

# **The Battle of Neighborhoods: Exploring Soccer Field for Seeking Business Opportunity**

Coursera Applied Data Science Capstone Project

Week 5 Submission – Full Report

I-Hsin Lin

4 July 2021

---

## **Contents**

1. Introduction .....	1
1.1. Background Information .....	1
1.2. Problem Statement.....	2
2. Data .....	2
3. Methodology.....	3
3.1. Regions, Postal Codes, and Geographical Coordinates .....	3
3.2. Exploring Venue .....	4
3.3. Most Common Venues Overall.....	5
3.4. One-Hot Encoding.....	5
3.5. Clustering Subdistricts Based on Venues Similarity.....	6
3.6. Examining Each Cluster .....	8
3.7. Visualizing Distribution of Soccer Field Locations in Kutai Kartanegara.....	9
4. Results and Discussion .....	9
5. Conclusion .....	10

---

## **1. Introduction**

### **1.1. Background Information**

The Government of Indonesia has announced that the nation's capital will be moved to the East Kalimantan region. The location of the new capital city is located in the part of North Penajam Paser Regency and partly in Kutai Kertanegara Regency, East Kalimantan. The government considered the location due to several factors such as infrastructure, supporting facilities, and being separated from the business center. Some time ago, Kompas.com summarized that in East Kalimantan there are three international standard football stadiums. Thus, it can also be said that the East Kalimantan region also has an attraction in the field of sports<sup>1</sup>.

---

1 .<https://bola.kompas.com/read/2019/08/26/19490918/ada-3-stadion-bertaraf-internasional-di-kawasan-kalimantan-timur?page=all>

## 1.2. Problem Statement

With the aforementioned prospect, the neighborhoods around the soccer fields may come as the priority for development. Various stakeholders (entrepreneurs, investors) may be interested to explore business opportunities in Kutai Kertanegara. This data science project is thus carried out to help them answer the following question: Which of the Kutai Kertanegara regions(around soccer fields) are strategic for investment?

The project will be of interest to fellow entrepreneurs.

## 2. Data

In order to explore the potential answers to the problems, the following data are required:

1. The names of administrative regions in East Kalimantan and their corresponding postal codes. The regions include three levels: city, district, subdistrict. The information was scraped from a directory on [indonesiapostcode.com](http://indonesiapostcode.com). The region names are useful to perform analysis across different sub-regions. The postal codes are needed to obtain the coordinates of each subdistrict.
2. Geographical coordinates of East Kalimantan and its subdistricts, which will, in turn, be needed to utilize Foursquare API in the subsequent step. Coordinates are obtained using GeoPy libraries, with Nominatim API.
3. Information about venues in East Kalimantan regions: the names, category, venue latitudes, venue longitudes. These are obtained using Foursquare API. The subdistricts of East Kalimantan will be clustered based on their surrounding venues to find the best location candidates..

### 3. Methodology

It begins with web scratching of Kutai Kartanegara city locations and postal codes and recovery of topographical arranges. Leveraging Foursquare API to investigate in the subdistricts.

One-hot encoding is to analyze and narrow down the most common venues in each of the subdistricts. Subdistricts are clustered utilizing K-means calculation. The number of ideal clusters is chosen to utilize the elbow method. Each cluster is separately analyzed to examine one discriminating venue that characterizes them.

Investigation of the clusters and visualization of soccer fields dissemination over Kutai Kartanegara will give insights as to where the strategic regions to set up the business are.

The following Python libraries and dependencies were used: pandas, NumPy, string, Requests, time.sleep, BeautifulSoup, GeoPy (Nominatim geocoder), JSON, Folium, Matplotlib, and scikit-learn.

#### 3.1. Regions, Postal Codes, and Geographical Coordinates

The city is further subdivided into districts (kecamatan) and then subdistricts (kelurahan). In total, there are 18 districts across Kutai Kertanegara

Using the scraped postal codes as inputs, the Nominatim geocoder were used to retrieve latitudes and longitudes of every subdistrict. Figure 1 displays the first 10 rows of the resulting dataframe: *post*.

512 venues collected.

