DATA VISUALIZATION AND ANALYSIS OF THE EFFECTS OF CORONAVIRUS IN CHINA

BY SHYAM SANKAR

FEB 20, 2020

1.INTRODUCTION

1.1 BACKGROUND

Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome. A novel coronavirus is a new strain that has not been previously identified in humans.

Coronavirus outbreak is an ongoing epidemic of coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2, which started in December 2019. It was first identified in Wuhan, capital of Hubei province China, after 41 people presented with pneumonia of no clear cause. It can spread between people, with the time from exposure to onset of symptoms generally between 2 and 14 days. Symptoms may include fever, cough, and shortness of breath. Complications may include pneumonia and acute respiratory distress syndrome. There is no vaccine or specific antiviral treatment, with efforts typically to manage symptoms and support functioning. Hand washing is recommended to prevent the spread of the disease. Anyone who is suspected of carrying the virus is advised to monitor their health for two weeks, wear a mask, and seek medical advice by calling a doctor rather than directly visiting a clinic.

1.2 PROBLEM

Coronavirus is considered to have a spreading rate that is similar to that of common-flu. The virus can spread easily through fluids. This fast spreading rate builds a curiosity in me to question a relation between the coronavirus confirmed rates in different provinces of China and the population density of the respective provinces.

1.3 INTEREST

Coronavirus is a fast spreading epidemic that has struck us hard as we enter into a new decade. We as data scientists are capable of contributing a lot in making people aware of its effects and providing analytic information to whoever that needs it, whether it be research or just to gain knowledge about it.

2. DATA ACQUISITION AND CLEANING

2.1 DATA SOURCES

The major data set that has been used for the analysis was acquired from kaggle datasets(here). However data regarding the population and population density of provinces in Mainland China had to be collected from other sources. The data on population of each province was collected from here. The population density of each province/state was collected from here.

2.2.DATA CLEANING

The initial csv file acquired from kaggle contained data on province,country, update date, confirmed cases of coronavirus, suspected, recovered, deaths. Information on population density and province names in Chinese were added into the csv file before loading into the notebook and converted into pandas dataframe.
The data frame was filtered to only keep rows containing information as acquired from the latest date ie , the 26th of January 2020.
The “Nan” values in the confirmed, recovered, suspected and deaths columns were updated with zero as an approximation for unhindered analysis.
Since analysis was only being carried out for provinces that are part of mainland china the other columns that had country value set to any other was removed.
The above mentioned second dataset was used to add the longitudes, latitudes and the population of each province in the country.
The final data set that was used for analysis and visualization.

	Province/State	Country	lat	lng	Confirmed	Suspected	Recovered	Deaths	Population_density	Chinese_province_names
0	Hubei	Mainland China	30.583333	114.266667	1058.0	0	42	52	2804.0	湖北省
1	Guangdong	Mainland China	23.116667	113.250000	111.0	0	2	0	3469.0	广东省
2	Zhejiang	Mainland China	30.293650	120.161419	104.0	0	1	0	2137.0	浙江省
3	Henan	Mainland China	34.683611	113.532500	83.0	3	0	1	4903.0	河南省
4	Chongqing	Mainland China	29.562778	106.552778	75.0	0	0	0	2026.0	重庆市
5	Hunan	Mainland China	28.200000	112.966667	69.0	0	0	0	3174.0	湖南省

