Image Shown Is For Illustration Purpose Only

IBM Applied Data Science Capstone Project

The Battle of Neighborhoods

Authored by Varianto Angga

Introduction

In this project, I take a role as an analyst that helps hotel managers or owners in deciding which part of Surabaya is the best location to establish a new luxury (5 stars) hotel. Location is the most crucial aspect for hotel owners, because different locations may have different surroundings and venues that also attract different types of travelers. The success of a hotel depends largely on the location of its site. For this reason, hotel owners must carefully decide which parts of the city to choose according to the type and price of the hotel. In order to decide the most appropriate location, there are several factors that should be considered in advance.

For this project, the factor to be considered is the surrounding location offering, such as restaurants, bus lines, banks, and so forth. In addition, this project will focus only on the city of Surabaya, as the second largest city in Indonesia and the capital of the province of East Java.

Data

The data required to solve the problem of deciding the best location to establish a new hotel in Surabaya area are:

1. **Basic Neighborhoods Information**
    - Data Source: Wikipedia Page (https://id.wikipedia.org/wiki/Daftar_kecamatan_dan_kelurahan_di_Kota_Surabaya)
    - Data Description: The Wikipedia page contains the neighborhood basic information, such as unique code and name in the form of HTML which will be scrapped to get the data

2. **Coordinates Of Neighborhoods**
    - Data Source: The neighborhoods data are provided in a CSV file named "indonesian_coordinates.csv" (https://raw.githubusercontent.com/vari8/applied-data-science-capstone/main/indonesian_coordinates.csv)

- Data Description: The CSV file contains a dataset of each Indonesia neighborhood's coordinate information. So, to obtain data only for the city of Surabaya, a portion of the dataset will be selected

3. **Venues Information For Each Neighborhood**
    - Data Source: Foursquare
    - Description: By using the Foursquare API, a list of venues can be obtained along with information, such as categories and coordinates

Methodology

This project is for audiences interested in discovering which neighborhood in Surabaya is the best for setting up a new 5-star luxury hotel. In this project, I will analyze the density in each neighborhood and predict which has the most venues nearby to set up a new 5-star luxury hotel.

There are several processes involved in this project:

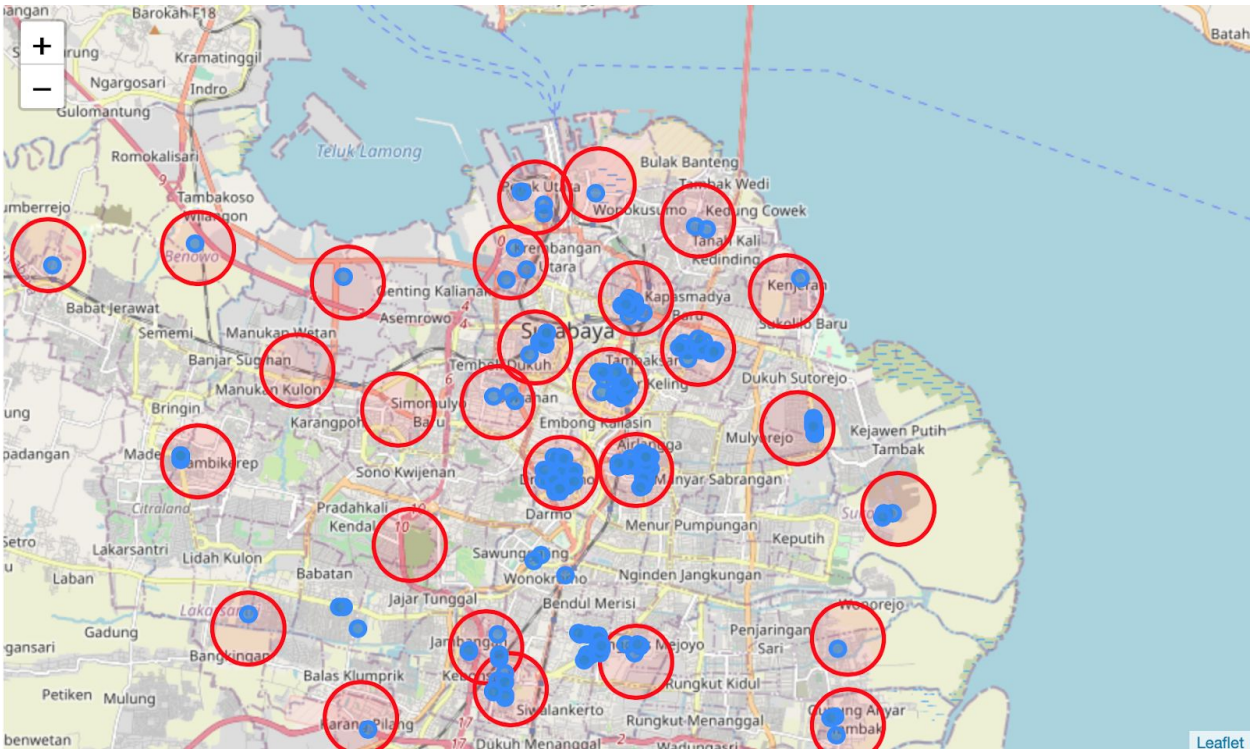
1. The data needed to answer the problem will be collected through various sources. The data collection process has been carried out in the data section above. After that, I will perform data wrangling to clean up the data so that it can be used for further analysis.

