**COURSERA CAPSTONE**

**IBM Applied Data Science Capstone**

***Opening a new Hospital in Delhi, India***

**PART - 2**



**By: Tanvi Jain**

**February 2020**

**Table of contents**

1. [Introduction: Business Problem](https://e-84fd004678.cognitiveclass.ai/lab#introduction)
2. [Data Description](https://e-84fd004678.cognitiveclass.ai/lab#data)
3. [Methodology](https://e-84fd004678.cognitiveclass.ai/lab#methodology)
4. [Analysis](https://e-84fd004678.cognitiveclass.ai/lab#analysis)
5. [Results](https://e-84fd004678.cognitiveclass.ai/lab#results)
6. Discussion and [Conclusion](https://e-84fd004678.cognitiveclass.ai/lab#conclusion)
7. **INTRODUCTION**

The detail analysis of Population Census 2011 published by Govt. of India for Delhi state reveal that population of Delhi has increased by 21.21% in this decade compared (2001-2011) to past decade (1991-2001). The density of Delhi state in the current decade is 29298 per sq mile.

* Delhi is a State of India with population of Approximate 1.68 Crores.
* The population of Delhi state is 16,787,941.
* The density of Delhi state is 11,320 per sq km.
* Delhi State is spread over 1,483 Sq Km.

Health sector infrastructure in Delhi comprises of 1,298 dispensaries, 1160 nursing homes, 230 maternity homes,178 polyclinics/special clinics, 88 hospitals and 17 medical colleges. Important vital indicators like Infant Mortality Rate, Neo-Natal Mortality Rate, Under Five Mortality Rate in respect of Delhi stand at lower levels i.e 18, 12 and 22 respectively in comparison to all-India levels at 34, 24 and 39 respectively. The total fertility rate of Delhi is 1.6 which is lowering among all states in India. Per capita expenditure on health in Delhi has increased from Rs 2116 in the year 2014-15 to Rs 2493 in 2017-18. The total expenditure in medical and public health sector of Delhi government has significantly increased from Rs 861.66 crore which accounts to 9.85 per cent of the total expenditure in 2007-08 to Rs 1912.42 crore which accounts to 13.28 per cent of the total outlay in 2017-18. To provide primary health care services at the doorstep of the citizens of Delhi, 189 Aam Aadmi Mohalla clinics have already been set up, it added.

The Delhi government has decided to remodel around 16 existing hospitals so as to enhance the number of existing beds as per Floor Area Ratio (FAR) norms. Around 7,000 new beds will be added as per planned remodelling of these existing 16 hospitals, it said.

**Business Problem**

The objective of this project is to take part of this remodelling process therefore, the proposal to make a new hospital having 500 beds. For this, the need is to analyse and select the best location in Delhi to open a new hospital. Using Data Science methodology and machine learning techniques as clustering, regression etc., this project aims to answer the business problem: In Delhi, India if government wants to open a hospital, where would the data science team will recommend building it?

**Target Audience of the Project**

This hospital will help to overcome the shortage of beds per 1000 population. Currently, the total number of hospital beds in the national capital increased from 48,096 in 2014-15 to 57,194 in the last fiscal, according to Delhi's Economic Survey report. This translated in a jump in beds per 1000 population from 2.68 to 2.99 in the corresponding periods. Delhi government need to further improve his number and provide the best health practices to its people and decline the death rate due to poor health care facilities in the country.

