Business Opportunities in Indonesian Urban Areas

Explored with Machine Learning in Python

A report by - Isaiah Fleming

1. Introduction

Figuring out what kind of business to open alone can be a daunting task, but factoring in location and competition can make this endeavor even more difficult. For this project, I wanted to imagine that I received a random email from someday asking where, and what sort of business they should open in a random part of the world. I ended up creating this “email” as my prompt:

“I am a young and enthusiastic business pioneer with their sights set on Indonesia! All I want is a business to call my own, I don’t care what kind it is. However, I don’t know anything about Indonesia. I want to know what kind of business I should start and where in Indonesia I should start it in order to be the most successful.”

Now, I personally know almost nothing about business, let alone what sort of business would be successful in any given location. But by leveraging locational API data from Foursquare and using a bit of machine learning, I was able to create something that knew more about businesses in Indonesia than I ever could alone. And using this information I am able to confidently suggest a business plan to anybody with a similar situation to the above email.

2. Data

Data used in this projects is as follows:

- Table of “Built-up urban areas” in Indonesia sourced from Wikipedia.[1] A Built-up urban area is described as “...according to Demographia's "World Urban Areas" study. Demographia defines an urban area (urbanised area agglomeration or urban centre) as a continuously built up land mass of urban development that is within a labor market”. I selected this data as it seemed most appropriate given the context of business information.

- Venue information sourced from the Foursquare API for each of these built-up urban areas. This information includes venue names and their categories (cafe/bowling alley/park/etc.)[3]

- Location data was also obtained for each urban area in order to properly plot a map for visualization. Data was sourced using area names and the Nominatim geocoder API.[4]

3. Methodology

For this project, all data was sourced and manipulated through a Python Jupyter notebook using various libraries.

To start, data from the “Built-up urban areas” in Indonesia table on Wikipedia was scraped into a Pandas dataframe using the Beautifulsoup library.

	Urban Area	Area (Sq. Km)	Estimated Population
0	Jakarta	3,540	34,540,000
1	Bandung	487	7,065,000
2	Surabaya	911	6,499,000
3	Medan	478	3,632,000
4	Semarang	259	1,992,000
5	Makassar	178	1,952,000
6	Palembang	221	1,880,000

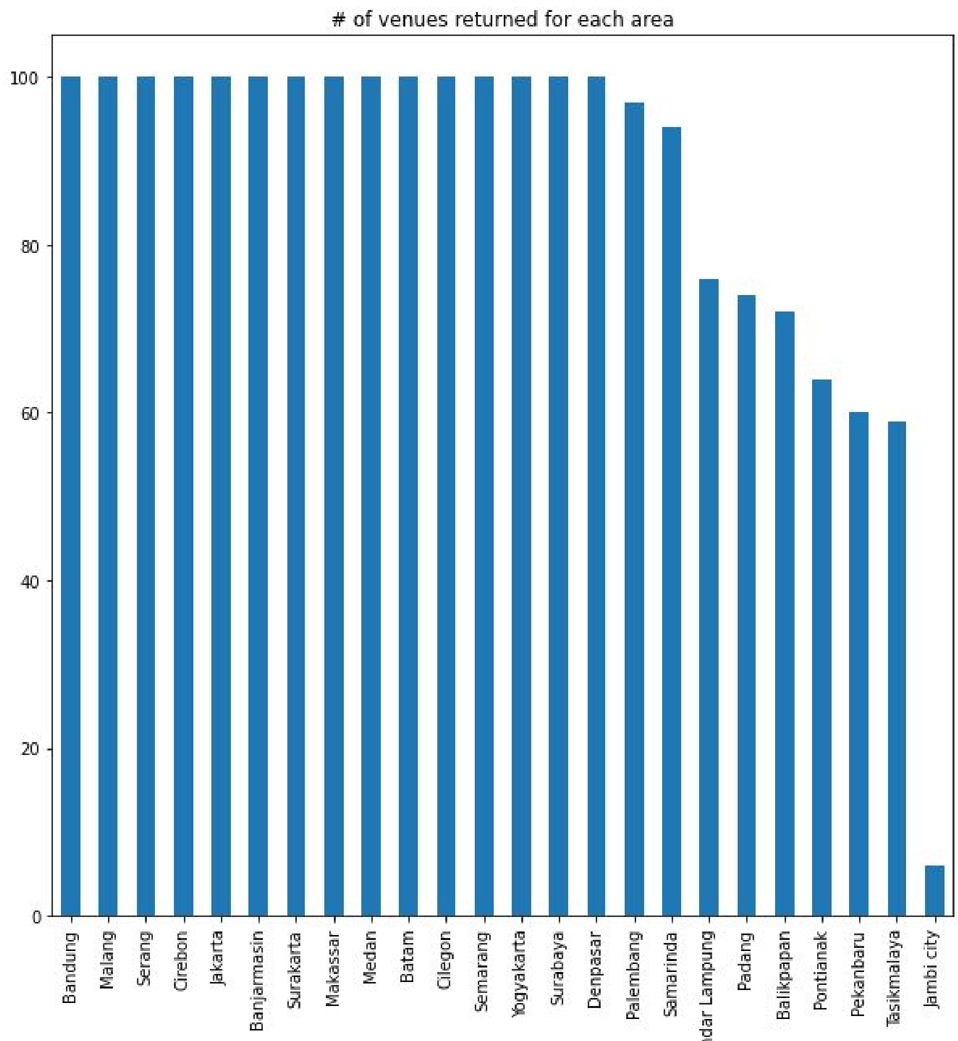
The Nominatim geocoder was then used to obtain latitude and longitude values for each of these areas, which were then appended to the dataframe.

