# Clustering model for Saudi Arabia cities according to its population, covid 19 cases and total of hospitals

Abdulwahed Salam

December 21, 2020

## 1. Introduction

### 1.1 Background

Saudi Arabic is a country located in middle east. it consists of many states and each state consist of many cities. KSA play a major role to control the covid 19 epidemic. Nowadays, the total number of cases around 250 up to 350. Now it is time of getting the vaccine and distributed to all people in different locations. As we know the vaccine is limited and to distributed it fairly is important. we have many cities with different features. Based on that we have to build a model to cluster the cities. Also, the vaccine needs to be store in specific conditions. one of them is cold and must to be minus 85. KSA like many countries, there are small cities or region does not have hospital and we want to consider this issue.

### 1.2 Problem

suppose we have just 1,00,0000 of the vaccine and we want to distribute it to all cities. we know that there are priorities to give it to the most effected people like old people and People with chronic diseases. but before considering this, we have to think about grouping the cities and states according to the population, historic information of covid 19 cases and the number of public hospitals to store the vaccine. we give a solution to group those cities according to their properties. It will facility the distribution of the vacancies. we have more than 1000 hospital, 200 cities and more than 30,000,000. The clustering model will help us to group the cities according to that number of cases, population and number of hospitals in easy and correct way.

### 1.3 Problem questions

- which group consist of big number of Cases and big number of populations to give them the biggest number of vacancies.

- which group consist big number population and small number Cases to give them less than the previous group.
- which group of cities does not have any cases as not our priority group?
- which group does not have hospital to store the vacancies and how we can give the vacancies? what is the solution?

### 1.4 Target audience

Target audiences for this project are:

- The government
- Ministry of Health
- Local governments
- Vaccine providers
- Interested people

## 2. Data acquisition and cleaning

### 2.1 Data sources

To build a clustering model to solve the problem of distribution of the vaccine, we need data and lots of data from different sources. Data can answer question which are unimaginable and difficult to be answered by humans. it is a tendency to analyses such large dataset and produce analytics to find a solution. This model based on population, covid cases and number of hospitals. In this work we use four different resource as follow:

- Population dataset gathering from Ministry of housing. This data was about the number of populations categories by state and cities but it lacks the location of the cities which we collect using Google Api by implementing the geopy library. The reader can find the data in this link.
- Hospitals dataset from Wiki, which is not enough, and we add latitude and longitude using geopy library. This is the link for data.
- Daily covid cases to know the situation in Saudi Arabia gather from Ministry of Health. The data is in this link
- Covid cases according to the cities gathering from Ministry of Health in this link. Also, we add latitude and longitude using geopy library.
- Cities hospitals gathering by using Foursquare API. The data from Wiki is not enough so we collect the data of hospital by using Foursquare API.

## 2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined and analyzed. There were a lot of missing values. We fill the needs values by scraping the data using geopy library special those depend on the latitude and longitude. The following tables shows the data sources after they became ready for works.

*Table 1: Daily cases of covid 19 in Saudi Arabia*

| | Day | Date | Case | Cure | Death |
|---|---|---|---|---|---|
| 0 | الخميس | 20-12-17 | 174 | 208 | 10 |
| 1 | الأربعاء | 20-12-16 | 181 | 173 | 11 |
| 2 | الثلاثاء | 20-12-15 | 180 | 199 | 11 |
| 3 | الاثنين | 20-12-14 | 142 | 201 | 10 |
| 4 | الأحد | 20-12-13 | 125 | 243 | 11 |

*Table 2: Covid 19 Cases according to KSA Cities*

| | City | State | population | lat | long |
|---|---|---|---|---|---|
| 0 | الرياض | الرياض | 5188286 | 24.713552 | 46.675296 |
| 1 | جدة | مكة المكرمة | 3430697 | 21.485811 | 39.192505 |
| 2 | مكه المكرمه | مكة المكرمة | 1534731 | 21.389082 | 39.857912 |
| 3 | المدينه المنوره | المدينة المنورة | 1100093 | 24.524654 | 39.569184 |
| 4 | الدمام | المنطقة الشرقية | 903312 | 26.420683 | 50.088794 |

*Table 3: KSA Cities Hospital according to Wiki Data*

| | Hospital | City | State | lat | long |
|---|---|---|---|---|---|
| 0 | مستشفى الملك خالد | نجران | منطقة نجران | 17.545267 | 44.233758 |
| 1 | مستشفى نجران العام | نجران | منطقة نجران | 17.495299 | 44.144086 |
| 2 | مستشفى الولاده والأطفال | نجران | منطقة نجران | 17.553579 | 44.272491 |
| 3 | مستشفى شروره | شروره | منطقة نجران | 17.484406 | 47.101262 |
| 4 | مستشفى حبونا العام | حبونا | منطقة نجران | 17.847307 | 44.019830 |

*Table 4: Hospital information according to Foursquare API*

| | categories | hasPerk | id | address | cc | city | country | crossStreet | distance | formattedAddress | labeledLatLngs | lat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Hospital | False | 4f5e9dd1e4b033732b5efaf6 | Ar-Rahmanyah | SA | الرياض | المملكة العربية السعودية | Takhassusi Street | 1961 | [Ar-Rahmanyah (Takhassusi Street), الرياض 1234... | [{'label': 'display', 'lat': 24.71924218810133... | 24.719242 |
| 1 | Hospital | False | 4bdf3036e75c0f47f033ca03 | Takhassusi St. | SA | الرياض | المملكة العربية السعودية | Khurais Rd. | 4661 | [Takhassusi St. (Khurais Rd.), الرياض 11211, ا... | NaN | 24.671692 |
| 2 | Hospital | False | 4df9d62ad22d964c6b9b133a | King Abdullah Rd. | SA | الرياض | المملكة العربية السعودية | King Khalid Rd. | 5422 | [King Abdullah Rd. (King Khalid Rd.), الرياض 1... | NaN | 24.712649 |
| 3 | Hospital | False | 4b9ba5c4f964a5207f1636e3 | King Fahd Rd. | SA | الرياض | المملكة العربية السعودية | NaN | 4383 | [King Fahd Rd., الرياض 12381, المملكة العربية ... | [{'label': 'display', 'lat': 24.74690924692862... | 24.746909 |
| 4 | Hospital | False | 4eecae9a29c28028ddd9bdcb | Khurais Road | SA | الرياض | المملكة العربية السعودية | NaN | 10089 | [Khurais Road, الرياض 11635, المملكة العربية ا... | [{'label': 'display', 'lat': 24.72124816759136... | 24.721248 |