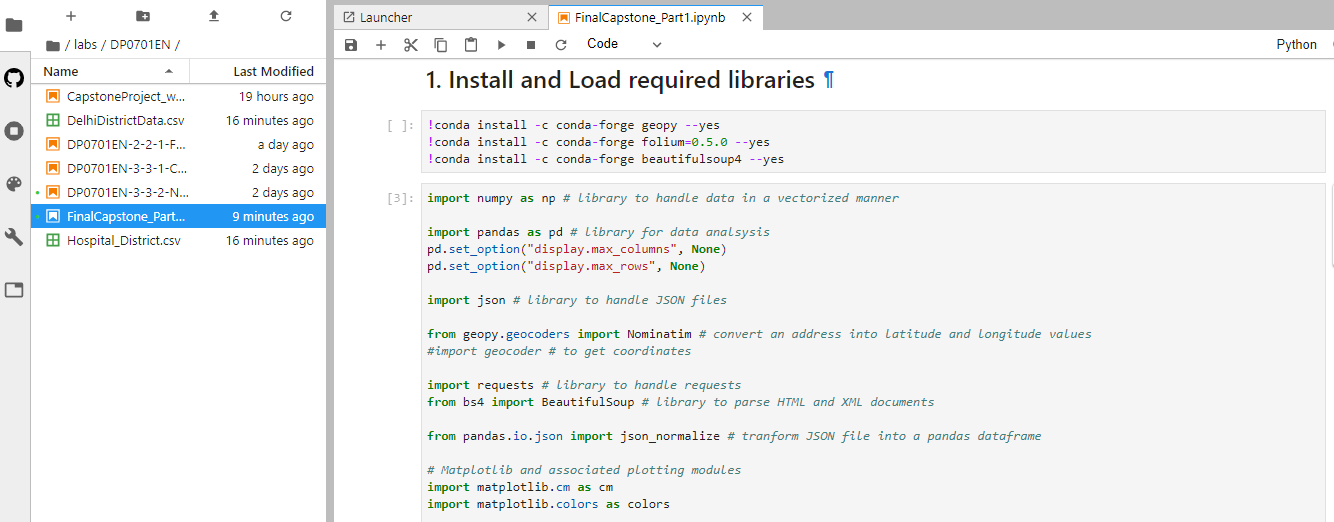
1. **DATA DESCRIPTION**

Data is prepared using:

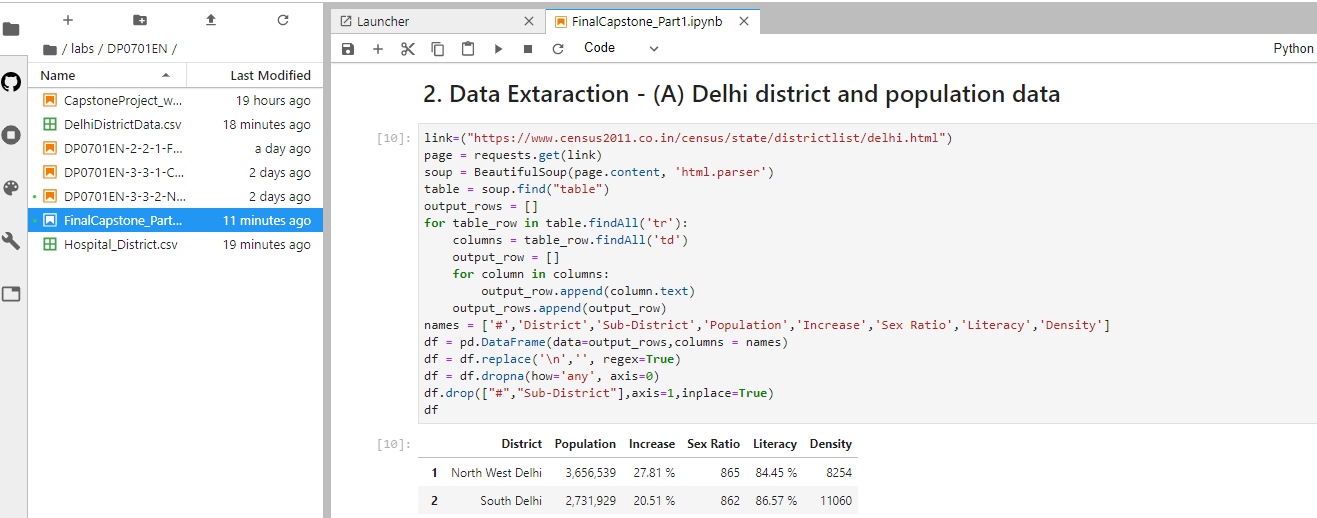
1. Data Extraction
2. Web Scraping from various sources.
3. Data Downloading from government website.
4. Data Preparation
5. Data Cleaning
6. Data merging
7. Data Definition