	Urban Area	Area (Sq. Km)	Estimated Population	Latitude	Longitude
0	Jakarta	3,540	34,540,000	-6.175394	106.827183
1	Bandung	487	7,065,000	-6.934469	107.604954
2	Surabaya	911	6,499,000	-7.245972	112.737827
3	Medan	478	3,632,000	3.589665	98.673826
4	Semarang	259	1,992,000	-6.990399	110.422910
5	Makassar	178	1,952,000	-5.134296	119.412428
6	Palembang	221	1,880,000	-2.988830	104.756857

This data was then used to obtain venue information for each area from the Foursquare API. Multiple dataframes and plots were created using the data gathered.

	Urban Area	Hotel	Coffee Shop	Shopping Mall	Indonesian Restaurant	Bakery	Sushi Restaurant	Multiplex	BBQ Joint	Clothing Store	...
0	Jakarta	16	12	6	4	3	3	3	2	2	...
1	Bandung	14	18	2	2	15	2	3	0	1	...
2	Surabaya	12	14	8	12	5	1	5	1	2	...
3	Medan	3	12	1	8	7	3	2	3	2	...
4	Semarang	6	6	0	9	0	0	2	2	0	...

A dataframe containing the number of each kind of venue for each area.

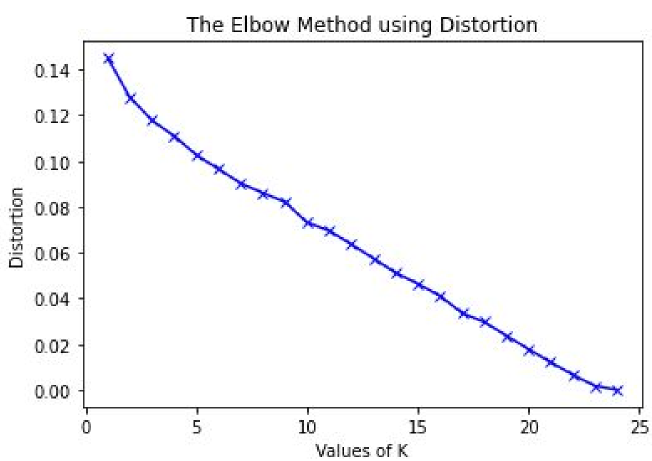


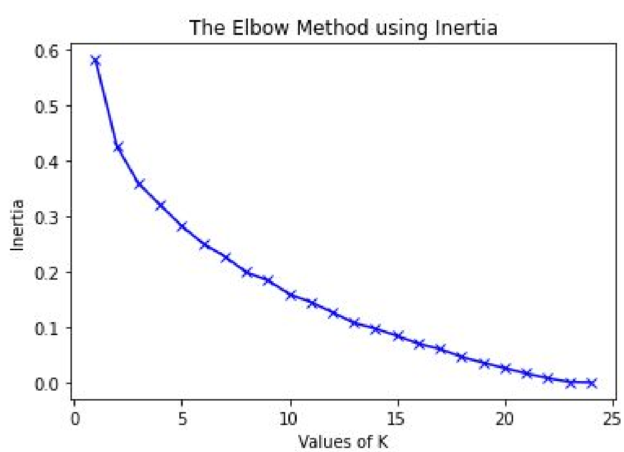
A plot showing the total number of venues retrieved for each area. (I learned later that Foursquare imposes a limit of 100 results for each API call. This is discussed in the discussion section.)

	Urban Area	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	Balikpapan	Coffee Shop	Seafood Restaurant	Bakery	Indonesian Restaurant	Park	Shopping Mall	Hotel	Café	Lounge	Chinese Restaurant
1	Bandar Lampung	Indonesian Restaurant	Beach	Coffee Shop	Noodle House	Hotel	Chinese Restaurant	Bakery	Snack Place	Fast Food Restaurant	Breakfast Spot
2	Bandung	Coffee Shop	Bakery	Hotel	Café	Japanese Restaurant	Sundanese Restaurant	Multiplex	Steakhouse	Shopping Mall	Udon Restaurant
3	Banjarmasin	Hotel	Indonesian Restaurant	Café	Diner	Coffee Shop	Breakfast Spot	Soup Place	Asian Restaurant	Food	Fast Food Restaurant
4	Batam	Hotel	Coffee Shop	Beach	Waterfront	Ice Cream Shop	Theme Park Ride / Attraction	Resort	Café	Shopping Mall	Botanical Garden
5	Cilegon	Indonesian	Asian	Café	Hotel	Beach	Resort	Diner	Soup Place	Fast Food	Food Truck

A dataframe containing the top 10 venues for each area.

Using the last dataframe, I used the K means machine learning algorithm to cluster all of the Urban areas into 3 different groups. 3 was determined as the ideal number of clusters using the elbow method.





While the distortion graph is a little difficult to find the elbow on, the inertia graph shows a somewhat more prominent elbow around a K(# of clusters) of 3. In both of these, it can be seen that values for distortion and inertia both start at very low values, which implies certain things about our data set. This is also discussed in the discussion section.

I then sliced the top 10 dataframe to show information for each cluster.

	Urban Area	Area (Sq. Km)	Cluster Labels	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
21	Jambi city	78.0	0	Convenience Store	River	Scenic Lookout	Trail	Pizza Place	Hotel	Farm	Food Court	Food & Drink Shop	Food

This is cluster 0, with only Jambi city.