	City	District	Subdistrict	Latitude	Longitude	Venue	Category	Venue_Lat	Venue_Lng
0	Kutai Kartanegara	Anggana	Anggana	-0.144614	115.569558	Stadion Santiago Berdebu	Soccer Field	-0.143411	115.571268
1	Kutai Kartanegara	Muara Badak	Badak Baru	-0.144614	115.569558	Stadion Santiago Berdebu	Soccer Field	-0.143411	115.571268
2	Kutai Kartanegara	Muara Badak	Badak Mekar	-0.144614	115.569558	Stadion Santiago Berdebu	Soccer Field	-0.143411	115.571268
3	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	Pasar Loa Janan	Flea Market	-0.567558	117.087335
4	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	mini market iwan	Department Store	-0.567858	117.088151
5	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	Bakso & Mie Ayam Uka Uka	Noodle House	-0.564819	117.086197
6	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	Ayam Bakar Lamongan	Javanese Restaurant	-0.562595	117.087168
7	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	RM Padang Bofet Salyo	Indonesian Restaurant	-0.560216	117.085548
8	Kutai Kartanegara	Tenggarong Seberang	Bangun Rejo	-0.439151	117.013672	Toko Haadannur	Convenience Store	-0.438995	117.006694
9	Kutai Kartanegara	Tenggarong Seberang	Bangun Rejo	-0.439151	117.013672	Keripik Singkong Shufia	Snack Place	-0.433249	117.008919

Figure 1. Dataframe *post*, containing postal codes and geographical coordinates of Kutai Kertanegara

### 3.2. Exploring Venues in the Subdistricts of Kutai Kertanegara

A total of 512 venues were collected through API calls to Foursquare that were made using a user-defined function. The result (Figure 2) is a dataframe containing a maximum of 100 venues within 1 km of a region (i.e., the center of a subdistrict), with the following details: venue name, venue category, venue latitude, venue longitude. On average, there are 2 venues per subdistrict (Figure 3).

512 venues collected.

	City	District	Subdistrict	Latitude	Longitude	Venue	Category	Venue_Lat	Venue_Lng
0	Kutai Kartanegara	Anggana	Anggana	-0.144614	115.569558	Stadion Santiago Berdebu	Soccer Field	-0.143411	115.571268
1	Kutai Kartanegara	Muara Badak	Badak Baru	-0.144614	115.569558	Stadion Santiago Berdebu	Soccer Field	-0.143411	115.571268
2	Kutai Kartanegara	Muara Badak	Badak Mekar	-0.144614	115.569558	Stadion Santiago Berdebu	Soccer Field	-0.143411	115.571268
3	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	Pasar Loa Janan	Flea Market	-0.567558	117.087335
4	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	mini market iwan	Department Store	-0.567858	117.088151
5	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	Bakso & Mie Ayam Uka Uka	Noodle House	-0.564819	117.086197
6	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	Ayam Bakar Lamongan	Javanese Restaurant	-0.562595	117.087168
7	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	RM Padang Bofet Salyo	Indonesian Restaurant	-0.560216	117.085548
8	Kutai Kartanegara	Tenggarong Seberang	Bangun Rejo	-0.439151	117.013672	Toko Haadannur	Convenience Store	-0.438995	117.006694
9	Kutai Kartanegara	Tenggarong Seberang	Bangun Rejo	-0.439151	117.013672	Keripik Singkong Shufia	Snack Place	-0.433249	117.008919

Figure 2. Dataframe *jakvenues*, containing venues information across Kutai Kertanegara.

	count	mean	std	min	25%	50%	75%	max
Venue Count	177.0	2.892655	7.101921	1.0	1.0	1.0	2.0	46.0

Figure 3. Statistical distribution of venue counts in every Kutai Kertanegara subdistricts.

### 3.3. Most Common Venues Overall

As shown in Figure 4, various kinds of restaurant top the list of most common venues in Kutai Kertanegara. Soccer Field, which is our venue of interest, comes in first. With almost 129 soccer fields, it is sure that business chances exist.

	Category	Count
0	Soccer Field	129
1	Café	43
2	Arcade	26
3	Convenience Store	26
4	Coffee Shop	25
5	Hotel	25
6	Indonesian Restaurant	19
7	Snack Place	19
8	Fast Food Restaurant	13
9	Restaurant	10

Figure 4. Top 10 most common venues in all of Kutai Kertanegara.