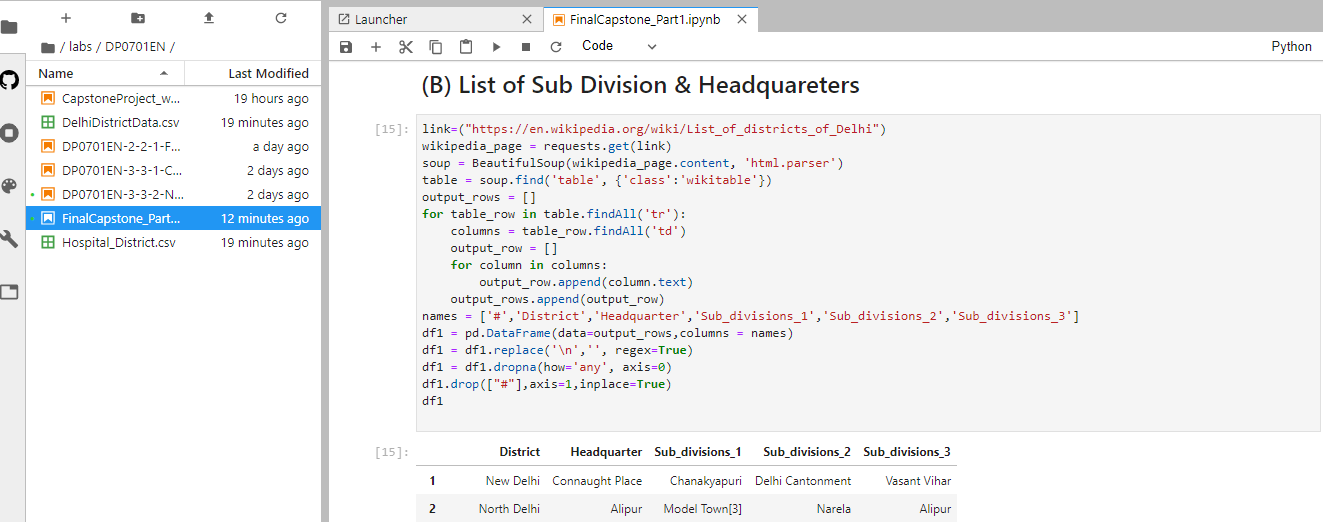
* 1. **Data Extraction**

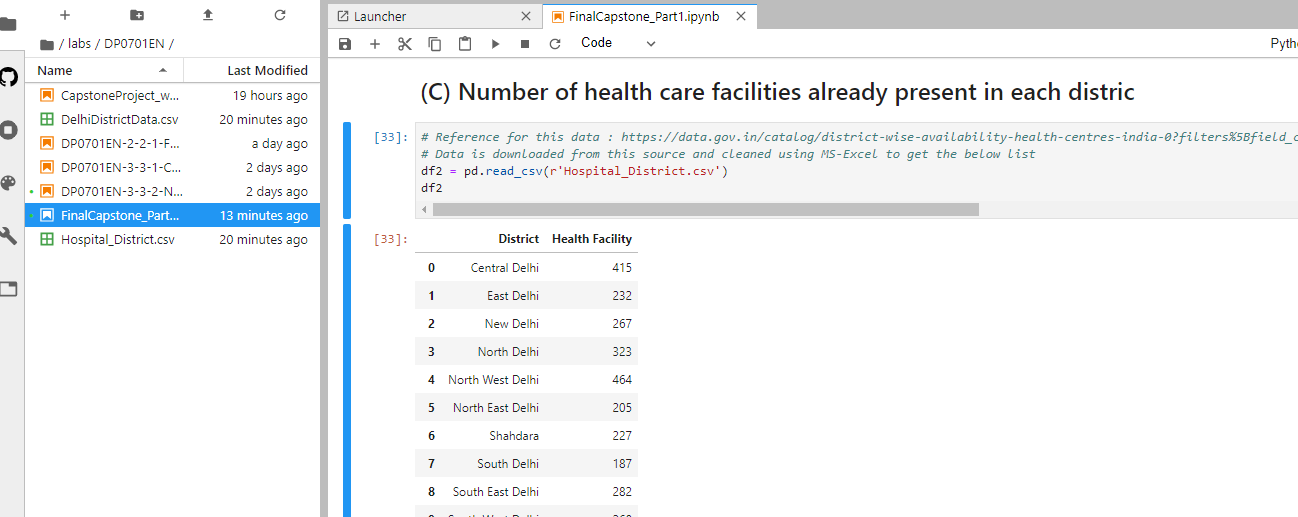
**Load the require libraries to start data extraction.**

