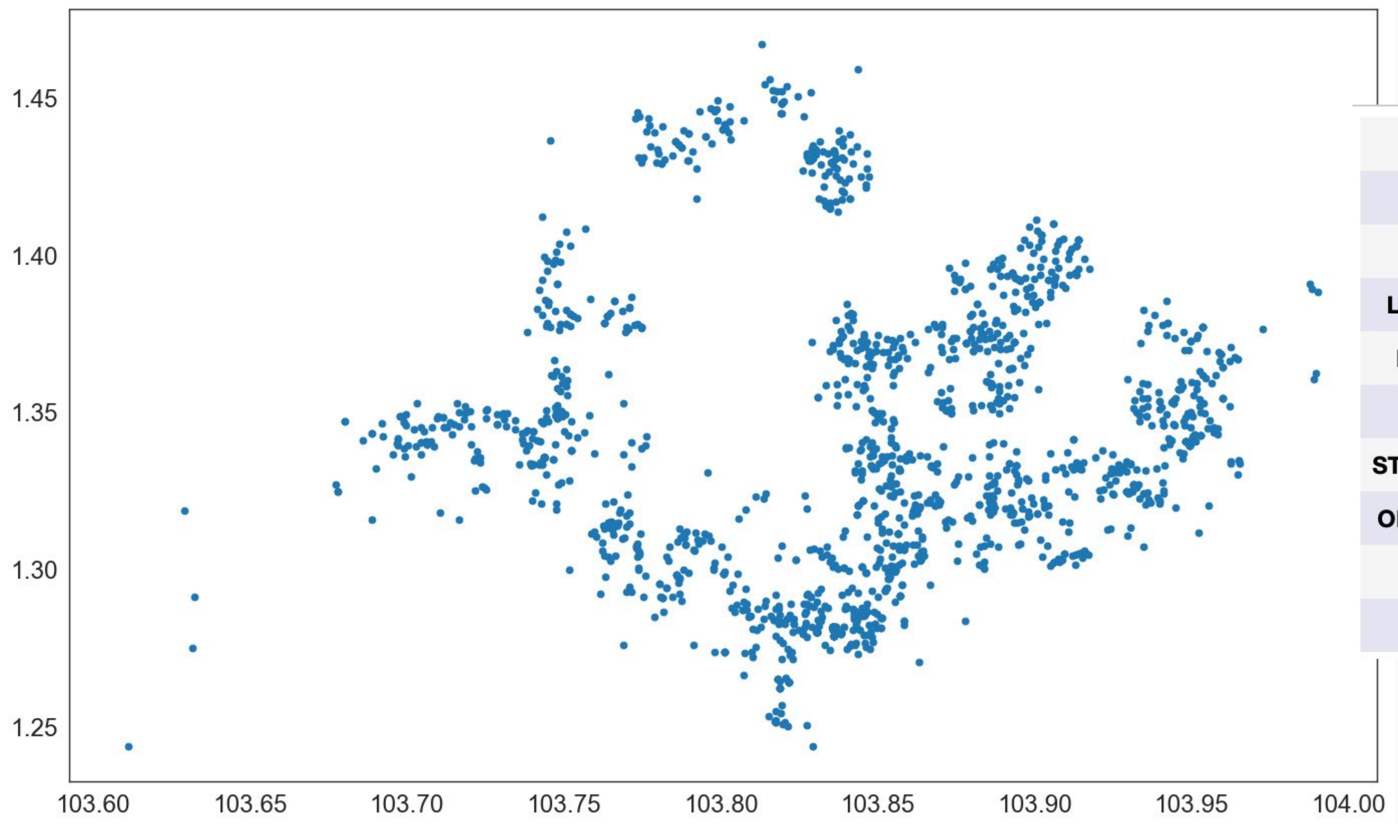
Build a multi-class classifier to predict 200 missing `location_type` values from the Wireless@SG hotspots dataset.