	Subdistrict	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Bakery	Board Shop	Boat or Ferry	Bookstore	Breakfast Spot	Buffet	Building	Café	Chinese Restaurant	Clothing Store	Coffee Shop
0	Anggana	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Badak Baru	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Badak Mekar	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Bakungan	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Bangun Rejo	0.333333	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	Baru (Kampung Baru)	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	Batu Batu	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	Batuah	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	Batua	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5. One-hot encoding, showing the mean frequency of venue occurrence in each subdistrict.

### 3.4. One-Hot Encoding

One-hot encoding converts categorical variables (i.e., venues) into numeric variables. In this case, a dummy of all the venues was made and the mean of the frequency of venue occurrence were calculated. The dataframe is then grouped by subdistrict, as

shown in Figure 5. The data were then filtered with a user-defined function to obtain 5 most common venues in each subdistrict, as displayed in Figure 6.

	Subdistrict	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Anggana	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
1	Badak Baru	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
2	Badak Mekar	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
3	Bakungan	Flea Market	Indonesian Restaurant	Javanese Restaurant	Department Store	Noodle House
4	Bangun Rejo	Arcade	Snack Place	Convenience Store	Chinese Restaurant	Food Court
5	Baru (Kampung Baru)	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
6	Batu Batu	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
7	Batuah	Flea Market	Indonesian Restaurant	Javanese Restaurant	Department Store	Noodle House
8	Batuq	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
9	Beloro	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market

Figure 6. Five most common venues in each of the Kutai Kertanegara subdistricts.

### 3.5. Clustering Subdistricts Based on Venues Similarity

The subdistricts were clustered based on a set of similar characteristics or features, i.e., their surrounding venues. *K*-Means clustering, which was used in this part of the analysis, is a machine learning algorithm that creates homogeneous subgroups/clusters from unlabeled data such that data points in each cluster are as similar as possible to each other according to a similarity measure (e.g., Euclidian distance).

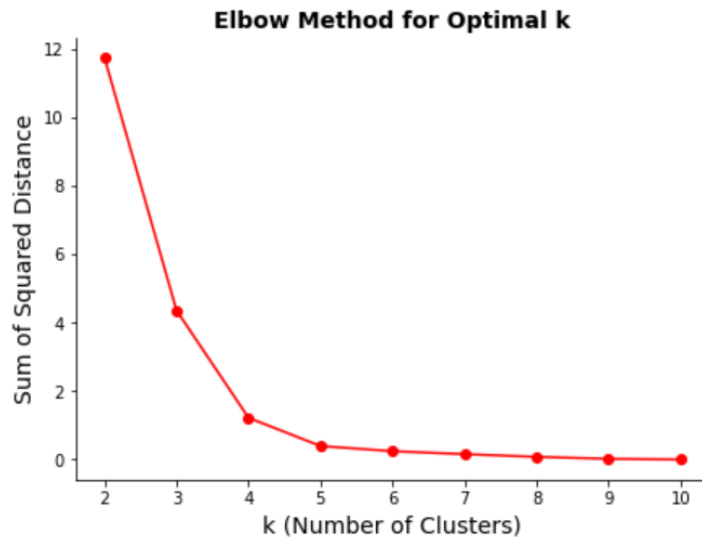


Figure 7. Elbow method: sum of squared distances for different *k* values.

	Postal_Code	City	District	Subdistrict	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	75381	Kutai Kartanegara	Anggana	Anggana	-0.144614	115.569558	0.0	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
3	75382	Kutai Kartanegara	Muara Badak	Badak Baru	-0.144614	115.569558	0.0	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
4	75382	Kutai Kartanegara	Muara Badak	Badak Mekar	-0.144614	115.569558	0.0	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
5	75391	Kutai Kartanegara	Loa Janan	Bakungan	-0.567465	117.087515	1.0	Flea Market	Indonesian Restaurant	Javanese Restaurant	Department Store	Noodle House
6	75572	Kutai Kartanegara	Tenggarong Seberang	Bangun Rejo	-0.439151	117.013672	3.0	Arcade	Snack Place	Convenience Store	Chinese Restaurant	Food Court
...	...	...	...	...	...	...	...	...	...	...	...	...
211	75553	Kutai Kartanegara	Muara Kaman	Tunjungan	-0.149484	116.917942	2.0	Coffee Shop	Café	Yoga Studio	Clothing Store	Food Court

Figure 8. A merged dataframe with cluster labels for every subdistrict.

