Data Engineering Final Project Report

Topic: Analysis of COVID-19 impact on Economy using Stock Market data and Unemployment rates

Name: Tirupal Rao Ravilla Net-Id: trr321 Group: Thursday - Group 6

Tasks:

1. To analyze the COVID-19 cases among all the countries of the world, using Clustering
2. To compare the impact of Covid-19 cases on stock indexes of 3 countries (US, UK, China)
3. To predict the stock indexes for the next 10 days after May 4th
4. To analyze the unemployment rate using the historical data for US, UK, and China

Data Sources:

Covid-19 data: https://github.com/CSSEGISandData/COVID-19 (JHU)

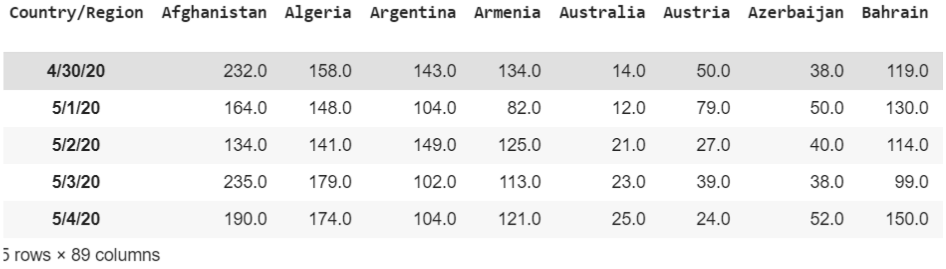
Historical Unemployment Rates: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Historical Stock Value: https://tradingeconomics.com

Methods:

Data Preparation:

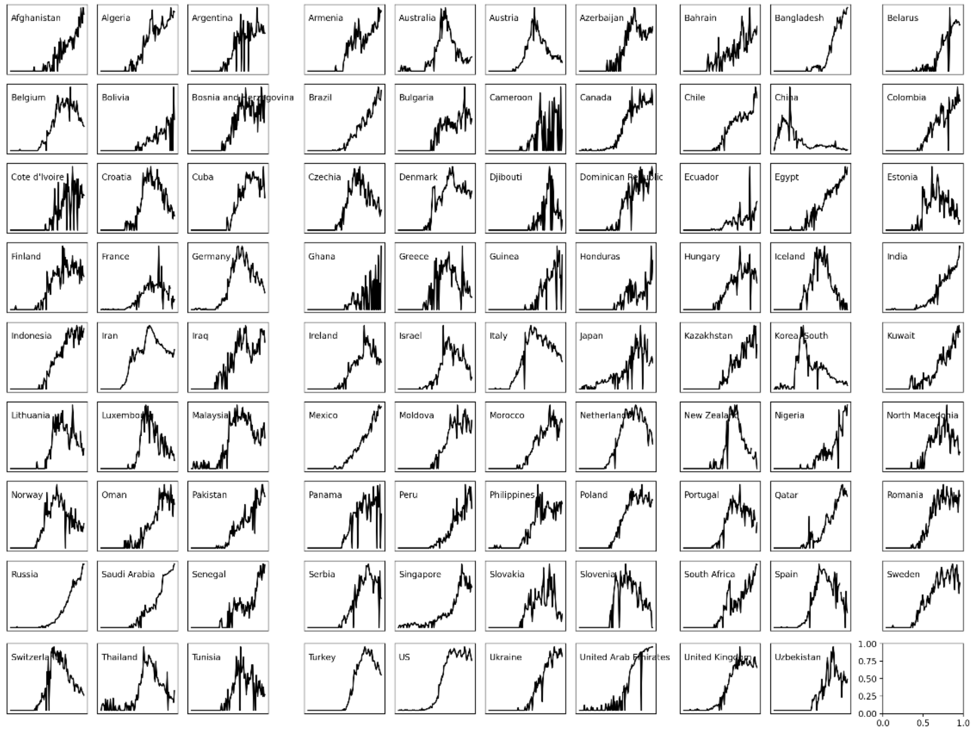
- Satiny checks were done, to see the initial condition of the data, filling the missing values, and understand the count, mean, std, min, max of the day-to-day values of COVID 19 cases, starting from Jan-22 2020 to May-04 2020
- Data is filtered to use only the relevant information for the analysis
- Only the countries which have more than 1000 cases in total were taken for analysis



Country/Region	Afghanistan	Algeria	Argentina	Armenia	Australia	Austria	Azerbaijan	Bahrain
4/30/20	232.0	158.0	143.0	134.0	14.0	50.0	38.0	119.0
5/1/20	164.0	148.0	104.0	82.0	12.0	79.0	50.0	130.0
5/2/20	134.0	141.0	149.0	125.0	21.0	27.0	40.0	114.0
5/3/20	235.0	179.0	102.0	113.0	23.0	39.0	38.0	99.0
5/4/20	190.0	174.0	104.0	121.0	25.0	24.0	52.0	150.0

