Epidemiology of COVID-19 Across Countries

Yohan Jhaveri, Haniel Paulos, Katherine Walton

Fall 2019

1 Collaboration statement

We consulted documentation for the various libraries used in the project. This source book was also referenced: Data Science from Scratch Python (Haniel).

2 Problem Description

With the COVID-19 pandemic severely affecting countries across the globe, we thought it would be interesting to find country-specific determinants that most significantly impact the incidence of the virus. This includes observing factors like spread, mortality and, reproductive rate (R_0)

2.1 Tasks:

1. Effect of demographic factors on incidence
2. Effect of healthcare factors on incidence
3. Finding highly vulnerable countries
4. Finding the most effective safety measures

2.2 Demographic Factors:

- GDP (PPP)
- GDP per capita (PPP)
- Population
- Population Density (per km^3)
- Population over 60 years (% of population)
- Urban Population (% of population)
- Median Age
- Life Expectancy
- Smokers (% of population)
- Hospital Beds (per 1000 people)
- Healthcare Expenditure per capita
- Access to Sanitation Facilities (% of population)

2.3 Epidemiological Factors:

- Cases
- Deaths
- Cases Daily Growth Rate (%)
- Deaths Daily Growth Rate (%)
- Mortality Rate (%)
- Cases (per 1000 people)
- Deaths (per 1000 people)

3 Data Description

We aggregated our dataset from various sources. The dataset has a total of 21 attributes and 183 tuples. Our combined dataset can be found at: `https://github.com/yohanjhaveri/cs470/blob/master/countries.csv`

3.1 COVID-19 Data

For COVID-19 specific data on confirmed cases and deaths, we used the *Johns Hopkins* Dataset on Github which sources its data from various trusted sources like the WHO, CDC and ECDC. Link: `https://github.com/CSSEGISandData/COVID-19`

3.2 Demographic Data

Demographic data was sourced from the website called *Gapminder*. This website has an abundant archive of country-specific information with datasets on education, healthcare, economy, environment, infrastructure and energy. Link: `www.gapminder.org/data`

3.3 Safety Measures Data

Safety measures taken by various governments in response to COVID-19 was collected by *The Humanitarian Data Exchange* along with the date on which it was implemented. Link: `https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset`

4 Data Pre-processing

The dataset had to be preprocessed in order to improve model and algorithm performance.

4.1 Grouping by Country

For certain countries, COVID-19 statistics were given by Province / State. We wanted to have a single tuple represent the entire country for greater consistency and we therefore grouped the tuples by country name and combining the tuples by taking the sum of their attribute values.

4.2 Data Extraction

Through the COVID-19 specific and demographic data, we were able to extract attributes through a combination of other attributes. Some extracted attributes were:

- GDP per capita (PPP) = GDP (PPP) / Population
- Cases (per 1000 people) = Cases * 1000 / Population
- Deaths (per 1000 people) = Deaths * 1000 / Population
- Mortality Rate (%) = Deaths * 100 / Cases
- Cases Daily Growth Rate (%)
- Deaths Daily Growth Rate (%)

$$r = (\sqrt[n]{\frac{P_n}{P_0}} - 1) * 100 \tag{1}$$

Equation 1: Daily Growth Rate Formula

Where:

1. P_0 is the number of cases on the day of the first case.
2. P_t is the number of cases on the current day
3. n is the number of days between the first case and the current day

4.3 Manual Pre-processing

Some of the data points under the country variable were not countries at all, but cruise ships such as Diamond Princess and MS Zaandam. We chose to purge these instances from our data completely because we want to look at demographic information for countries alone. We also decided to get rid of the Latitude and Longitude attributes that came along with the COVID-19 data.

A challenge we faced that had to be solved manually was the different naming conventions for countries between datasets. For example, "US" vs "United States". As a result, we had to create dictionaries to represent the conversions between the naming conventions of the different datasets for consistency purposes.