*Table 5:Data set for the Clustering model after combining all datasets and filtering and filling null values*

| | City | State | population | lat_x | long_x | disp_x | Case | disp_y | lat_y | long_y | Hospital |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | الرياض | الرياض | 5188286.0 | 24.713552 | 46.675296 | ض.ايرلا | 59805.0 | ض.ايرلا | 24.713552 | 46.675296 | 10.0 |
| 1 | جدة | مكة المكرمة | 3430697.0 | 21.485811 | 39.192505 | ةدج | 34600.0 | ةدج | 21.485811 | 39.192505 | 9.0 |
| 2 | مكة المكرمة | مكة المكرمة | 1534731.0 | 21.389082 | 39.857912 | ةمركملا ةكم | 34839.0 | ةمركملا ةكم | 21.389082 | 39.857912 | 6.0 |
| 3 | المدينة المنورة | المدينة المنورة | 1100093.0 | 24.524654 | 39.569184 | ةرونملا ةنيدملا | 22563.0 | ةرونملا ةنيدملا | 24.524654 | 39.569184 | 0.0 |
| 4 | الدمام | المنطقة الشرقية | 903312.0 | 26.420683 | 50.088794 | ماملدا | 19851.0 | ماملدا | 26.420683 | 50.088794 | 2.0 |

## 3. Exploratory Data Analysis

Saudi Arabia is one of the best countries control the virus. It made and early curfew. Nowadays the number of cases is decreased and become less and less. Figure below shows the cases and cure and death numbers of Covid 19 since the beginning of the epidemic.

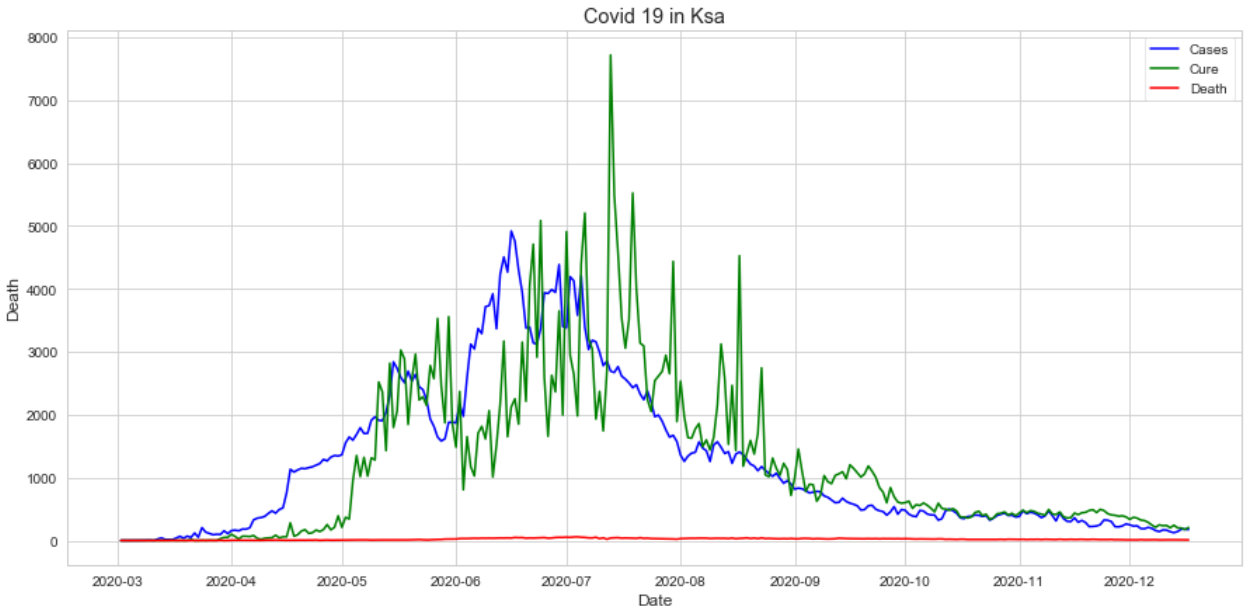


*Figure 1: Cases of the Covid 19 with Cure and death cases*

Also the table below show the total numbers of the Covid in KSA

| Case | Cure | Death |
|---|---|---|
| 360690 | 351573 | 6101 |

As we explain at the begging that we are in the phase of giving vacancy and it important to know the public hospital which will be responsible for distributing the vacancies to the people.