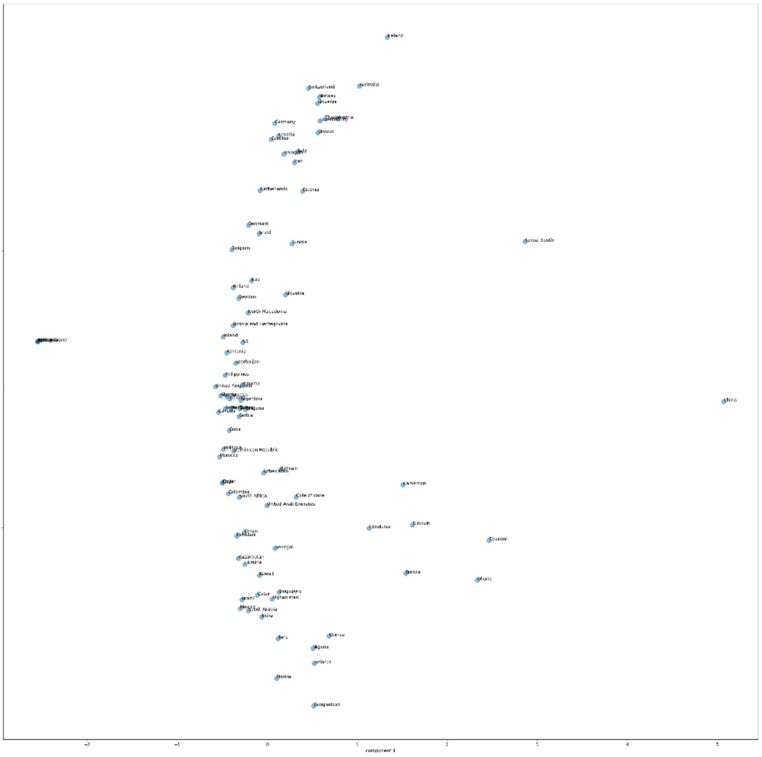
5 rows × 89 columns

- For each country, a plot of cases was done to see the trends



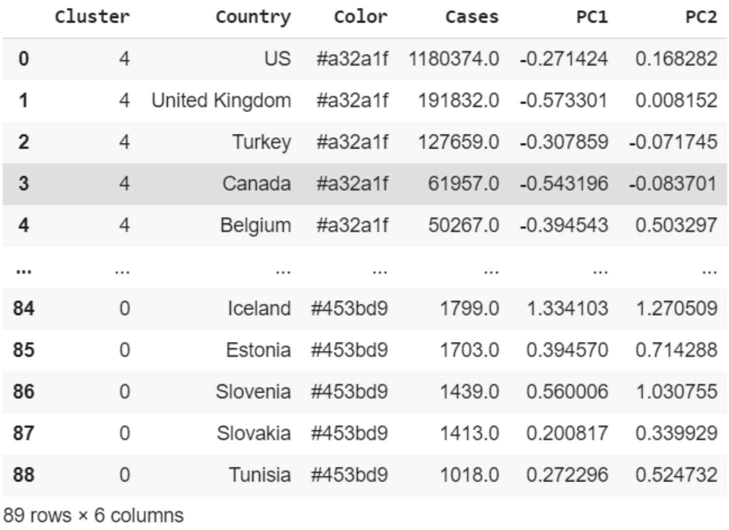
PCA:

- Distance Matrix is calculated for each country, by calculating the Bray-Curtis distance for each country from all others using the day-to-day confirmed cases, to see how close the trend of one country is from another.
- We have 89 features for each country using this method, so PCA is used to compress them into 2 components (variance: [1.20720431 0.37731189]
- Scatter plot is made between the components



Clustering:

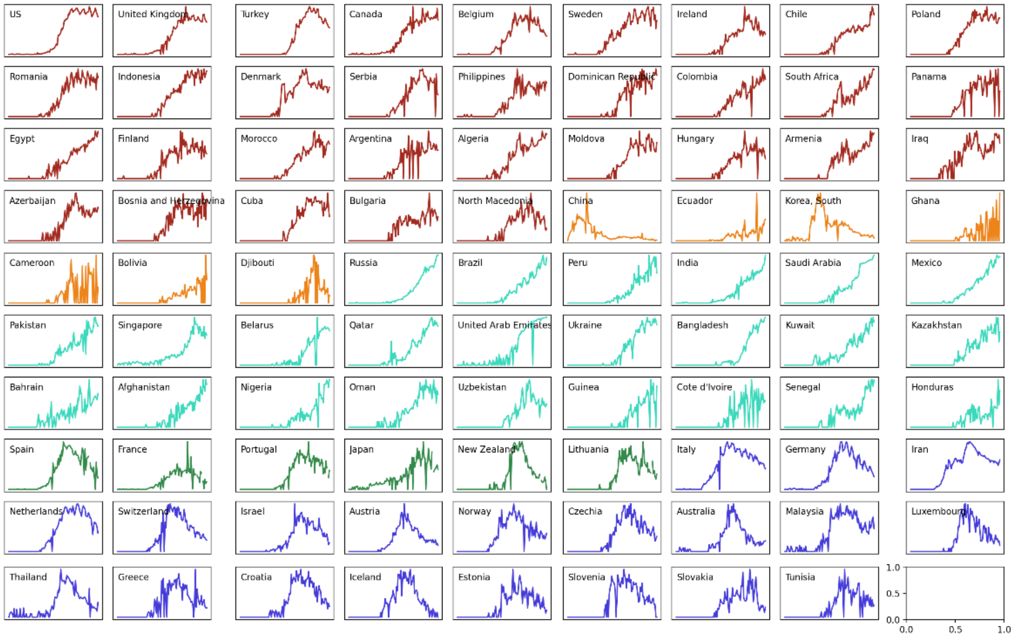
- We used Kmeans++ clustering algorithm to make 5 clusters of the data.
- We assigned each cluster a color, so that it can be used to re-plot the PCAs and the trends

