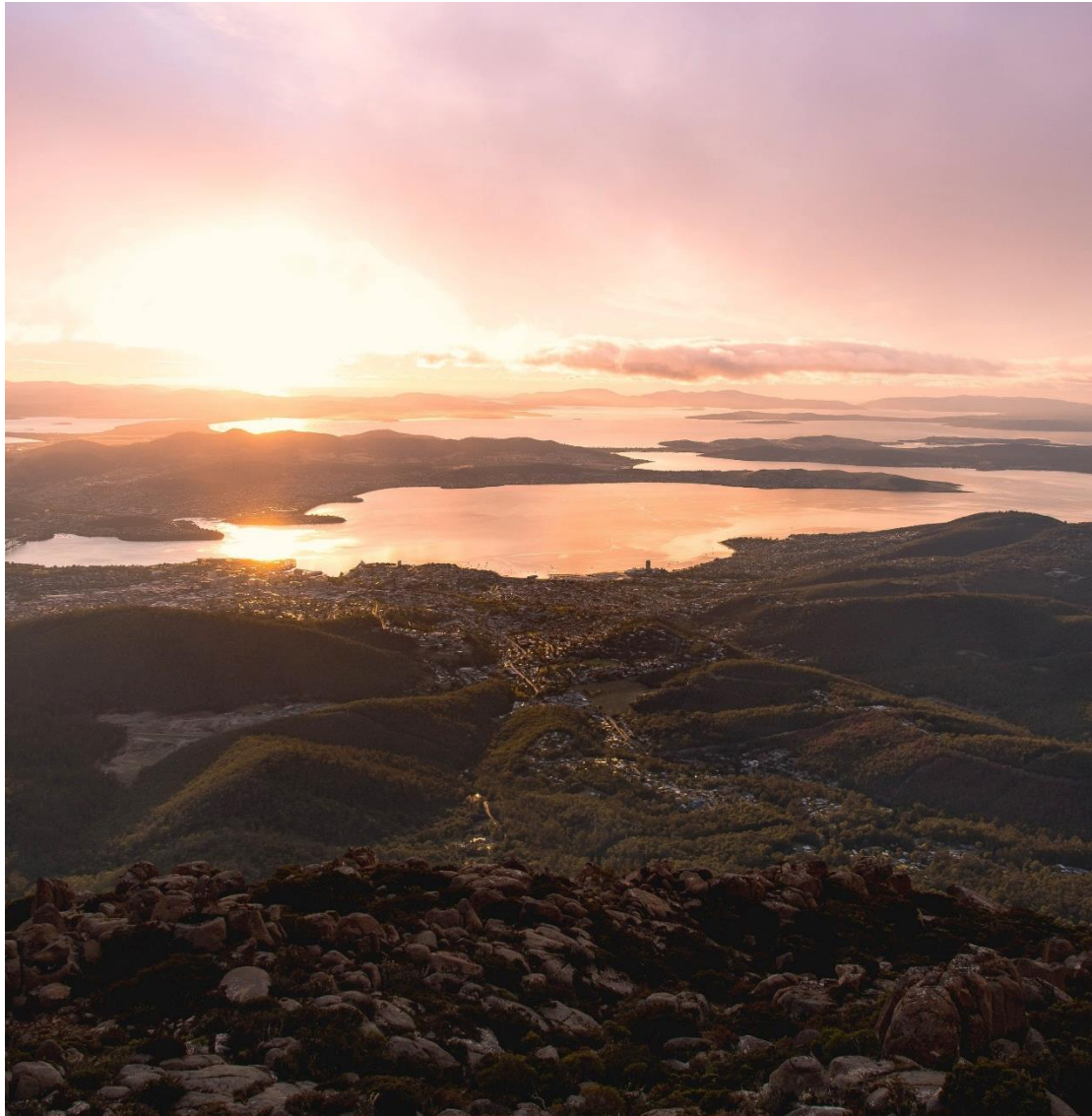


# **An Exploratory Data Analysis: Where to Start a Restaurant Business in Hobart**



# Table of Contents

1. Introduction .....	1
1.1 Description of the problem .....	1
1.2 Discussion of the background.....	1
2. Data requirements.....	2
3. Methodology.....	3
3.1 Data Preparation .....	3
3.2 Exploratory Data Analysis.....	5
4. Results and Discussion .....	10
5. Conclusion .....	12

# **1.Introduction**

## **1.1 Description of the problem**

The business problem we are currently posing is: Hobart is a small Australian city with daring art, a dynamic food scene, and a wealth of natural attractions. The opening of which type of restaurants in which neighborhood of Hobart can be the most profitable choice for investors?

