Exploring Data to Open a New Coffee Shop in Surabaya, Indonesia

Vejryn Shaviero
July, 2020

1. Introduction

1.1. Background

Surabaya is one of the biggest metropolitan city in Indonesia, for various reason, people in Surabaya like to spend their leisure time discussing hobbies and many other things with coffee as their main beverage to be with, hence the term 'Ngopi' is popular among the ears of the people.

When we think about it from investor perspective, we should take advantage of this trend to develop a Coffee Shop, but of course, we need to do it with serious consideration. The location of the Coffee Shop is one of the most important things to consider to determine the success rate of the Coffee Shop itself, but, **how can we determine the best possible location?**

When will do Data Analysis to help us find the best possible location for our Coffee Shop.

2. Data and Data Srouces

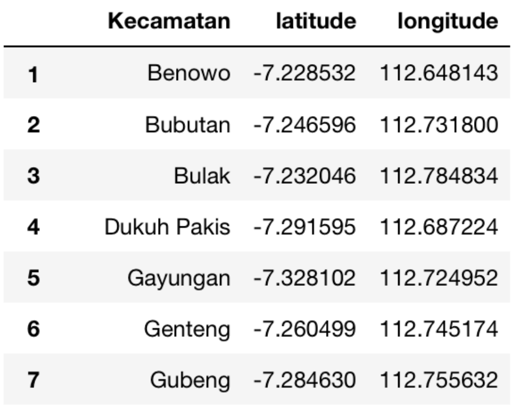
2.1. Data Sources

To solve the problem, we need the following data:

- List of Kecamatan in Surabaya, we need to correctly determine the actual data of Kecamatan in Surabaya to perfectly analyze our data. (Kecamatan sometimes would be described as neighborhood, in our notebooks). We retrieve the data from Wikipedia page [1] using built in pandas function.
- Geodata of Kecamatan in Surabaya, we need the geodata of Kecamatan in Surabaya to help us process the Foursquare API. We retrieve the data using geopy library in pandas.
- List of Coffee Shop in Surabaya. We retrieve the data using Foursquare API. [2]

3. Methodology

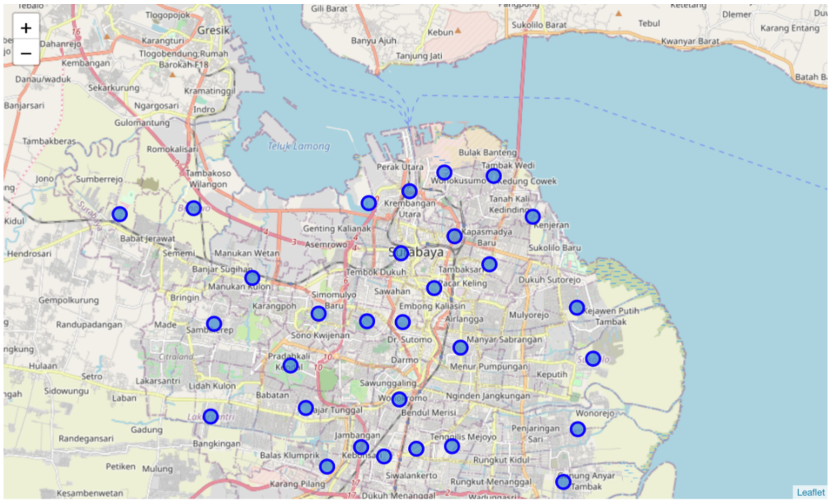
There are many methods to collect data from around the internet, fortunately, pandas has a built in function to collect html table data into dataframe. We use this method to collect List of Kecamatan in Surabaya from Wikipedia page [1]. However, mentioned data doesn't have a longitude and latitude(location) that we need to visualise into map and retrieve coffee shop venues around Surabaya. We need to retrieve the longitude and latitude separately from our data, to do so, we use Geopy library. Here is some of Kecamatan in Surabaya and it's location.



	Kecamatan	latitude	longitude
1	Benowo	-7.228532	112.648143
2	Bubutan	-7.246596	112.731800
3	Bulak	-7.232046	112.784834
4	Dukuh Pakis	-7.291595	112.687224
5	Gayungan	-7.328102	112.724952
6	Genteng	-7.260499	112.745174
7	Gubeng	-7.284630	112.755632

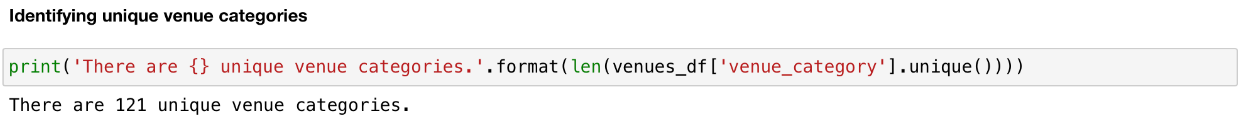
Table 1. List of Kecamatan in Surabaya, Indonesia

After gathering the list of Kecamatan in Surabaya and its location, we are able to visualise it using Folium library.



