Exploring Most Popular Venues for Three Cities in Saudi Arabia

Feras Alshaheen

February 7, 2021

1. Introduction

a. Background

Saudi Arabia has started working on its vision for 2030 since 2016 to strengthen and diversify its economy, and develop public sectors. One of the main themes of its 2030 vision is *Thriving Economy*, under which we have three goals. One of those goals is to have the country open for business and investment. This goal aims to encourage both local and foreign business owners to invest their money and start their business in Saudi Arabia.

b. Business problem

In order to start a business, a person would have to look for both the right type of business to invest in and where to invest. That person would have to ask the following questions: *Where do people spend most of their time? And what type of places they are looking for?*

By asking these two questions, he will be able to know where to search and what kind of service can he offer to attract. Moreover, an opportunity can be found by looking for specific sector that are not fully developed to invest in it.

c. Cities of interest

In this project, three cities were selected to be explored. Those are the three major cities in Saudi Arabia that are located in different areas. Those cities are: Riyadh, Jeddah, and Dammam.

Riyadh is the capital city of the country and located in the central region of Saudi Arabia. Jeddah is the main and the largest harbor on the western region. Likewise, Dammam is the major harbor on the eastern region of the country.

2. Data gathering and preparing

a. Data source

The main source of data is chosen to be *Foursquare*. All required information needed such as the name of the venue, venue category, location, and map coordination are available and ready to be used.

b. Data preparing

The obtained data has been through three main phases of cleaning and refining. First phase is filling missing data and having value consistency. The picture below (figure 1) is an example that shows how the city's name where the venues are located

is written differently and sometimes a "NaN" value was found, which means there is no value.

lat	lng	city
24.636793	46.714460	الرياض
24.642024	46.717322	NaN
24.631262	46.713420	Riyadh
24.630015	46.713228	الرياض
24.630936	46.712279	الرياض

FIGURE 1

The followed procedure here is to replace all values with one value (figure 2).

lat	lng	city
24.636793	46.714460	Riyadh
24.642024	46.717322	Riyadh
24.631262	46.713420	Riyadh
24.630015	46.713228	Riyadh
24.630936	46.712279	Riyadh

FIGURE 2

Second phase is extracting a meaningful value out of the given one. For example, by looking in the 3rd figure below, under "venue categories" column, the values have no meaning.

.location.country	venue.location.formattedAddress	venue.categories	venue.photos.count	venue.photos.groups
المملكة العربية	[Dammām, المملكة العربية السعودية]	[{'id': '4bf58dd8d48988d1c4941735', 'name': 'R...	0	[]
المملكة العربية	[الدمام سوق الحُب (شارع الملك فهد, Dammām 324...	[{'id': '4bf58dd8d48988d111951735', 'name': 'J...	0	[]
المملكة العربية	[المملكة العربية السعودية]	[{'id': '4bf58dd8d48988d143941735', 'name': 'B...	0	[]
المملكة العربية	[21st St, Dammām 32242, المملكة العربية السعودية]	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	0	[]
المملكة العربية	[المملكة العربية السعودية]	[{'id': '4bf58dd8d48988d1ce941735', 'name': 'S...	0	[]

FIGURE 3

The followed procedure was to extract the meaningful value out using a pre-defined function in the Python code. The result is shown in figure 4.

categories	lat	lng
Restaurant	26.434712	50.098633
Jewelry Store	26.445071	50.103030
Breakfast Spot	26.443876	50.103850
Hotel	26.427482	50.097544
Seafood Restaurant	26.431684	50.115110

FIGURE 4

Third and last phase is filtering the columns and limit them only to the required ones. For example, there were columns such as venue ID, country name, or postal codes and those information were not needed. The followed procedure was limiting our inputs to five main columns: venue name, category, latitude, longitude, and city's name. Figure 5 shows the final data frame.