It is important to keep it in degree below 85 to be save, so the this study suggest that we have to consider the cities which have public hospital or the government find away to give it to private hospitals in cities which do not have. Figure below show the number of hospitals according to the cities

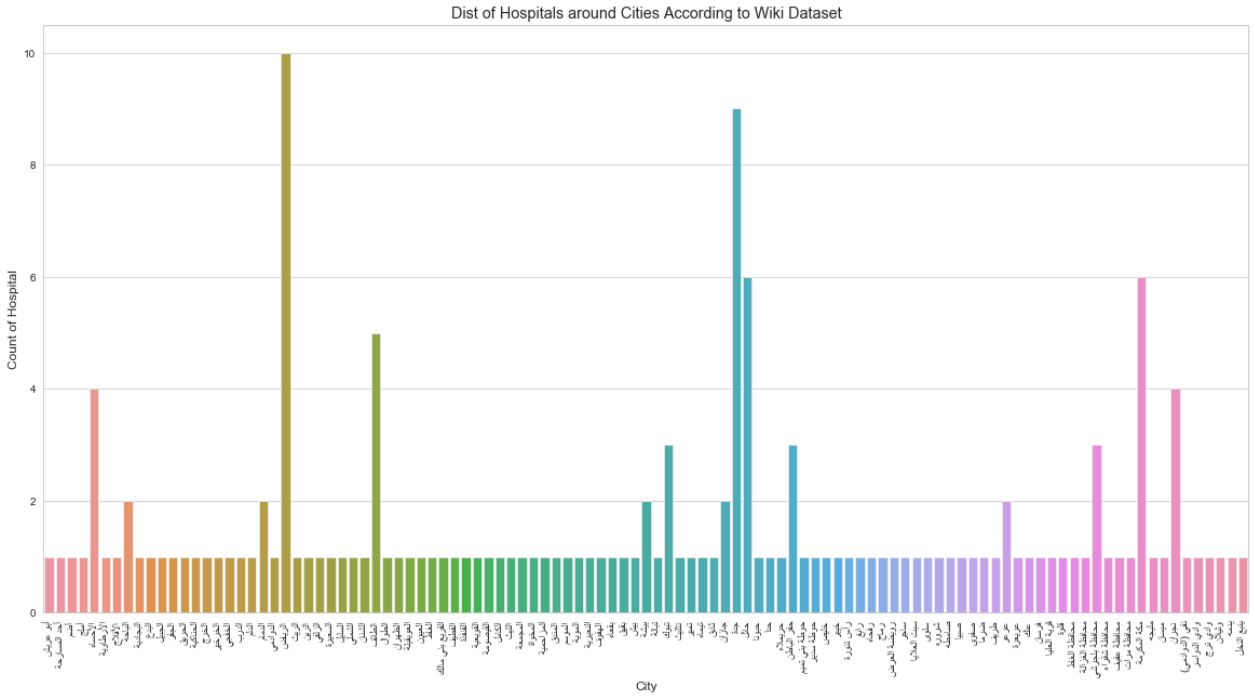


*Figure 3:Number of hospitals according to the cites.*