Picture 1. Kecamatan in Surabaya visualised using Folium

Next, we need to be able to locate Coffee Shop within 3000 meters radius around Kecamatan in Surabaya. We use Foursquare API to gather venues data, after we do Exploratory data Analysis, there are 121 unique venues categories in our data.



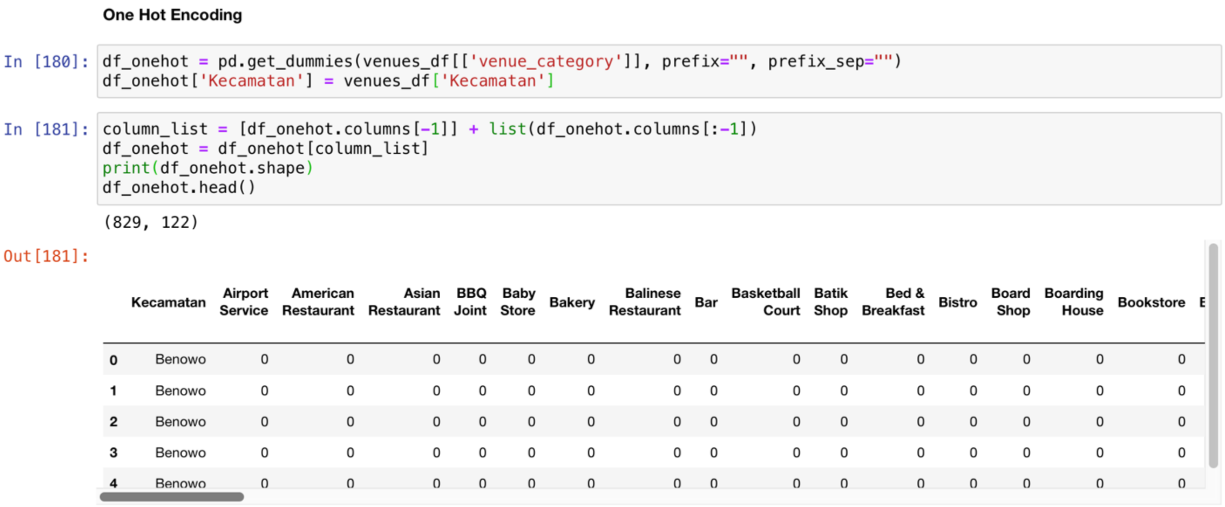
Identifying unique venue categories

```
print('There are {} unique venue categories.'.format(len(venues_df['venue_category'].unique())))
```

```
There are 121 unique venue categories.
```

Picture 2. Sum of unique venue categories

We then apply **One Hot Encoding** to process the data



One Hot Encoding

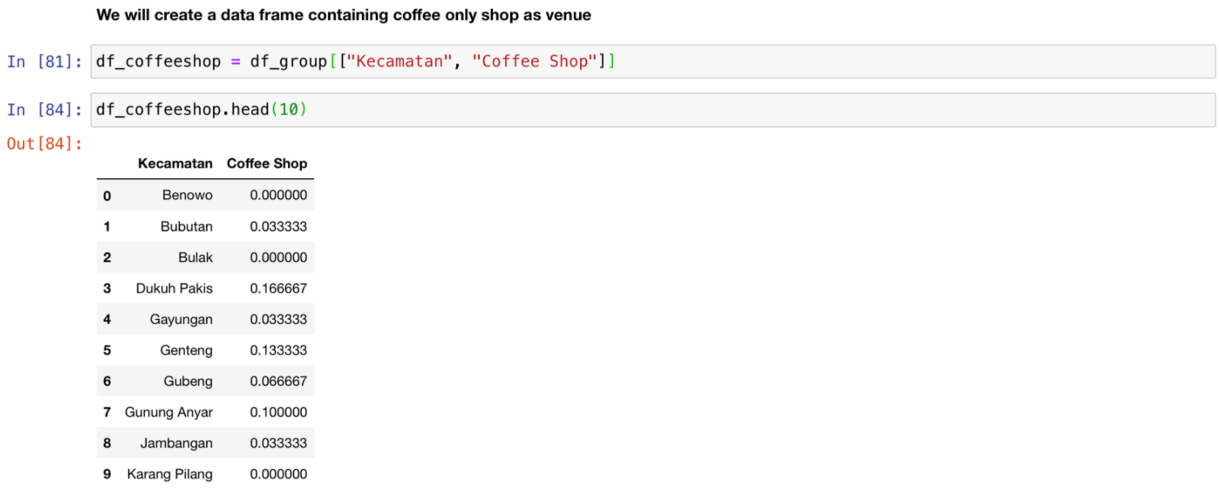
```
In [180]: df_onehot = pd.get_dummies(venues_df[['venue_category']], prefix="", prefix_sep="")
          df_onehot['Kecamatan'] = venues_df['Kecamatan']
```

```
In [181]: column_list = [df_onehot.columns[-1]] + list(df_onehot.columns[:-1])
          df_onehot = df_onehot[column_list]
          print(df_onehot.shape)
          df_onehot.head()

          (829, 122)

Out[181]:
```

	Kecamatan	Airport Service	American Restaurant	Asian Restaurant	BBQ Joint	Baby Store	Bakery	Balinese Restaurant	Bar	Basketball Court	Batik Shop	Bed & Breakfast	Bistro	Board Shop	Boarding House	Bookstore	E
0	Benowo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Benowo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Benowo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Benowo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Benowo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Picture 3. Applying One Hot Encoding to our Data

and group our Kecamatan data based on the presence of Coffee Shop.



