OPENING COFFE SHOP IN JAKARTA

COURSERA CAPSTONE FINAL PROJECT

Budi Irawan Saleh

Abstract

- The coffee shop has become an emerging business, especially in a growing business city. Jakarta is one of the growing capital business city in Indonesia, and opening a coffee shop will be a very profitable business. However, the business of coffee shops has been trending, and a new coffee shop has opened in this city. It will require a good analysis of where to locate the coffee shop not just to avoid a highly competitive market but also to reach a higher maximum profit. This research will use the data analysis techniques, which have been thought in the Coursera capstone course. The data of the neighborhood was collected using a web scraping technique to Wikipedia. Then the geo-location of the neighborhood was obtained by using Google API (google maps) and mapped to the Folium map. The venues in the neighborhood collected using Foursquare API and clustered (using K-mean) and analyzed based on the numbers of coffee shops in each cluster. K-mean cluster using the elbow method was found for 2 clusters. Based on the data analysis, the coffee shop can be open in the second cluster because the data analysis showed that numbers of coffee shops in the area are still small and can reduce the competition. However, it also found that the second cluster has the smallest number of area and have a potential smaller profit if the coffee shop is open in the first cluster

BACKGROUND

- Indonesia has become one of the attractive business markets in the world. It has been forecasted by the end of 2024 will be among the highest GDP country in the world
- SME's in Indonesia is dominated by a culinary business (i.e : restaurants, coffee shops, bar, etc.).
- Indonesia's most populated city is in Jakarta, where people in the archipelago are gathering in this city for business and government
- Coffee shops have become trends, especially in middle-income workers, to hang out after the office and to await a time to go home. It was also used as a place to do a business transaction, working and spending time with family
- Data analysis has been used significantly to assist business decision and planning for years. Research using these methods has increased significantly. More business has used the methods both supervised or unsupervised learning to help their business.

OBEJECTIVE

- to find the best locations for coffee shops in Jakarta and expecting less competition, among others

METHODOLOGY

1. Data on the locations of Jakarta, especially its borough area, need to be collected from the trusted place, Wikipedia is the place to gather the data. Web scraping technique is used to gather the borough data and placed it into a panda data frame
2. Geospatial location was then obtained using the Google API methods. And to check the location accuracy by plotting the data using the Folium map
3. The venue then collected using Foursquare API and placed into the panda data frame.
4. The K-mean then used to cluster the location based on geospatial location. The cluster number was obtained using the elbow methods.
5. Data then analyzed to obtain conclusions for the location of the coffee shops

METHODOLOGY

Importing the necessary libraries

In [1]:
```
# import necessary libraries
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup
import requests
import urllib.request
```

extracting table from wikipedia for borough in the City

In [2]:
```
data = requests.get("https://en.wikipedia.org/wiki/List_of_Jakarta_Subdistricts").text
```

In [3]:
```
# parse data from the html into a beautifulsoup object
soup = BeautifulSoup(data, 'html.parser')
```

In [4]:
```
# create a list to store neighborhood data
Borough = []
```

In [5]:
```
# append the data into the list
for row in soup.find_all("div", class_="mw-parser-output")[0].findAll("li"):
    Borough.append(row.text)
```

In [6]:
```
# create a new DataFrame from the list
df = pd.DataFrame({"Borough":Borough})

df
```

In [14]:
```
import googlemaps
```

In [15]:
```
gmaps_key = googlemaps.Client(key = "AIzaSyAxPJUAU4xQsSBF9UJtJrCDCSRNPG57W3k")
```

In [16]:
```
dataset["Latitude"] = None
dataset["Longitude"] = None
for i in range (0, len(dataset),1 ):
    geocode_result =gmaps_key.geocode(df.iat[i,0])
    try:
        lat = geocode_result[0]["geometry"]["location"]["lat"]
        long = geocode_result[0]["geometry"]["location"]["lng"]
        dataset.iat[i, dataset.columns.get_loc("Latitude")] = lat
        dataset.iat[i, dataset.columns.get_loc("Longitude")] = long
    except:
        lat = None
        long = None
```