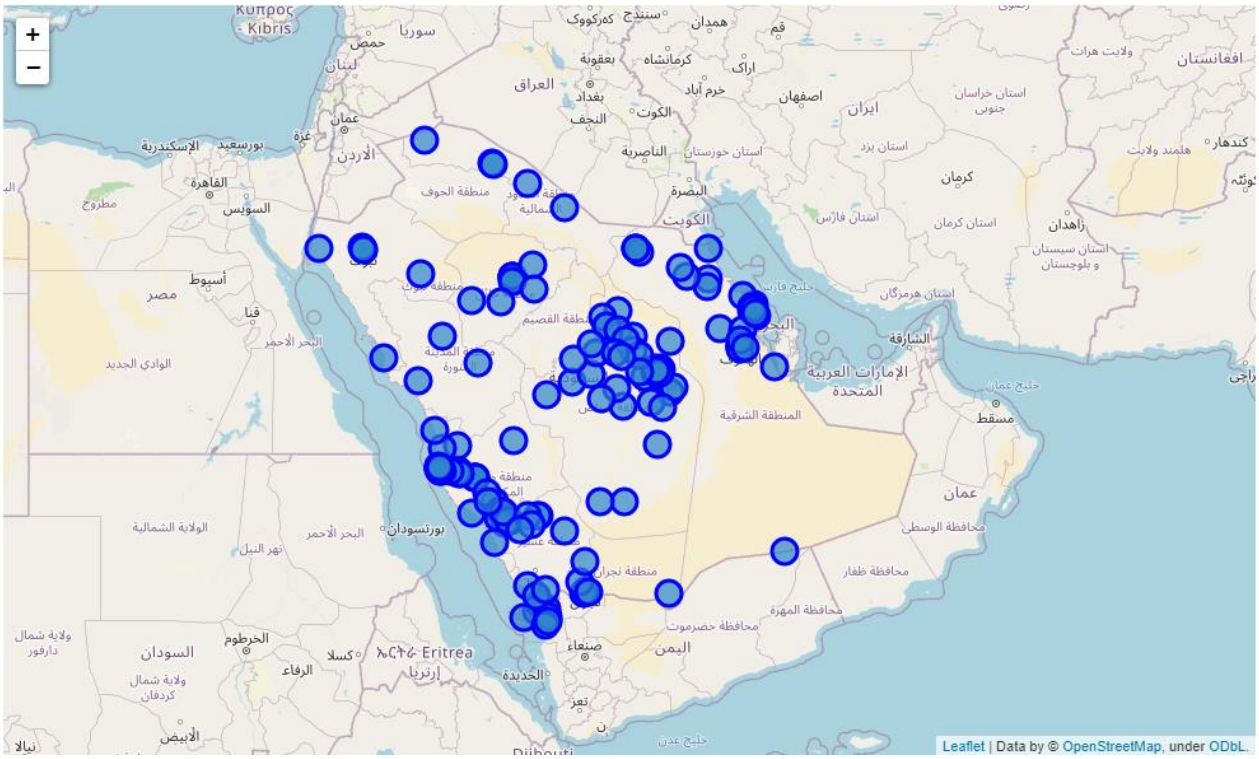


*Figure 2: Locations of the public hospitals*

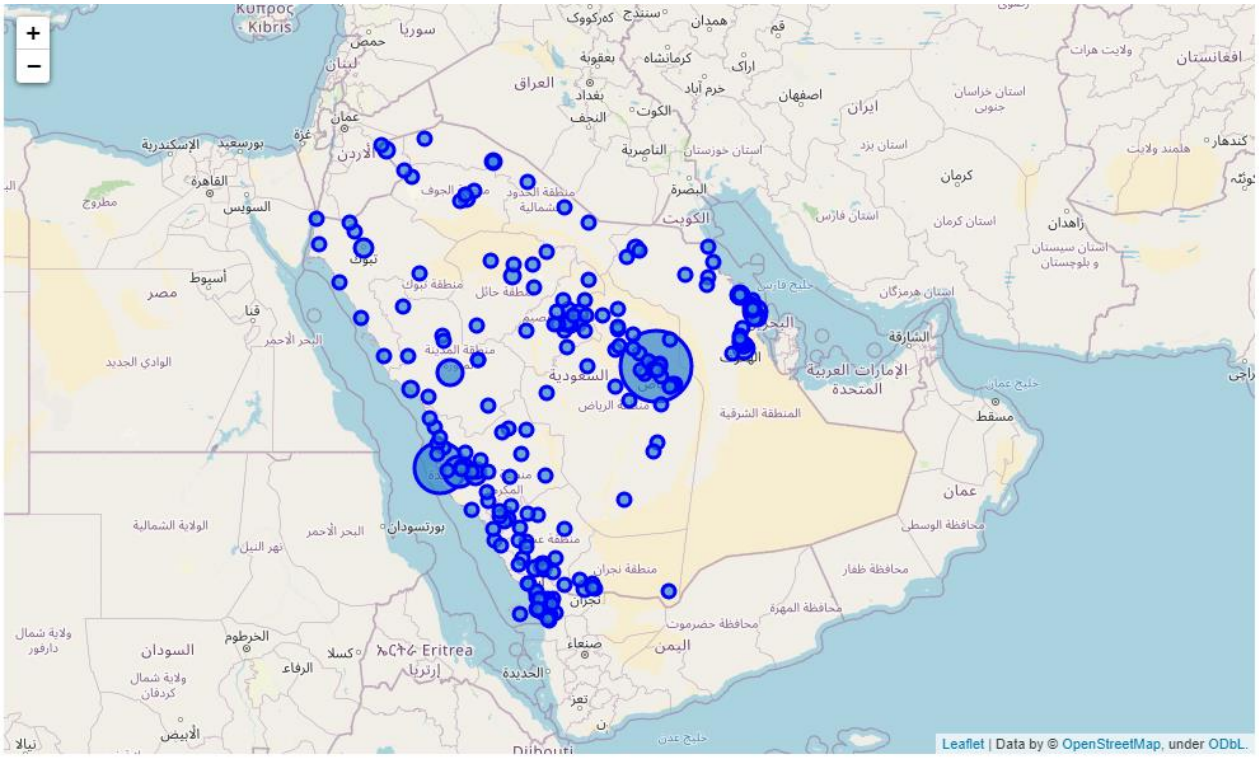


*Figure 4: Distribution of populatin according to the KSA Cities*

To know the number of the Covid 19 cases comparing to the states and the cities to have a general idea about relationship the figures below show the information and relation.

