Exploring Riyadh City neighborhoods

Applied Data Science Capstone

Hichem Dridi

3/27/21

IBM Data Science Professional Certificate

1 Introduction:

1.1 Background:

During the last course of the professional certificate, we've been using different tools to explore famous cities like *New York & Toronto*. The process was simple: gather data about the city, use *Foursquare API* to explore the venues in the neighborhoods and then group similar neighborhoods into different clusters.

In order to accomplish the final assignment, I've decided to explore also one of the popular cities in the world: ***Riyadh City***.

1.2 About the city:

Riyadh ('The Gardens') is the capital and main financial hub of *Saudi Arabia,* and the largest city on the Arabian Peninsula. Located in the center of the an-Nafud desert, on the eastern part of the Najd plateau, the city sits at an average of 600 meters (2,000 ft) above sea level, and receives around ***5 million tourists*** each year, making it the ***forty-ninth*** most visited city in the world and the ***6th*** in the Middle East.

Riyadh had a population of 7.6 million people in 2019, making it the most-populous city in Saudi Arabia, 3rd most populous in the Middle East, and 38th most populous in Asia.

1.3 Objective:

Considering the mentioned characteristics about the city, the main goal of this project is to explore the city and answer the following questions:

- What are the most common venues in the city?
- Which neighborhoods in the city are similar?

The answers to these questions could be exploited by:

- Tourists that would like to come and explore the city.
- People moving to the city for a job offer, or looking for a better neighborhood to live in.
- Small investors who would like to open a restaurant, a market etc.

2 Data Acquisition and wrangling:

The pre-step before starting to acquire data about the neighborhoods is to familiarize myself with the country rules and concepts, how they divide states and cities etc.

After that, in order to acquire a clean, well-structured data, the following steps were made:

2.1 Collect data about the city:

The main information about the city are: **Municipality, Neighborhood and Postal Code**. These three features are necessary to identify each neighborhood in the city and to acquire further information in the third step.

After a little research, I found all the mentioned data in this web page. The page has the name of each municipality in a header format and under each one land a table that contains neighborhoods' names and their corresponding postal code. Scrapping that page gives the results showed in Figure 1:

	Municipality_ar	Neighborhood_ar	Postal_code
0	بلدية العليا	حي العليا	12222
1	بلدية العليا	حي الازدهار	12486
2	بلدية العليا	حي المصيف	12466
3	بلدية العليا	حي التعاون	12475
4	بلدية العليا	حي الورود	12252

Figure 1. Web scrapping results for information about Riyadh city neighborhoods

2.2 Data wrangling:

2.2.1 Data transformation:

The above figure shows that municipalities and neighborhoods names are in Arabic. So, the first was translating these names to English.

To accomplish this task, I used *deep-translator* module, it's one of the modules that makes it straightforward to translate texts, and it provides support for multiple famous translators.

The next step was to remove 'municipality', 'neighborhood' and 'district' from the translated names.

2.2.2 Data cleaning:

The inspection of the data showed there is 203 neighborhoods in total and no nulls.

Checking for nulls gives:

- No duplicate values in general terms.
- Two neighborhoods appeared in two different municipalities.
- Some neighborhoods share the same postal code.

To deal with the duplicates in the data set I followed these instructions:

- In *Saudi Arabia* two municipalities could share the same neighborhood. So, we'll merge the municipalities names.
- Neighborhoods with the same postal codes and the same municipality will be merged together.
- If a postal code has different neighborhoods and different municipalities, we'll keep one item of that postal code with merged names.
- For the exception (postal code number 14522), one neighborhood had the correct name (merged names of two neighborhoods), we'll keep it and drop the second one.

The cleaning process left us with 185 entry in the data set.

2.3 Data preparation:

The goal of this project is to group neighborhoods based on their similarities. The main factor to accomplish that is the most common venues in each neighborhood and to explore venues we'll use the *Foursquare API*. This API takes the latitude and the longitude of each neighborhood and return a descriptive list about the venues in it.

To prepare our data set for further work, we need the geo-coordination of each neighborhood. The results in Figure 2 were acquired using *Geocoder API*, it's an API that finds you the geo-coordinates of a known address, place, locality or administrative area, even if the query is incomplete or partly incorrect.

	Municipality_ar	Neighborhood_ar	Postal_code	Municipality_en	Neighborhood_en	Latitude	Longitude
0	بلدية العليا	حي العليا	12222	Olaya	Al Alia	24.69502	46.69004
1	بلدية العليا	حي الازدهار	12486	Olaya	Al-Izdihar	24.77904	46.72446
2	بلدية العليا	حي المصيف	12466	Olaya	Al-Masif	24.76386	46.68849
3	بلدية العليا	حي التعاون	12475	Olaya	Al-Taawon	24.76724	46.69640
4	بلدية العليا	حي الورود	12252	Olaya	Al Worood	24.72213	46.68584

Figure 2. Data collection final results

All the entries have their corresponding latitude and longitude values.