Introduction

This project wants to describe which Lombardy province is the best equipped to fight the Covid-19 pandemic,to determine the effcience of the response we are going to use the “number of ICUs bed vs region population “ parameter.

Data

For this project we will gather data from the following sources

1.Lombardy city data that contains Province , City , popultation data and geospatial data.
https://github.com/MatteoHenryChinaski/Comuni-Italiani-2018-Sql-Json-excel
2. Hospital positions from FourSquare
3. Hospital ICUs beds from `Letti_per_struttura_sanitaria_di_ricovero.csv`

Approach

In order to get information of the best province :

- Collect the New York city dataset.
- Collect population data for each province by scraping Wikipedia.
- Using Foursquare API we will get hospitals for each province.
- Collect hospital by analyzing hospitals data
- Analyzing using Clustering (Specially K-Means).
- Find the best value of K
- Visualize clusters on a scatter plot

Methodology

- Import the hospital data and filter only the ICUs

In [3]:
```
#import Hospital data and filter only ICUs beds
df_hosp = pd.read_csv(r'C:\Users\giaco\Documents\GitHub\AnalysisOfItalianICUs\Letti_per_struttura_sanitaria_di_r

df_ICU=df_hosp.loc[df_hosp['DESCR_DISCIPLINA'] == "TERAPIA INTENSIVA"].reset_index(drop=True)
df_ICUclean=df_ICU.drop(['ANNO', 'ASL', 'COD_ENTE', 'COD_STRUTTURA', 'COD_SUB',
        'COD_TIPO_STR', 'DESCR_TIPO_STR', 'PUBBL_PRIV',
       'COD_DISCIPLINA',  'AREA', 'TIPO_DEGENZA',
       'MESI_DEG_ORD', 'MEDIA_LETTI_DEG_ORD',
       'MESI_DH', 'MEDIA_LETTI_DH_DS','Posti_letto_DH_DS_attivati_al_31_12'], axis='columns')
df_ICUclean
```

Out[3]:

	PROVINCIA	DENOM_STRUTTURA	DESCR_DISCIPLINA	Posti_letto_ORD_attivati_al_31_12
0	MI	OSPEDALE CIVILE DI LEGNANO	TERAPIA INTENSIVA	34.0
1	CO	OSPEDALE S. ANNA - COMO	TERAPIA INTENSIVA	35.0
2	MI	CASA DI CURA IGEA - MILANO	TERAPIA INTENSIVA	4.0
3	CO	OSPEDALE S. ANTONIO ABATE - CANTU'	TERAPIA INTENSIVA	10.0
4	MB	PRESIDIO OSPEDALIERO DI CARATE	TERAPIA INTENSIVA	5.0
...	...	...	...	...
86	MN	OSP.CIVILE DESTRA SECCHIA-PIEVE CORIANO	TERAPIA INTENSIVA	4.0
87	BS	OSPEDALE DI MANERBIO	TERAPIA INTENSIVA	11.0
88	MI	OSPEDALE S. MARIA DELLE STELLE MELZO	TERAPIA INTENSIVA	5.0
89	MI	CENTRO CARDIOLOGICO "FOND. MONZINO" - MILANO	TERAPIA INTENSIVA	0.0
90	BS	PRESIDIO OSPEDALIERO DI CHIARI	TERAPIA INTENSIVA	13.0

- Clean the dataframe and change the province abbrevation with full name and then get all beds for each province

In [5]:
```
#Now We need to get the total number of ICU beds for each region
df_Beds=df_ICUclean.drop(['DENOM_STRUTTURA','DESCR_DISCIPLINA'],axis='columns')
df_Beds=df_Beds.groupby(['PROVINCIA']).size().reset_index(name='ICUBEDS')
df_Beds.columns = ['Province', 'IcuBeds']
df_Beds
```

Out[5]:

	Province	IcuBeds
0	Bergamo	6
1	Brescia	13
2	Como	7
3	Cremona	3
4	Lecco	2
5	Lodi	2
6	Mantova	3
7	Milano	33
8	Monza e Brianza	5
9	Pavia	5
10	Sondrio	2
11	Varese	7

- Scrape the wikipedia page to get population data of each province and put them into a dataframe

In [6]:
```
#Get all region population data
url = "https://it.wikipedia.org/wiki/Lombardia#Suddivisione_amministrativa"
req = requests.get(url)
soup = BeautifulSoup(req.content, 'html.parser')

table_contents=[]
tab=soup.find_all("table",{"class":"wikitable sortable"})
tab_pop=tab[5]
df_pop = pd.read_html(str(tab_pop))[0]
```

In [7]:
```
df_pop=df_pop.drop(['Superficie(km2)','Densità(ab./km2)','Comuni(n.)','Mappa'],axis='columns')
df_pop.columns = ['Province', 'Population']
df_pop
```

Out[7]:

	Province	Population
0	Bergamo	1110457
1	Brescia	1262135
2	Como	599637
3	Cremona	358578
4	Lecco	337256
5	Lodi	229946
6	Mantova	411959
7	Milano	3233541
8	Monza e Brianza	871523
9	Pavia	545611
10	Sondrio	181249