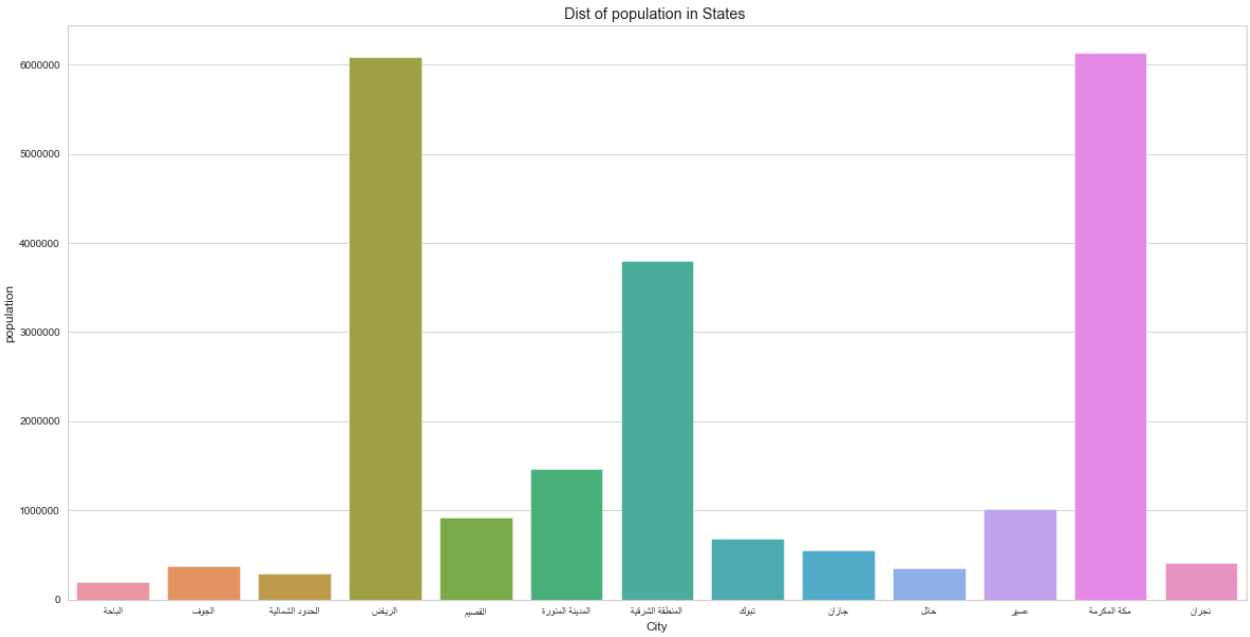


***Figure 5:Distribution of population according to the KSA states***

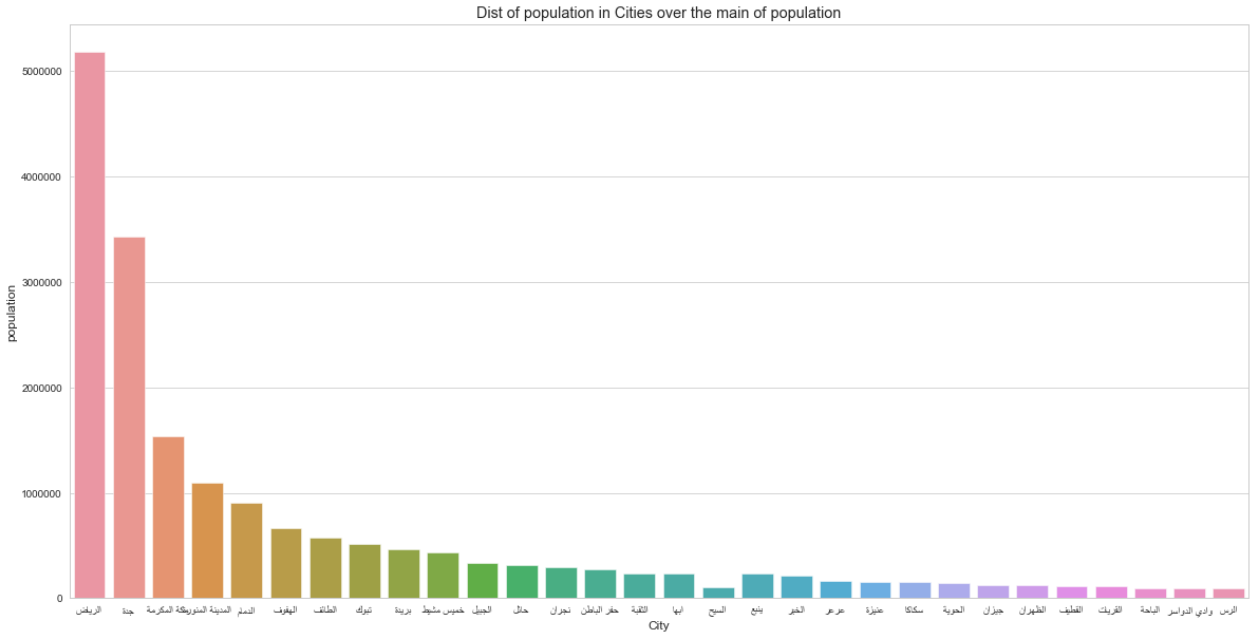


*Figure 6:Distribution of the population according to the Cities*

Because the number of cities is more than 200, so in the figure above we just show the population of the cities greater than mean of all city's population.

