## Coursera Applied Data Science Capstone

Ahmad Keewan
ahmedkeewan@gmail.com
https://github.com/ahmedkeewan

# The battle of neighborhoods

Week 5

**April 06, 2020**

## 1. Introduction

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.(wikipedia,Big_data).

**Geometry** and topology are very natural tools for analysing massive amounts of **data** since **geometry** can be regarded as the study of distance functions. Mathematical formalism, which has been developed for incorporating **geometric** and topological techniques, deals with point cloud **data** sets, i.e. finite sets of points.

Amman is one of the most crowded cities in the middle east, given its size and population, almost 4 million people live in the area of 1,680 km² with tons and tons of restaurants and venues, Amman is famous for its diversity, which makes opening a new small business kind of risky

1.1. Problem:
I need to compare different areas within the city of Amman,Jordan to recommend the hottest spots for an imaginary contactor who is planning to open a restaurant or any other business in a popular area.

1.2. Interest:
Anyone with small money capital planning to open a new business, would be very interested in checking the hottest or dense places before choosing where their business is going to take place, to avoid competition and possible loss.

## 2. Data

2.1. Required data:

The data that i need for the project is a geographical dataset that contains the different types of venues located in the city of Amman.

The dataset should consist of the Latitudes,Longitudes,Categories,Neighborhoods and the names of the venues that i will be exploring.

2.2. Data source:

To apply what I have learned during the last couple of weeks while preparing for the capstone, I will use Foursquare api to obtain my data, because it’s free and kind of provides me with what I need.

2.3. How will it solve the problem:

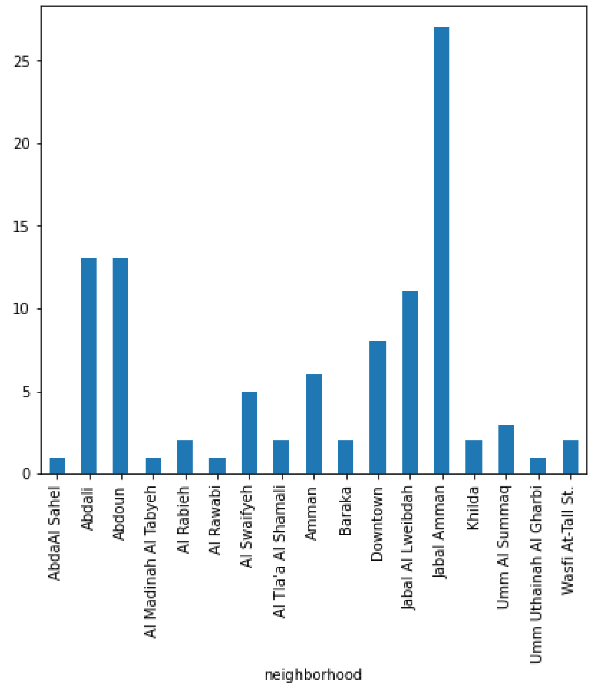
Using the features extracted from the dataset which are the ones mentioned in point 1, we can compute the frequency of the different types of venues in each neighborhood, and accordingly use these frequencies to cluster and recommend the right spots, and by that we solve the problem.

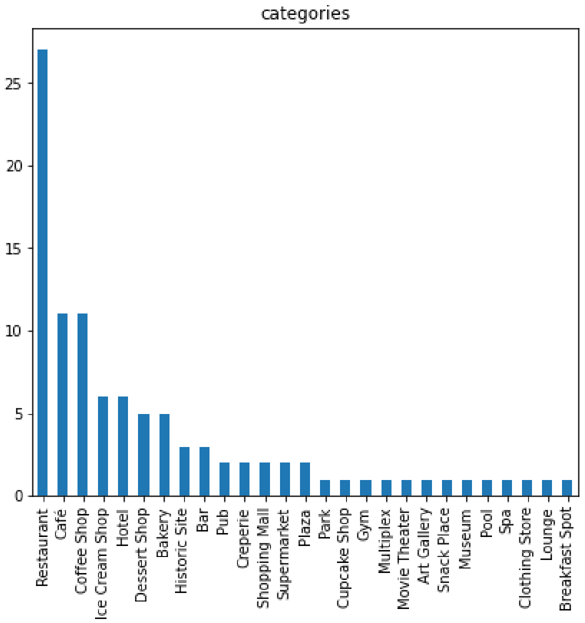
2.4. Data Cleaning:

The collected data from foursquare api was missing a lot of values and contained duplicate or miss-classified categories, so i started with normalizing the json file, then i created a data-frame consisting of the columns i need for my project, i then passed one of the columns to a function to extract the categories of each venue, after that i started merging the categories, where each category that consisted of the word Restaurant was converted to the word restaurant only regardless of its type, same thing was repeated for the neighborhoods where i fixed the ones that were consisting of missing values or miss-classified values, using google maps and the latitude,longitude of each venue.