2. I will perform an Exploratory Data Analysis (EDA) on the cleaned data set to see the main characteristics of the data. The analysis I will do is as follows:
    - Summary statistics in the form of pivot tables to summarize and group neighborhoods and venues data
    - Bar graph to compare the number of venues for each neighborhood and boxplot to see how the values in the data are spread out
    - Map to locate venue on Surabaya map

3. A cluster model that uses k-Means algorithm will be used to cluster each different neighborhood based on the similarity of venues. k-Means is used because it guarantees convergence and generalizes to clusters of different shapes and sizes. I will choose 5 as the k value, which means that the neighborhood will be clustered to 5 different groups.

Results

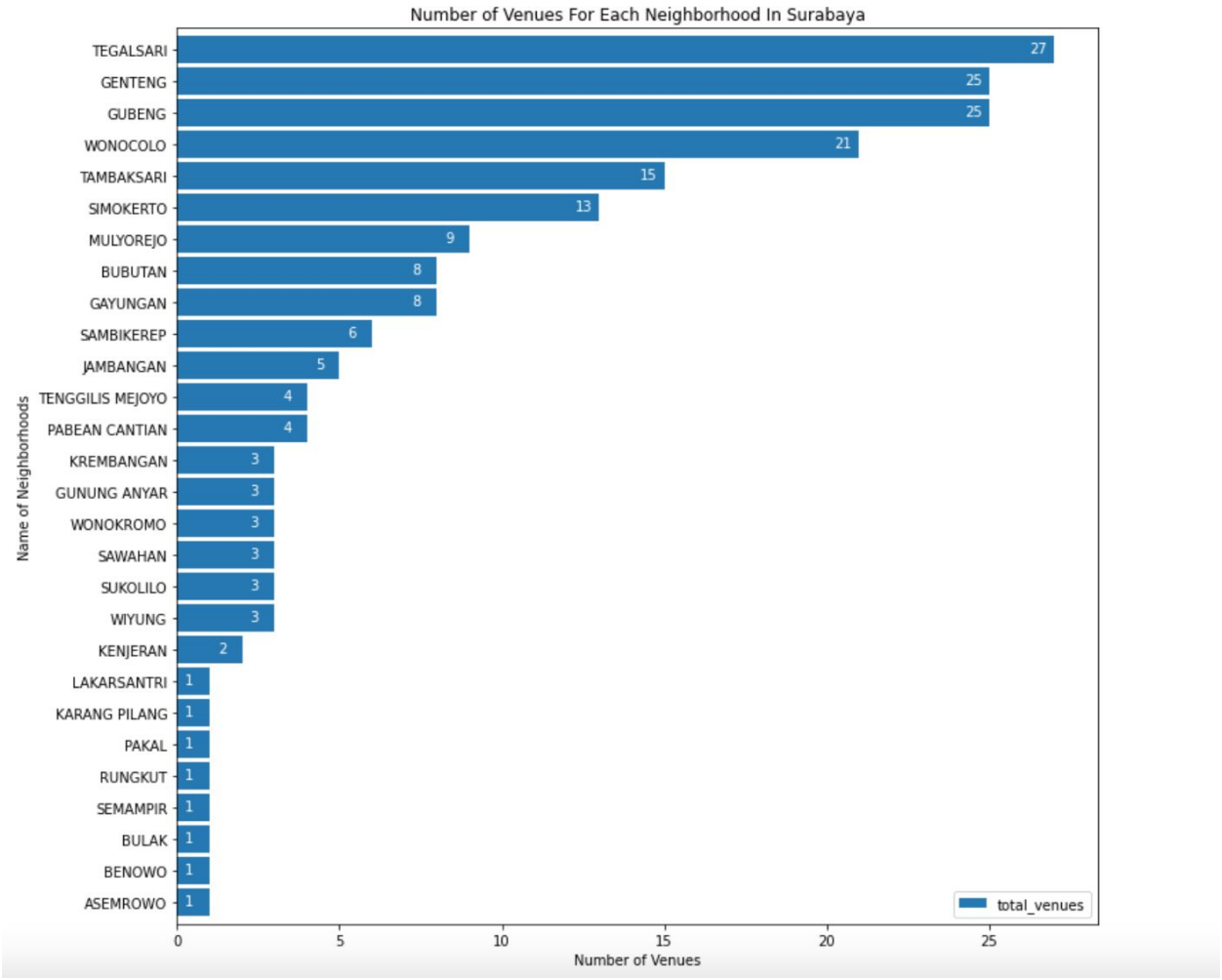