A value of  $k$  (number of clusters) needs to be defined before proceeding with the clustering. The “Elbow Method” was used, which calculates the sum of squared distances of data points to their closest centroid (cluster center) for different values of  $k$ . The optimal value of  $k$  is the one after which there is a plateau (no significant decrease in sum of squared distances). Given that there is a peak at  $k = 4$  (Figure 7), the K-Means clustering was proceeded with that value.

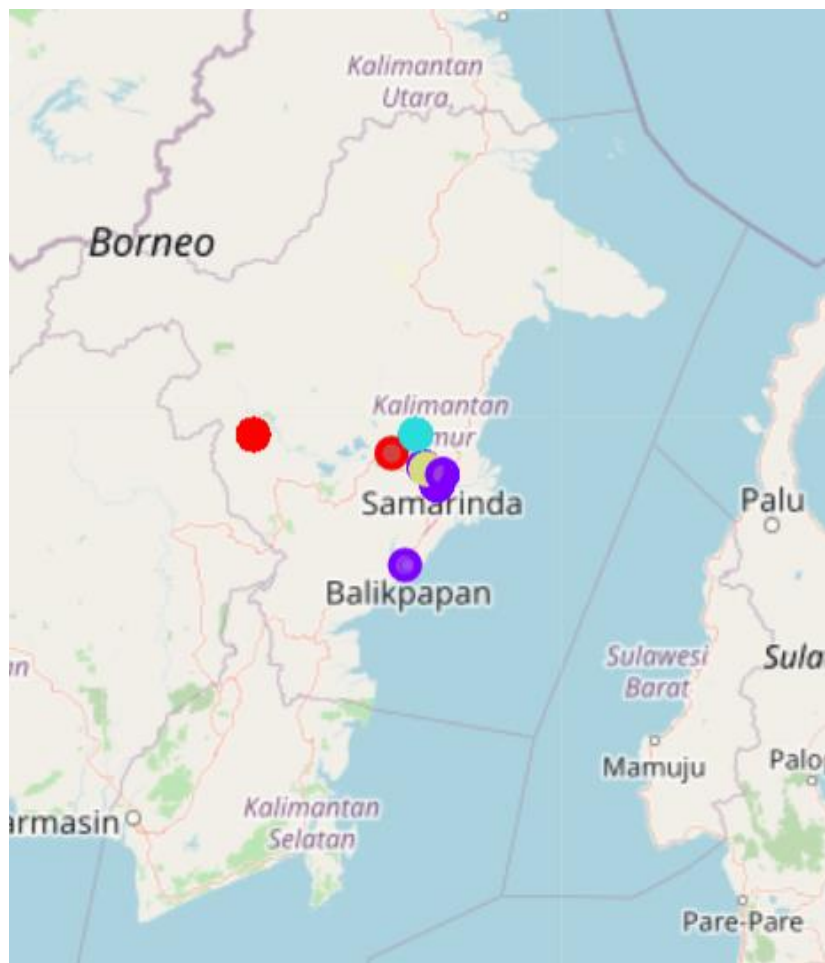


Figure 9. Clusters of Kutai Kertanegara subdistricts based on similarity of venues.

After each subdistrict had been assigned a cluster label (Figure 8), the clusters were color-coded and visualized on a map of Kutai Kertanegara (Figure 9) to understand how they are distributed across the regions.

### 3.6. Examining Each Cluster

Each cluster was filtered from the dataframe previously created in the clustering stage. The clusters were separately analyzed in order to gain an understanding of a discriminating venue that characterize each of them. The number one most common venue categories from each cluster, as well as the regions (cities) in which a particular cluster is highly concentrated were singled out.