4.4 Missing Data

Some values in the dataset were missing. The countries with missing feature values can be seen below:

Feature	Number Missing
Population	9
Life Expectancy	14
Median Age	16
GDP (PPP)	10
Smokers	51
Population over 60 years	16

Due to the fact that each of these variables are continuous, we decided to impute with the mean of the column. For continuous variables without outliers, imputing with the mean makes the most sense.

4.5 Scaling

Due to the fact that K-Means Clustering is a distance-based learner, it can large variations can heavily skew the distance metric leading to the formation of worse clusters. To tackle this problem, we chose to standardize each attribute using standard scaling:

$$z = \frac{x - \mu}{\sigma}$$

.

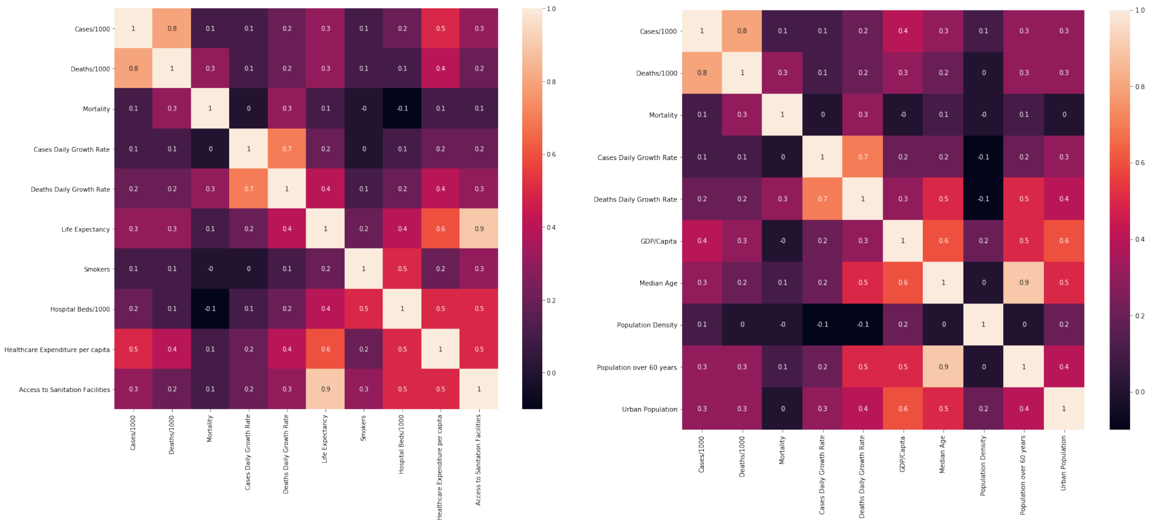
4.6 Visualizing Data

Data was visualized after processing in Python and Tableau. Below in the appendix, I have included various cuts of the data we looked at. Once we applied the K-means algorithm, we also visually explored some of the relations between attributes with the additional class layer.

5 Data Mining

5.1 Correlation Analysis

To find the effects of demographic and healthcare factors on the incidence of COVID-19, we decided to use a correlation analysis on the dataset to find attributes that highly correlated with epidemiological factors. Figure 1 represents the correlation matrices for healthcare and demographic factors.



(a) Healthcare

(b) Demographic

Figure 1: Data Correlation Matrix

5.2 K Means Clustering

To find countries that were the most vulnerable to COVID-19 we decided to use K Means Clustering to cluster countries according to epidemiological data and recognize the demographic factors that contributed to worse spread. By detecting these high-risk factors, we evaluated the vulnerability of different countries based on these factors.

5.3 Multi-Level Regression Models with K Means

We created growth,death, and mortality rate models with respect to the attributes that interested us. Then we included identical models with the inclusion of the K-means cluster class to gauge whether the fit of the models are improved upon.