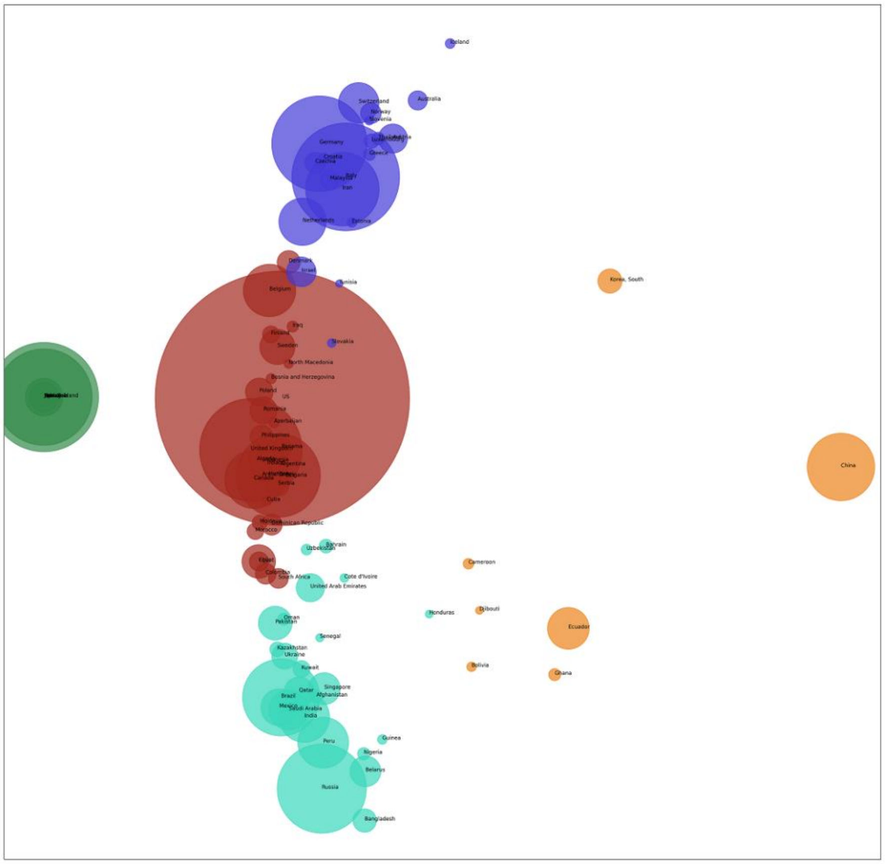


	Cluster	Country	Color	Cases	PC1	PC2
0	4	US	#a32a1f	1180374.0	-0.271424	0.168282
1	4	United Kingdom	#a32a1f	191832.0	-0.573301	0.008152
2	4	Turkey	#a32a1f	127659.0	-0.307859	-0.071745
3	4	Canada	#a32a1f	61957.0	-0.543196	-0.083701
4	4	Belgium	#a32a1f	50267.0	-0.394543	0.503297
...	...	...	...	...	...	...
84	0	Iceland	#453bd9	1799.0	1.334103	1.270509
85	0	Estonia	#453bd9	1703.0	0.394570	0.714288
86	0	Slovenia	#453bd9	1439.0	0.560006	1.030755
87	0	Slovakia	#453bd9	1413.0	0.200817	0.339929
88	0	Tunisia	#453bd9	1018.0	0.272296	0.524732

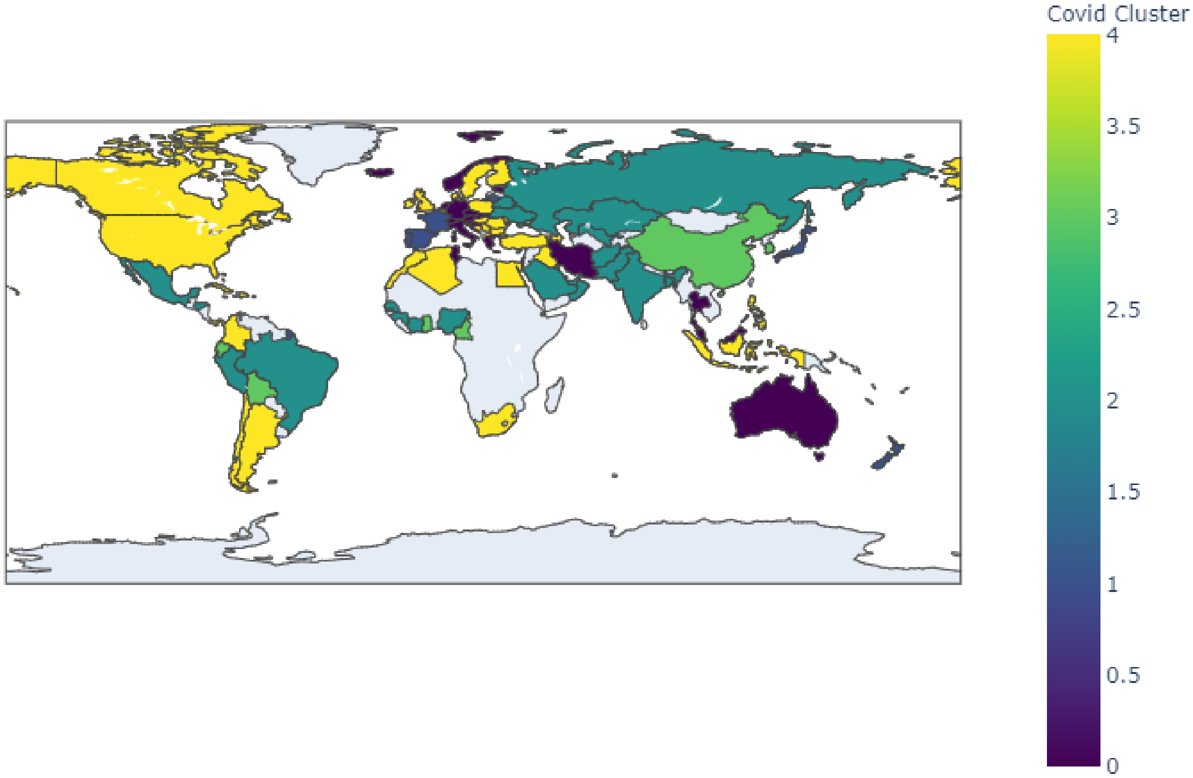
89 rows × 6 columns

- We replotted the PCAs and trends as follows:





- We plotted a geographical map with the cluster information

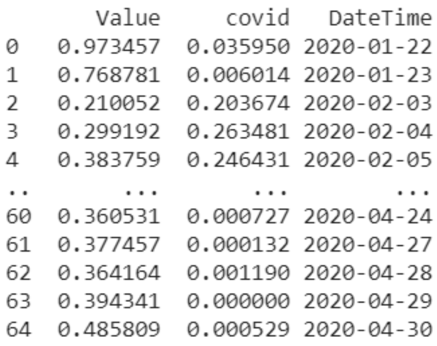


Classification:

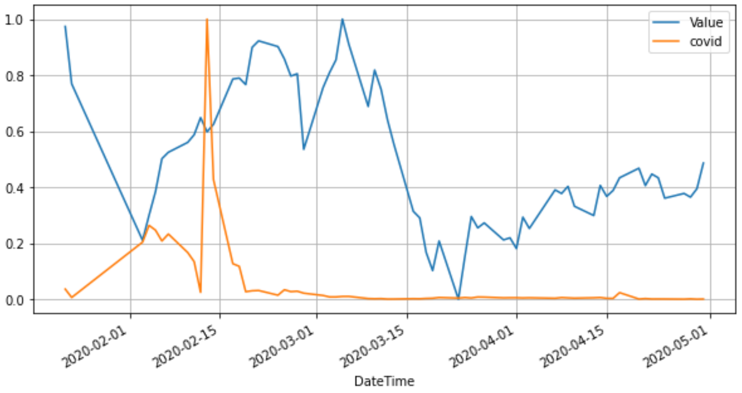
- We used SVM and KNN algorithms over the PCA data, using clusters as labels to perform classification, and figured that SVM performs better
  SVC with test size 0.4 : Accuracy: 97.2 || KNN with test size 0.4: Accuracy: 94.4

Stock Market Analysis*:*

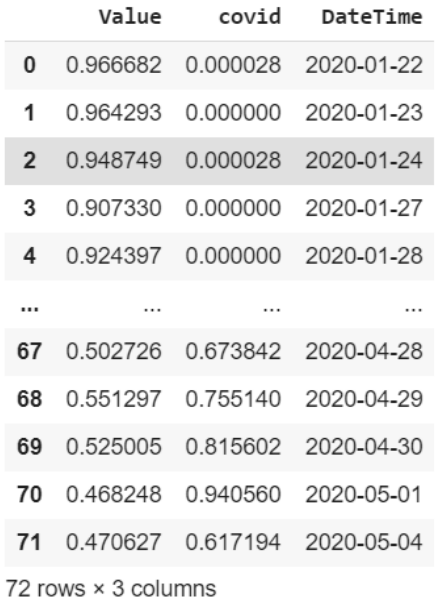
China:



```
       Value     covid    DateTime
0   0.973457  0.035950  2020-01-22
1   0.768781  0.006014  2020-01-23
2   0.210052  0.203674  2020-02-03
3   0.299192  0.263481  2020-02-04
4   0.383759  0.246431  2020-02-05
..       ...       ...         ...
60  0.360531  0.000727  2020-04-24
61  0.377457  0.000132  2020-04-27
62  0.364164  0.001190  2020-04-28
63  0.394341  0.000000  2020-04-29
64  0.485809  0.000529  2020-04-30
```

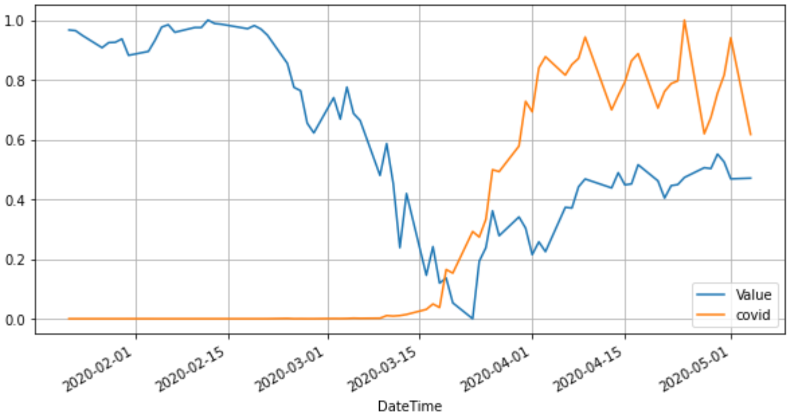


Unites States:

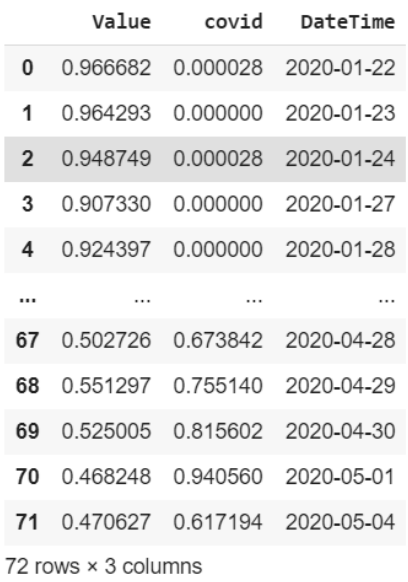


	Value	covid	DateTime
0	0.966682	0.000028	2020-01-22
1	0.964293	0.000000	2020-01-23
2	0.948749	0.000028	2020-01-24
3	0.907330	0.000000	2020-01-27
4	0.924397	0.000000	2020-01-28
...	...	...	...
67	0.502726	0.673842	2020-04-28
68	0.551297	0.755140	2020-04-29
69	0.525005	0.815602	2020-04-30
70	0.468248	0.940560	2020-05-01
71	0.470627	0.617194	2020-05-04

72 rows × 3 columns

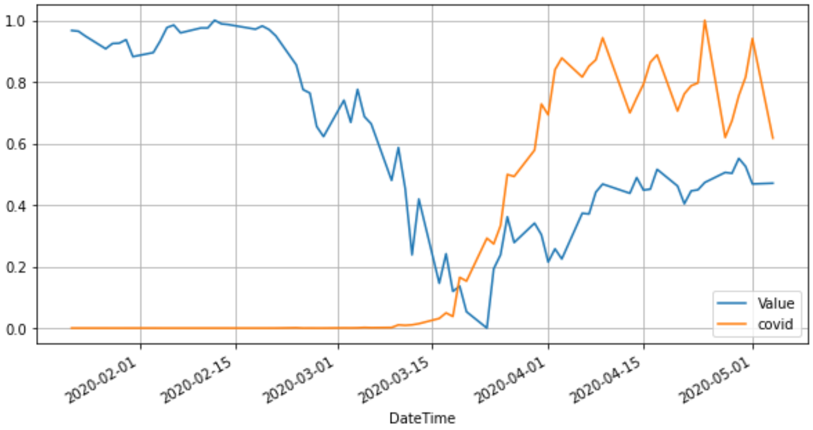


United Kingdom:



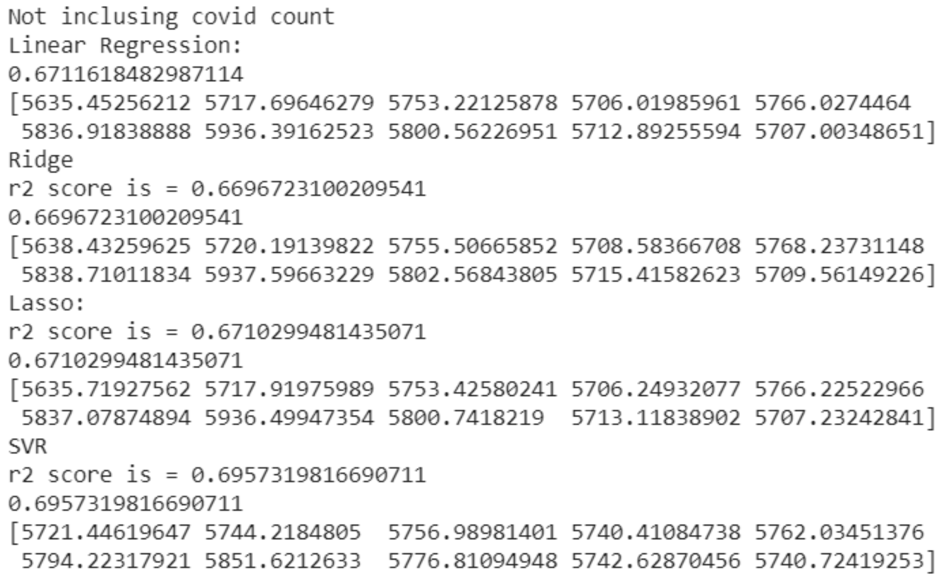
	Value	covid	DateTime
0	0.966682	0.000028	2020-01-22
1	0.964293	0.000000	2020-01-23
2	0.948749	0.000028	2020-01-24
3	0.907330	0.000000	2020-01-27
4	0.924397	0.000000	2020-01-28
...	...	...	...
67	0.502726	0.673842	2020-04-28
68	0.551297	0.755140	2020-04-29
69	0.525005	0.815602	2020-04-30
70	0.468248	0.940560	2020-05-01
71	0.470627	0.617194	2020-05-04

72 rows × 3 columns

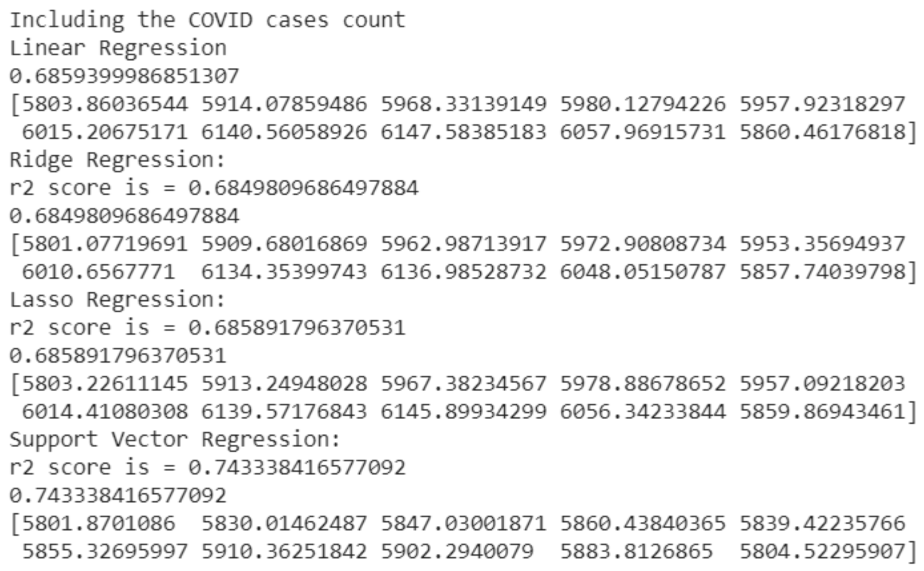


Regression on Stock Market data and Forecasting:

- We used Liner Regression, Ridge Regression, Lasso Regression, and Support Vector Regression on the stock data of US, UK, and China to forecast next 10 days of stock values.
- We did the forecasting using the Covid-19 cases and without using the cases, and in the case where we used the Covid-19 cases as none of the features, the forecasting worked better and the results are as shown below:



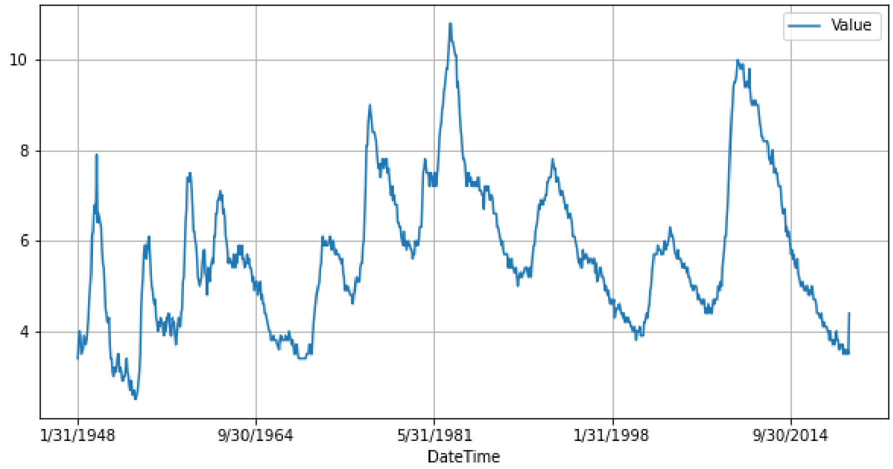
```
Not inclusing covid count
Linear Regression:
0.6711618482987114
[5635.45256212 5717.69646279 5753.22125878 5706.01985961 5766.0274464
 5836.91838888 5936.39162523 5800.56226951 5712.89255594 5707.00348651]
Ridge
r2 score is = 0.6696723100209541
0.6696723100209541
[5638.43259625 5720.19139822 5755.50665852 5708.58366708 5768.23731148
 5838.71011834 5937.59663229 5802.56843805 5715.41582623 5709.56149226]
Lasso:
r2 score is = 0.6710299481435071
0.6710299481435071
[5635.71927562 5717.91975989 5753.42580241 5706.24932077 5766.22522966
 5837.07874894 5936.49947354 5800.7418219  5713.11838902 5707.23242841]
SVR
r2 score is = 0.6957319816690711
0.6957319816690711
[5721.44619647 5744.2184805  5756.98981401 5740.41084738 5762.03451376
 5794.22317921 5851.6212633  5776.81094948 5742.62870456 5740.72419253]
```