## **1.2 Discussion of the background**

Hobart is the capital city of the Australian island state of Tasmania, and it is Australia's second oldest capital city after Sydney, New South Wales. Currently, Hobart is the financial and administrative hub of Tasmania, it is considered to be an ideal tourist destination, in 2016, 1.8 million visitors came to Hobart to explore untouched wilderness including World Heritage-Listed Tarkine cool temperate rainforest and its unique geological features. In addition, Hobart is a city with history and culture, boosted by convict-era architecture, Salamanca Market, and the Museum of Old and New Art (MONA), the Southern Hemisphere's largest private museum as well as food and wine, music festivals, markets, and art galleries.


Hobart has a good reputation for tourism, and it attracts domestic and international travelers every year, and these potential consumers may bring considerable revenue for tourism practitioners, and the restaurant is one of

an important venue for visitors to taste scrumptious food after a grueling day of travel, hence choosing an optimal location for restaurants seems to be extremely crucial for the stakeholders. Furthermore, the Tasmania government encourages business investors to make a significant contribution to Tasmania's economy and create job opportunities for the local people. Since running a restaurant is much more cost control and it doesn't require technological and scientific backgrounds in comparison with other key sectors such as Aquaculture, Agriculture and Agribusiness, Information and Communication Technology, Renewal Energy as well as Bio Renewal/Waste Recycle, etc. Therefore, many new immigrants plan to start a restaurant business in Hobart.

Based on the above analysis, running a restaurant in Hobart is profitable and is consistent with the interests of the local government. In this project, I would like to propose a strategy using Foursquare location data and clustering analysis to divide regions into different groups by their restaurant venues information.

## 2.Data requirements

For this project, the necessary data are listed below:

 **Hobart data contains the list of Boroughs and neighborhoods as well their latitudes and longitudes.**

Data source:

<https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3218.02017-18> and <https://latitude.to/articles-by-country/au/australia/>

Description: Boroughs and neighborhoods data can be downloaded in the format of Excel, since Geopy client can't accurately transform Boroughs and neighborhoods in Hobart into correct GPS coordinates, which includes Cambridge, Claremont (Tas.), Mount Wellington, etc. In this post, GPS coordinates will be confirmed manually.

#### **Restaurants in each neighborhood of Hobart**

Data source: Foursquare API

Description: All venues in each community will be fetched by Foursquare API, and the venues with categories that only contain restaurants will be retained.

## **3.Methodology**

### **3.1 Data Preparation**

The information of boroughs and neighborhoods in Hobart can be obtained on the website of Australian Bureau of Statistics, and these data can be directly downloaded in Excel format, which was followed by loading data in Jupyter Notebook with the assistance of pandas through transforming Excel into DataFrame that contains 35 boroughs and neighborhoods, the population as well as population density. After the data

manipulation, we will observe the DataFrame as below:

```
import pandas as pd
import numpy as np

Tas=pd.read_excel('Population_Estimates_by_Statistical_Area_Level_2.xls',sheet_name="Table_6",skiprows=7)
Tas.dropna(subset=['S/T name'],inplace=True)
Tas.rename(columns={'no..1':'population','km2':'Area_(km^2)','persons/km2':'Population_density_persons/km2'},inplace=True)
Tas=Tas.loc[1:,['S/T name','GCCSA name','SA4 name','SA3 name','SA2 name','population','Area_(km^2)','Population_density_persons/km2']]
Hobart=Tas[Tas['SA4 name']=='Hobart']
Hobart
```