Dataset(s) / source(s):

1. Wireless Hotspots (link)
2. National Map Polygon (link)

Section 2: Question 2 (Classification)

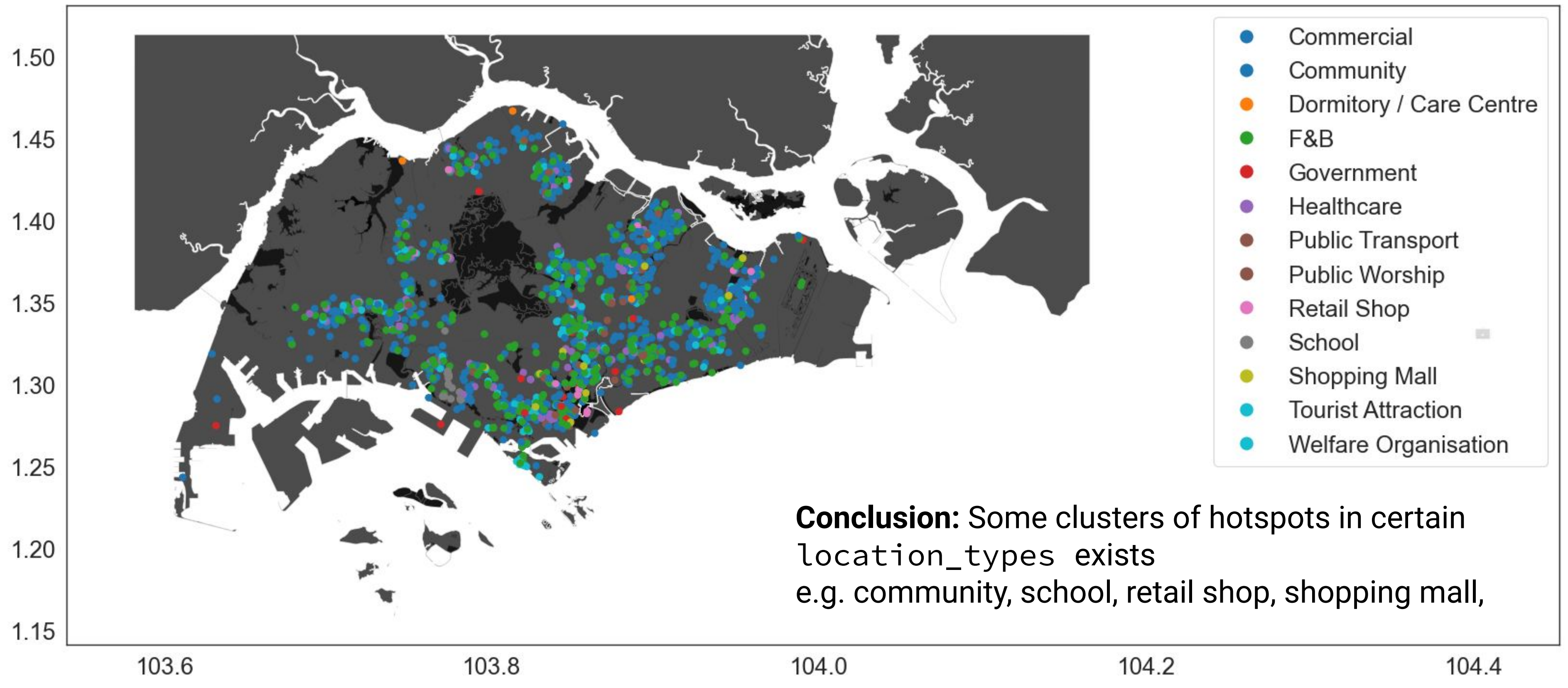
Raw data



Attributes	
Y	36922.92412
X	18450.95232
LOCATION_NAME	Bukit Batok CC
LOCATION_TYPE	Community
POSTAL_CODE	659959
STREET_ADDRESS	21 Bukit Batok Central
OPERATOR_NAME	M1
INC_CRC	9BB55356462956EF
FMEL_UPD_D	20200318162531

Section 2: Question 2 (Classification)