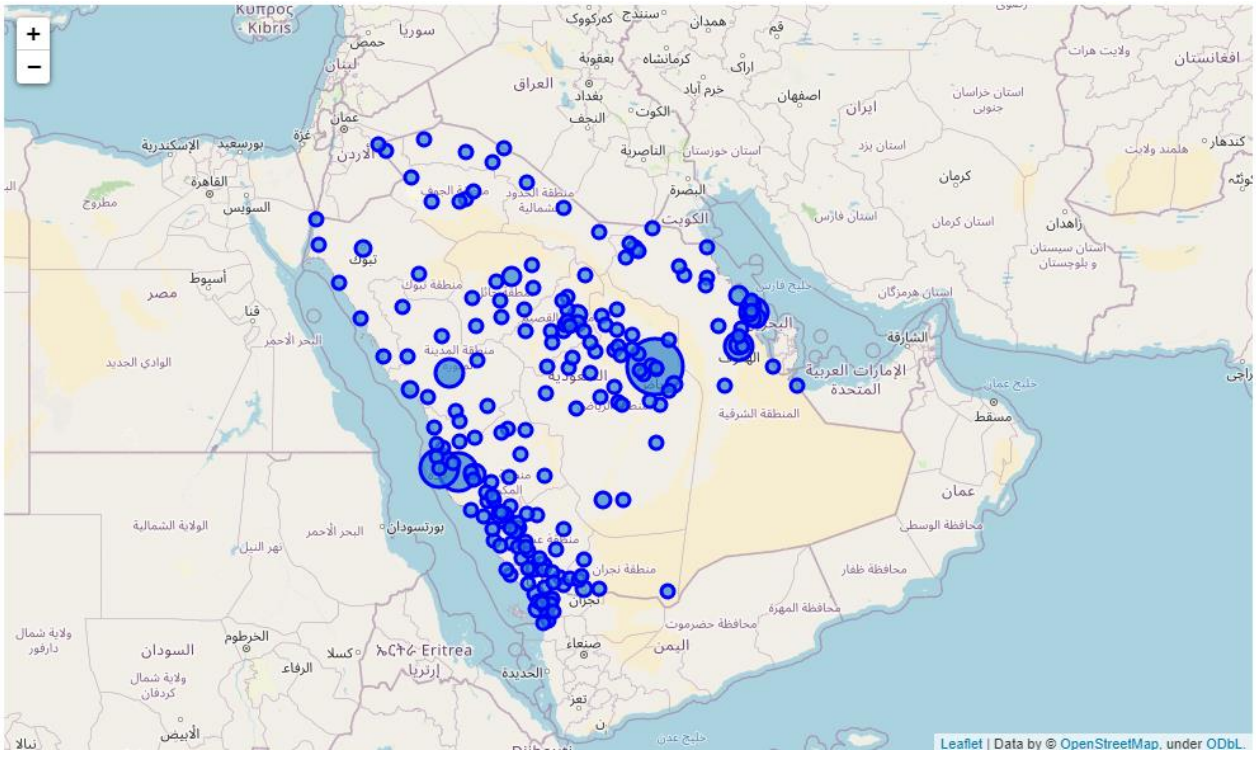


*Figure 7: The Distribution of Covid 19 cases according to the KSA cities*

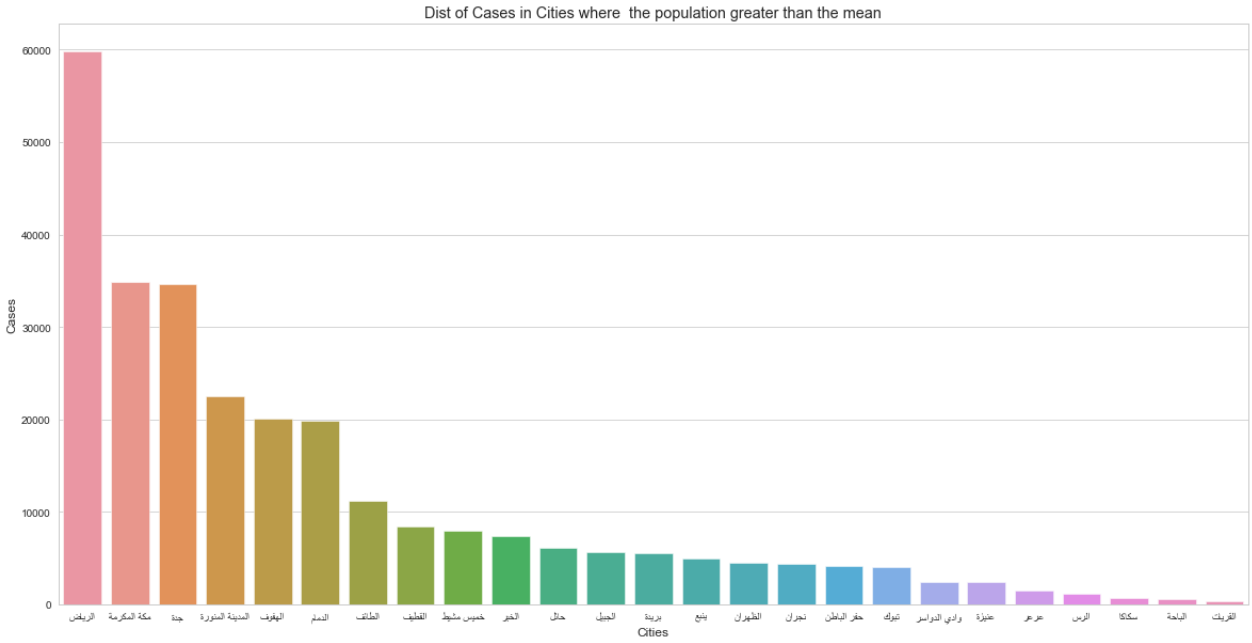


*Figure 9:The histogram of Covid 19 cases according to the KSA cities which their population greater than the mean of the populations*

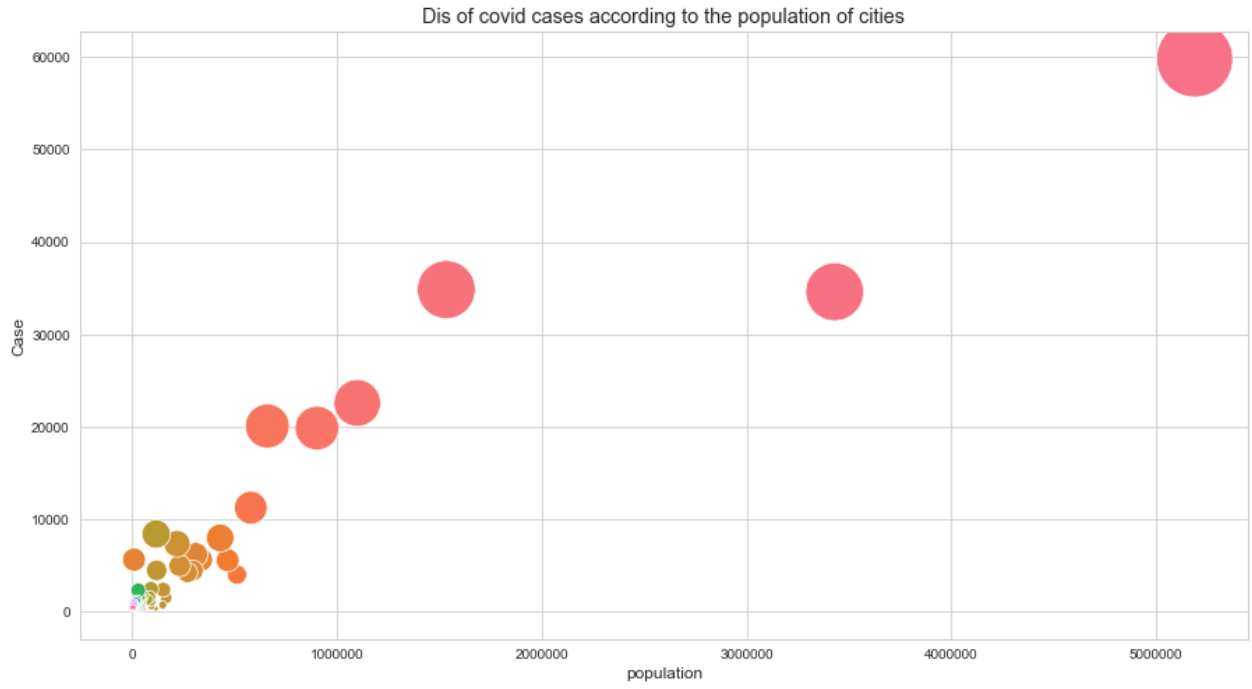


*Figure 8: Relationship between the population and the number of Covid 19 Cases*


*Figure 10: Relationship between number of hospitals and the population*

## 4. Clustering Modeling

After gathering the data, cleaning and filling the missed data. Now it time to build a clustering model which group the data according to their properties. We use Kmean cluster method to do that. It depends of the three different inputs Covid 19 Cases, number of hospitals and population.