	City	Venue Name	Category	Latitude	Longitude
0	Jeddah	starbuck-drive-thru | ستاربكس	Coffee Shop	21.579160	39.166585
1	Jeddah	Talah Aljood (تالة الجود)	Dessert Shop	21.579704	39.161000
2	Jeddah	Fitness Time Ladies	Gym / Fitness Center	21.585359	39.161615
3	Jeddah	Perks	Coffee Shop	21.574460	39.164485
4	Jeddah	Ma'asob Alsultan (معصوب السلطان)	Breakfast Spot	21.589741	39.165862

FIGURE 5

3. Data Analysis Methodology

The methodology used to analyze the data is fairly simple and straight forward. Since the aim was to find the trend and what are popular venues' categories in each city, the relationship between the cites and venues categories is changed from just a categorical variable like what was shown in figure 5, to numerical ones. Those numerical values are basically zeros and ones, where one indicates this type of venue such as art gallery exists in Riyadh. While zero indicates that this type of venue doesn't exist under a Jeddah city for example (figure 6).

	City	African Restaurant	American Restaurant	Art Gallery	Athletics & Sports	Bakery	Bank	Big Box Store	Bookstore	Boxing Gym	Breakfast Spot
0	Jeddah	0	0	0	0	0	0	0	0	0	0
1	Jeddah	0	0	0	0	0	0	0	0	0	0
2	Jeddah	0	0	0	0	0	0	0	0	0	0
3	Jeddah	0	0	0	0	0	0	0	0	0	0
4	Jeddah	0	0	0	0	0	0	0	0	0	1

FIGURE 6

After that, the frequency of appearance for each kind of venue is calculated by summing all ones for specific category related to each city. For example, if the category *Breakfast Spot* appeared five times in the city of Jeddah, it will be divided by the total numbers of venues related to Jeddah. It worth mentioning that we limited our search for only 100 venue. This means that *Breakfast Spot* in Jeddah represents only 5% of the total numbers of venues in the city of Jeddah (figure 7).

	City	African Restaurant	American Restaurant	Art Gallery	Athletics & Sports	Bakery	Bank	Big Box Store	Bookstore	Boxing Gym	Breakfast Spot
0	Dammam	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.02	0.01	0.05
1	Jeddah	0.01	0.01	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.06
2	Riyadh	0.00	0.00	0.03	0.01	0.02	0.01	0.01	0.00	0.00	0.08

FIGURE 7

The finding of the above analysis will be presented in the below section. However, since we have around 85 categories under each city, it won't be feasible to show all of them on the charts because they will be congested, so only the highest 10 categories were selected.

4. Findings and Results

a. Visualizing the findings

After converting the relationship as mentioned in the analysis section to numerical type, and finding the popularity (or the frequency) of each category; the findings were visualized using bar charts. Below are three figures to sum up and explains the findings of our analysis (figures 8 – 10). There are four main observations over what is shown below

1) Coffee shops are the most popular across all three cities.
2) Riyadh has some categories that are not available in Jeddah and Dammam. Those categories are among the top 10 popular places: historic sites, history museums, and art galleries. This can be a main contribution in having hotels as one of the top 5 popular venues. More attractions means that people are willing to stay and spend more time exploring the city.
3) Jeddah and Dammam are very similar to each other from the point of view that they have 6 common categories (even though the ranking may differ).
4) Finally, from below charts it can be seen that most attractive city to stay in is Jeddah, and the lowest is Dammam. This is shown by having *Hotels* ranked forth for Jeddah, fifth in Riyadh, and tenth in Dammam.