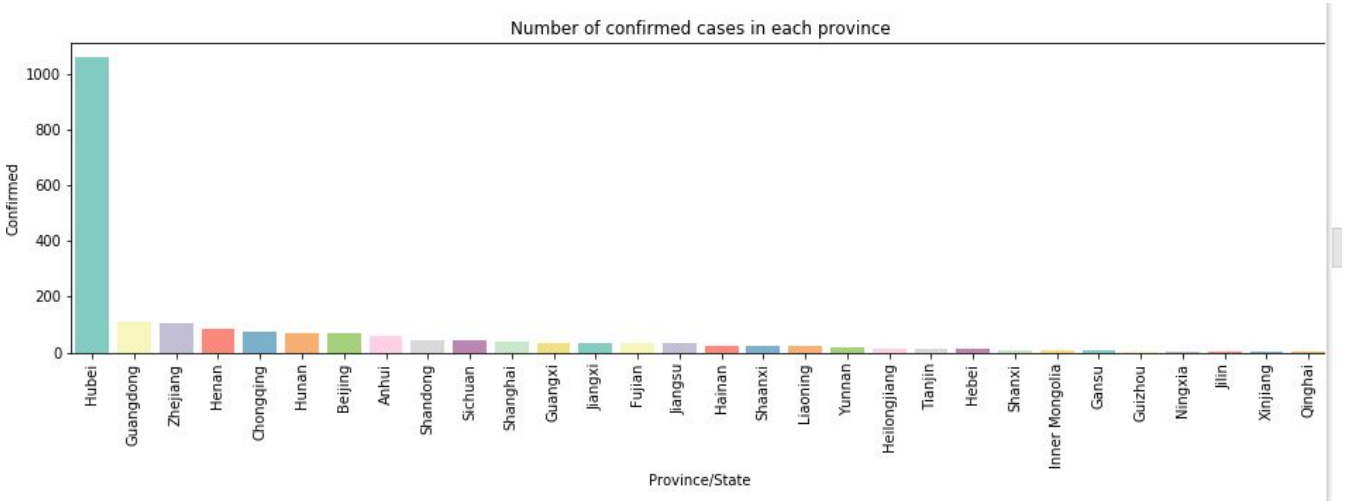
2.3. FEATURE SELECTION

Information from wikipedia and sources provided by WHO clearly indicated that the coronavirus is one of the fastest spreading virus and can be transmitted from one person to the other via fluids or molecules that are dispersed into the air when a person sneezes or coughs. This idea became the foundation of comparing the confirmed cases of coronavirus in a province and the population density of that province.
Also the symptoms and early indicators that result in the suspicion of coronavirus is very much similar with that of common flu, hence taking the suspected cases as a parameter of study would render meaningless.
Also coronavirus is seen to show drastic effects in only those people who have a very weak immune system. The deaths due to coronavirus are mostly among men and women of old age

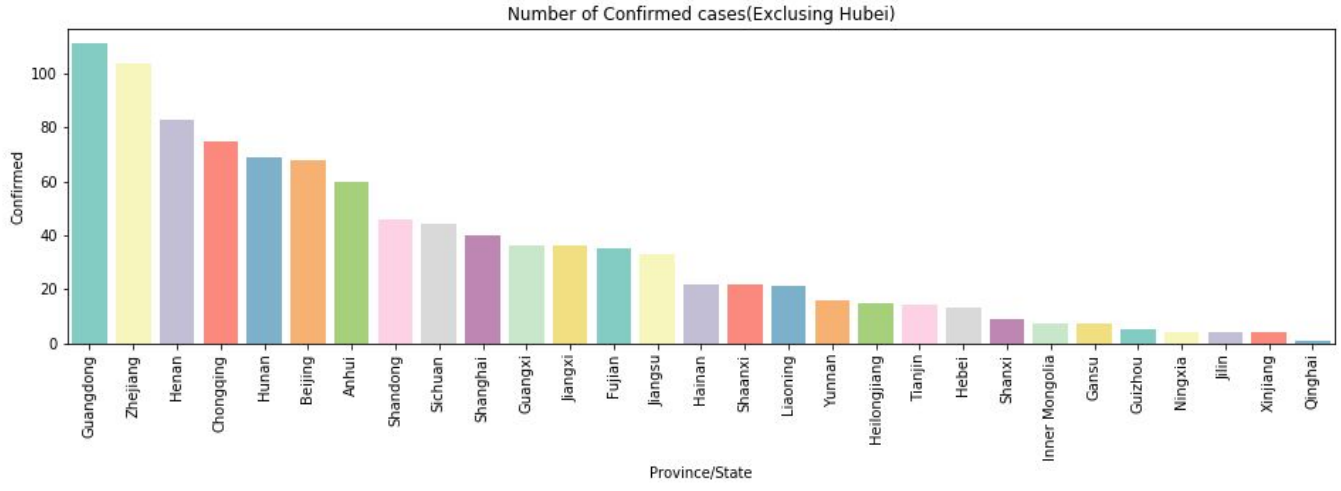
or already suffering from other diseases. This means that population density of regions and deaths would not show much of a useful trend on performing analysis.
This brought me to the conclusion that the parameters that has to be used for the segmentation and clustering are population density and confirmed cases of coronavirus.