	S/T name	GCCSA name	SA4 name	SA3 name	SA2 name	population	Area (km^2)	Population density persons/km2
1	Tasmania	Greater Hobart	Hobart	Brighton	Bridgewater - Gagebrook	7543.0	55.7	135.3
2	Tasmania	Greater Hobart	Hobart	Brighton	Brighton - Pontville	6213.0	88.8	70.0
3	Tasmania	Greater Hobart	Hobart	Brighton	Old Beach - Otago	4975.0	31.6	157.6

Geopy client is a powerful library to get GPS coordinates of specific locations, however, Cambridge in Hobart was incorrectly encoded as the county town of Cambridgeshire, England. Besides, Mount Wellington can be considered as a mountain in Canada or New York or Victoria, Australia, or in Auckland, New Zealand, and I can't get accurate coding results via the current Geopy client. In addition to these limitations, some neighborhoods can't even be encoded such as Bridgewater-Gagebrook. To solve this issue, I got GPS coordinates of boroughs and neighborhoods in Hobart through the Google Search, and these datasets were combined with Hobart Boroughs and neighborhoods table as follow:

```
Location=pd.read_csv('Hobart_Location.csv')
Hobart=pd.merge(Hobart,Location,how='left',on='SA2 name')
Hobart
```

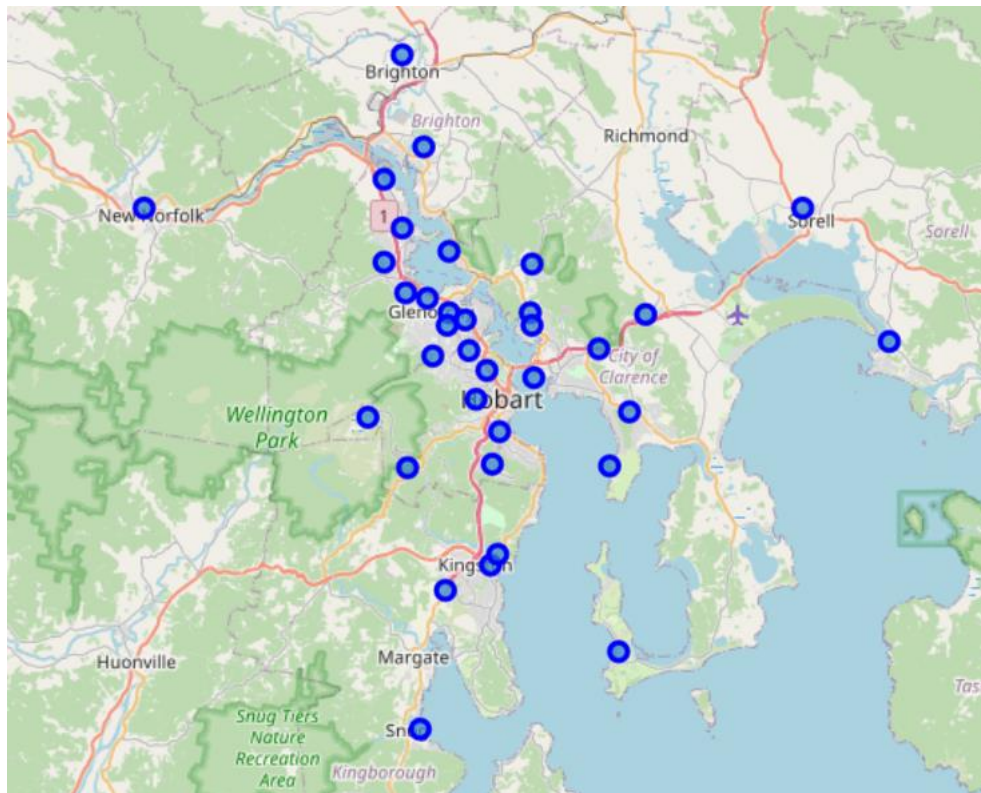
	S/T name	GCCSA name	SA4 name	SA3 name	SA2 name	population	Area (km^2)	Population density persons/km2	Latitude	Longitude
0	Tasmania	Greater Hobart	Hobart	Brighton	Bridgewater - Gagebrook	7543.0	55.7	135.3	-42.742330	147.268999
1	Tasmania	Greater Hobart	Hobart	Brighton	Brighton - Pontville	6213.0	88.8	70.0	-42.690997	147.252832
2	Tasmania	Greater Hobart	Hobart	Brighton	Old Beach - Otago	4975.0	31.6	157.6	-42.800497	147.287999
3	Tasmania	Greater Hobart	Hobart	Hobart - North East	Bellerive - Rosny	6092.0	4.7	1291.2	-42.870663	147.352832
4	Tasmania	Greater Hobart	Hobart	Hobart - North East	Cambridge	8367.0	101.7	82.3	-42.835330	147.437998
5	Tasmania	Greater Hobart	Hobart	Hobart - North East	Geilston Bay - Risdon	3314.0	8.7	380.3	-42.835163	147.350665
6	Tasmania	Greater Hobart	Hobart	Hobart - North East	Howrah - Tranmere	11295.0	9.4	1200.1	-42.920330	147.409832