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	125	125	125	125	125
unique	1	1	1	1	1
top	Soccer Field	Yoga Studio	Chinese Restaurant	Food Court	Flea Market
freq	125	125	125	125	125

Figure 10. Most common venue, Cluster 0(Red): Soccer Field

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	18	18	18	18	18
unique	6	6	6	6	6
top	Flea Market	Indonesian Restaurant	Javanese Restaurant	Department Store	Noodle House
freq	8	8	8	8	8

Figure 11. Most common venue, Cluster 1(Purple): Flea Market

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	21	21	21	21	21
unique	1	1	1	1	1
top	Coffee Shop	Café	Yoga Studio	Clothing Store	Food Court
freq	21	21	21	21	21

Figure 12. Most common venue, Cluster 2(Cyan): Coffee Shop

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	19	19	19	19	19
unique	1	2	2	2	2
top	Arcade	Snack Place	Convenience Store	Chinese Restaurant	Food Court
freq	19	17	17	17	17

Figure 13. Most common venue, Cluster 3(Yellow): Arcade



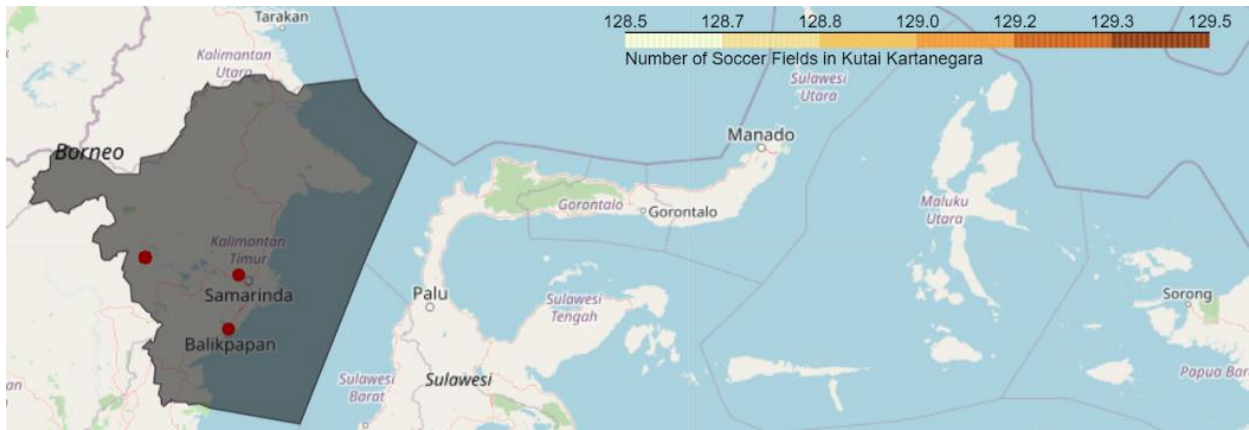


Figure 14. Concentrations of soccer fields across Kutai Kartanegara

### 3.7. Visualizing Distribution of Soccer Fields Locations in Kutai Kartanegara

Based on data in the *kalvenues* dataframe, the total number of soccer fields were calculated to examine the distribution of soccer fields businesses and to help figure out strategic locations. This distribution was visualized in the choropleth map (Figure 14).

## 4. Results and Discussion

A total of 512 venues from Kutai Kartanegara city were returned at the time the API call was made. There are on average 2 venues within a kilometer of a subdistrict center, where two of the most common categories overall is soccer fields which we are interested.

After deciding on an optimal *k* value of 4, K-Means algorithm was run to cluster the subdistricts based on their most common surrounding venues. Each of the four clusters, labeled 0–3, is characterized by a dominant venue as follow:

Cluster	Member	#1 Common Venues	Highly Concentrated District
0	125	Soccer Field	Tabang
1	18	Flea Market	Loa Janan
2	21	Coffee Shop	Muara Kaman
3	19	Arcade	Tenggarong Seberang

A considerable number of soccer fields can be found within Cluster 0 (125 soccer fields). In fact, it is the most common venues in that cluster. Choropleth map of soccer field locations shows that Kutai Kartanegara has a very high concentration of the business: 129 soccer fields are located there.

Looking at the visualizing distribution of soccer field locations, we can find out that there are three locations having most soccer fields. The districts in Kutai Kartanegara for suggestion will be Tenggarong and Balikpapan, because Tabang's location is more inland. They can be the options for developing and business investing.

## **5. Conclusion**

Regions in Kutai Kartanegara city were explored and then clustered based on the similarity of their surrounding venues using the K-Means algorithm. Analysis results show that Tenggarong and Balikpapan are the best candidate for developing business when moving to the new capital