Exploratory Data Analysis for Neighbourhoods in Auckland

A Capstone Project - The Battle of Neighbourhoods

1 Introduction

In this project, `K-Means` clustering methodology is used to cluster neighbourhoods in Auckland. `Foursqaure` api is used to get relevant venues in the neighbourhoods for better exploratory data analysis.

1.1 Background

Auckland(Tāmaki Makaurau) is the biggest city in New Zealand, with the population growing, the crimes happening frequency is increasing too. This project is to help people who migrate to Auckland to know more about the neighbourhoods in Auckland.

1.2 Current Problems

In Auckland, neighbourhood and neighbourhood are very different, some neighbourhood is quiet, some are busy. With the population growing, the crime frequency and safety of neighbourhood are changing too. For example, Albany was the place for trash land filling, but now more and more houses are built, at the same time land price increased too. New Zealand Police published the crime dataset on its website, but the problem there is no easy and direct way to know about the neighbourhoods' safety. This project is to help to solve this problem to let the residents and new migrants in Auckland to know about their neighbourhoods.

2 Data and Dataset

New Zealand goverment provides crime data which is public, to build this project, there are 2 data sets required.

- `NZ Crimes Data`
- `neighbourhood Coordinate Data`

The crime dataset can be easily gotten from NZ Police Website, but they do not offer a neighbourhood and coordinate dataset. To get the coordinate data, we can use `geopy`'s Nominatim function to call Open Street Map geocode api. There might be a problem to use this method that there is a limit for the api usage. In this project, we generated a JSON file to build a dictionary of neighbourhood and coordinates.

2.1 How to Solve the Problem

1. Clustering neighbourhoods in Auckland
2. Calling `foursquare` api to explore the neighbourhood

For example, the sample below is a venue information getting through `foursqaure explore` api.


```
{'reasons': {'count': 0,
   'items': [{'summary': 'This spot is popular',
     'type': 'general',
     'reasonName': 'globalInteractionReason'}]},
  'venue': {'id': '4cafb344562d224ba94b1088',
   'name': '茶顔觀色 Bubble Tea Cafe',
   'location': {'address': '198-200 SH 1',
    'crossStreet': 'Albany',
    'lat': -36.72617803610858,
    'lng': 174.6972947295745,
    'labeledLatLngs': [{'label': 'display',
      'lat': -36.72617803610858,
      'lng': 174.6972947295745}],
    'distance': 111,
    'cc': 'NZ',
    'city': 'North Shore',
    'state': 'Auckland',
    'country': 'New Zealand',
    'formattedAddress': ['198-200 SH 1 (Albany)',
     'North Shore',
     'New Zealand']},
   'categories': [{'id': '52e81612bcbc57f1066b7a0c',
     'name': 'Bubble Tea Shop',
     'pluralName': 'Bubble Tea Shops',
     'shortName': 'Bubble Tea',
     'icon': {'prefix':
'https://ss3.4sqi.net/img/categories_v2/food/bubble_',
      'suffix': '.png'},
     'primary': True}],
   'photos': {'count': 0, 'groups': []}},
  'referralId': 'e-0-4cafb344562d224ba94b1088-1'},
```

We can simply get category, address and location for the venue.

2.2 Neighbourhood Data Preparation

The first data can be easily got through `police.govt.nz`. There is more than 400 neighbourhood in Auckland. It is easy to get time out error when getting location through `geopy` geocoder. To fix this problem, the `locations_2020.sjon` file is used. This json file contains all location data required for this project.

To generate this file, each API calling of the geocoder is waited by 1 or 2 seconds, and every about 60 to 90 api calls are run separatelly and manuelly.

3 Methodology

3.1 Clustering

Before explore each neighbourhood in Auckland, cluster the neighbourhoods will help to get a basic understanding of Auckland. There are many methodologies in Sklearn, such as K-Means, Mean-shift and Rich, implementing all these methodologies on our neighbourhoods is not required here. In this project `K-Means` is used for the clustering.

3.1.1 Data Preparation for Clustering

Pandas and Numpy are 2 popular data science libraries for Python users. To get the clean and correct data before clustering, in sum, these step below are implemented.