- Merge the two dataframes and calculate the Beds/Population ratio

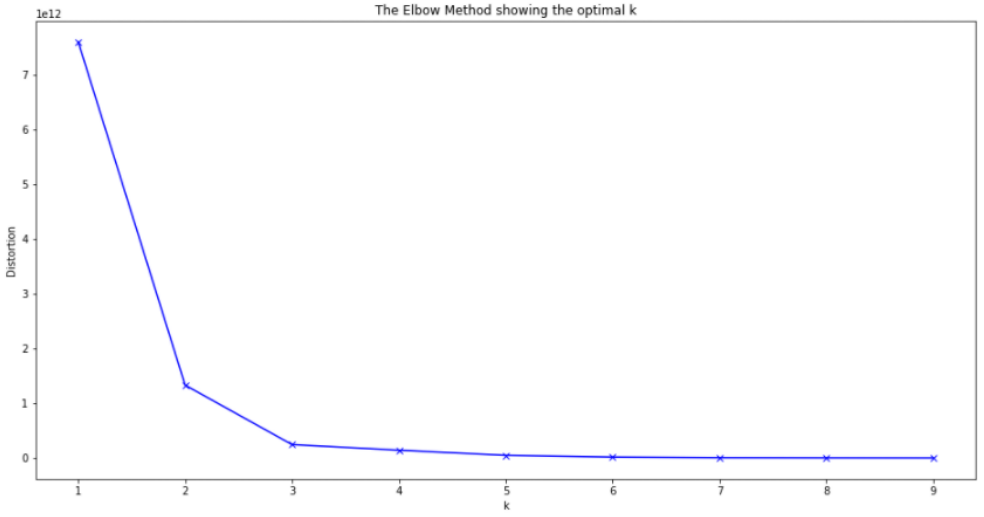
In [26]:
```
#Merge of population data and ICU BEDS by region
data_final=pd.merge(df_pop, df_Beds, on=['Province'])
data_final.columns = ['Province','Population' ,'ICU_Beds']

# using apply function to create a new column
data_final['BedPopulationRatio'] = data_final.apply(lambda row: (row.ICU_Beds/row.Population) , axis=1)

# Print the DataFrame after addition
# of new column
data_final
```

Out[26]:

	Province	Population	ICU_Beds	BedPopulationRatio
0	Bergamo	1110457	6	0.000005
1	Brescia	1262135	13	0.000010
2	Como	599637	7	0.000012
3	Cremona	358578	3	0.000008
4	Lecco	337256	2	0.000006
5	Lodi	229946	2	0.000009
6	Mantova	411959	3	0.000007
7	Milano	3233541	33	0.000010
8	Monza e Brianza	871523	5	0.000006
9	Pavia	545611	5	0.000009
10	Sondrio	181249	2	0.000011
11	Varese	890418	7	0.000008

- Find the best K for kmeans clustering with the elbow method

In [28]:
```
#Find the best k for kmeans
ICUclustering=data_final.drop(['Province'],axis='columns')
distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(ICUclustering)
    distortions.append(kmeanModel.inertia_)
#Print distortions to find the best k with elbow method
plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



- Get KMeans clusters and scatter plot visualization

k

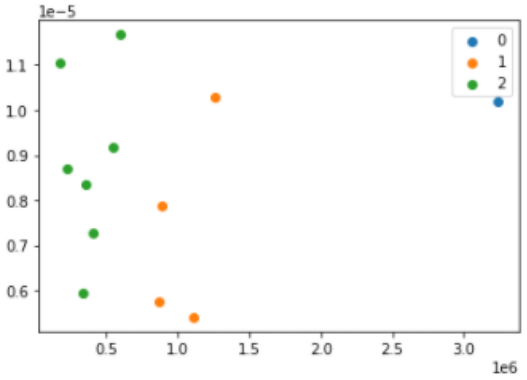
In [29]:
```
#We use 3 as K to define clusters
kmeanModel = KMeans(n_clusters=3)
kmeanModel.fit(ICUclustering)
kmeanModel.labels_[0:10]
#Add labels to datas
data_final.insert(0,'Cluster_Labels',kmeanModel.labels_)
data_final
```

Out[29]:

	Cluster_Labels	Province	Population	ICU_Beds	BedPopulationRatio
0	1	Bergamo	1110457	6	0.000005
1	1	Brescia	1262135	13	0.000010
2	2	Como	599637	7	0.000012
3	2	Cremona	358578	3	0.000008
4	2	Lecco	337256	2	0.000006
5	2	Lodi	229946	2	0.000009
6	2	Mantova	411959	3	0.000007
7	0	Milano	3233541	33	0.000010
8	1	Monza e Brianza	871523	5	0.000006
9	2	Pavia	545611	5	0.000009
10	2	Sondrio	181249	2	0.000011
11	1	Varese	890418	7	0.000008

In [31]:
```
data_C1=data_vis.loc[data_vis['Cluster_Labels'] == 0]
data_C2=data_vis.loc[data_vis['Cluster_Labels'] == 1]
data_C3=data_vis.loc[data_vis['Cluster_Labels'] == 2]


u_labels = data_vis['Cluster_Labels'].unique()


#plotting the results:

plt.scatter(data_C1['Population'] , data_C1['BedPopulationRatio'], label = 0)
plt.scatter(data_C2['Population'] , data_C2['BedPopulationRatio'],label = 1)
plt.scatter(data_C3['Population'] , data_C3['BedPopulationRatio'], label = 2)
plt.legend()
plt.show
```

Out[31]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Results and Discussion

During the analysis we found that the population is a very important factor to define the effectiveness of the response . We have an outlier (Cluster 0) that is the Lombardy province that has the largest population among all the provinces and the largest number of beds.
The other two clusters are made of medium populated province (Cluster 1) and low populated province (Cluster 2). A general discussion can be done about how low is in general the Beds / Population ratio in all of province. This low number is one of the guilty of the large number of deaths in this part of Italy.

Conclusion

With this brief analysis we can conclude that in Italy we have a very low amount of ICU beds , it would be nice to extend the analysis to all Italy in order to get a better understanding of this problem, but its so difficult to get all hospital datas from each italian region because the Healthcare system in Italy is independent in each region.