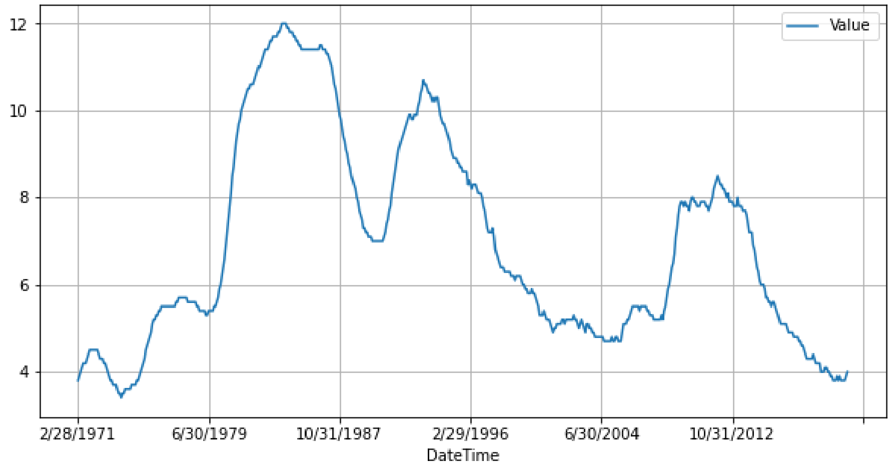
```
Including the COVID cases count
Linear Regression
0.6859399986851307
[5803.86036544 5914.07859486 5968.33139149 5980.12794226 5957.92318297
 6015.20675171 6140.56058926 6147.58385183 6057.96915731 5860.46176818]
Ridge Regression:
r2 score is = 0.6849809686497884
0.6849809686497884
[5801.07719691 5909.68016869 5962.98713917 5972.90808734 5953.35694937
 6010.6567771  6134.35399743 6136.98528732 6048.05150787 5857.74039798]
Lasso Regression:
r2 score is = 0.685891796370531
0.685891796370531
[5803.22611145 5913.24948028 5967.38234567 5978.88678652 5957.09218203
 6014.41080308 6139.57176843 6145.89934299 6056.34233844 5859.86943461]
Support Vector Regression:
r2 score is = 0.743338416577092
0.743338416577092
[5801.8701086  5830.01462487 5847.03001871 5860.43840365 5839.42235766
 5855.32695997 5910.36251842 5902.2940079  5883.8126865  5804.52295907]
```

Unemployment Data Analysis:

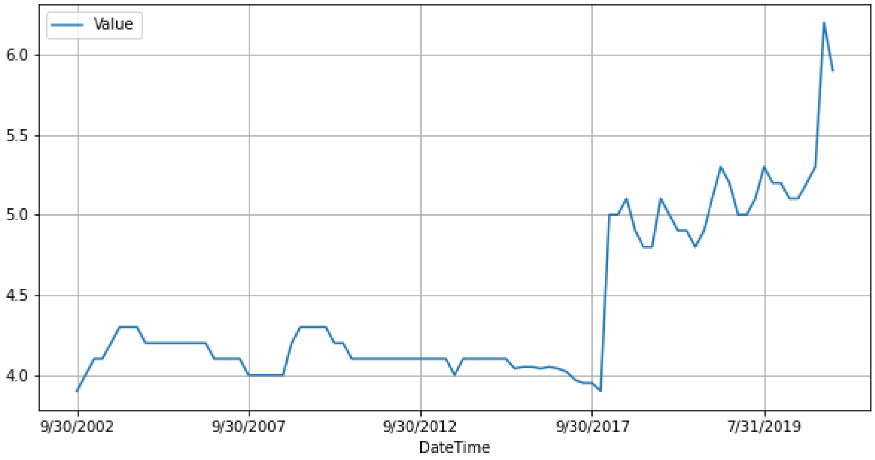
United States:



United Kingdom:



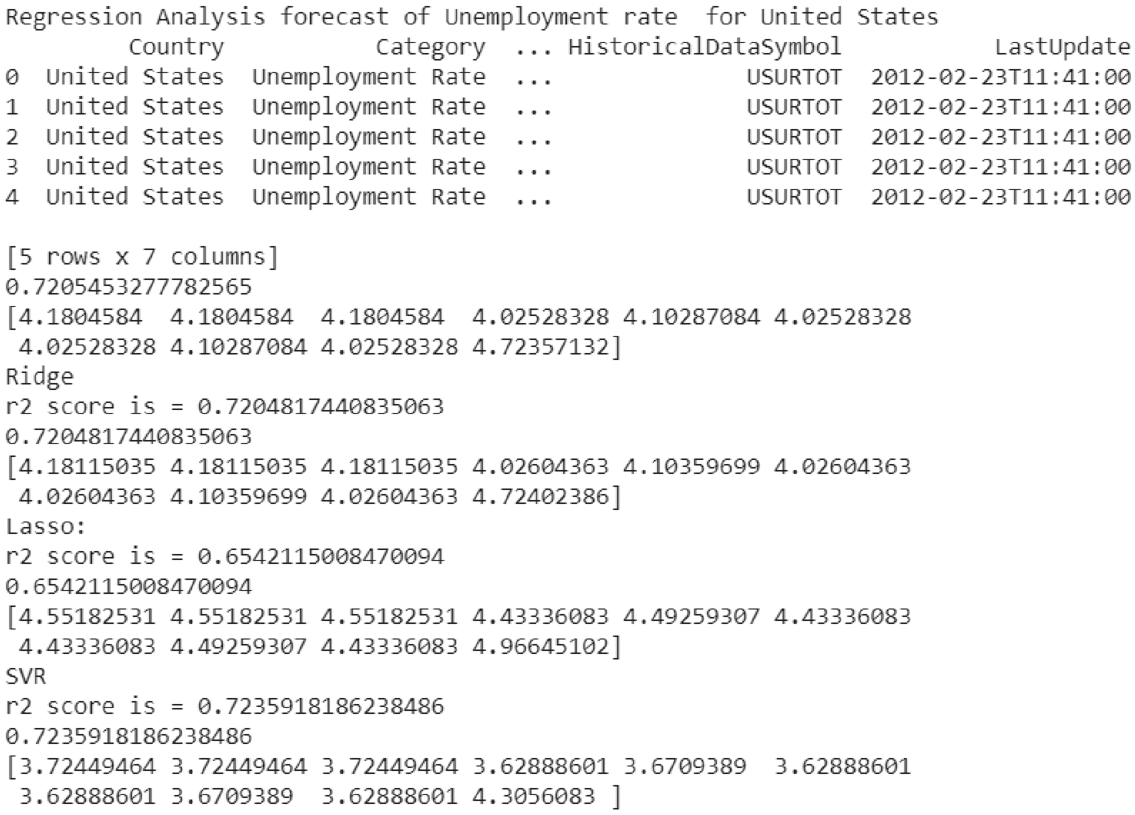
China:



Regression on Historical Unemployment data and Forecasting:

- We used the unemployment data available for US, UK, and China to perform Regression analysis and forecast for the next 10 months
- We used Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression for each country to forecast. Results can be seen below:

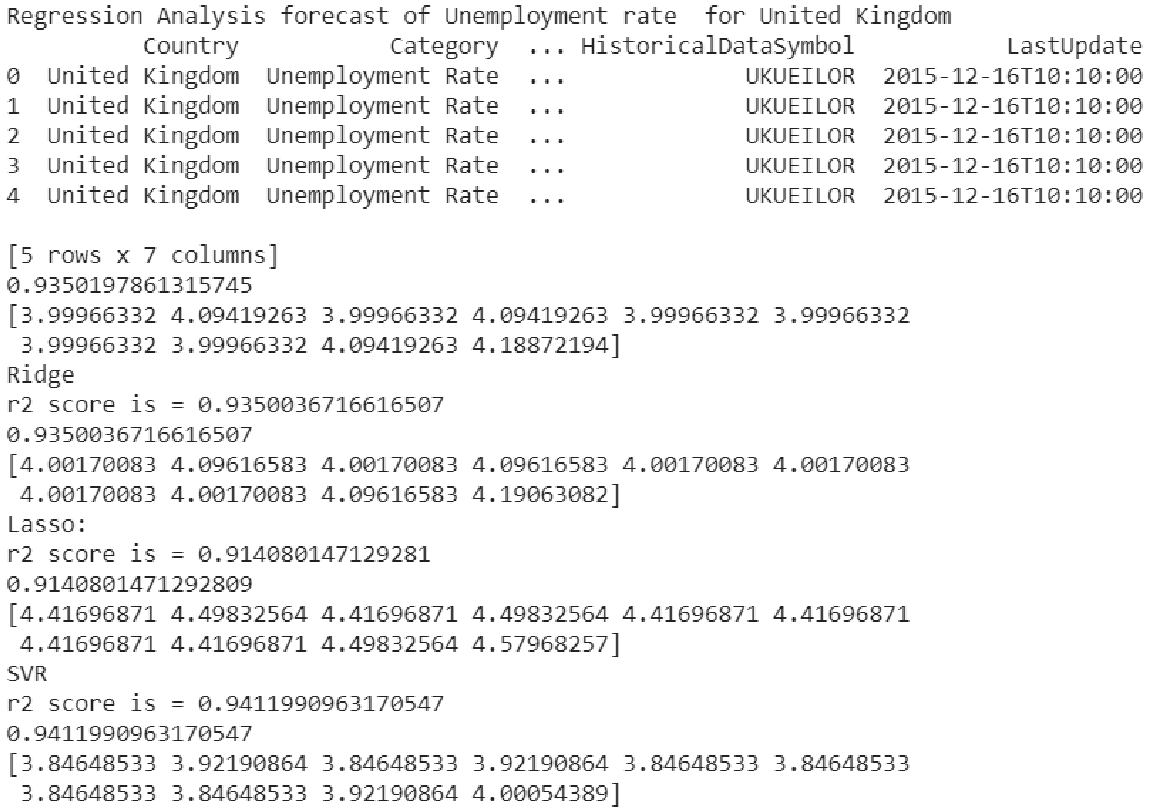
United States:



```
Regression Analysis forecast of Unemployment rate  for United States
         Country         Category  ... HistoricalDataSymbol           LastUpdate
0  United States  Unemployment Rate  ...              USURTOT  2012-02-23T11:41:00
1  United States  Unemployment Rate  ...              USURTOT  2012-02-23T11:41:00
2  United States  Unemployment Rate  ...              USURTOT  2012-02-23T11:41:00
3  United States  Unemployment Rate  ...              USURTOT  2012-02-23T11:41:00
4  United States  Unemployment Rate  ...              USURTOT  2012-02-23T11:41:00

[5 rows x 7 columns]
0.7205453277782565
[4.1804584  4.1804584  4.1804584  4.02528328 4.10287084 4.02528328
 4.02528328 4.10287084 4.02528328 4.72357132]
Ridge
r2 score is = 0.7204817440835063
0.7204817440835063
[4.18115035 4.18115035 4.18115035 4.02604363 4.10359699 4.02604363
 4.02604363 4.10359699 4.02604363 4.72402386]
Lasso:
r2 score is = 0.6542115008470094
0.6542115008470094
[4.55182531 4.55182531 4.55182531 4.43336083 4.49259307 4.43336083
 4.43336083 4.49259307 4.43336083 4.96645102]
SVR
r2 score is = 0.7235918186238486
0.7235918186238486
[3.72449464 3.72449464 3.72449464 3.62888601 3.6709389  3.62888601
 3.62888601 3.6709389  3.62888601 4.3056083 ]
```