3. EXPLORATORY DATA ANALYSIS

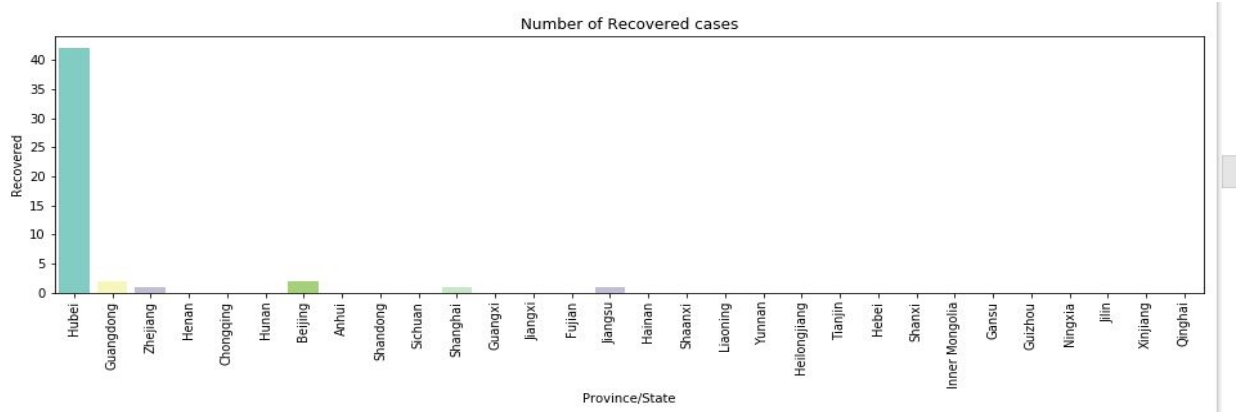
3.1.1 Distribution of confirmed cases of coronavirus in each province:



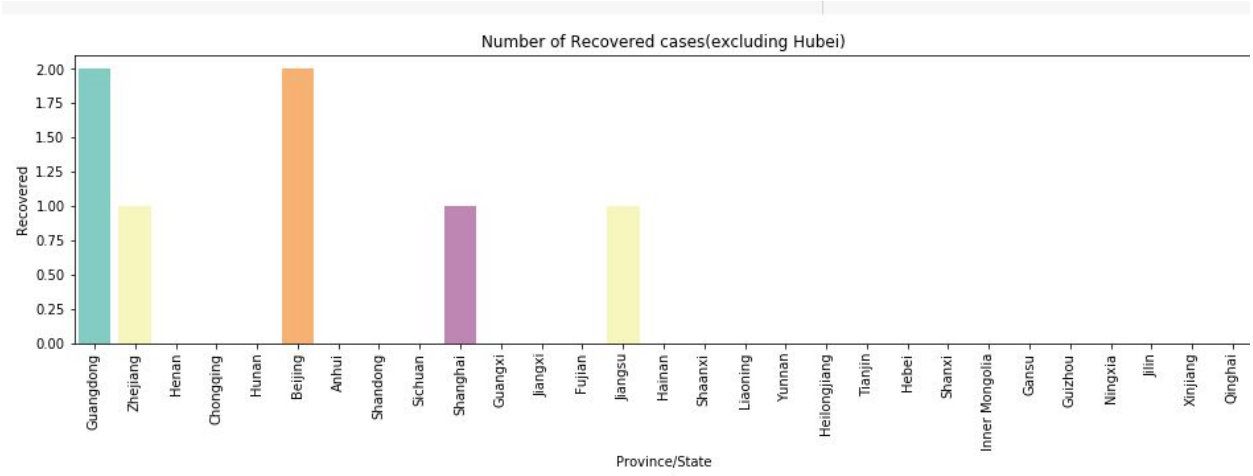
3.1.2 Excluding Hubei(Most affected region)