## 3. Exploratory data analysis

In this part of the project I calculated the total number of venues in each neighborhood along with the number of venues in general in the second plot, as we can see the most dense area in Amman is Jabal Amman, followed by Abdoun, Abdali, and Jabal Al Weibdeh.





After that i calculated the frequency of each venue in each neighborhood using the one-hot encoding for each venue in each neighborhood, and using the list of frequencies i created another table consisting of the top 10 venues in each neighborhoods according to their frequencies, and this can be seen in the table shown down below

| | Cluster Labels | neighborhood | 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 | 0 | AbdaAl Sahel | Lounge | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop | Creperie |
| 1 | 1 | Abdali | Hotel | Restaurant | Bakery | Shopping Mall | Plaza | Coffee Shop | Movie Theater | Dessert Shop | Supermarket | Gym |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | Abdoun | Restaurant | Supermarket | Gym | Shopping Mall | Pub | Café | Multiplex | Coffee Shop | Bakery | Bar |
| 3 | 4 | Al Madinah Al Tabyeh | Museum | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop | Creperie |
| 4 | 1 | Al Rabieh | Ice Cream Shop | Café | Hotel | Bakery | Bar | Breakfast Spot | Clothing Store | Coffee Shop | Creperie | Cupcake Shop |
| 5 | 2 | Al Rawabi | Creperie | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop | Cupcake Shop |
| 6 | 1 | Al Swaifyeh | Bakery | Ice Cream Shop | Bar | Coffee Shop | Hotel | Breakfast Spot | Café | Clothing Store | Creperie | Cupcake Shop |
| 7 | 5 | Al Tla'a Al Shamali | Restaurant | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop | Creperie |
| 8 | 1 | Amman | Coffee Shop | Café | Pool | Cupcake Shop | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Clothing Store |
| 9 | 1 | Baraka | Restaurant | Café | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Clothing Store | Coffee Shop | Creperie |
| 10 | 1 | Downtown | Historic Site | Café | Dessert Shop | Restaurant | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Clothing Store |
| 11 | 1 | Jabal Al Lweibdah | Café | Restaurant | Art Gallery | Snack Place | Plaza | Park | Clothing Store | Coffee Shop | Gym | Bakery |
| 12 | 1 | Jabal Amman | Restaurant | Ice Cream Shop | Coffee Shop | Bakery | Bar | Hotel | Pub | Breakfast Spot | Café | Dessert Shop |
| 13 | 2 | Khilda | Creperie | Dessert Shop | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop |
| 14 | 1 | Umm Al Summaq | Ice Cream Shop | Café | Coffee Shop | Hotel | Bakery | Bar | Breakfast Spot | Clothing Store | Creperie | Cupcake Shop |
| 15 | 3 | Umm Uthainah Al Gharbi | Spa | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop | Creperie |
| 16 | 5 | Wasfi At-Tall St. | Restaurant | Supermarket | Hotel | Bakery | Bar | Breakfast Spot | Café | Clothing Store | Coffee Shop | Creperie |

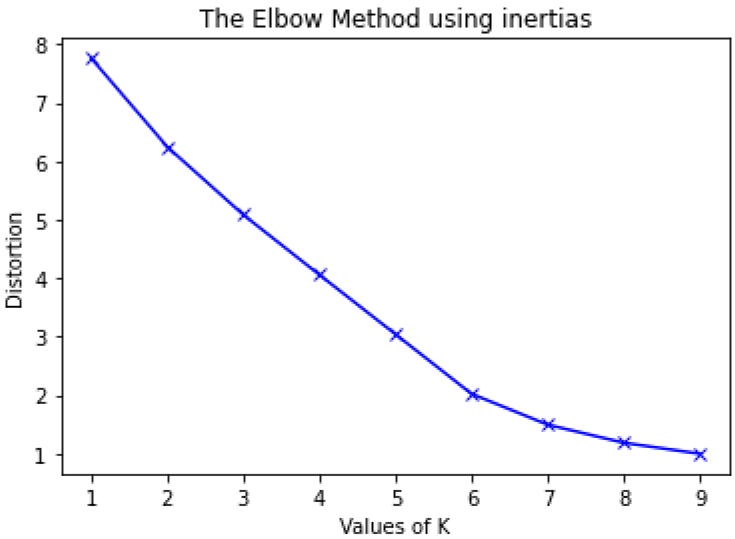
## 4. Modeling

In the modeling part of the project i used 2 different methods to model my dataset, Starting from normal clustering, to plotting the results in 4 different maps.