Data Cleaning

1. Filter out all non-Auckland records. The dataset here for the New Zealand Crimes, it includes all crime records for each neighbourhoods in whole country in 2019, so or better performance of the implementation all non-Auckland records are dropped.
2. Drop unused and duplicated columns. Many columns are not used for the clustering, for example "Meshblock". The more detail about the dropped columns can be found in Jupiter Notebook attached.
3. Drop dirty rows in the data frame. some neighbourhood names is not correct, for example, one row in the data set, the neighbourhood name of it is `-29`, which does not exist as a neighbourhood or suburb name in Auckland.

[5]:

	ANZSOC Division	ANZSOC Group	ANZSOC Subdivision	Area Unit	Locn Type Division	Meshblock	Occurrence Day Of Week	Occurrence Hour Of Day	Weapon	YearMonth
24	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Albany.	Other Location	178900	Saturday	15	Not Applicable	2019-01-01
25	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Auckland Central West.	Other Location	433501	Saturday	15	Not Applicable	2019-01-01
26	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Auckland Central West.	Other Location	433600	Saturday	15	Not Applicable	2019-01-01
27	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Auckland Central West.	Other Location	434000	Saturday	15	Not Applicable	2019-01-01
28	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Mt Wellington South.	Other Location	637400	Saturday	15	Not Applicable	2019-01-01

Fig. 1: Data Sample After Cleaning

There are 74156 rows in the dataset, after cleaning there are 69507 rows.

Data Preparation

As you can see in the Fig.1, Crime data set itself does not contain coordinate data for the neighbourhood. which is required to the further clustering. In this step, `geopy` is used to get the coordinate data for each neighbourhood.

	ANZSOC Division	ANZSOC Group	ANZSOC Subdivision	Area Unit	Locn Type Division	Meshblock	Occurrence Day Of Week	Occurrence Hour Of Day	Weapon	YearMonth	latitude	longitude
24	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Albany.	Other Location	178900	Saturday	15	Not Applicable	2019-01-01	-36.727058	174.698055
25	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Auckland Central West.	Other Location	433501	Saturday	15	Not Applicable	2019-01-01	-36.848911	174.765226
26	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Auckland Central West.	Other Location	433600	Saturday	15	Not Applicable	2019-01-01	-36.848911	174.765226
27	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Auckland Central West.	Other Location	434000	Saturday	15	Not Applicable	2019-01-01	-36.848911	174.765226
28	Theft and Related Offences	Theft From Retail Premises	Theft (Except Motor Vehicles)	Mt Wellington South.	Other Location	637400	Saturday	15	Not Applicable	2019-01-01	-36.904793	174.835118

Fig. 2: Data Sample after inserting Coordinate Data

There is one tricky step here is that it is not allowed to batch to call the Open Street Map api to geocode the location in a short period. So the API is called manually and a cached coordinate JSON file is generated in this step.

3.1.2 Clustering Implementation

K-Means methodology is used here to do the clustering. K's value is set to 10, basically it can help us to cluster Auckland to 10 categories.

Clustering

```
from sklearn.cluster import KMeans
# set number of clusters
kclusters = 10

df_to_cluster =
df_borough.drop(['Neighborhood','CrimeCount'], 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters,
random_state=0).fit(df_to_cluster)
```

Group by "name" to Get the Crime Count

	Cluster Labels	Neighborhood	latitude	longitude	CrimeCount
0	9	Akarana.	-36.904895	174.745517	226
1	3	Albany.	-36.727058	174.698055	829
2	6	Algies Bay.	-36.432677	174.736948	10
3	0	Ambury.	-36.949533	174.770149	51
4	0	Aorere.	-36.980988	174.832658	111

Fig. 3: Data Sample after Clustering

4 Results and Data Visualisation

4.1 Results

After clustering, all data will be clustered to 10 categories.