	Urban Area	Area (Sq. Km)	Cluster Labels	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
3	Medan	478.0	1	Coffee Shop	Indonesian Restaurant	Bakery	Chinese Restaurant	Restaurant	Seafood Restaurant	Noodle House	Spa	Pizza Place	BBQ Joint
4	Semarang	259.0	1	Indonesian Restaurant	Asian Restaurant	Coffee Shop	Hotel	Food Truck	Chinese Restaurant	Snack Place	Steakhouse	Seafood Restaurant	Food Court
5	Makassar	178.0	1	Hotel	Coffee Shop	Indonesian Restaurant	Café	Seafood Restaurant	Soup Place	Noodle House	Snack Place	Pizza Place	Clothing Store
6	Palembang	221.0	1	Indonesian Restaurant	Noodle House	Asian Restaurant	Café	Multiplex	Coffee Shop	Restaurant	Donut Shop	Shopping Mall	Hotel
8	Malang	212.0	1	Hotel	Soup Place	Coffee Shop	Café	Indonesian Restaurant	Supermarket	Chinese Restaurant	Indonesian Meatball Place	Snack Place	Asian Restaurant
9	Denpasar	177.0	1	Café	Hotel	Restaurant	Resort	Coffee Shop	Indonesian Restaurant	Beach	Bakery	Chinese Restaurant	Surf Spot
11	Pekanbaru	239.0	1	Café	Indonesian Restaurant	Asian Restaurant	Coffee Shop	Hotel	Seafood Restaurant	Chinese Restaurant	Diner	Dessert Shop	Fried Chicken Joint
12	Surakarta	477.0	1	Indonesian Restaurant	Hotel	Coffee Shop	Asian Restaurant	Snack Place	Shopping Mall	Pizza Place	Café	Fried Chicken Joint	Soup Place
13	Cirebon	105.0	1	Indonesian Restaurant	Hotel	Café	Asian Restaurant	Coffee Shop	Bakery	Gift Shop	Restaurant	Supermarket	Sundanese Restaurant
14	Bandar Lampung	107.0	1	Indonesian Restaurant	Beach	Coffee Shop	Noodle House	Hotel	Chinese Restaurant	Bakery	Snack Place	Fast Food Restaurant	Breakfast Spot
15	Samarinda	102.0	1	Café	Hotel	Asian Restaurant	Convenience Store	Seafood Restaurant	Fast Food Restaurant	Coffee Shop	Chinese Restaurant	Soup Place	Japanese Restaurant
16	Padang	99.0	1	Padangnese Restaurant	Indonesian Restaurant	Seafood Restaurant	Asian Restaurant	Café	Coffee Shop	Hotel	Donut Shop	Bakery	Convenience Store
17	Banjarmasin	65.0	1	Hotel	Indonesian Restaurant	Café	Diner	Coffee Shop	Breakfast Spot	Soup Place	Asian Restaurant	Food	Fast Food Restaurant
18	Tasikmalaya	62.0	1	Indonesian Restaurant	Sundanese Restaurant	Indonesian Meatball Place	Hotel	Juice Bar	Coffee Shop	Café	Department Store	Diner	Snack Place
19	Pontianak	62.0	1	Indonesian Restaurant	Chinese Restaurant	Asian Restaurant	Seafood Restaurant	Coffee Shop	Hotel	Café	Arcade	Athletics & Sports	Food Truck
20	Balikpapan	124.0	1	Coffee Shop	Seafood Restaurant	Bakery	Indonesian Restaurant	Park	Shopping Mall	Hotel	Café	Lounge	Chinese Restaurant
22	Serang	65.0	1	Indonesian Restaurant	Asian Restaurant	Resort	Café	Hotel	Food Truck	Beach	Fast Food Restaurant	Diner	Soup Place
23	Cilegon	122.0	1	Indonesian Restaurant	Asian Restaurant	Café	Hotel	Beach	Resort	Diner	Soup Place	Fast Food Restaurant	Food Truck

This is cluster 1. It is the largest cluster.