Subsequently, the folium library was leveraged to visualize geographic details of Hobart and its 35 neighborhoods, and I created a map of Hobart with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:

```
# create map of Tokyo using latitude and longitude values
map_Hobart = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, label in zip(Hobart['Latitude'], Hobart['Longitude'], Hobart['SA2_name']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_Hobart)

map_Hobart
```



### 3.2 Exploratory Data Analysis

To begin with, I would like to adopt exploratory data analysis (EDA) to reveal hidden patterns of neighborhood data sources and then provide



informative and valuable insights to the readers and future investors.

The next step is to get the top 100 venues in Bridgewater-Gagebrook which is the first neighborhood in DataFrame within a radius of 1000 meters by employing the Foursquare API, the result has been shown in the following Figure. We found that only 2 unique venue categories including supermarket and brewery were returned by Foursquare, the reason why this neighborhood had a small number of venues is probably Hobart is not a big city compared to other global metropolises such as Sydney and Melbourne in Australia, and Hobart is relatively less congestive and population density is smaller than international cities, thus the need for market demand is limited.

```
venues = results['response']['groups'][0]['items']
nearby_venues = json_normalize(venues).#.flatten().JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
nearby_venues = nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1)

# clean columns
nearby_venues.columns = [col.split(".")[1] for col in nearby_venues.columns]

nearby_venues.head()
```

	name	categories	lat	lng
0	IGA X-Press Gagebrook	Supermarket	-42.74869	147.266000
1	MONA Beer & Wine Tastings	Brewery	-42.73640	147.278059

```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
2 venues were returned by Foursquare.

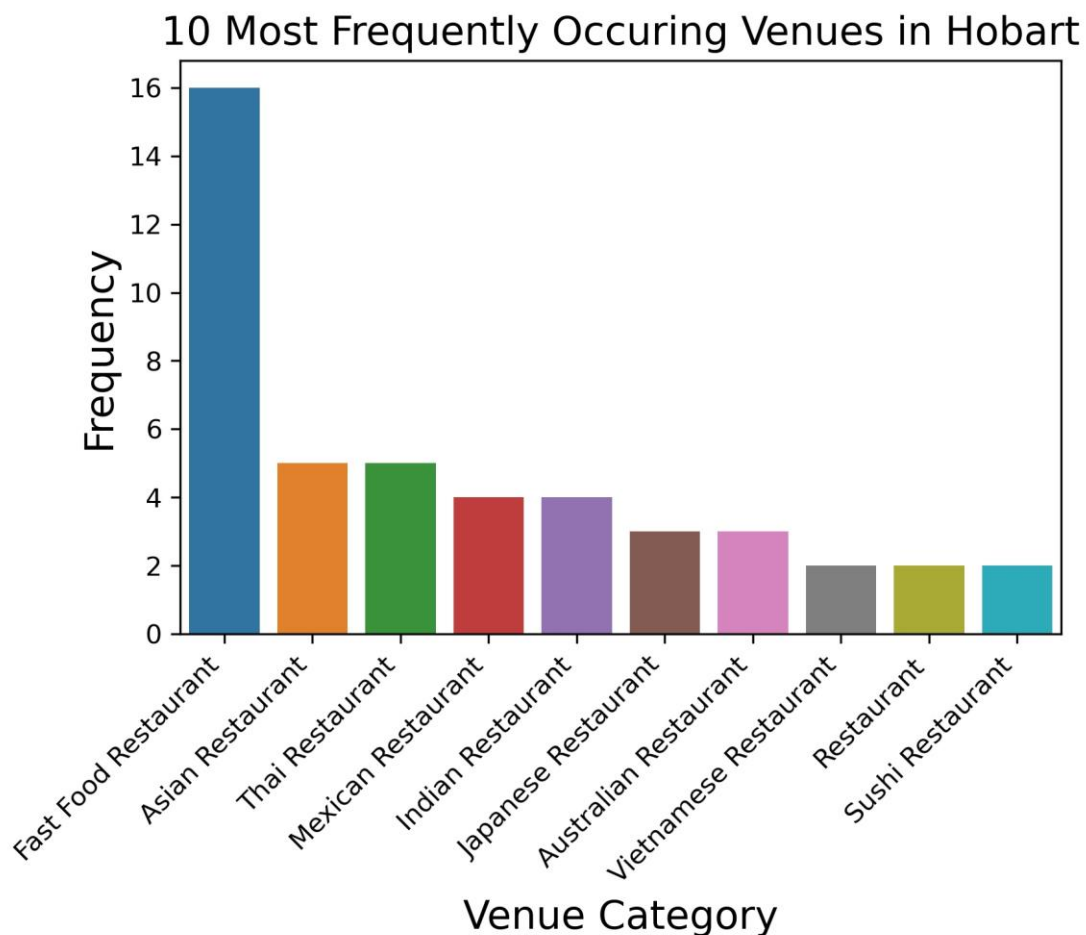
print('{} unique categories in Bridgewater - Gagebrook'.format(nearby_venues['categories'].value_counts().shape[0]))
2 unique categories in Bridgewater - Gagebrook

print(nearby_venues['categories'].value_counts()[0:10])
Supermarket    1
Brewery        1
Name: categories, dtype: int64
```