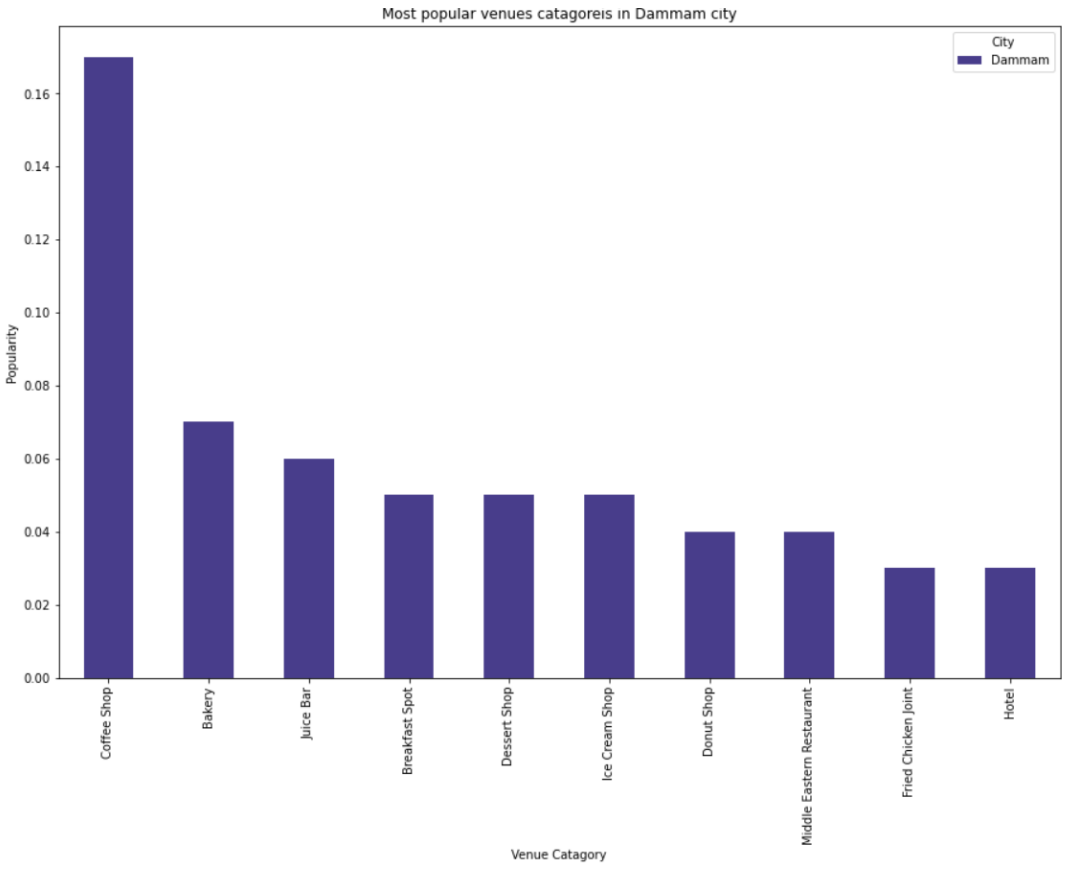

FIGURE 8

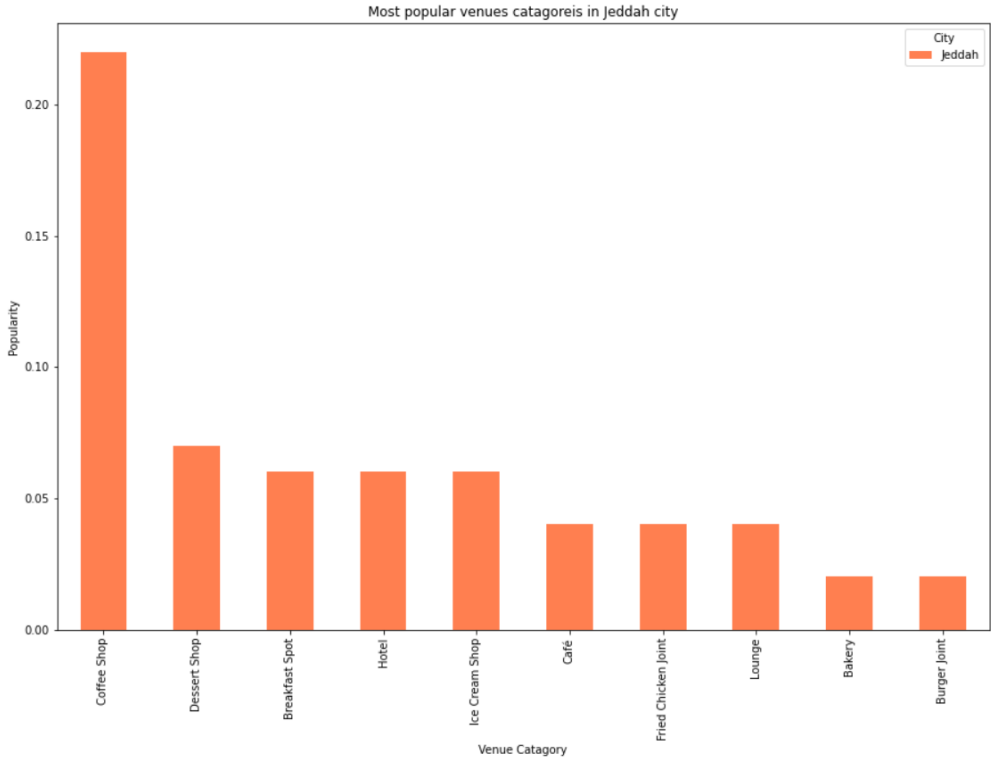

FIGURE 9

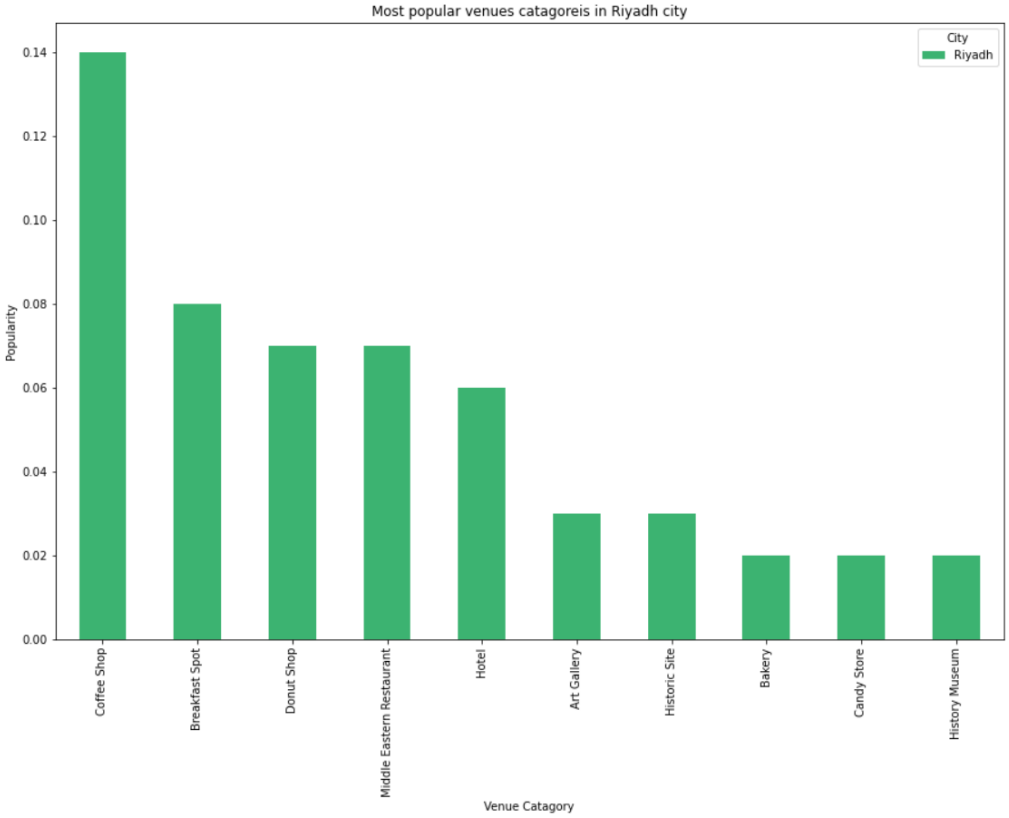

FIGURE 10

b. Sample representing over the map

The last step is to find where are most of the venues located in each city. In figure 11, it is shown that there are 100 venues shown fall under each city.