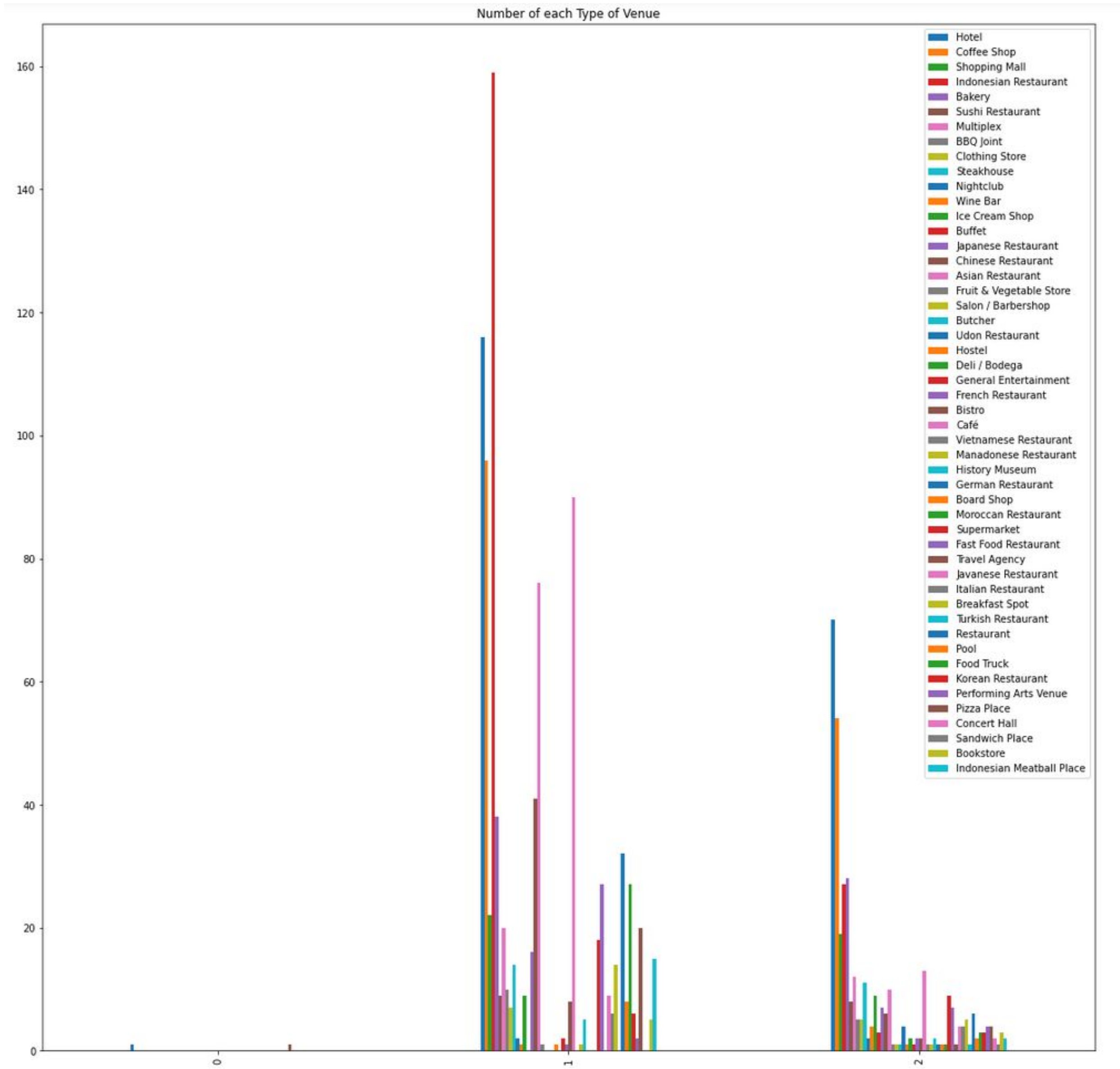
	Urban Area	Area (Sq. Km)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Jakarta	3540.0	2	Hotel	Coffee Shop	Shopping Mall	Indonesian Restaurant	Multiplex	Sushi Restaurant	Bakery	Nightclub	Clothing Store	Chinese Restaurant
1	Bandung	487.0	2	Coffee Shop	Bakery	Hotel	Café	Japanese Restaurant	Sundanese Restaurant	Multiplex	Steakhouse	Shopping Mall	Udon Restaurant
2	Surabaya	911.0	2	Coffee Shop	Hotel	Indonesian Restaurant	Shopping Mall	Multiplex	Bakery	Steakhouse	Supermarket	Movie Theater	Seafood Restaurant
7	Yogyakarta	230.0	2	Hotel	Indonesian Restaurant	Coffee Shop	Asian Restaurant	Bakery	Pizza Place	Breakfast Spot	Javanese Restaurant	Food Truck	Fast Food Restaurant
10	Batam	243.0	2	Hotel	Coffee Shop	Beach	Waterfront	Ice Cream Shop	Theme Park Ride / Attraction	Resort	Café	Shopping Mall	Botanical Garden

This is cluster 2.