****

**Data is extracted from various sources to get the district level information and number of hospitals in each district.**

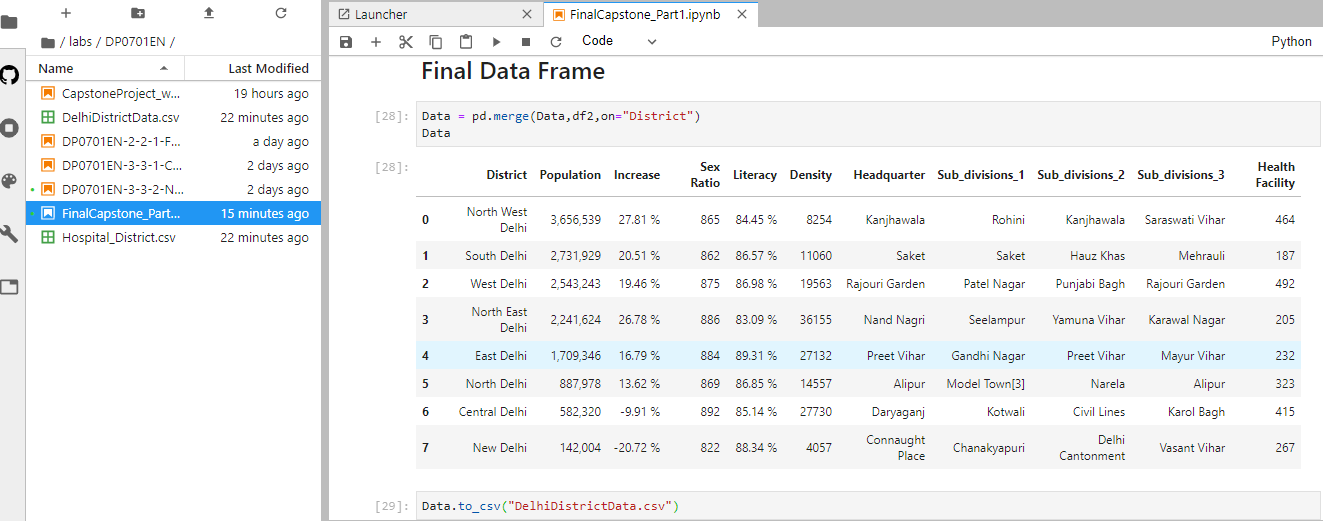
****

****

****

* 1. **Data Preparation**

**All the data is extracted now a common data frame is formed to get the final required data which will be further processed to solve the business problem.**

****

* 1. **Data Definition**

The final Data frame contains 11 columns:

1. District: Delhi is divided into various districts; this column represents the districts in Delhi.
2. Population: Each district populations is represented in this column.
3. Increase: Increase in population since last decade.
4. Sex Ration: Female, Male ratio is represented in this column.
5. Literacy: represents literacy rate in the respective districts.
6. Density: People density in respective districts.
7. Headquarter: Headquarter of the respective districts lies in these areas.
8. Subdivision: Divisions under each district.
9. Health Facility: Number of hospitals under each district.
10. **METHODOLOGY**

In this project we will direct our efforts on detecting areas of Delhi that have low hospital density, particularly those with low number of government hospitals. We will limit our analysis to area ~2km around city centre.