This project analyzes venues within a 500m radius of each neighborhood in the city of Surabaya. Venues data that is provided by Foursquare, and please note that no other location data from other parties is considered in this project. To see the distribution of venues and neighborhoods in the city of Surabaya, a Folium package is used to draw a map, center the map to Surabaya, and plot the neighborhoods and venues as markers on the map. Here is the image of the distribution of places. The small blue circles are the markers for venues, while the big red circles are the markers for neighborhoods.



A Surabaya City Map With Its Neighborhoods and Venues

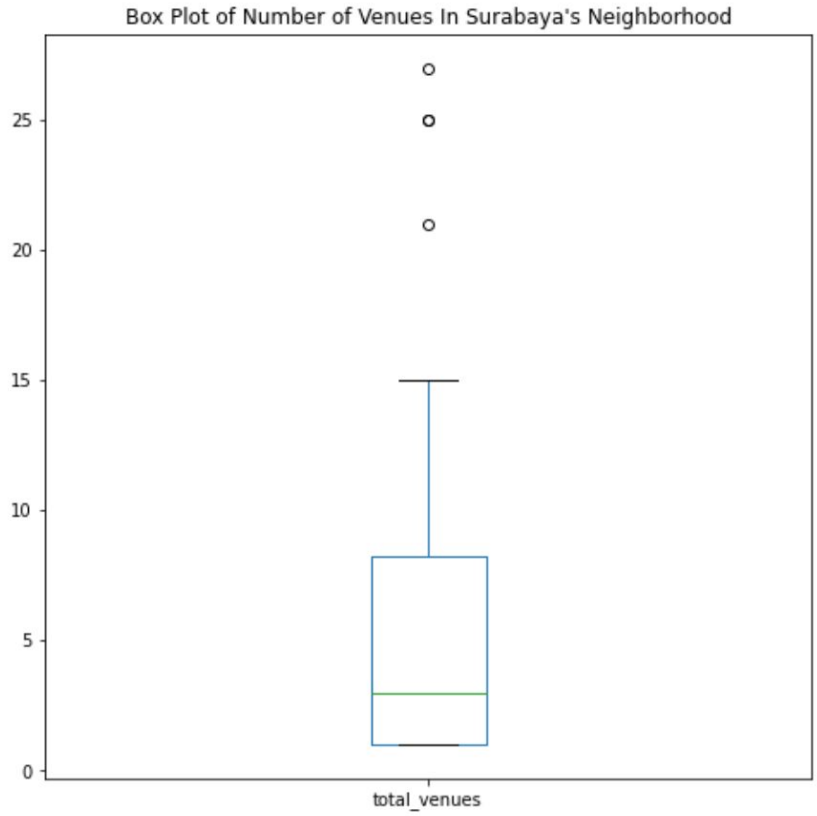
By constructing the bar graph to visualize the number of venues for each neighborhood, we can compare the total number of venues among neighborhoods in Surabaya. The bar graph tells us that there is an uneven distribution of venues among the neighborhoods in Surabaya. There are three neighborhoods that stand out the most in the number of venues, namely