Using the dataframe containing the number of each kind of venue for each area shown earlier, I manipulated it to show the total number of each kind of venue for each cluster.

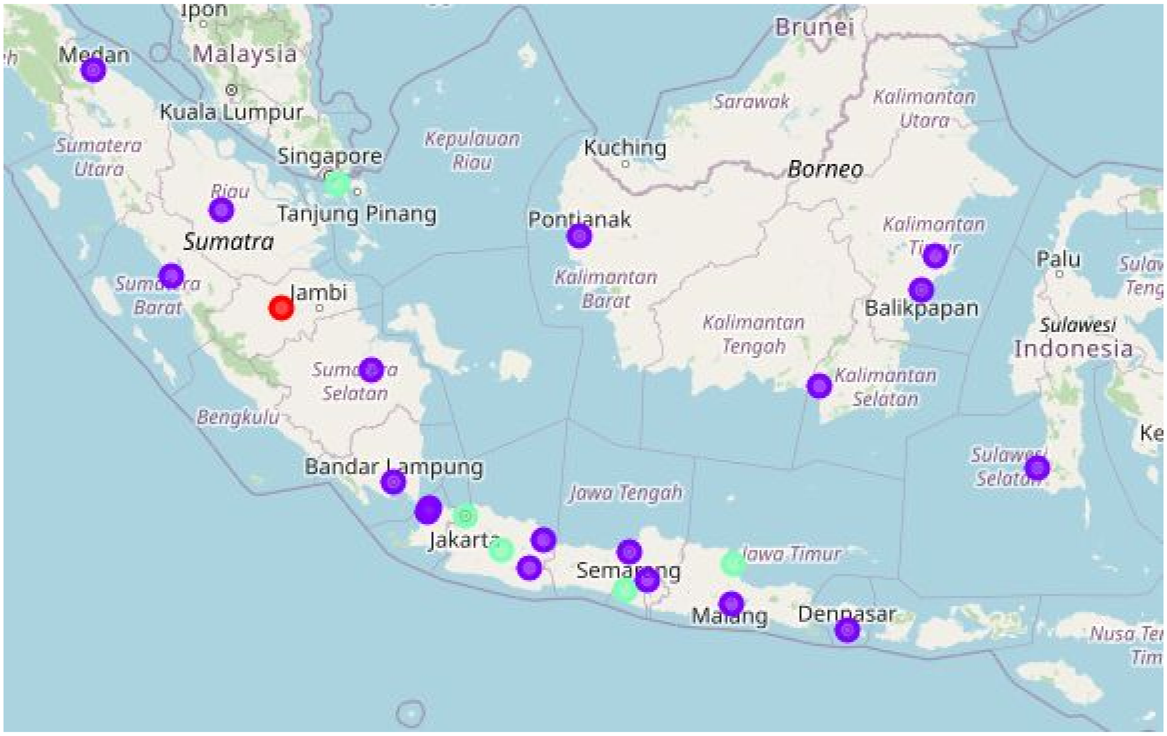
	Hotel	Coffee Shop	Shopping Mall	Indonesian Restaurant	Bakery	Sushi Restaurant	Multiplex	BBQ Joint	Clothing Store	Steakhouse	...
0	1	0	0	0	0	0	0	0	0	0	...
1	116	96	22	159	38	9	20	10	7	14	...
2	70	54	19	27	28	8	12	5	5	11	...

Using this dataframe I created a plot to help visualize the overall makeup of each cluster.



4. Results

After using our clustering algorithm to group each urban area, I then used the location data from earlier to plot a map of each urban area colored respectively based on their cluster groups.



It can be seen that almost all areas got clustered into two groups (purple/green), with just one urban area being grouped into its own cluster (red). Red is cluster 0, Purple is cluster 1, and Green is cluster 2.

5. Discussion

Looking at the cluster map we can immediately make some basic observations. While areas that fall into cluster 1 (purple) can be found throughout Indonesia, areas that fall into cluster 2 (green) are mostly located on the same island.