Explore `latitude` **and** `longitude` **relationship with** `location_type`

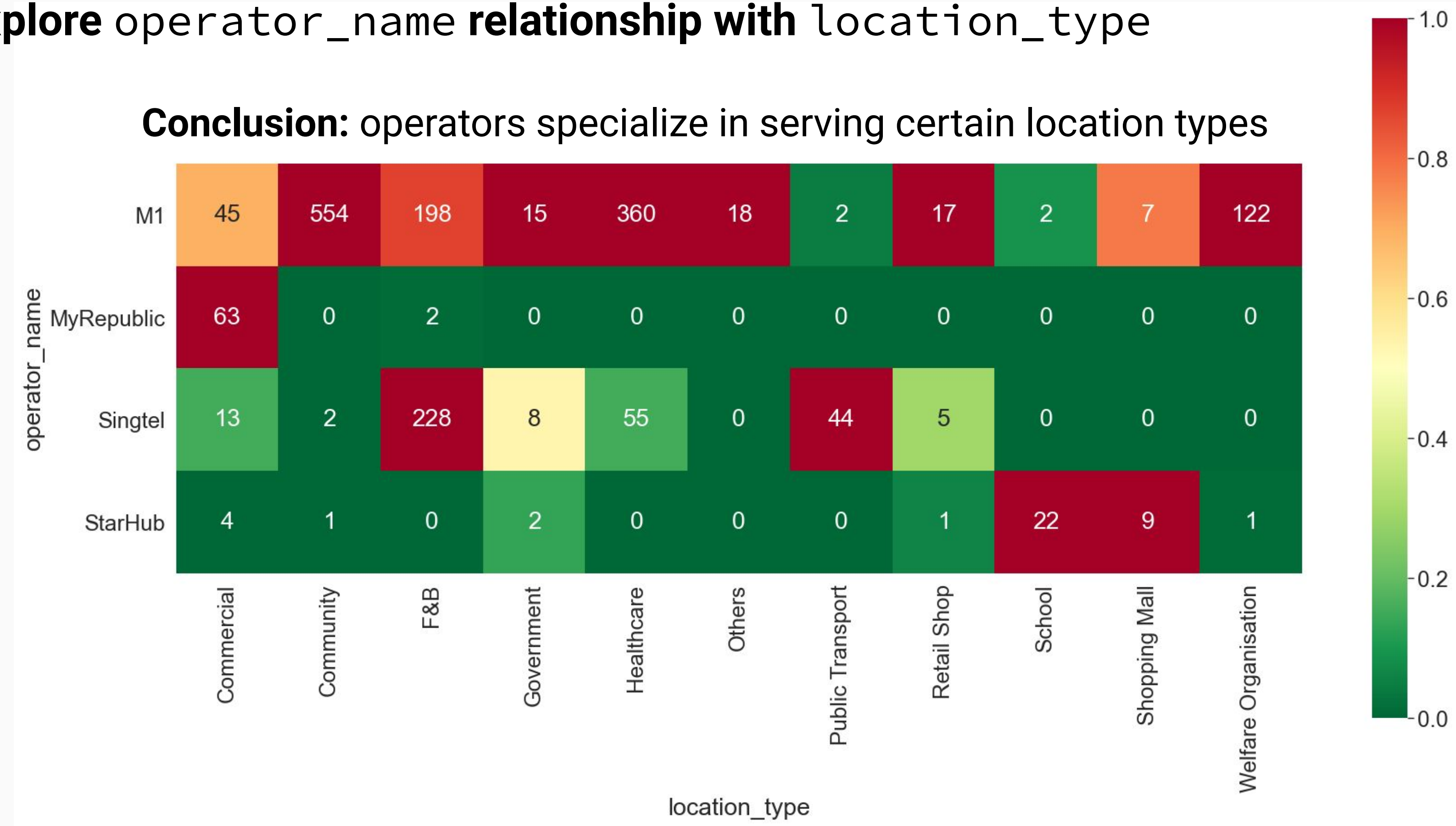


Conclusion: Some clusters of hotspots in certain `location_types` exists
e.g. community, school, retail shop, shopping mall,