Next, I will filter the categories that only contains restaurant in all 35 neighborhoods in Hobart. We found that 16 unique restaurant categories and Fast Food Restaurant seems to be the most popular genre accounting



for the majority of the entire restaurants, an elegant data visualization was presented in the Figure below. To make a more in-depth analysis of this phenomenon, I read an interesting post written by Hannah McDonough, she believes that fast food is affordable, inexpensive, convenient, and delicious, most importantly, it is fast, once you place your order, you usually don't have to wait very long until your food comes out unless it is very busy. In addition, Asian restaurants such as Thai, Indian, Japanese, and Vietnamese restaurants prevail in the local community. To sum up, running a fast food or Asian restaurant in Hobart might be a wise strategic decision for investors.



Let's follow the instructions below:

Firstly, creating a DataFrame with pandas one-hot encoding for the venue categories.

```
# one hot encoding
Hobart_onehot = pd.get_dummies(Hobart_Venues_only_restaurant[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
Hobart_onehot['Neighborhood'] = Hobart_Venues_only_restaurant['Neighborhood']..

# move neighborhood column to the first column
fixed_columns = [Hobart_onehot.columns[-1]] + list(Hobart_onehot.columns[:-1])
Hobart_onehot = Hobart_onehot[fixed_columns]

Hobart_onehot.head()
```

	Neighborhood	Asian Restaurant	Australian Restaurant	Chinese Restaurant	Fast Food Restaurant	French Restaurant	Indian Restaurant	Indonesian Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Mexican Restaurant
1	Cambridge	0	0	0	1	0	0	0	0	0	0	0
2	Geilston Bay - Risdon	0	0	0	1	0	0	0	0	0	0	0
3	Lindisfarne - Rose Bay	0	0	0	1	0	0	0	0	0	0	0
4	Claremont (Tas.)	0	0	0	1	0	0	0	0	0	0	0
5	Derwent Park - Lutana	0	0	0	1	0	0	0	0	0	0	0

Secondly, using the built-in “groupby” function in the pandas library on the neighborhood column and then calculate the mean of the frequency of occurrence for each venue category.

```
Hobart_grouped = Hobart_onehot.groupby('Neighborhood').mean().reset_index()
Hobart_grouped
```