In first step we have collected the required district wise.

Second step in our analysis will be calculation and exploration of '**hospital density**' across different areas of Delhi.

In third and final step we will focus on most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders: we will be using **k-means clustering** of those locations to identify general zones / neighbourhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

**4. ANALYSIS**

**4.1 Data Visualization**

**In data visualization first we converted the object type columns in integer and float.**

District object

Population int64

Increase float64

Sex Ratio int64

Literacy float64

Density int64

Headquarter object

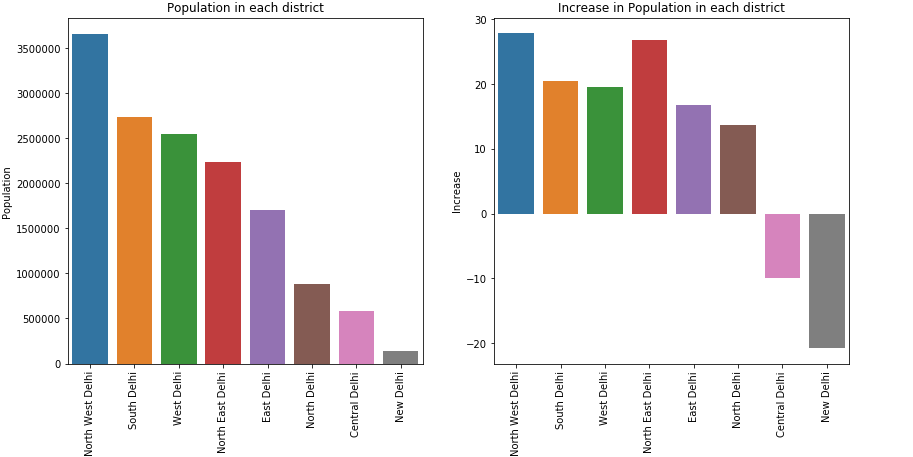
Sub\_divisions\_1 object

Sub\_divisions\_2 object

Sub\_divisions\_3 object

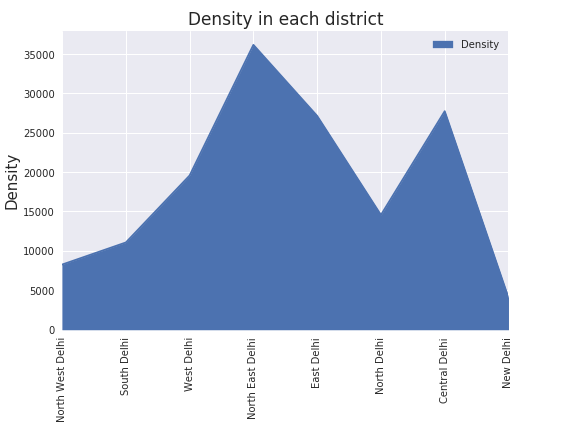
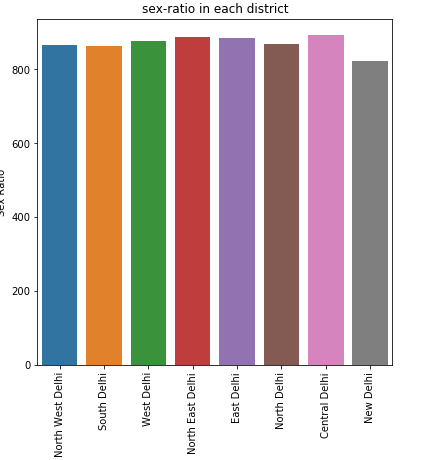
Health Facility int64