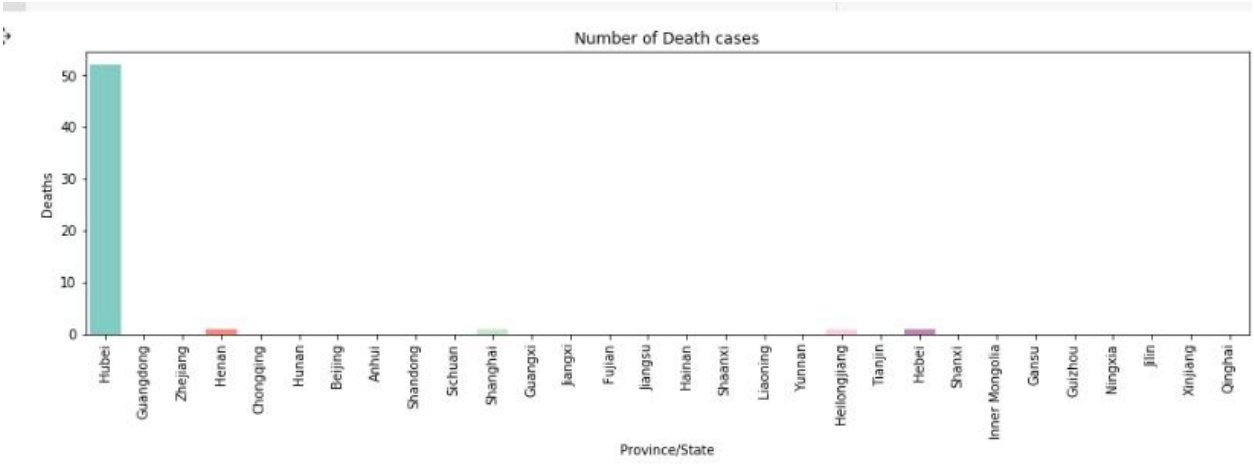
3.2.1 Distribution of recovered cases of coronavirus in each province