TEGALSARI, GENTENG, and GUBENG. On the other hand, there are 8 neighborhoods in Surabaya that each has only one venue.



Bar Graph of Total Number of Venues For Each Neighborhood in Surabaya

Boxplot is constructed so that we can observe the statistical distribution of total venues across all neighborhoods in Surabaya. There are some neighborhoods that have a high number of venues and are classified as outliers.



Box Plot of Total Number of Venues In Surabaya's Neighborhood

We can observe a few key observations from the box plot:

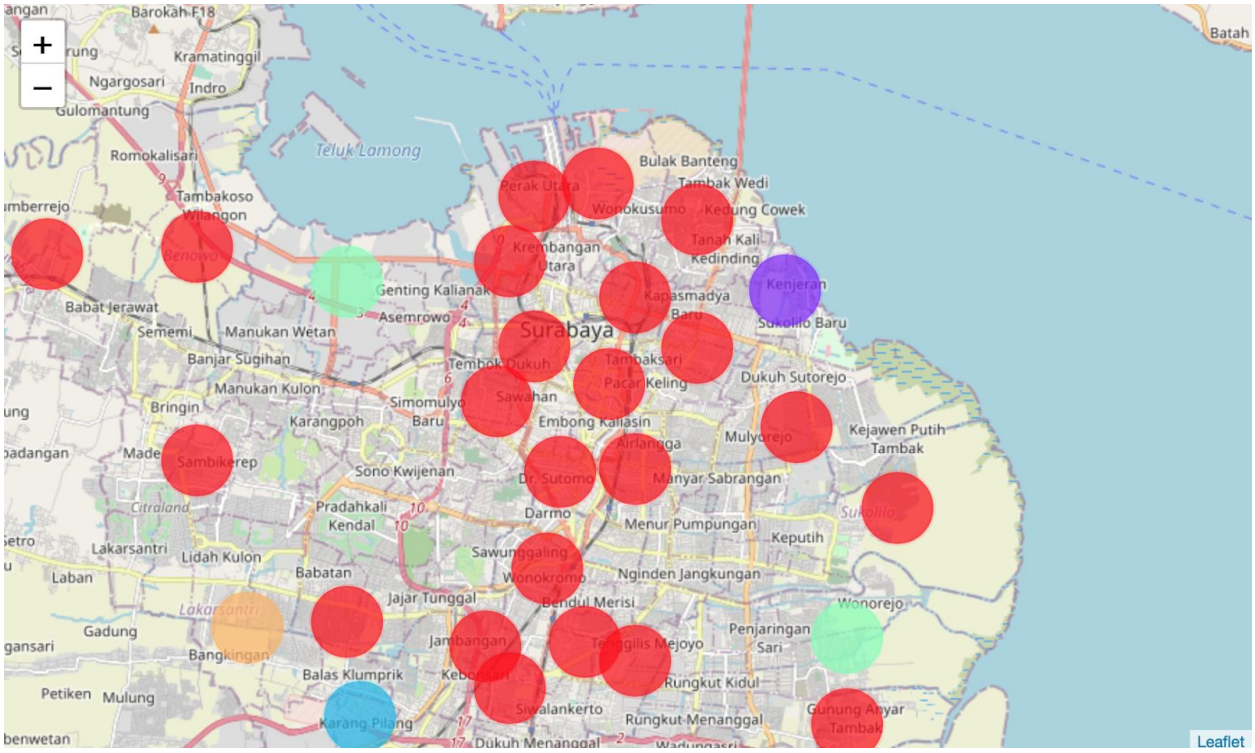
- The minimum number of total venues is 1 and the maximum number (Q3 + 1.5*IQR) is 15.
- 25% of neighborhoods have 1 or fewer venues (First quartile).
- 50% of neighborhoods have 3 venues (Median number).
- 75% of neighborhoods have around 8 or fewer venues (Third quartile).
- There are 3 outliers, and we can confirm the neighborhoods as TEGALSARI, GENTENG, and GUBENG by looking at the bar graph.
- The data are right skewed (positively skewed) since the median is closer to the first quartile than to the third quartile.