We will create a data frame containing coffee only shop as venue

```
In [81]: df_coffeeshop = df_group[["Kecamatan", "Coffee Shop"]]
```

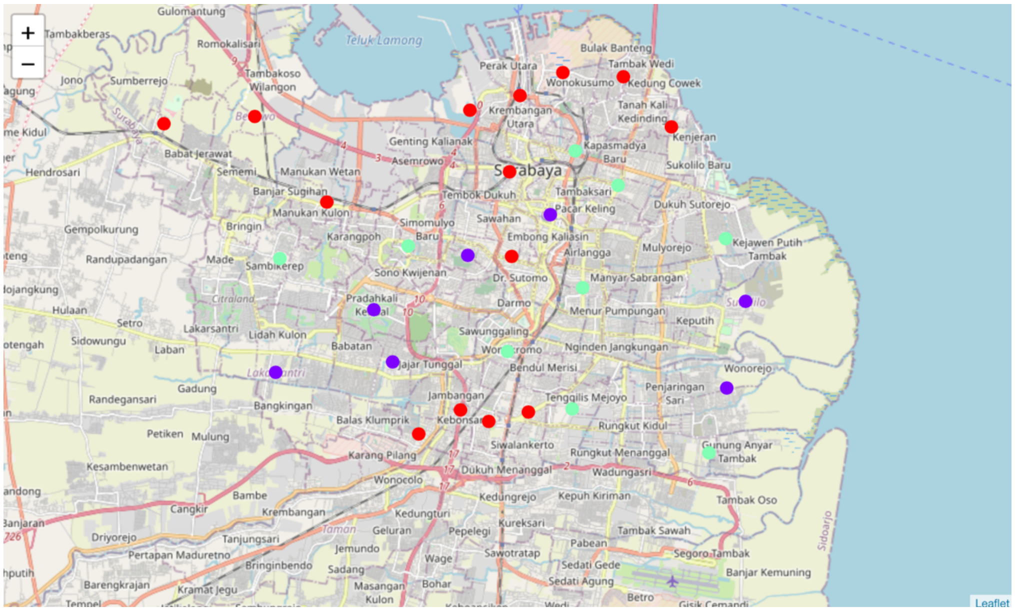
```
In [84]: df_coffeeshop.head(10)

Out[84]:
```

	Kecamatan	Coffee Shop
0	Benowo	0.000000
1	Bubutan	0.033333
2	Bulak	0.000000
3	Dukuh Pakis	0.166667
4	Gayungan	0.033333
5	Genteng	0.133333
6	Gubeng	0.066667
7	Gunung Anyar	0.100000
8	Jambangan	0.033333
9	Karang Pilang	0.000000

Picture 4. Group our data by Kecamatan and Coffee Shop

Using the dataset, we then apply unsupervised machine learning algorithm which is **K-means algorithm** to cluster our data. We will cluster the Kecamatan into **3 clusters** based on the presence of Coffee Shop, it allow us to identify which Kecamatan have higher concentration of Coffee Shop.

D. Results



Picture 5. Visualisation of Clustering Applied on Maps

We visualise the result of K-means clustering algorithm for each Kecamatan so that we can see three different cluster:

- Cluster 0 (Red marker) : Kecamatan with high number Coffee Shops.
- Cluster 1 (Red marker) : Kecamatan with low number Coffee Shops.
- Cluster 2 (Red marker) : Kecamatan with moderate number Coffee Shops.

E. Discussion and Suggestion for Future Research

While the clustering of Kecamatan based on the concentration or the presence of Coffee Shop may help us in identifying which location is good for building or developing a new Coffee Shop from it, we may use another variable in our future research such as ratings, price, and trends of venues from Foursquare API to better determine our success rate in building a new Coffee Shop.

F. Conclusion

The purpose of this project was to explore Coffee Shop concentration or presence in Kecamatan in Surabaya to answer our business question that was raised in our introduction section, and the answer is to use machine learning clustering algorithm. Our visualisation of the data reveals that Cluster 1 have the lowest concentration of Coffee Shop, the findings of this project will help investors in determining the location for increasing its success rate of building a Coffee Shop.

G. References

- [1] Daftar Kecamatan dan Kelurahan di Kota Surabaya
- [2] Foursquare API