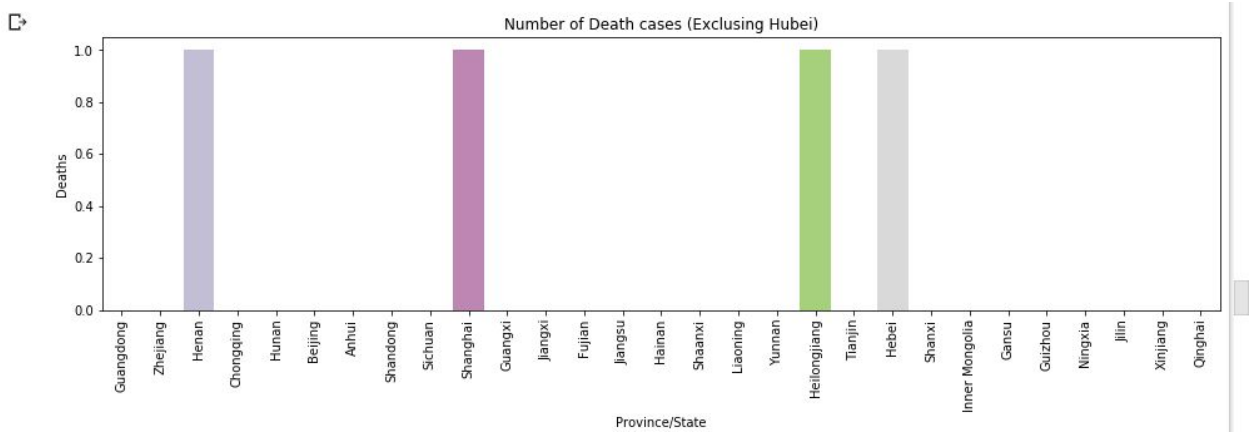
3.2.2 Excluding Hubei



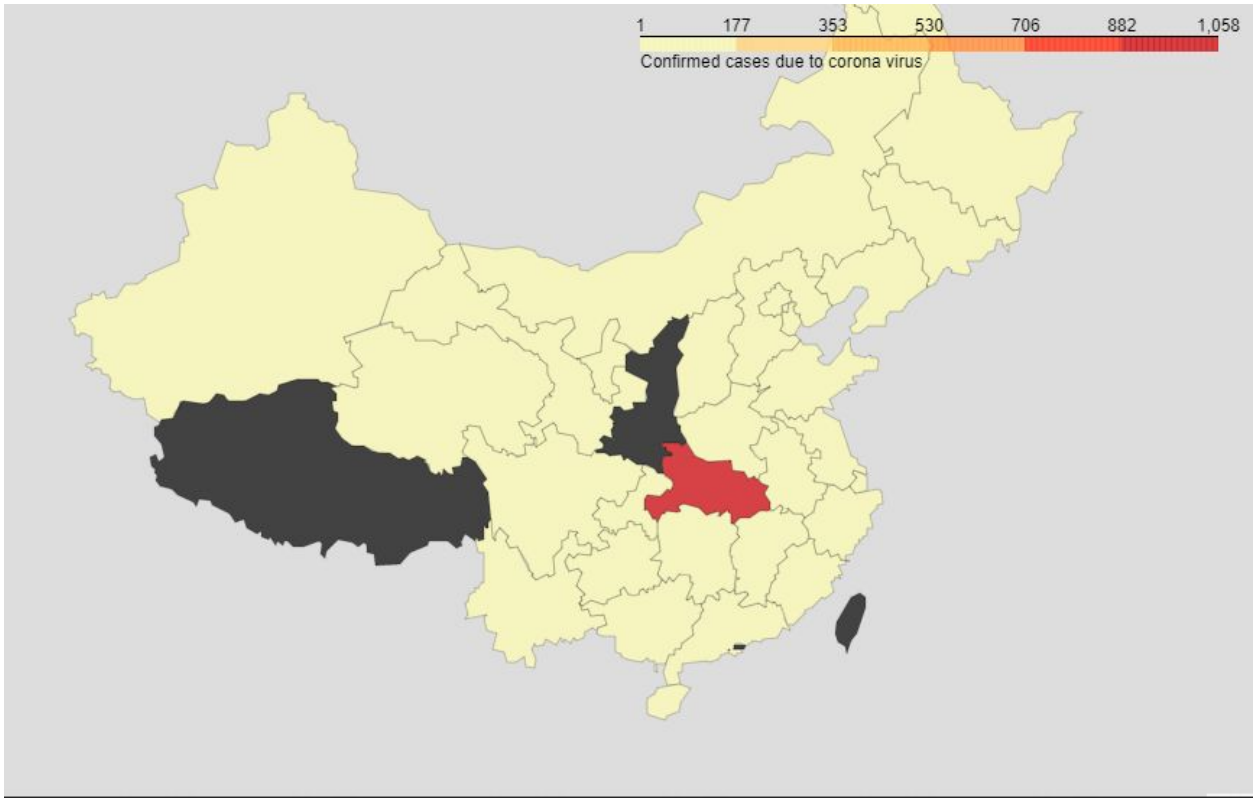
3.3.1 Distribution of deaths due to coronavirus in each province