A statistical summary of the total venues is generated to help understand the axis values on the box plot.

	total_venues
count	28.000000
mean	7.071429
std	8.123713
min	1.000000
25%	1.000000
50%	3.000000
75%	8.250000
max	27.000000

Summary Statistic of Total Number of Venues In Surabaya's Neighborhood

The neighborhoods are clustered into 5 different groups. The clustering process performed by the kMeans model takes into account the similarity of venue categories, so that neighborhoods with the same venue are clustered to the same group. The neighborhoods that have been clustered are then plotted on the map to show the location of the cluster area above the city Surabaya city. Below is the image for a map that shows location of the clustered neighborhoods.



A Surabaya City Map With Its Clustered Neighborhoods

Here is the cluster results:

- **Cluster 1 = Red circles**
- **Cluster 2 = Purple circles**
- **Cluster 3 = Blue circles**
- **Cluster 4 = Green circles**
- **Cluster 5 = Orange circles**

Cluster 1 consists of 23 neighborhoods. On the other hand, cluster 4 has two neighborhoods, and other clusters have only one neighborhood. Cluster 1 was chosen because it contains the neighborhoods with the most number of venues compared to any other clusters. After that, I narrowed my focus to cluster 1 to select a few neighborhoods as potential candidates to be selected as a new luxury hotel location.

Cluster 1

	neighborhood_name	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	total_venues
23	TEGALSARI	0	Indonesian Restaurant	Bakery	Indonesian Meatball Place	Karaoke Bar	Coffee Shop	Health & Beauty Service	Bookstore	Gluten-free Restaurant	27
5	GENTENG	0	Coffee Shop	Indonesian Restaurant	Fried Chicken Joint	Soup Place	Furniture / Home Store	Bed & Breakfast	Japanese Restaurant	Multiplex	25
6	GUBENG	0	Indonesian Restaurant	Food Truck	Bakery	Chinese Restaurant	Multiplex	Kids Store	Convenience Store	Dumpling Restaurant	25
26	WONOCOLO	0	Bakery	Hotel	Pool Hall	Restaurant	Convenience Store	Coffee Shop	Café	Bubble Tea Shop	21
22	TAMBAKSARI	0	Convenience Store	Indonesian Restaurant	Asian Restaurant	Noodle House	Food Truck	Food	Bakery	Balinese Restaurant	15
20	SIMOKERTO	0	Seafood Restaurant	Food Truck	Cosmetics Shop	Arcade	Convenience Store	Soup Place	Bakery	Basketball Court	13
13	MULYOREJO	0	Indonesian Restaurant	Convenience Store	Bakery	Coffee Shop	Photography Lab	Fast Food Restaurant	Mobile Phone Shop	Supermarket	9
4	GAYUNGAN	0	Boutique	Convenience Store	Steakhouse	Café	Bakery	Food Truck	Indonesian Restaurant	Seafood Restaurant	8
2	BUBUTAN	0	Convenience Store	Bookstore	Donut Shop	Market	Fast Food Restaurant	Food Truck	Shopping Mall	Food Court	8
17	SAMBIKEREP	0	Bistro	Miscellaneous Shop	Housing Development	Market	Indonesian Restaurant	Flea Market			6