	Neighborhood	Asian Restaurant	Australian Restaurant	Chinese Restaurant	Fast Food Restaurant	French Restaurant	Indian Restaurant	Indonesian Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Mexican Restaurant	Portuguese Restaurant
0	Cambridge	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Claremont (Tas.)	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Derwent Park - Lutana	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Geilston Bay - Risdon	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Glenorchy	0.000000	0.000000	0.000000	0.666667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Hobart	0.071429	0.071429	0.000000	0.000000	0.000000	0.142857	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429
6	Kingston Beach - Blackmans Bay	0.000000	0.500000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Lindisfarne - Rose Bay	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Moonah	0.166667	0.000000	0.000000	0.166667	0.000000	0.000000	0.000000	0.166667	0.000000	0.333333	0.000000	0.000000
9	New Norfolk	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

After the above procedures, I will obtain each neighborhood along with the top 5 most common venues.

```

num_top_venues = 5

for hood in Hobart_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = Hobart_grouped[Hobart_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue', 'freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')

----Cambridge----
           venue  freq
0  Fast Food Restaurant  1.0
1    Asian Restaurant  0.0
2 Australian Restaurant  0.0
3   Chinese Restaurant  0.0
4    French Restaurant  0.0

----Claremont (Tas.)----
           venue  freq
0  Fast Food Restaurant  1.0
1    Asian Restaurant  0.0
2 Australian Restaurant  0.0
3   Chinese Restaurant  0.0
4    French Restaurant  0.0

```

I will attempt to use a commonly used machine model, namely K-Means clustering help an investor to decide an optimal location and Venue Category to start a restaurant. Finally, I would like to cluster these 15 neighborhoods based on the venue categories using K-Means clustering. So, our expectation would be based on the similarities of venue categories, these neighborhoods will be clustered. The scripts are shown below:

```

# set number of clusters
kclusters = 5

Hobart_grouped_clustering = Hobart_grouped.drop('Neighborhood', 1)

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

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]...

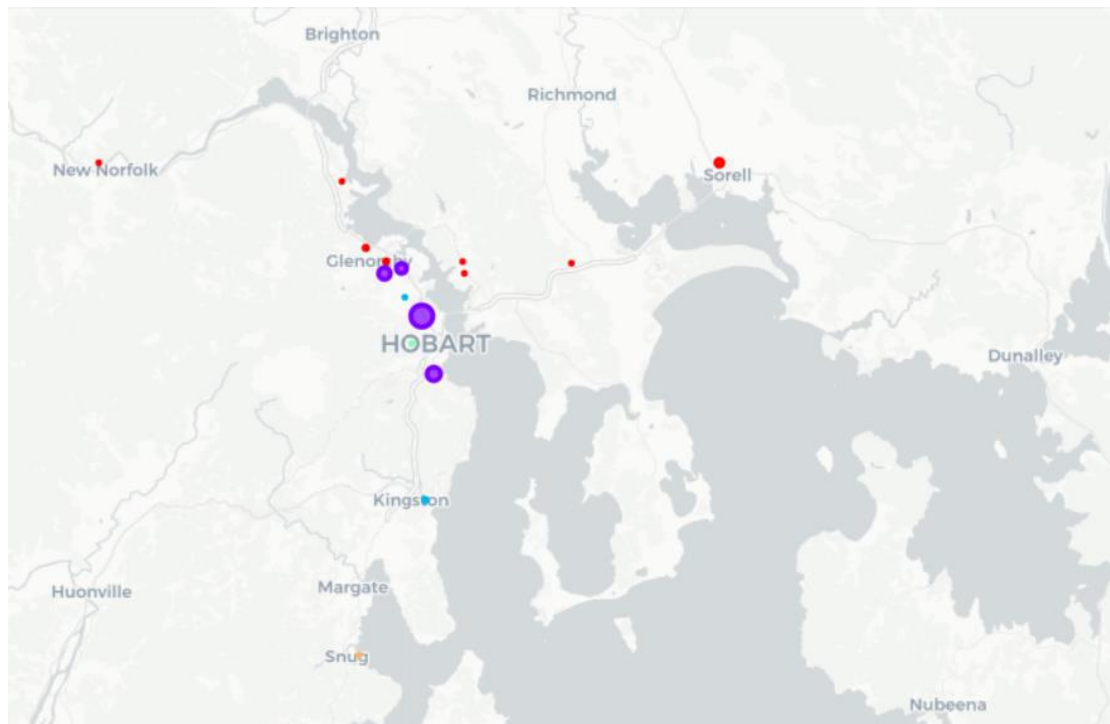
array([0, 0, 0, 0, 0, 3, 2, 0, 3, 0], dtype=int32)