In [43]:
```
#  center map on mean of Latitude/Longitude
map_world = folium.Map(location=[-6.22818 , 106.934385], zoom_start = 11)

#  add Locations to map
for lat, lng, label in zip(dataloc.Latitude, dataloc.Longitude, dataloc.Borough):
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        fill=True,
        color='Blue',
        fill_color='Yellow',
        fill_opacity=0.6
        ).add_to(map_world)

#  display interactive map
map_world
```

Use the Foursquare API to explore the neighborhoods

In [25]:
```
import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe
```

In [26]:
```
# define Foursquare Credentials and Version
CLIENT_ID = 'JB3N4QWKPKI14RWWLY3R0GWKF3X04NTN4KNB04R3J10G2V0X' # your Foursquare ID
CLIENT_SECRET = 'W5JL4J5T0JS2WEZOMWHK1KSKUMYURD44VCCXIXAQBHOTZJEB' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version

print('Your credentials:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentials:
CLIENT_ID: JB3N4QWKPKI14RWWLY3R0GWKF3X04NTN4KNB04R3J10G2V0X
CLIENT_SECRET:W5JL4J5T0JS2WEZOMWHK1KSKUMYURD44VCCXIXAQBHOTZJEB
```

In [27]:
```
radius = 2000
LIMIT = 100

venues = []

for lat, long, neighborhood in zip(dataloc['Latitude'], dataloc['Longitude'], dataloc['Borough']):

    # create the API request URL
    url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".form
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)

    # make the GET request
    results = requests.get(url).json()["response"]['groups'][0]['items']

    # return only relevant information for each nearby venue
    for venue in results:
        venues.append((
            neighborhood,
            lat,
            long,
            venue['venue']['name'],
            venue['venue']['location']['lat'],
            venue['venue']['location']['lng'],
```

DISCUSSION

- From the web scrapping technique from Wikipedia, the borough data is stored in the data frame as follow :

Out[12]:

	Borough
0	Angke
1	Cakung
2	Cempaka Putih
3	Cengkareng
4	Cilandak
5	Cipayung
6	Ciracas
7	Duren Sawit
9	Duri Kepa
10	East Cengkareng
11	Gambir, Jakarta
12	Grogol

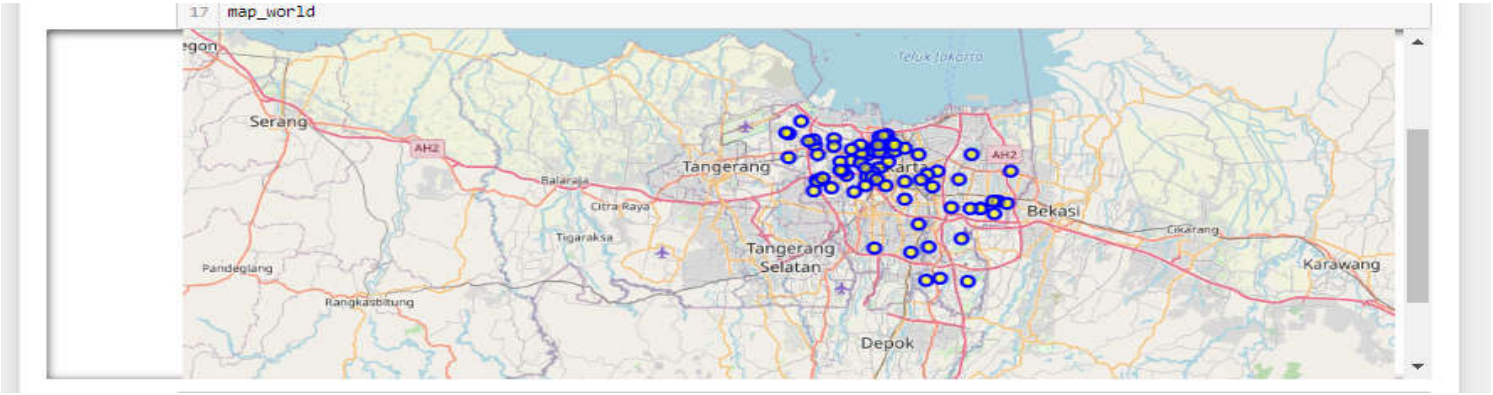
- The data of geospatial location then obtained using Google Map API as follow :