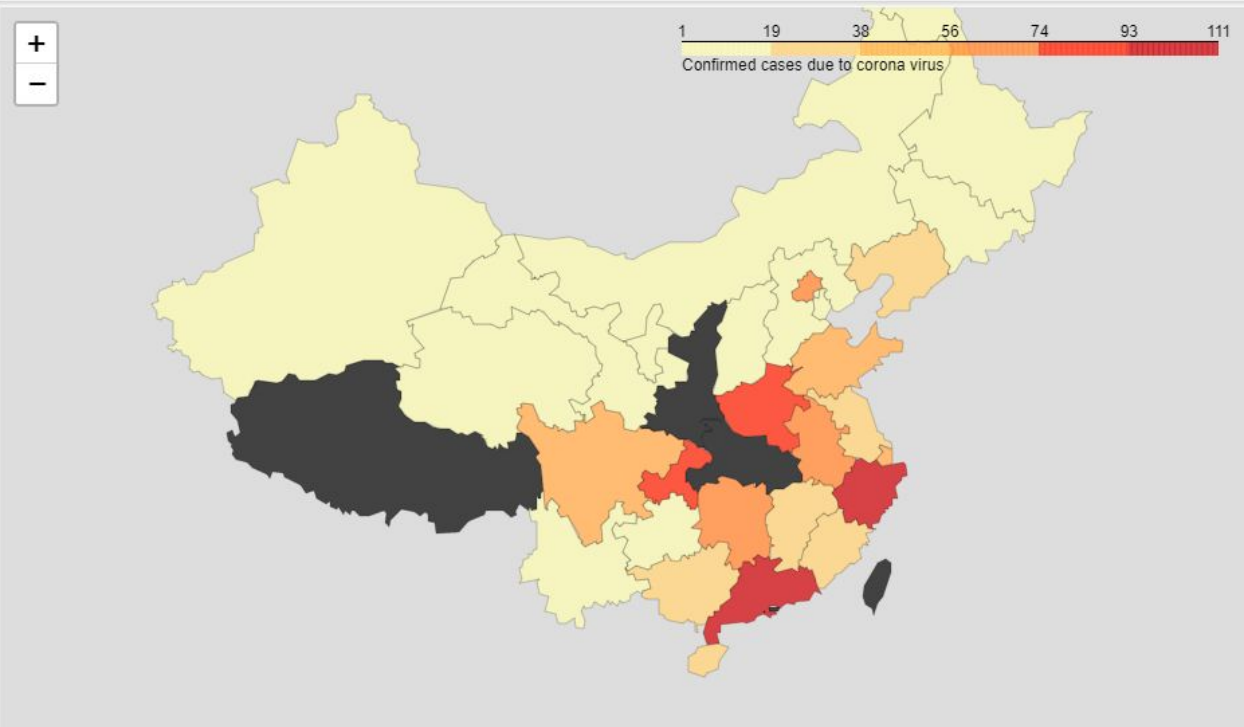
3.3.2 Excluding Hubei:



3.4.1 Choropleth map of confirmed cases due to coronavirus



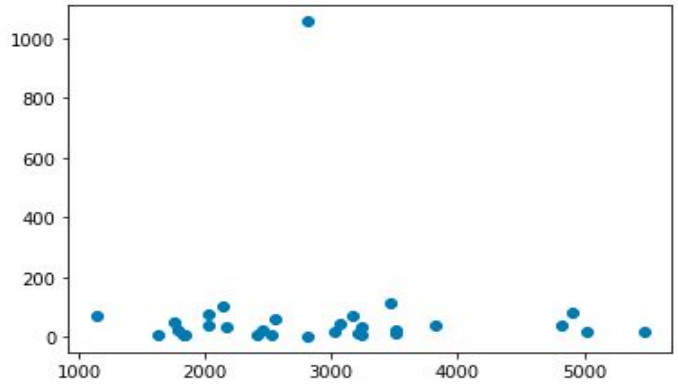
3.4.2 Excluding hubei:



4.CLUSTERING AND SEGMENTATION USING K-MEANS

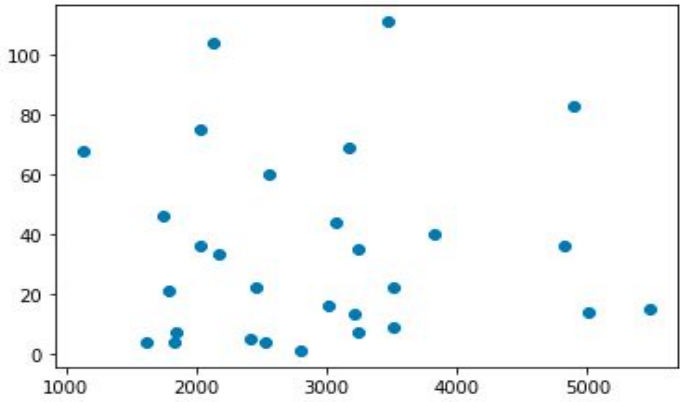
K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