### 4.1. Clustering:

In the clustering part, i used the famous partitional clustering algorithm (kmeans) to generate cluster for the areas of Amman using the dataset that i prepared in the previous sections

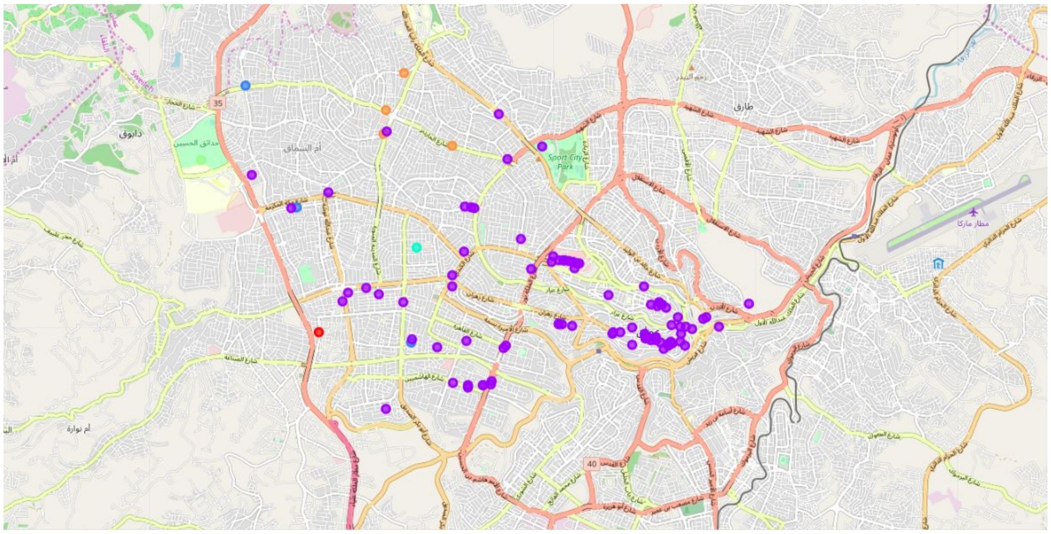
To choose the right number of clusters, i used the Elbow-method by plotting the inertias extracted from the kmeans algorithm and matplotlib plot function



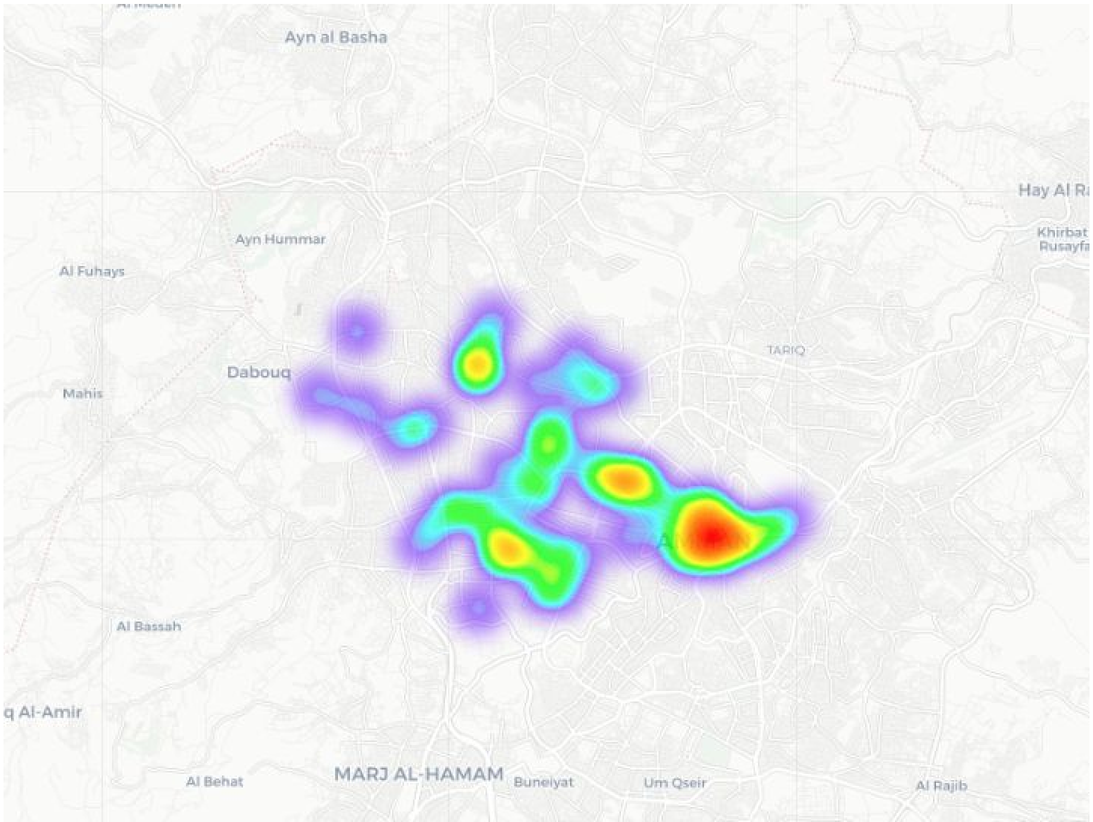
As we can see in the plot the perfect K value is either 6 or 7, and after testing both values, i clustered the final results using 6 as the value of K