Out[21]:

	Borough	Latitude	Longitude
0	Angke	41.493662	-81.704854
1	Cakung	-6.182629	106.947666
2	Cempaka Putih	-6.182671	106.867990
3	Cengkareng	-6.148665	106.735258
4	Cilandak	-6.284528	106.800140
5	Cipayung	-6.327251	106.900447
6	Ciracas	-6.323116	106.870940
7	Duren Sawit	-6.232191	106.915202
8	Duri Kepa	-6.232191	106.915202
9	East Cengkareng	-6.169185	106.774860
10	Gambir, Jakarta	-6.142332	106.733784
11	Grogol	-6.170340	106.814805

DISCUSSION

- Plotted using folium map to check the locations, and it was found that no data error and all area is covered

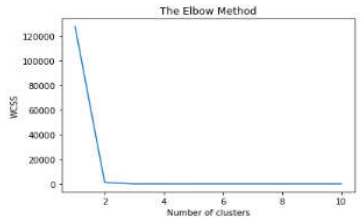


- The foursquare methods then used to obtain the venues in the location and put into the data frame, and it was found that a total of 272 uniques venues is at Jakarta. The K-mean clustering methods are used to cluster based on the location of the borough. The elbow method then used to obtain a sufficient clustering number. Elbow methods found that 2 clusters need to be used for clustering.

DISCUSSION

Using the elbow method to get the K mean clustering value

In [56]:
```
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

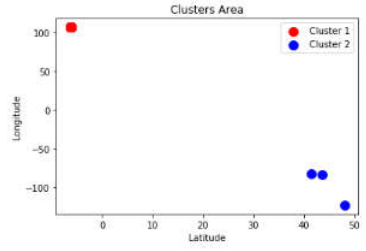


From the elbow methods it was found that the Kcluster value is - 2

- The data then plotted based on the cluster to check the logic of the cluster number. And have found no error in the clustering number.

Analyzing the cluster value

In [59]:
```
plt.scatter(X[y_kmeans1 == 0, 0], X[y_kmeans1 == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans1 == 1, 0], X[y_kmeans1 == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.title('Clusters Area')
plt.xlabel('Latitude')
plt.ylabel('Longitude')
plt.legend()
plt.show()
```



DISCUSSION

- From the data on the table, the best place to start the business for a coffee shop is the second cluster, where the number of coffee shops is less than the first cluster and can avoid the highly competitive market. However, because the number of the area in the second cluster is also less than the first cluster, it was expected that in the second cluster will have a potential of less profit.

CONCLUSION

- The data analysis technique was used in this research to find the best location for opening the coffee shop in the Jakarta area. It was found by using the K-mean clustering that a total of clusters area in Jakarta divided into two clusters. The clusters were analyzed to seek the best location based on the numbers of coffee shops venue, and it was found that the second cluster that has fewer coffee shops venues is the most appropriate location to avoid high competition. However, because of the number of locations of the second cluster far less then the first clusters, it can potentially impact to the profit of the business.
- This research limitation is being conducted in the Jakarta area using K-mean clustering for analyzing. Furter clustering methods can be used to analyze the data, and also regression can be used to predict the future growth of the business.
- Detail data and code are available in the following link:
-
- https://github.com/budisaleh/github-courseracapst/blob/master/Week%204/coursera_caps.ipynb