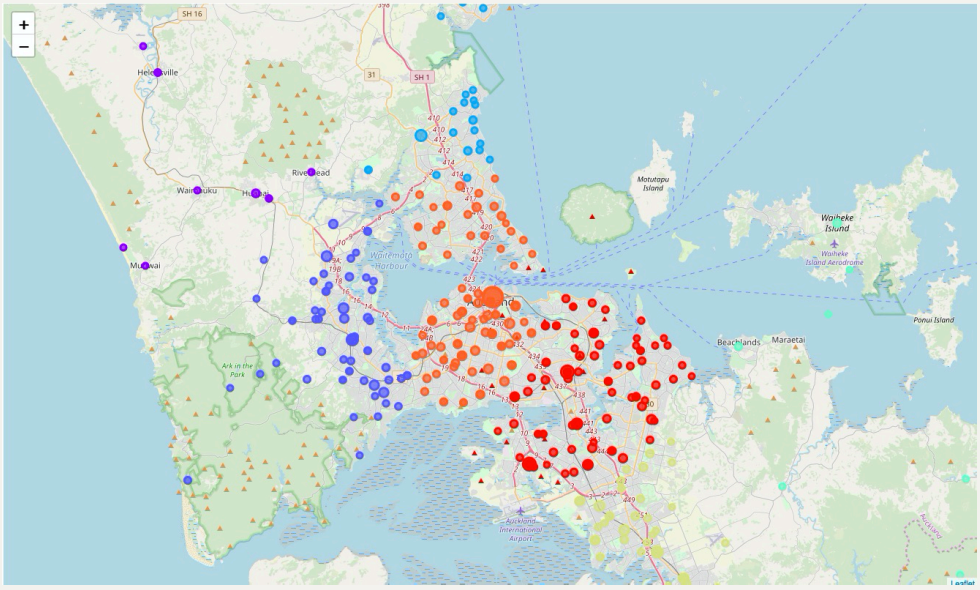

Fig. 4: Neighbourhood Crime Markers on Map

To visualise the clustering neighbourhoods more clearly, different colour are used for different category, and the markers' radius is set with normalised crime count.

```
radius= 4+10*scaler, # change the marker size here
```

The base radius is 4, the scaler is a number between 0 and 1. As you can see in the figure below, Crime Count in Albany is 829 and the scaler is 0.360784.

	CrimeCountScaler	Cluster Labels	Neighborhood	latitude	longitude	CrimeCount
0	0.098039	9	Akarana.	-36.904895	174.745517	226
1	0.360784	3	Albany.	-36.727058	174.698055	829
2	0.003922	6	Algies Bay.	-36.432677	174.736948	10
3	0.021786	0	Ambury.	-36.949533	174.770149	51
4	0.047930	0	Aorere.	-36.980988	174.832658	111

Fig. 5: Data Frame with Crime Count Scaler

Other Clustering Results during Implementation

Before clustering by location to 10 categories, other implementation is tried too, for example, clustering by Crime Count in each neighbourhood.