Section 2: Question 2 (Classification)

Explore `operator_name` **relationship with** `location_type`

Conclusion: operators specialize in serving certain location types



Section 2: Question 2 (Classification)

Explore `location_name` **relationship with** `location_type`

Community

```
Bukit Batok CC
1 Northpoint Drive
20 Upper Pickering St
ACE The Place CC
Acacia RC
...
Tampines North Zone 6 RC
Tampines North Zone 7 RC
Keppel Club
PASIR RIS ZONE '7' RESIDENTS COMMITTEE
People's Association
```

Public Transport

```
SBS Kampong Bahru Bus Terminal
SBS Shenton Way Bus Terminal
City Hall MRT Station - EWL
City Hall MRT Station - NSL
Clake Quay MRT - NEL
Clementi - EWL
```

Healthcare

```
Ng Teng Fong Hospital - Ward_Tower(RH)_L7
Ng Teng Fong Hospital - Ward_Tower(RH)_L8
Ng Teng Fong Hospital - Ward_Tower(RH)_L9
Ng Teng Fong Hospital - Ward_Tower(RH)_LIFT
Queenstown Polyclinic
```

F&B

```
Upper Boon Keng Market & Food Centre
Whampoa Drive Makan Place
Whampoa Market
Yuhua Place
Yuhua Village
```

Commercial

```
West Coast Ferry Terminal
S K Yap Construction
Sembcorp Marine Ltd
The Swatch Group
Webnatics Singapore Pte Ltd
```

School

```
Grace Orchard School
Ngee Ann Polytechnic
CAPT
CAPT/RC4 Dining Area
Cinnamon
Cinnamon Dining Area
Eusoff Hall (EH)
Food Court / Canteen
```

Welfare Organization

```
@27 FSC
AMP
AWWA Senior Activity Centre
Comnet SAC (Sin Ming)
Covenant FSC
```

Conclusion: `location_name` is a key feature to determine `location_type`!

Section 2: Question 2 (Classification)

Features Engineering

- ➢ Categorical data: `operator_name` → One-Hot Encoding
- ➢ Numerical data: `lat`, `long` → MinMax Scaling
- ➢ Text data: `location_name` → Count Vectorization with fixed vocabulary

```
# handle location_name feature
keywords = [
    "rc", "cc", "rn", "zone", "nlb", "residents", "committee", "cafe", "hawker",  # community
    "kfc", "mcdonald", "pizza", "food", "coffee", "market", "makan",  "hotel", # f&b
    "nel", "mrt", "ccl", "ewl", "nsl", "bus",    # public transport
    "pte", "ltd", "limited", "group", "branch", "company", "industrial", "tower", "holding", # commercial
    "housing", "development", "board", "hdb", "ministiry", "national", "hub", "singapore", # government
    "hospital", "singhealth", "nfk", "sgh", "polyclinic", "medicine", "academia",
    "heart", "eye", "dental", "care",  # healthcare
    "boutique", "orchard",  # retail
    "hall", "canteen", "polytechnic", "school",  # school
    "mall", "plaza", "shopping", "square",  # shopping mall
    "home", "children", "outreach", "fsc", "sac", "senior", "seniors", "activity",   # welfare
]
```

Section 2: Question 2 (Classification)

Multiclass Classification Models

- **Binary Classifier Transformation**
  - **One vs. Rest:** Logistic Regression
  - **One vs. One:** Support Vector Classification, SGD Classifier

- **Native Multiclass Classifiers**
  - **Naive Bayes**: Multinomial, Complement
  - Decision Tree Classifier
  - k-Nearest Neighbour Classifier
  - **Ensemble:** Random Forest, Gradient Boosting
  - **Neural Networks:** Multilayer Perceptron

Metrics:

- Accuracy
- Precision
- Recall
- F1-score
  - By class
  - Macro average
  - Micro average

Section 2: Question 2 (Classification)