dtype: object



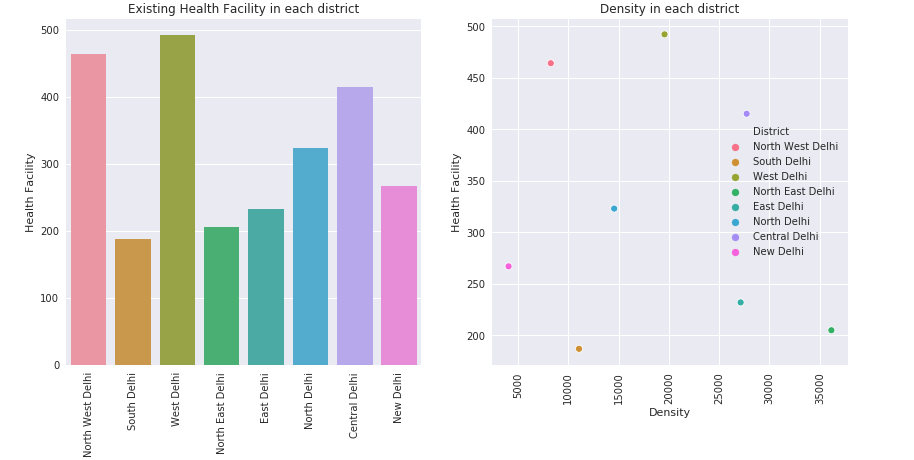
1. The Maximum population is of "North-west Delhi" having Headquarters at "Kanjhawala"

2. During the years, The population has grown much in the same "North west delhi" while in some places like "Central Delhi" and "North Delhi", it has reduced.



1. The sex-ratio is almost equal in all districts and doesnot matters much.

2. While the density (total number of people per land area) of "North East Delhi" having Headquarter:"Nand Nagri" is the highest.

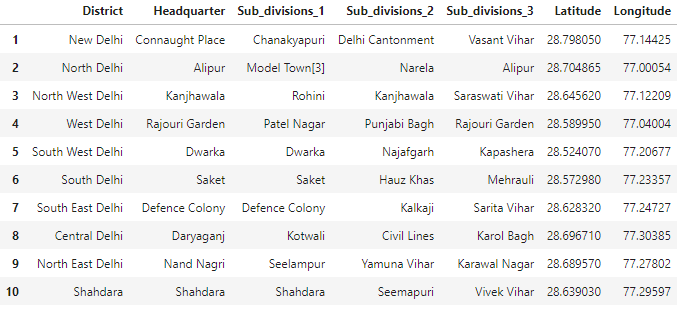


1. Number of hospitals in South Delhi,North East Delhi,ast Delhi are very less.

2. As we can see in density and health facility chart that all the points are not lying on a single increasing order line according to density this shows some districts have less number of health facility as the population is.

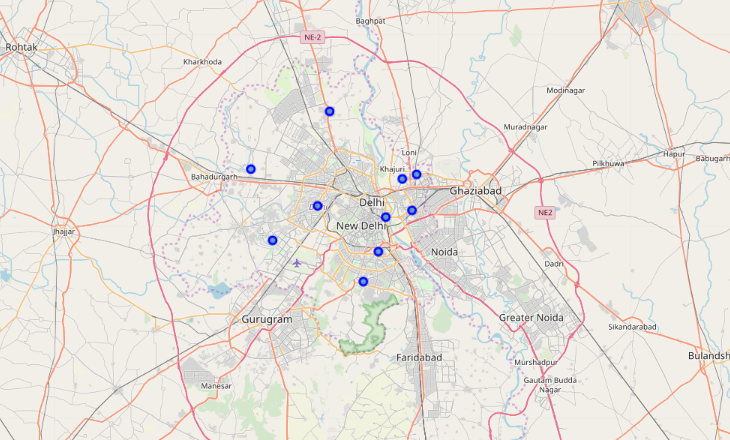
**4.2. Getting Latitudes and Longitudes**

In obtaining the location data of the Headquarters of the Districts, the Geocoder package is used with the arcgis\_geocoder to obtain the latitude and longitude of the needed locations.