4.1 Scatter plot of population density vs confirmed cases(including Hubei)

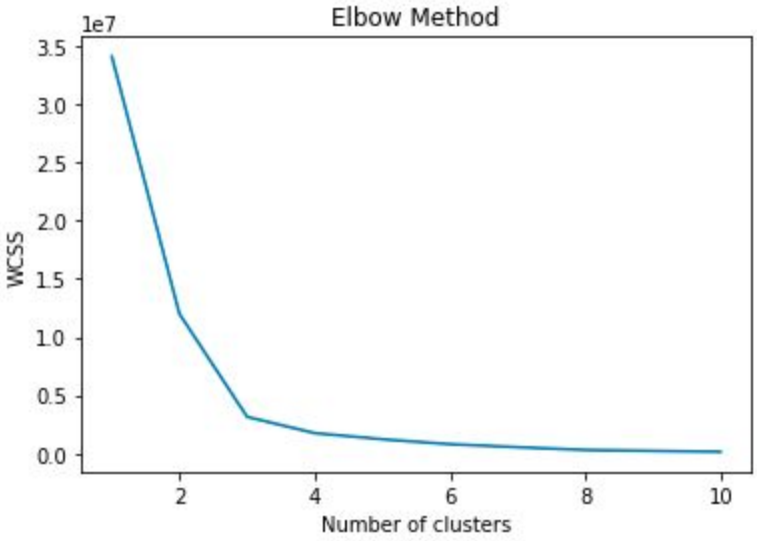


The above plot shows the scatter plot of population density vs confirmed cases in provinces including Hubei. The point corresponding to approximately 3000 in the x-axis and above the 1000 mark in the y axis is the most affected province of Hubei.
Clustering the above data would only share insights that can be drawn manually as Hubei would be standing out in every means.
Thus the aim of this segmentation and clustering module is to look at all the provinces except that of Hubei.

4.2 Scatter plot of population density vs confirmed cases excluding Hubei



4.3. Elbow method:
The elbow method is a heuristic method of interpretation and validation of consistency within cluster analysis designed to help find the appropriate number of clusters in a dataset.
Elbow method applied on the dataset gave the following plot:



The above plot of the elbow method clearly indicates that the appropriate number of clusters in the dataset is 3.

5.Results

The following table shows some of the provinces with their corresponding cluster labels

	Province/State	Cluster value	Country	lat	lng	Confirmed	Suspected	Recovered	Deaths	Population_density	Chinese_province_names
1	Guangdong	0	Mainland China	23.116667	113.250000	111.0	0	2	0	3469.0	广东省
2	Zhejiang	1	Mainland China	30.293650	120.161419	104.0	0	1	0	2137.0	浙江省
3	Henan	2	Mainland China	34.683611	113.532500	83.0	3	0	1	4903.0	河南省
4	Chongqing	1	Mainland China	29.562778	106.552778	75.0	0	0	0	2026.0	重庆市
5	Hunan	0	Mainland China	28.200000	112.966667	69.0	0	0	0	3174.0	湖南省
6	Beijing	1	Mainland China	39.928819	116.388869	68.0	0	2	0	1136.0	北京市
7	Anhui	1	Mainland China	31.863889	117.280833	60.0	4	0	0	2559.0	安徽省
8	Shandong	1	Mainland China	36.790556	118.063333	46.0	0	0	0	1750.0	山东省

5.1 Examining the first cluster

	Province/State	Cluster value	Country	lat	lng	Confirmed	Suspected	Recovered	Deaths	Population_density	Chinese_province_names
1	Guangdong	0	Mainland China	23.116667	113.250000	111.0	0	2	0	3469.0	广东省
5	Hunan	0	Mainland China	28.200000	112.966667	69.0	0	0	0	3174.0	湖南省
9	Sichuan	0	Mainland China	30.666667	104.066667	44.0	4	0	0	3068.0	四川省
10	Shanghai	0	Mainland China	31.222222	121.458056	40.0	72	1	1	3823.0	上海市
13	Fujian	0	Mainland China	24.513333	117.655556	35.0	20	0	0	3238.0	福建省
16	Shaanxi	0	Mainland China	34.258479	108.924205	22.0	0	0	0	3514.0	山西省
18	Yunnan	0	Mainland China	25.038889	102.718333	16.0	36	0	0	3021.0	云南省
21	Hebei	0	Mainland China	38.041389	114.478611	13.0	0	0	1	3210.0	河北省
22	Shanxi	0	Mainland China	37.869444	112.560278	9.0	0	0	0	3514.0	山西省
24	Gansu	0	Mainland China	36.057006	103.839868	7.0	0	0	0	3237.0	甘肃省
29	Qinghai	0	Mainland China	36.625541	101.757390	1.0	0	0	0	2804.0	青海省