Model	**Train accuracy (%)**	**Test accuracy (%)**
Logistic Regression (LR)	91.3	90.5
Support Vector Classifier (SVC)	91.7	91.0
SGD Classifier (SGD)	92.9	**92.0**
Multinomial Naive Bayes (MNB)	88.1	89.5
Complement Naive Bayes (CNB)	88.4	88.5
Decision Tree Classifier (DT)	99.9	89.0
k-Nearest Neighbours Classifier (kNN)	92.9	91.5
Random Forest Classifier (RF)	99.9	91.5
Gradient Boosting Classifier (GB)	98.6	**92.0**
Multilayer Perceptron Neural Network (MPNN)	90.0	91.0

Section 2: Question 2 (Classification)

Classification Report for Gradient Boosting Classifier

	precision	recall	f1-score	support
Commercial	0.79	0.73	0.76	15
Community	0.99	0.97	0.98	68
F&B	0.96	0.98	0.97	52
Government	0.00	0.00	0.00	3
Healthcare	0.86	1.00	0.92	37
Others	0.00	0.00	0.00	0
Public Transport	1.00	1.00	1.00	4
Retail Shop	0.50	0.50	0.50	4
School	1.00	1.00	1.00	2
Shopping Mall	0.50	0.50	0.50	2
Welfare Organisation	0.90	0.69	0.78	13
accuracy			0.92	200
macro avg	0.68	0.67	0.67	200
weighted avg	0.91	0.92	0.91	200

Section 2: Question 2 (Classification)

Further potential improvements:

- Dataset:
  - Consider resampling to cope with class imbalance
  - Use `address` to reverse geocode location information

- Features engineering:
  - Improve keyword vocabulary selection
  - Use word embeddings instead of One Hot Encoding for `location_name`

- Modeling:
  - Try out more complicated neural networks with deeper layers
  - Tune hyper-parameters

Section 2: Question 2 (Classification)

Multiclass Classification Performance Metrics

Metric	Pros	Cons
$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$	• Easy to interpret	• Ignore class balance
$$\text{Precision}_{\textit{Class A}} = \frac{TP_{\textit{Class A}}}{TP_{\textit{Class A}} + FP_{\textit{Class A}}}$$ $$\text{Recall}_{\textit{Class A}} = \frac{TP_{\textit{Class A}}}{TP_{\textit{Class A}} + FN_{\textit{Class A}}}$$	• Class-specific • Consider class imbalance	• Non-aggregated

$$\text{Precision}_{\textit{Macro-average}} = \frac{\text{Precision}_{\textit{Class A}} + \text{Precision}_{\textit{Class B}} + \dots \text{Precision}_{\textit{Class N}}}{N}$$

$$\text{Recall}_{\textit{Macro-average}} = \frac{\text{Recall}_{\textit{Class A}} + \text{Recall}_{\textit{Class B}} + \dots \text{Recall}_{\textit{Class N}}}{N}$$

$$\text{Precision}_{\textit{Micro-average}} = \frac{TP_A + TP_B + \dots TP_N}{TP_A + FP_A + TP_B + FP_B + \dots TP_N + FP_N}$$

$$\text{Recall}_{\textit{Micro-average}} = \frac{TP_A + TP_B + \dots TP_N}{TP_A + FN_A + TP_B + FN_B + \dots TP_N + FN_N}$$

Section 2: Question 3

Section 2: Question 3 (Data Visualization)

Problem statement:

Present data in a summary table and its main insight through visualizations.

Dataset(s) / source(s): from the question

Summary table:

Job Nature	Industry	Student Group X		Student Group Y	
		Median Salary	Count	Median Salary	Count
Closely related to course of study	A	3150	83	3000	23
	B	3300	53	3100	9
	C	2650	47	2600	32
	D	2400	12	2400	15
Somewhat related to course of study	E	4100	30	3900	3
	F	3400	23	3150	7
	G	2800	12	2600	22
	H	2300	8	2200	11
Unrelated to course of study	Others	2900	21	1900	28

Section 2: Question 3 (Data Visualization)