One main thing we have to consider in building KMean is choosing the suitable number of k. we test the algorithm in range of 10 k. Figure below show the experiments of choosing k.

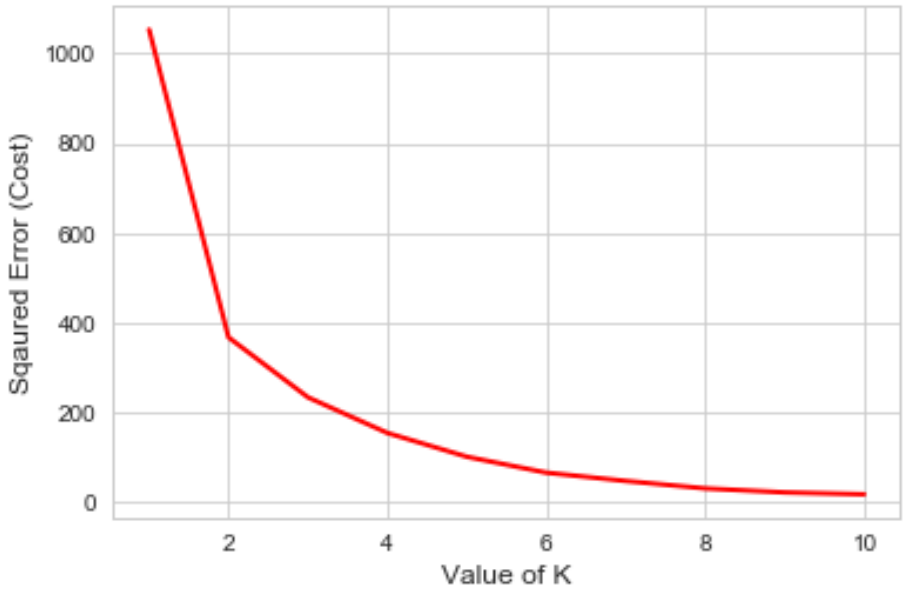


*Figure 11:experment of choosing the k number of kmean model*

We set the parameters of the Kmean model as follow:

- Init:k-means++
- n_clusters:4
- n_init:12

Table below show the properties(mean) of each cluster of the model

*Table 6: the mean of input parameters according to each cluster*

| cluster | population | Case | Hospital |
|---|---|---|---|
| 0 | 2.279911e+04 | 328.020661 | 0.000000 |
| 1 | 4.309492e+06 | 47202.500000 | 9.500000 |
| 2 | 9.557788e+05 | 21709.400000 | 2.800000 |
| 3 | 3.313381e+04 | 819.950980 | 1.205882 |