```

```
final_data=pd.merge(result,neighborhoods_venues_sorted,how='left',on='Neighborhood')
final_data
```

	Neighborhood	Label	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	Cambridge	0	Fast Food Restaurant	Vietnamese Restaurant	Thai Restaurant	Sushi Restaurant	Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant	Italian Restaurant
1	Claremont (Tas.)	0	Fast Food Restaurant	Vietnamese Restaurant	Thai Restaurant	Sushi Restaurant	Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant	Italian Restaurant
2	Derwent Park - Lutana	0	Fast Food Restaurant	Vietnamese Restaurant	Thai Restaurant	Sushi Restaurant	Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant	Italian Restaurant
3	Geilston Bay - Risdon	0	Fast Food Restaurant	Vietnamese Restaurant	Thai Restaurant	Sushi Restaurant	Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant	Italian Restaurant
4	Glenorchy	0	Fast Food Restaurant	Sushi Restaurant	Vietnamese Restaurant	Thai Restaurant	Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant	Italian Restaurant
5	Hobart	3	Thai Restaurant	Indian Restaurant	Vietnamese Restaurant	Sushi Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant	Italian Restaurant	Indonesian Restaurant
6	Kingston Beach - Blackmans Bay	2	Indian Restaurant	Australian Restaurant	Vietnamese Restaurant	Thai Restaurant	Sushi Restaurant	Restaurant	Portuguese Restaurant	Mexican Restaurant	Kebab Restaurant	Japanese Restaurant

We can represent these 5 clusters in a leaflet map using the Folium library as below:



## 4.Results and Discussion

We got a glimpse of the Restaurants in Hobart and were able to find out some useful insights that are beneficial to those investors who plan to start a restaurant in Hobart. Here are our findings:

- ✚ Fast Food Restaurants top the charts of most common venues in 35 neighborhoods.
- ✚ Hobart neighborhood (14) has the maximum number of restaurants, which was followed by Sandy Bay (8), West Moonah (7), and Moonah (6). Since the clustering was based only on the category of restaurants in each neighborhood, and we surprisingly found out these neighborhoods are all classified as cluster 2, indicating each of these neighborhoods presents a similar profile to investors in terms of food category.
- ✚ Cambridge, Claremont (Tas.), Geilston Bay-Risdon, Lindisfarne-Rose Bay, Margate-Snug, New Norfolk, New Town, and West Hobart just only have the least number of the restaurant (only 1 for all).

In this post, the k-means clustering algorithm is completely based on the most common venues obtained from Foursquare data. However, in our current data analysis, we didn't take other factors like yearly average income for each neighborhood, the competitors in that location, and whether the markets are untapped or saturated, etc. into our consideration, since it would be difficult to conduct this survey by individuals and the objective of this project is to leverage different tools to analyze geographical data. Therefore, our analysis could only help investors to

have an overview of restaurants distribution by categories in the 35 neighborhoods of Hobart, a more sufficient investigation for the local market is still needed.

## **5. Conclusion**

In the data-driven world, many real-life problems/scenarios can be explained or solved with the assistance of data science by analyzing data using appropriate tools. In this post, the data analysis is mainly focused on clustering neighborhoods in Hobart based on the most common food venues (Restaurants) in its 15 neighborhoods using diverse python libraries such as Pandas, Numpy, Matplotlib, Folium, sklearn, and Foursquare API, etc. This analysis is performed on very limited data, it would come up with better and robust results when a good amount of data is available. In general, the findings could help investors to decide about an optimal location to start a restaurant business to some extent, and a more comprehensive and integrated analysis is still needed in the future.