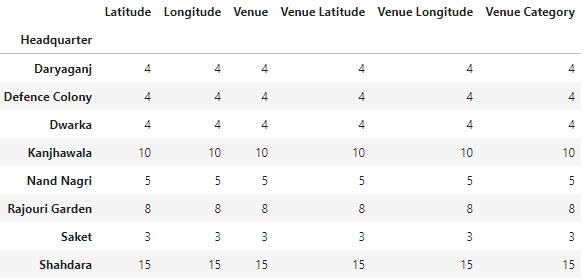
**4.3 Map Visualization**

Using the geopy library, the latitude and longitude values of Delhi is obtained. The geographical coordinate of Delhi are 28.6517178, 77.2219388.



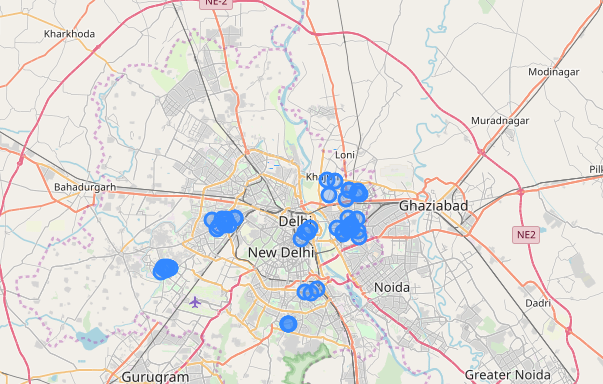
Four Square API

Using foursquare API we will pull the hospital information from Delhi Venues



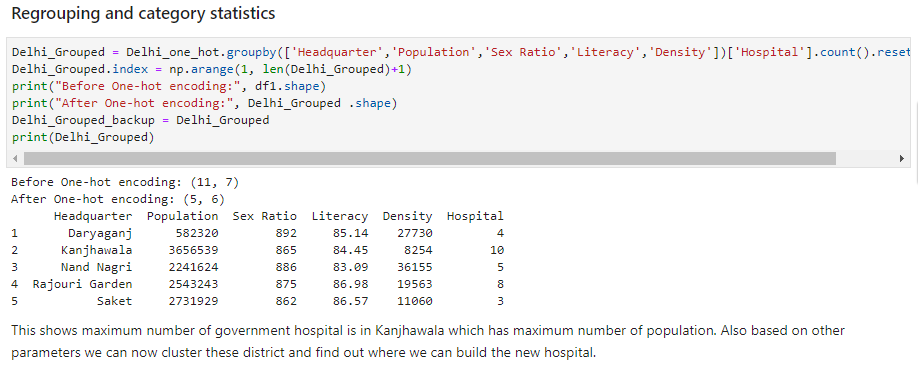
Here, we observe that maximum number of Hospitals are located in Shahdara.

Also, we will see headquarters having hospitals superimposed on a map.

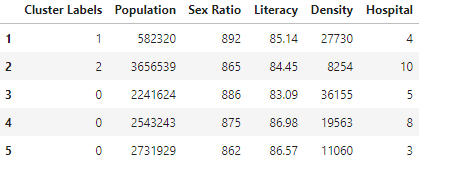


**4.4 Clustering**

**The venues will be clustered based on the processed data obtained above.**



After encoding Clustering is performed :



In clustering the data points will maximum inter distance and minimum intra distance is will be clustered in a single cluster.

Now, how to decide how many cluster we should choose:

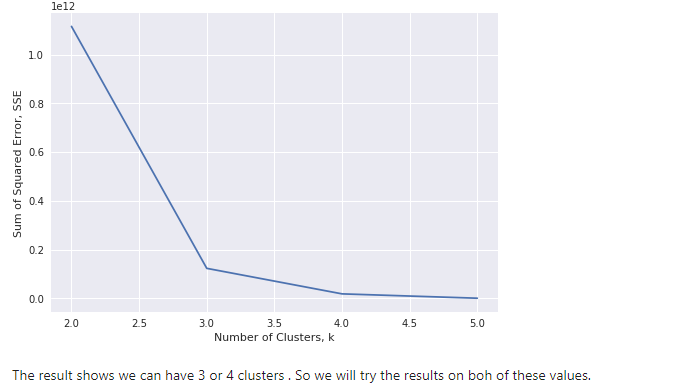
**Optimal Number of Clusters for K-means**