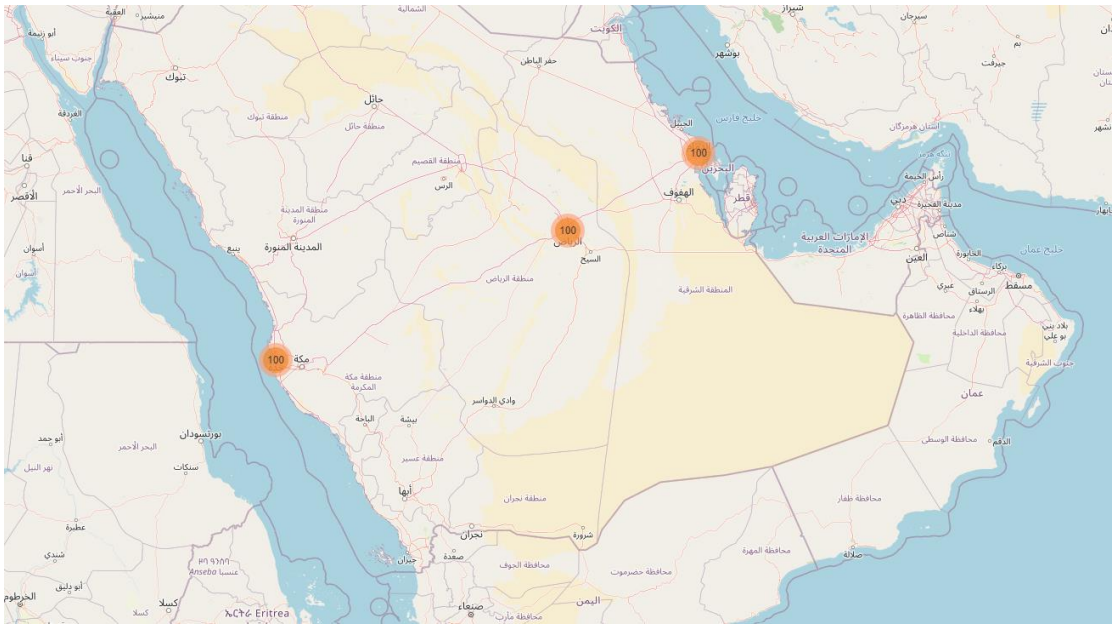

FIGURE 11

Most of the venues are located at the center of those cities. In figure 12 as an example, we are zooming in into Riyadh city. By using map clustering, we can see that we have three groups. One that contains 63 venues, the second has 19 venues and the last contains 16 venues. All located around the center.