Data Frame of Neighborhoods in Cluster 1

Cluster 1 is considered to be the best cluster for establishing a new luxury 5-star hotel, as this cluster is the most congested in terms of venues. Then, I would recommend hotel owners or hotel managers to choose TEGALSARI, GENTENG, or GUBENG neighborhoods as the best neighborhoods for establishing a new luxury hotel. In the 1st Cluster, TEGALSARI neighborhood has the most number of venues in cluster 1, followed by GENTENG and GUBENG, and others. All the top 3 neighborhoods have food & beverages category as their first to fourth most common venues, and other entertainment categories as their fifth to eight most common venues.

TEGALSARI, GENTENG, and GUBENG have many venues for tourists to explore. Tourists have many options to try the local food from the nearby restaurants, and visit other places, such as karaoke bars and convenience stores. Various categories of venues, both culinary and entertainment venues will attract the attention of middle to high income tourists.

Cluster 2

	neighborhood_name	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	total_venues
3	BULAK	1	Gift Shop								1

Data Frame of Neighborhoods in Cluster 2

Cluster 2 contains only one neighborhood and that neighborhood has only one venue. This cluster is not the best location for setting up a new luxury hotel, as the small number of venues is unlikely to attract high budget travelers.

Cluster 3

	neighborhood_name	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	total_venues
9	KARANG PILANG	2	Toll Booth								1

Data Frame of Neighborhoods in Cluster 3

Cluster 3 contains only one neighborhood. This cluster is not the best location for setting up a new luxury hotel, as the small number of venues is unlikely to attract high budget travelers.

Cluster 4

	neighborhood_name	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	total_venues
0	ASEMROWO	3	Food Truck								1
16	RUNGKUT	3	Food Truck								1

Data Frame of Neighborhoods in Cluster 4

Cluster 4 contains two neighborhoods, each of which has only one venue, which is a food truck. This cluster is not the best location for setting up a new luxury hotel, as high-budget travelers may not be interested only in food trucks.

Cluster 5

	neighborhood_name	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	total_venues
12	LAKARSANTRI	4	Pool								1

Data Frame of Neighborhoods in Cluster 5

Cluster 5 contains only one neighborhood. This cluster is not the best location for setting up a new luxury hotel, as the small number of venues is unlikely to attract high budget travelers.

Discussion

TEGALSARI, GENTENG, and GUBENG are recommended as they are the top 3 neighborhoods in terms of total number of venues. Moreover, the three neighborhoods both have many culinary spots for deep-pocketed tourists to explore. The recommended neighborhoods should not be directly considered as a major factor in choosing which area in Surabaya is the best place to set up a new luxury hotel. Recommendations are made based on the similarity of venues between neighborhoods in Surabaya. There are still many other factors, such as competitors, hotel concepts, and prices that must be considered before choosing the best location to establish a 5-star hotel in Surabaya.

Conclusion

This project aims to assess hotel owners or hotel managers in deciding which neighborhood(s) in Surabaya to choose to establish a new 5-star hotel. By obtaining relevant data, pre-processing, and conducting Exploratory Data Analysis (EDA) on the data, we can see the characteristics of each neighborhood in terms of venues. In the final step, clustering is carried out to group similar neighborhoods that have the same venues categories, and the results are plotted on a map of Surabaya. Therefore, insights can be found to recommend specific neighborhoods in which the cluster should be highly considered when deciding the final location for setting up a luxury hotel.

The venue-based similarity recommendation from this project should be used only as a consideration. The final decision made by the hotel owners or hotel managers are on their own, and the outcome of the decision is beyond the responsibility of the author on this project.