Models:
VarsOfInterest ⇒ $\beta_0 + \beta_1 * Population + \beta_2 * Populationdensity \beta_3 * UrbanPopulation + \beta_4 * LifeExpectancy + \beta_5 * MedianAge + \beta_6 * Smokers + \beta_7 * PopOver60 + \beta_8 * HospitalBedsPer100 + \beta_9 * HealthCareExpenditure + \beta_{10} * AccessToSanitationFacilities$

(Cases Standard Model) ⇒ $CasesDailyGrowthRate = VarsOfInterest + u$

(Death Standard Model) ⇒ $DeathDailyGrowthRate = VarsOfInterest + u$

(Morality Standard Model) ⇒ $MortalityRate = VarsOfInterest + u$

(DG)Rate K-Models = "Given" Standard model + $\beta_{11} * KMeansCluster$

5.4 Apriori

To find the best combination of safety measures to slow down the spread, we decided to use Apriori to find frequent combinations of safety measures present in countries with the lowest growth and mortality rates.

6 Results

6.1 Statistically Significant Correlations:

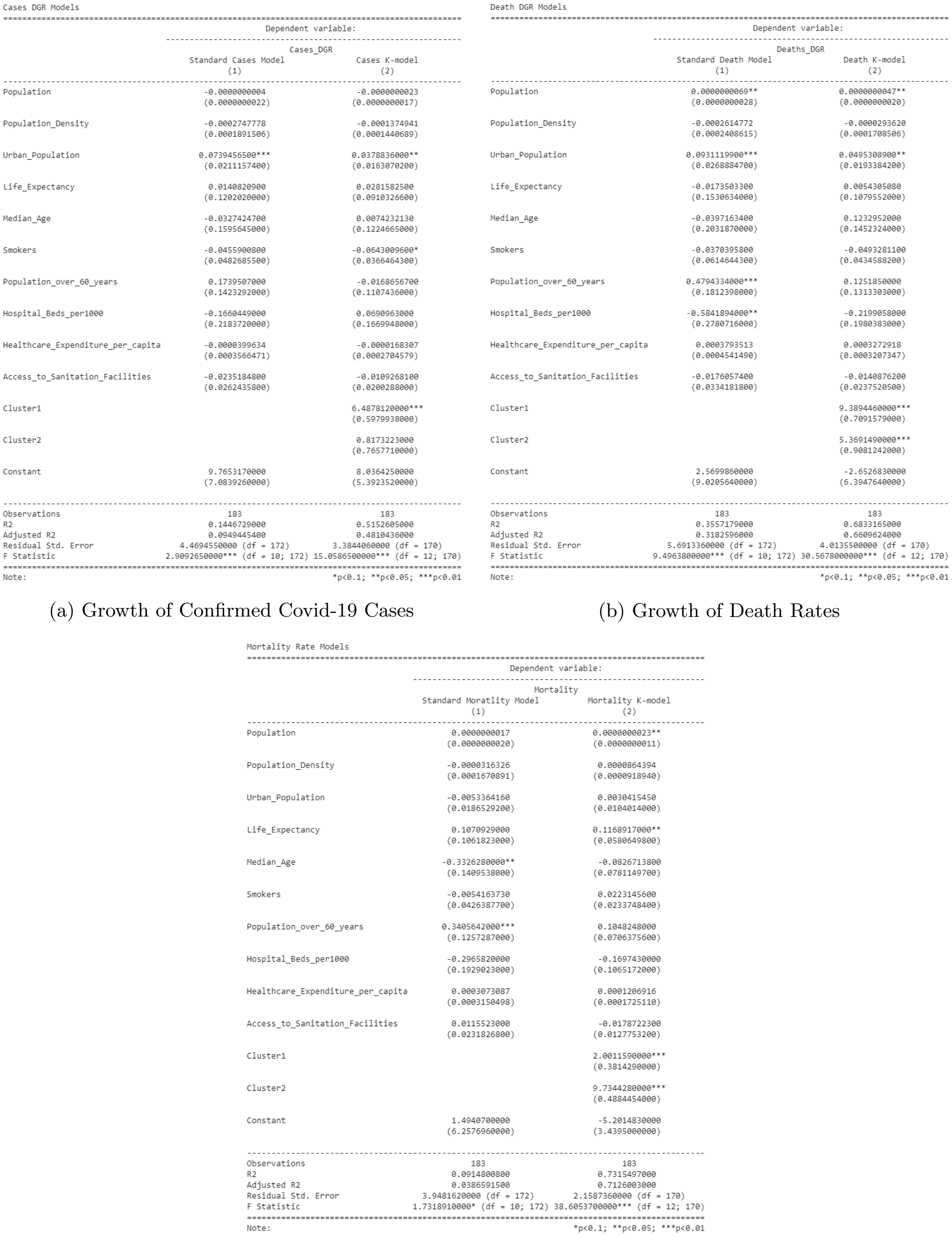
1. Cases/1000 vs Healthcare Expenditure per capita: 0.5
2. Deaths Daily Growth Rate vs Life Expectancy: 0.4
3. Cases/1000 vs GDP per capita: 0.4
4. Deaths Daily Growth Rate vs Population over 60 years: 0.5
5. Deaths Daily Growth Rate vs Urban Population: 0.4