*Table 7:number of clusters*

| Cluster | Number |
|---|---|
| 0 | 242 |
| 1 | 2 |
| 2 | 5 |
| 3 | 102 |

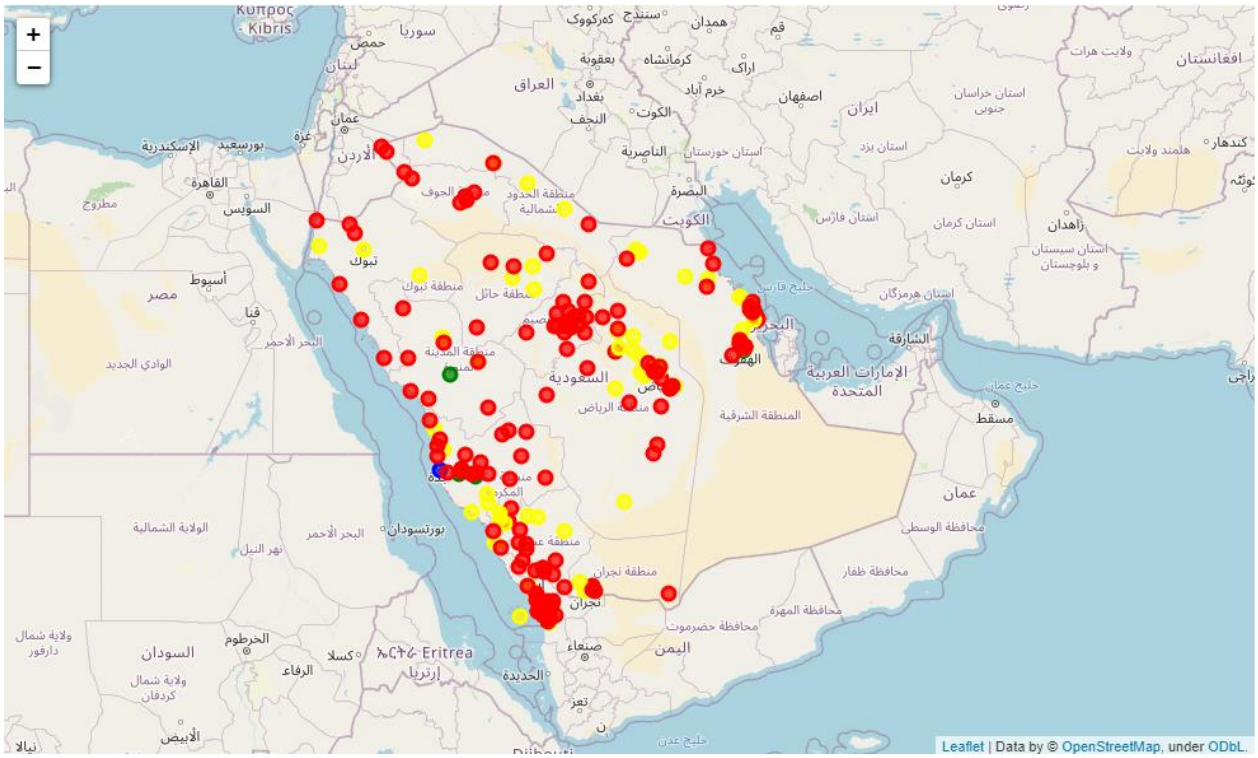


*Figure 12: Distribution of Clusters*

*Table 8: Cluster 1 information*

| | City | State | population | lat_x | long_x | disp_x | Case | disp_y | lat_y | long_y | Hospital | cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | الرياض | الرياض | 5188286.0 | 24.713552 | 46.675296 | ض‌ير‌ا | 59805.0 | ض‌ير‌ا | 24.713552 | 46.675296 | 10.0 | 1 |
| 1 | جده | مكة المكرمة | 3430697.0 | 21.485811 | 39.192505 | ةدج | 34600.0 | ةدج | 21.485811 | 39.192505 | 9.0 | 1 |

*Table 9:Cluster 2 information*

| | City | State | population | Case |
|---|---|---|---|---|
| 2 | مكة المكرمة | مكة المكرمة | 1534731.0 | 34839.0 |
| 3 | المدينة المنوره | المدينة المنوره | 1100093.0 | 22563.0 |
| 4 | الدمام | المنطقة الشرقية | 903312.0 | 19851.0 |
| 5 | الهفوف | المنطقة الشرقية | 660788.0 | 20066.0 |
| 6 | الطائف | مكة المكرمة | 579970.0 | 11228.0 |

*Table 10: Top 20 of cluster 3*

| | City | State | population | Case |
|---|---|---|---|---|
| 7 | تبوك | تبوك | 512629.0 | 3978.0 |
| 10 | الجبيل | المنطقه الشرقيه | 337778.0 | 5600.0 |
| 11 | الجبيل | المنطقه الشرقيه | 10241.0 | 5600.0 |
| 12 | حائل | حائل | 310897.0 | 6141.0 |
| 13 | نجران | نجران | 298288.0 | 4423.0 |
| 14 | حفر الباطن | المنطقه الشرقيه | 271642.0 | 4192.0 |
| 20 | عرعر | الحدود الشماليه | 167057.0 | 1474.0 |
| 25 | الظهران | المنطقه الشرقيه | 120521.0 | 4447.0 |
| 26 | القطيف | المنطقه الشرقيه | 118327.0 | 8377.0 |
| 28 | الباحه | الباحه | 95089.0 | 558.0 |
| 29 | وادي الدواسر | الرياض | 93036.0 | 2429.0 |
| 31 | بيشه | عسير | 86201.0 | 1571.0 |
| 43 | رابغ | مكه المكرمه | 55304.0 | 274.0 |
| 45 | رفحاء | الحدود الشماليه | 52712.0 | 0.0 |
| 46 | صفوى | المنطقه الشرقيه | 50447.0 | 1683.0 |
| 48 | طريف | الحدود الشماليه | 48108.0 | 209.0 |
| 49 | المجمعه | الرياض | 47743.0 | 1013.0 |
| 54 | الدلم | الرياض | 40114.0 | 428.0 |
| 56 | بقيق | المنطقه الشرقيه | 36207.0 | 1487.0 |
| 57 | العيون | المنطقه الشرقيه | 33042.0 | 679.0 |

*Table 11: Top 20 of cluster 0*

| | City | State | population | Case |
|---|---|---|---|---|
| 8 | بريده | القصيم | 467410.0 | 5532.0 |
| 9 | خميس مشيط | عسير | 430828.0 | 7956.0 |
| 15 | الثقبه | المنطقه الشرقيه | 238066.0 | 0.0 |
| 16 | ابها | عسير | 236157.0 | 0.0 |
| 17 | السيح | الرياض | 103216.0 | 0.0 |
| 18 | ينبع | المدينه المنوره | 233236.0 | 4948.0 |
| 19 | الخبر | المنطقه الشرقيه | 219679.0 | 7325.0 |
| 21 | عنيزه | القصيم | 152895.0 | 2357.0 |
| 22 | سكاكا | الجوف | 150257.0 | 699.0 |
| 23 | الحويه | مكه المكرمه | 148151.0 | 0.0 |
| 24 | جيزان | جازان | 127743.0 | 0.0 |
| 27 | القريات | الجوف | 116162.0 | 304.0 |
| 30 | الرس | القصيم | 92501.0 | 1098.0 |
| 32 | داروت | المنطقه الشرقيه | 77757.0 | 0.0 |
| 33 | سيهات | المنطقه الشرقيه | 75794.0 | 14.0 |
| 34 | شروره | نجران | 75237.0 | 1286.0 |
| 35 | بحره | مكه المكرمه | 75213.0 | 0.0 |
| 36 | الخفجي | المنطقه الشرقيه | 67012.0 | 0.0 |
| 37 | صبياء | جازان | 63143.0 | 0.0 |
| 38 | الدوادمي | الرياض | 61834.0 | 0.0 |

## 5. Conclusions

In this work, I analyzed the Covid 19 of Saudi Arabia and the situation. Also, I analyze and show how Saudi Arabia control the epidemic. In this study we collect the data from many resources like ministries, Wiki, foursquare api and, etc. we combine the data and manipulate it and fill it with required data using different python libraries like geopy, foursquare and Beautiful Soup libraries. Finally, we build a cluster model using Kmean algorithm and set its parameters in correct way. The input data of the model were population, Covid 19 cases and number of hospitals. The mean reason of this study is to help Stakeholders to distribute the vacancies to the people in fairly way.

## 6. Future work

I plan to build a model of classification algorithm to help the authorities to give priorities to people according to the people in need more than authors according to the study of the affection of people.