mining the second cluster

5.2 Examining the second cluster

	Province/State	Cluster value	Country	lat	lng	Confirmed	Suspected	Recovered	Deaths	Population_density	Chinese_province_names
2	Zhejiang	1	Mainland China	30.293650	120.161419	104.0	0	1	0	2137.0	浙江省
4	Chongqing	1	Mainland China	29.562778	106.552778	75.0	0	0	0	2026.0	重庆市
6	Beijing	1	Mainland China	39.928819	116.388869	68.0	0	2	0	1136.0	北京市
7	Anhui	1	Mainland China	31.863889	117.280833	60.0	4	0	0	2559.0	安徽省
8	Shandong	1	Mainland China	36.790556	118.063333	46.0	0	0	0	1750.0	山东省
11	Guangxi	1	Mainland China	23.002700	109.840000	36.0	0	0	0	2025.0	广西壮族自治区
14	Jiangsu	1	Mainland China	32.061667	118.777778	33.0	0	1	0	2176.0	江苏省
15	Hainan	1	Mainland China	20.045833	110.341667	22.0	0	0	0	2460.0	海南省
17	Liaoning	1	Mainland China	41.792222	123.432778	21.0	0	0	0	1782.0	辽宁省
23	Inner Mongolia	1	Mainland China	40.652222	109.822222	7.0	0	0	0	1846.0	内蒙古自治区
25	Guizhou	1	Mainland China	26.583333	106.716667	5.0	0	0	0	2412.0	贵州省
26	Ningxia	1	Mainland China	38.468056	106.273056	4.0	0	0	0	1622.0	宁夏回族自治区
27	Jilin	1	Mainland China	43.880000	125.322778	4.0	0	0	0	1831.0	吉林省
28	Xinjiang	1	Mainland China	43.807347	87.630506	4.0	0	0	0	2525.0	新疆维吾尔自治区

5.3 Examining the third cluster

	Province/State	Cluster value	Country	lat	lng	Confirmed	Suspected	Recovered	Deaths	Population_density	Chinese_province_names
3	Henan	2	Mainland China	34.683611	113.532500	83.0	3	0	1	4903.0	河南省
12	Jiangxi	2	Mainland China	28.655758	115.905049	36.0	0	0	0	4818.0	江西省
19	Heilongjiang	2	Mainland China	45.750000	126.650000	15.0	0	0	1	5476.0	黑龙江省
20	Tianjin	2	Mainland China	39.142222	117.176667	14.0	0	0	0	5016.0	天津市

From examining the above clusters we can see that there is a trend in the number of confirmed cases of coronavirus and the population density of the province.

CONCLUSION

Coronavirus has already started spreading to the neighbouring countries and regions of China. The virus has even started showing its presence in Europe and American countries. More than 2,000 people have already died because of coronavirus. Thus the coronavirus epidemic could become one of the devastating diseases we have ever encountered. It is also notable that China as a nation has taken care of this very well, secluding patients in isolated areas, taking care of them with all the resources they can garner.
In the above data science project I have tried my best to visualise and analyse the data set on coronavirus. It is important that we as data scientists contribute in educating the people about this epidemic and perform analysis to draw insights that could in some way help the governments and the World Health Organisation who are fighting with all their resources to protect the people from the virus and helping the affected in recovering from the same.It has already been found out that only people with weak immune system due to old age or other diseases have resulted in death due to coronavirus. However the fact that any person can act as a carrier of coronavirus, thus we should never ignore the warnings and should ensure that us and those dear to us abide by the guidelines released by the WHO and the respective government agencies for preventing Coronavirus

Thus as citizens of the world and as a family of human beings, let's pray that our brothers and sisters recover fast…..

THANK YOU….