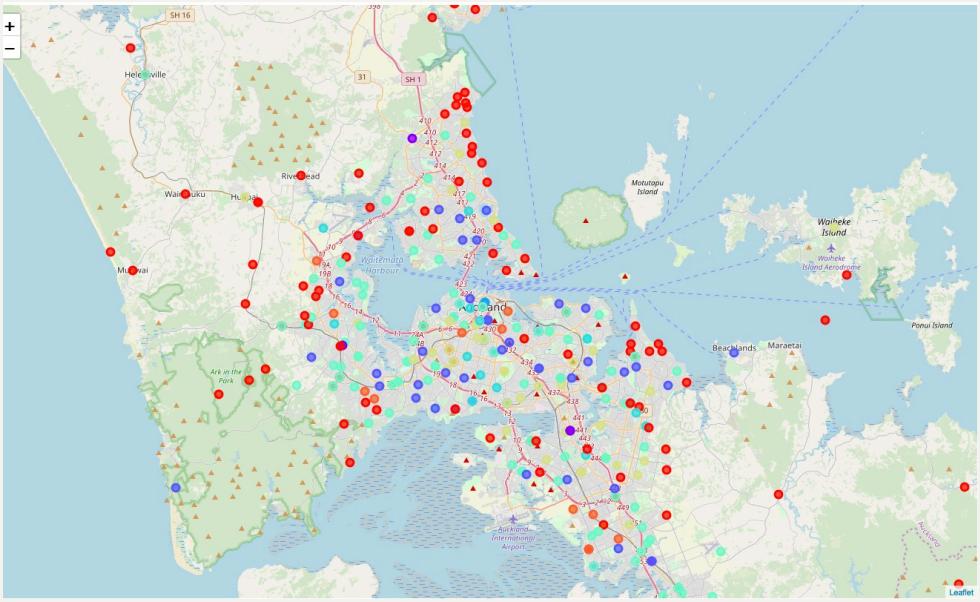

Fig .6: Clustering Result by Crime Count

The figure above is a basic result to cluster the neighbourhoods in Auckland by Crime Count.

4.2 Exploratory Data Analysis

Foursqaure API is used here to get the related venues information for the neighbourhoods.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Akarana.	Grocery Store	Asian Restaurant	Kids Store	Tea Room	Bakery	Farmers Market	Food Court	Food & Drink Shop	Food	Flower Shop
1	Albany.	Stadium	Café	Bar	Restaurant	Bubble Tea Shop	Playground	Tennis Court	Thai Restaurant	Convenience Store	Pizza Place
2	Algies Bay.	Nature Preserve	Yoga Studio	Farmers Market	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Fishing Store	Fish Market
3	Ambury.	Sandwich Place	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Flea Market	Fishing Store	Fish Market	Fish & Chips Shop
4	Arch Hill.	Café	Pizza Place	Bakery	Fish & Chips Shop	Brewery	Yoga Studio	Fast Food Restaurant	Food Court	Food & Drink Shop	Food

Fig. 7: Top Venue Categories for Neighbourhoods

	CrimeCountScaler	Cluster Labels	Neighborhood	latitude	longitude	CrimeCount	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0.098039	9	Akarana.	-36.904895	174.745517	226	Grocery Store	Asian Restaurant	Kids Store	Tea Room	Bakery	Farmers Market	Food Court	Food & Drink Shop	Food	Flower Shop
1	0.360784	3	Albany.	-36.727058	174.698055	829	Stadium	Café	Bar	Restaurant	Bubble Tea Shop	Playground	Tennis Court	Thai Restaurant	Convenience Store	Pizza Place
2	0.003922	6	Algies Bay.	-36.432677	174.736948	10	Nature Preserve	Yoga Studio	Farmers Market	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Fishing Store	Fish Market
3	0.021786	0	Ambury.	-36.949533	174.770149	51	Sandwich Place	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Flea Market	Fishing Store	Fish Market	Fish & Chips Shop
4	0.047930	0	Aorere.	-36.980988	174.832658	111	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 8: Join Clustering Data Fram and Venue Category Data Frame

In this data frame, each neighbourhood can be easily explored, for example, to explore `Auckland Central West` .

	CrimeCountScaler	Cluster Labels	Neighborhood	latitude	longitude	CrimeCount	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
10	1.0	9	Auckland Central West.	-36.848911	174.765226	2296	Café	Burger Joint	Vegetarian / Vegan Restaurant	Restaurant	Indian Restaurant	Japanese Restaurant	Sushi Restaurant	Coffee Shop	Pizza Place	Lounge

Fig. 8: Top Venue Categories for Neighbourhood Auckland Central West

5 Discussion & Conclusion

5.1 Discussion

As we can see in the Fig 4. `Auckland Central West` is the area which has the highest crime rate among all neighbourhoods in Auckland, and the most common Venue category is `Cafe`.

5.2 Conclusion

For the people who want to choose a place to live, if security is the most important concern, North West Auckland is a good option, South and Central Auckland are not good options.

6 References

- Jupyter Notebook Repo
- Police Stat for the crime records data
- Foursquare API for the venues data
- Coursera IBM Data Science Professional Certificate

6.1 Machine Learning Data Science Python Packages

- geopy for the corrdinate data
- sklearn

- numpy
- pandas