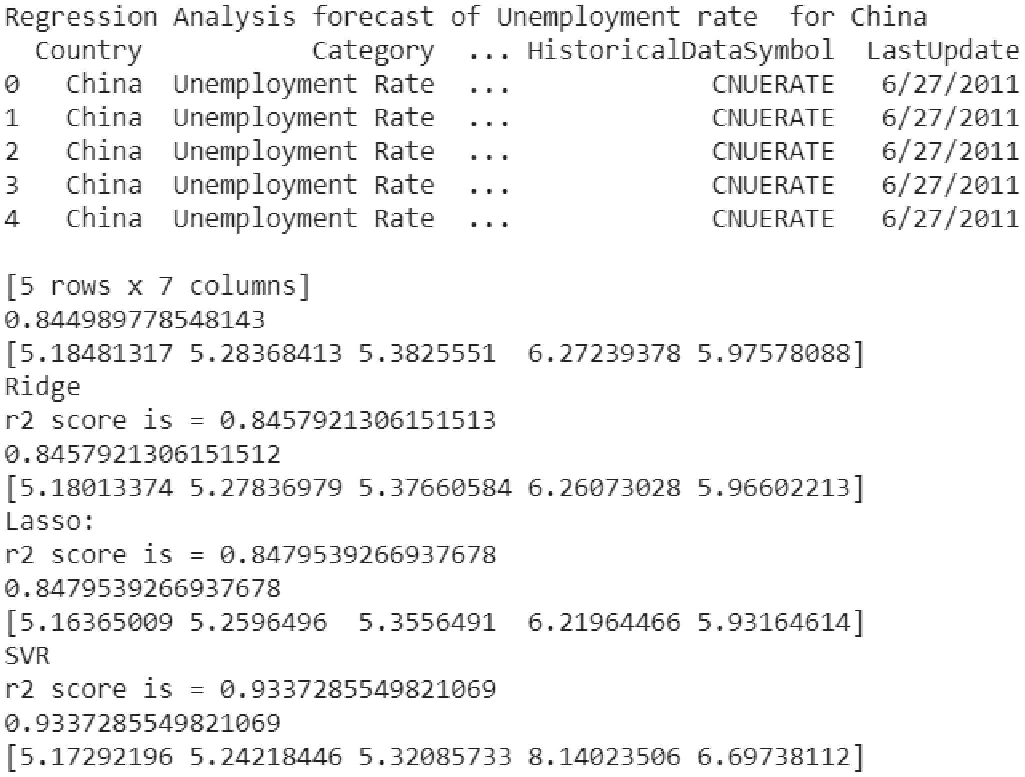
United Kingdom:



```
Regression Analysis forecast of Unemployment rate  for United Kingdom
          Country           Category  ... HistoricalDataSymbol           LastUpdate
0  United Kingdom  Unemployment Rate  ...             UKUEILOR  2015-12-16T10:10:00
1  United Kingdom  Unemployment Rate  ...             UKUEILOR  2015-12-16T10:10:00
2  United Kingdom  Unemployment Rate  ...             UKUEILOR  2015-12-16T10:10:00
3  United Kingdom  Unemployment Rate  ...             UKUEILOR  2015-12-16T10:10:00
4  United Kingdom  Unemployment Rate  ...             UKUEILOR  2015-12-16T10:10:00

[5 rows x 7 columns]
0.9350197861315745
[3.99966332 4.09419263 3.99966332 4.09419263 3.99966332 3.99966332
 3.99966332 3.99966332 4.09419263 4.18872194]
Ridge
r2 score is = 0.9350036716616507
0.9350036716616507
[4.00170083 4.09616583 4.00170083 4.09616583 4.00170083 4.00170083
 4.00170083 4.00170083 4.09616583 4.19063082]
Lasso:
r2 score is = 0.914080147129281
0.9140801471292809
[4.41696871 4.49832564 4.41696871 4.49832564 4.41696871 4.41696871
 4.41696871 4.41696871 4.49832564 4.57968257]
SVR
r2 score is = 0.9411990963170547
0.9411990963170547
[3.84648533 3.92190864 3.84648533 3.92190864 3.84648533 3.84648533
 3.84648533 3.84648533 3.92190864 4.00054389]
```

China:



```
Regression Analysis forecast of Unemployment rate  for China
  Country           Category  ... HistoricalDataSymbol  LastUpdate
0   China  Unemployment Rate  ...             CNUERATE   6/27/2011
1   China  Unemployment Rate  ...             CNUERATE   6/27/2011
2   China  Unemployment Rate  ...             CNUERATE   6/27/2011
3   China  Unemployment Rate  ...             CNUERATE   6/27/2011
4   China  Unemployment Rate  ...             CNUERATE   6/27/2011

[5 rows x 7 columns]
0.844989778548143
[5.18481317 5.28368413 5.3825551  6.27239378 5.97578088]
Ridge
r2 score is = 0.8457921306151513
0.8457921306151512
[5.18013374 5.27836979 5.37660584 6.26073028 5.96602213]
Lasso:
r2 score is = 0.8479539266937678
0.8479539266937678
[5.16365009 5.2596496  5.3556491  6.21964466 5.93164614]
SVR
r2 score is = 0.9337285549821069
0.9337285549821069
[5.17292196 5.24218446 5.32085733 8.14023506 6.69738112]
```

Conclusion:

Covid-19 has clearly made a huge impact on the economies of all the countries of the world. It is no doubt crucial to understand the respective impacts and take measures to bounce back. Our analysis of Stock data and Unemployment rates of three major economies of the world showed a great impact of the pandemic, to steer us towards making betters decisions. Similar analyses can be done on other indexes and this can also be done for other countries as well to understand the country level impacts with respect to these indicators.