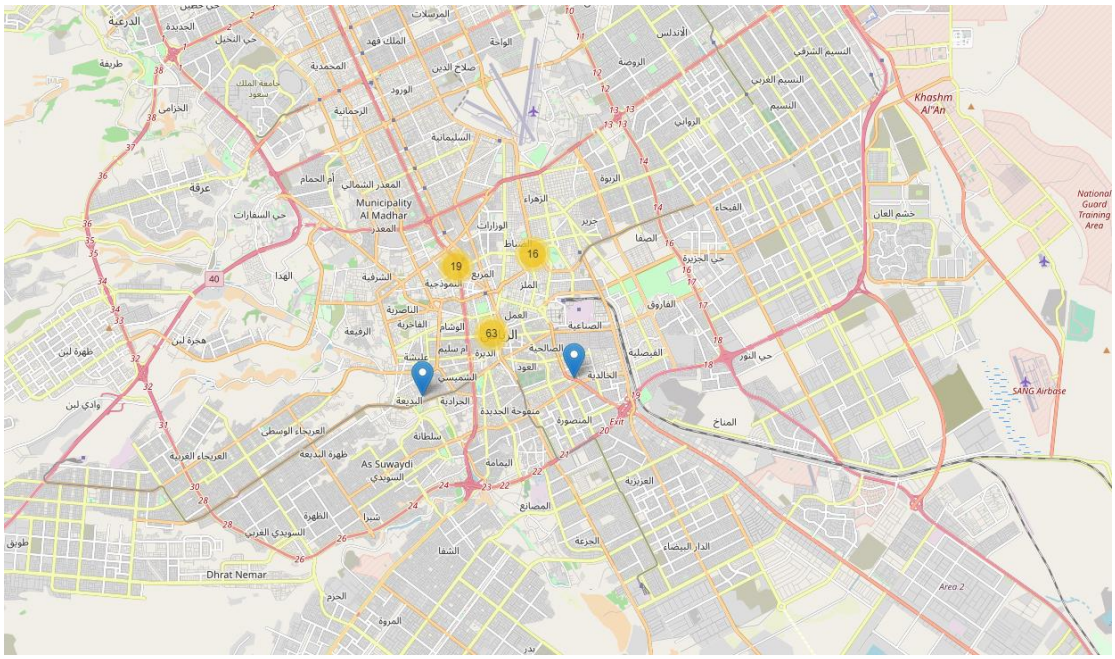

FIGURE 12

In figure 13 we zoom even further by choosing the group that contains 63 venues. Once we do, it can be seen how those venues are distributed and there are even smaller groups. The largest of them is the one that has 15 venues. This means there are 15 out of the 63 located nearby to each other.

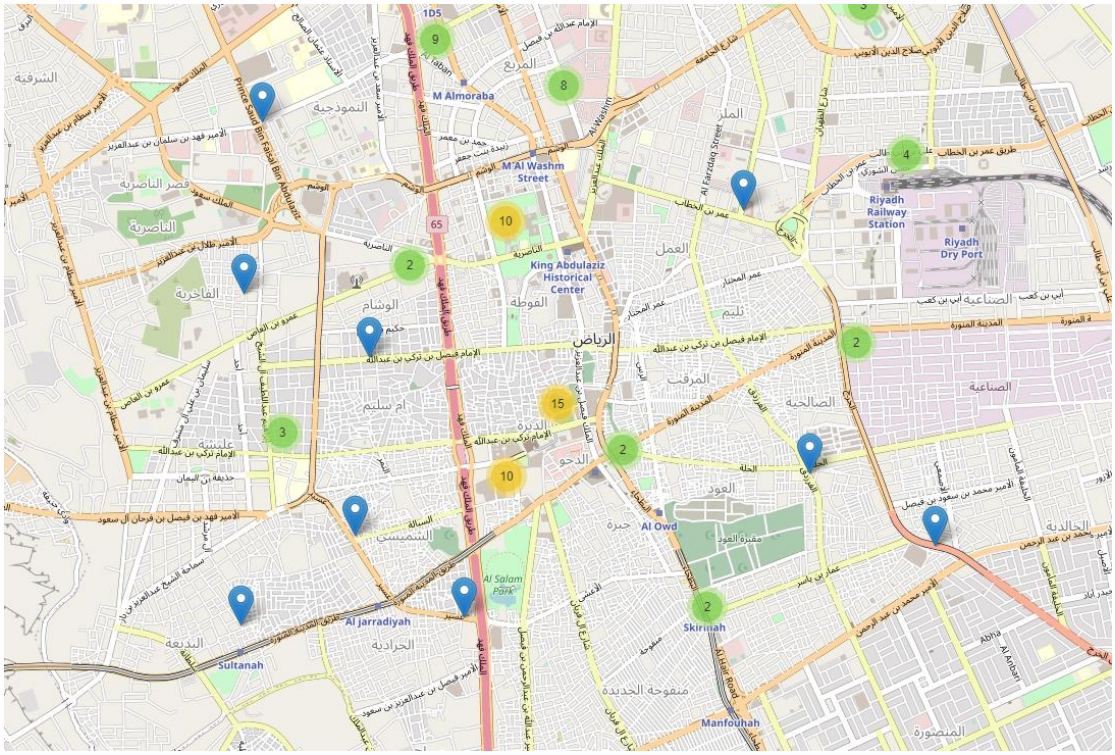

FIGURE 13

5. Conclusion

In conclusion, we extracted the data related for 100 venues for each of the three cities. The data were refined and prepared to be analyzed by extracting meaningful values, substituting missing ones, and filtering required columns. Then, to analysis the data the categorical values were replaced with numerical indicators that were summed and divided by the total number of values to get the percentage of each category we have. After, that the results were visualized in order to extract the findings and the outputs from our analysis. Finally, we mapped the results in order to know the focus areas where most of the venues are located.

Those findings should give the investors an idea on where do people spend their times and what are they looking for. Also, those findings can be utilized by finding what is missing and what can be introduced to people as a new place to spend their time and money such as amusement parks and national museums.