To get the optimal number of clusters to be used for the K-mean, there are a number ways possible for the evaluation. Therefore, in this task, the following are used:

1. Elbow (Criterion) Method
2. Silhouette Coefficient
3. Elbow Method

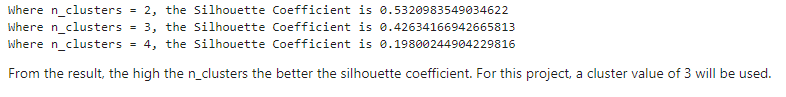
The elbow method is used to solve the problem of selecting k. Interestingly, the elbow method is not perfect either but it gives significant insight that is perhaps not top optimal but sub-optimal to choosing the optimal number of clusters by fitting the model with a range of values for k.

The approach for this is to run the k-means clustering for a range of value k and for each value of k, the Sum of the Squared Errors (SSE) is calculated., calculate sum of squared errors (SSE). When this is done, a plot of k and the corresponding SSEs are then made. At the elbow (just like arm), that is where the optimal value of k is. And that will be the number of clusters to be used. The whole idea is to have minimum SSE.

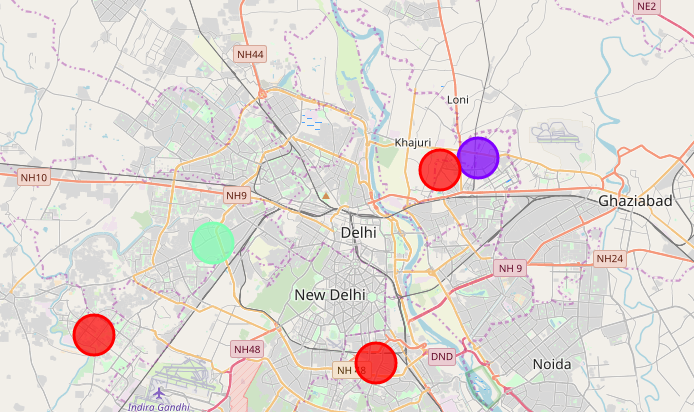


1. Silhouette Coefficient

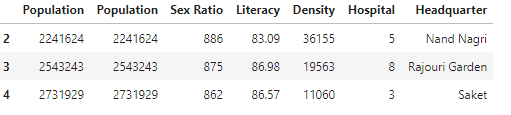
To find the optimal value of the number of clusters, k, the number of clusters is iterated corresponding Silhouette Coefficients calculated for each of the k-values used. The highest Silhouette Coefficient gives the best match to its own cluster. Please see below:



Visualizing the resulting clusters



Cluster 1



Cluster 2



Cluster 3



**5. RESULT**

The following are the highlights obtained from above 3 clusters and the data analysis:

It is surprising to know that despite being the most populated area of Delhi, we are have only 10 hospitals in Kanjhawala.

whereas daryaganj has 4 hospitals for 0.5 million people.

Although the clusters have variation, but as compared to population each district is short of hospitals.

The number of number of hospitals is less in cluster 1 as compared to the population they are catering to.

**6. Discussion and Conclusion**

Although if we see from population density point of view each area requires more number of hospitals to provide health care facilities to more number of people.

But, to be specific it will be beneficial to open a hospital in at Saket South Delhi and Nand Nagri, North East Delhi as this will cater to more population and the provide more health care facilities as there are more people in these areas and less facilities. We were not able to find any venues nearby Kanjhawala , the reason is that Khanjawala is a village and has very less amenities within 1kms.

Some drawback of these analyses is that clustering is done on only the most common venues which are obtained using four square location. Also, we have taken into consideration one city under each district . Also the result can vary if we use other clustering techniques like DBSCAN.

In conclusion, this project would have had better results if there were more data in terms of basic facilities like water etc. within the area, price of the land, traffic access and allowance of more venues exploration with the Foursquare (limited venues for free calls). Also, getting the speciality of hospital within the clusters would have helped in providing more insight into the best location.