Using the sliced tables as well as the graph showing the total number of each kind of venue per cluster, I classified these clusters in a way that makes more sense. Cluster 1 (purple), has a **significant** abundance of Indonesian restaurants. Considering this is Indonesia, this makes sense. However, the other popular venues are very similar to cluster 2 (green)'s top 10 venues, including things mostly like hotels, cafes, and restaurants for cuisine other than Indonesian. Keeping this in mind, to me, cluster 2 (green) appears to be much more tourism oriented. Considering most areas in this cluster appear in the same region, that probably means most tourists stay in that general area, and travel between those urban areas to see different things. The most common venues in cluster 2 are hotels and cafes. Hotels are necessary to house tourists while cafes are convenient places to stop in for a quick bite while out exploring. The remaining venues in cluster 2's most common venues are things like beaches and malls and other comfort foods, which are also popular with tourists anywhere.

Going back to cluster one (purple), while many of the popular venues are similar (many hotels and cafes), the abundance of Indonesian restaurants makes me think that these areas are built much more on business within Indonesia. There are still many hotels as people traveling for business still need places to stay, but the food surrounding those hotels is much more local. This tells me that the people staying in those hotels must also be more local as well. As a whole, the overall most common kinds of venues in this cluster are food related. Which makes sense as most people working tend to go out for lunch.

This all brings us to cluster 0 (red), with only one area. Comparing this cluster to the last two makes it fairly apparent that this one area is very different than the others. The most common venues there are convenience stores, with the runners up being outdoors type things like rivers and trails. Hotels are only the 7th most common for this area. This tells me that this area is much more local, and does not involve much travel at all.

However, there are some caveats with the data and methodology here. As shown previously in a graph, the Foursquare API caps out at 100 results for every call. Looking at that graph, the majority of the areas hit that limit. This means that I was potentially missing large numbers of additional venues for each of those areas, which could certainly prove significant if those missing venues would change the orders of the top 10 venues for each area. If I were to do this project again, this is something that would need to be addressed.

Additionally, when looking at the graphs used in the elbow method for determining the ideal number of clusters for my K means algorithm, it can be seen that even with a K of 1, the inertia and distortion are very low. This tells me that venue information for these areas of Indonesia in particular are already very similar when using this particular model. This is evident when you consider how similar clusters 1 and 2 are, and the fact the cluster 0 only has one area. In the future, I might consider using a DBSCAN model in an attempt to find more specifically shaped clusters.

6. Conclusion

Despite the potential issues discussed above, I believe I can still give a fairly concrete answer to our original email.

If you are looking to open a new business in Indonesia, consider the following:

- There are many, very developed areas containing many hotels.
    - Considering this, you could venture to open your own hotel, or perhaps some sort of business that caters to hotels or the people using them. (HVAC/Dry Cleaning/Etc.)
- In these areas with many hotels, many appear to be focused on tourism. Having many cafes and restaurants with food from all over. (Cluster 2)
    - Considering this, in these areas, a successful business strategy could just be starting another cafe or restaurant.
    - But starting something with high tourist appeal might also be a good move. Things like surf/scuba experiences for areas with beach venues, tours, or even transportation to help tourists get around.
- On the other hand, there are even more areas full of hotels which seem to focus more on big business. (Cluster 1)
    - These areas already appear to be dominated by local restaurants, however, if you have a particularly juicy family recipe up your sleeves, opening your own restaurant might prove to be very successful. Everyday people are going to be going out to lunch.
    - However, starting a business that caters to businesses and business people might also work well. Things like print shops, office stores, general IT support would work well in these areas.
- In this situation I would not recommend to start a business in Cluster 3.
    - It's main businesses are convenience stores and outdoor activities. It seems like maybe the people there don't have a lot of money to spend on things like restaurants or other businesses.

Overall this project is not perfect, and many things could be tweaked and improved. However, I still believe it provides useful and practical information about current, real world data.

7. References

1. List of Indonesian cities by population. (2020, June 25). Retrieved July 10, 2020, from https://en.wikipedia.org/wiki/List_of_Indonesian_cities_by_population
2. Foursquare.com
3. Nominatim.org