After clustering, I added the cluster labels to the dataset, then I merged the first dataset with the dataset that consisted of the most common values.

In the next phase of clustering, I created a map using Folium with all of the venues colored according to their clusters label and marked as circles, the map below shows the clustered venues:

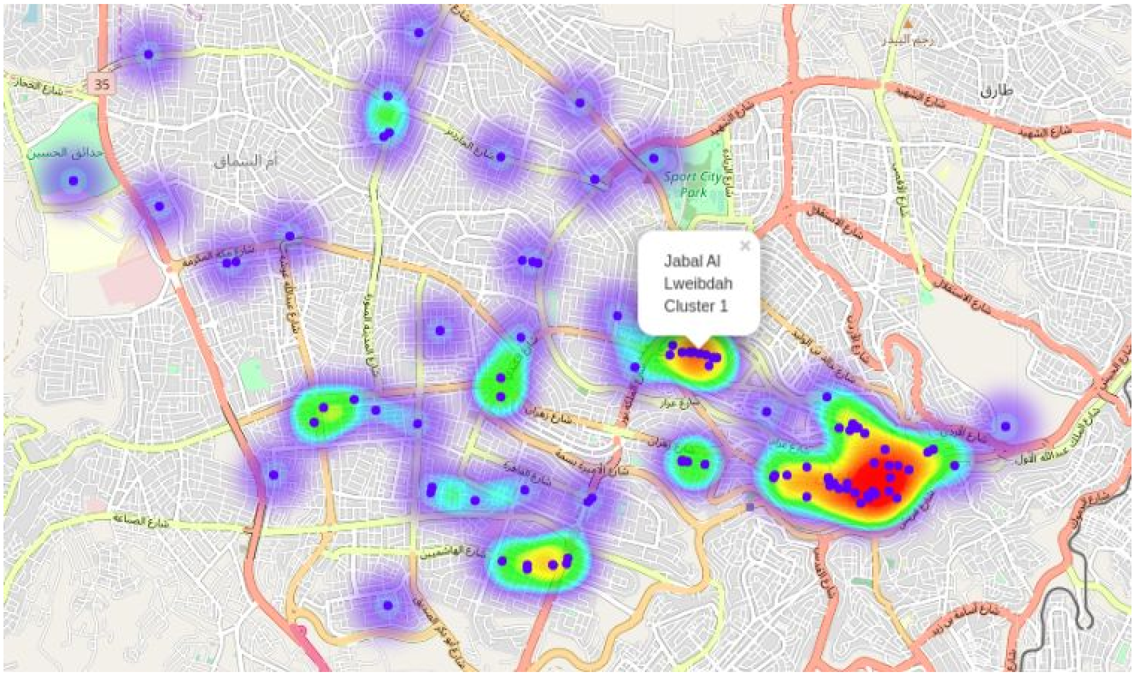


In the third phase, I used the same clustered dataset, to create another map to create a heat map without the venues, to see confirm that our clustering was able to collect the density of the area of Jabal Amman, as can shown in the following map:



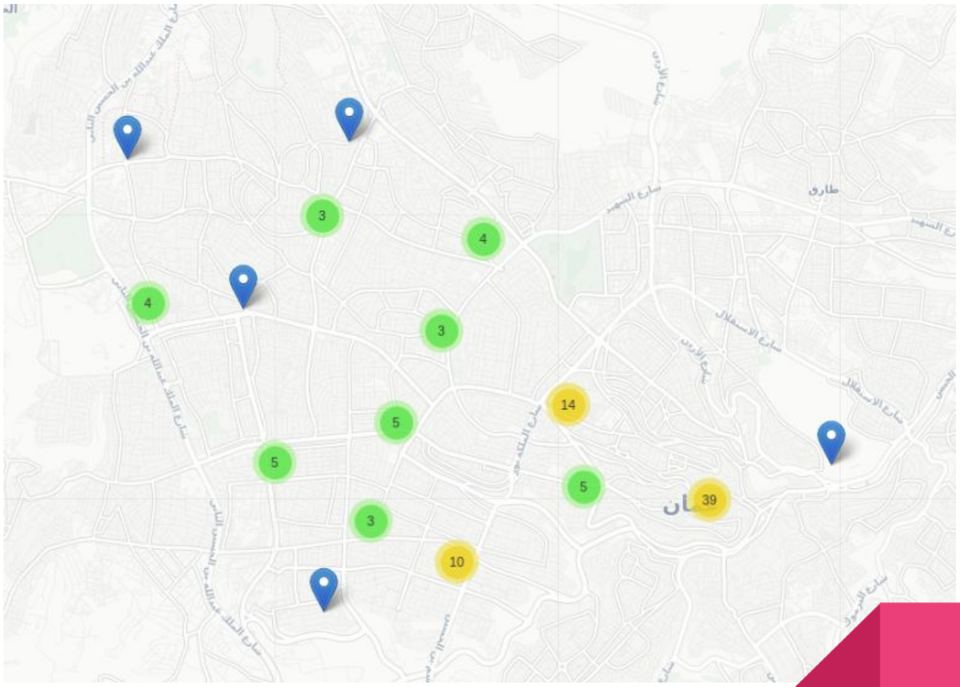
As we can see in the map above, our clustering was able to detect the density of Jabal Amman correctly.

In the final phase, i merged the two maps to see the point within the heatmap, which can be seen in the map down below:



4.2. Folium marker cluster

In this method i used the marker cluster method provided by folium to cluster the original dataset i cleaned first, to check whether i will get the same results or not:



## 5. Results and discussion

Our analysis shows that although Amman is a small city, but it has a large number of (Restaurants,Cafes, and Coffee houses), there are pockets of low restaurant density fairly close to(Umm Al Summaq, Al-Rawabi,Al-Swaifyeh,and Al Madinah Al Tabyeh). Highest concentration of restaurants was detected in Jabal Amman, so we wouldn't recommend opening a new restaurant in this area because it's dense and highly competitive.

(khilda, Umm Al Summaq, and Umm Uthainah Al Gharbi) are the areas which offer a combination of popularity among locals, closeness to west amman, strong socio-economic dynamics *and* a number of pockets of low restaurant density, so our recommendation for this area is to open a Restaurant not another type of venue.

After directing our attention to this more narrow area of interest (Amman) I first created a one-hot matrix of location candidates along with the types of venues in each neighborhood; those locations were then filtered so that I can get a better idea about the popular type of venues in each neighborhood.

Those location candidates were then clustered to create zones of heats or trends according to the frequency of each activity in each neighborhood,after that I created 4 maps,2 of which are heat maps, one is a clustering map, and one with the colored circles according to the number of clusters each point follows .

Result of all this is the area of Jabal Amman is the most dense/crowded area in Amman, and according to the analysis; Restaurants are the highest frequency venues in the whole city, so as motivating this may look, it's actually not, because opening a new Restaurant the provides the same type of dishes as the ones near it in a dense area would result in a total loss, so whenever someone needs to have an idea of the density and types of venues in the most popular neighborhoods in Amman, they can refer to this study.

6. Conclusion

Purpose of this project was to identify Amman areas that are popular with a low number of restaurants, in order to aid stakeholders in narrowing down the search for the optimal location for a restaurant. By calculating venue frequency distribution from Foursquare data, I have first collected the venues and neighborhoods that strictly fall in Amman, and then manually fixed the neighborhoods that were filled with null values, using google maps and the lat,long of each venue.

After preparing the data, I created a one-hot matrix for each neighborhood and the venues that are located in it, using the one-hot matrix, I calculated the frequency of each venue and used to the top 5 in each neighborhood, going from that, I then sorted the venues according to their common in a new matrix.

Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zones were used for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like frequency of each venue in each location and its density.