6.2 Multilevel Regression Analysis

In the standard Cases DGR model, the general attributes (population, life expectancy, etc...) in the model are not statistically significant except for Urban Population (at 1 % significance). In the Cases DGR K-model, class 1 is statistically significant. This means cluster 1 is significantly different from cluster 0 in terms of growth.

In the standard Deaths DGR model, population, urban population and population-over-60years are significant at 1 percent level and population and hospital beds per 1000 are at 5 percent level. In the deaths DGR K-model, population and urban-population are both significant at the 5% level. Both class 1 and class 2 are significant at the 1% level. This may mean that the clusters are dissimilar in terms of population over 60 and hospital beds.

In the standard mortality model, we found Median Age to be significant at the 5% level and population-over-60 at the 1% level. In the mortality K-model, neither Median Age and Population over 60 are significant anymore. Instead, population and life expectancy are significant at the 5% level. Within the mortality k-model class 1 and 2 are significant at the 1% level. This may show why the clusters are more separated in terms of mortality.



```
Cases DGR Models
===============================================================================================
                                                       Dependent variable:
                                        -------------------------------------------------------
                                                            Cases_DGR
                                          Standard Cases Model               Cases K-model
                                                  (1)                             (2)
-----------------------------------------------------------------------------------------------
Population                                   -0.0000000004                  -0.0000000023
                                            (0.0000000022)                 (0.0000000017)

Population_Density                           -0.0002747778                  -0.0001374941
                                            (0.0001891506)                 (0.0001440689)

Urban_Population                            0.0739456500***                0.0378836000**
                                            (0.0211157400)                 (0.0163070200)

Life_Expectancy                               0.0140820900                   0.0281582500
                                            (0.1202020000)                 (0.0910326600)

Median_Age                                   -0.0327424700                   0.0074232130
                                            (0.1595645000)                 (0.1224665000)

Smokers                                      -0.0455900800                 -0.0643009600*
                                            (0.0482685500)                 (0.0366464300)

Population_over_60_years                      0.1739507000                  -0.0168656700
                                            (0.1423292000)                 (0.1107436000)

Hospital_Beds_per1000                        -0.1660449000                   0.0690963000
                                            (0.2183720000)                 (0.1669948000)

Healthcare_Expenditure_per_capita            -0.0000399634                  -0.0000168307
                                            (0.0003566471)                 (0.0002704579)

Access_to_Sanitation_Facilities              -0.0235184800                  -0.0109268100
                                            (0.0262435800)                 (0.0200288000)

Cluster1                                                                   6.4878120000***
                                                                           (0.5979938000)

Cluster2                                                                     0.8173223000
                                                                           (0.7657710000)

Constant                                      9.7653170000                   8.0364250000
                                            (7.0839260000)                 (5.3923520000)

-----------------------------------------------------------------------------------------------
Observations                                       183                            183
R2                                            0.1446729000                   0.5152605000
Adjusted R2                                   0.0949445400                   0.4810436000
Residual Std. Error                     4.4694550000 (df = 172)        3.3844060000 (df = 170)
F Statistic                     2.9092650000*** (df = 10; 172) 15.0586500000*** (df = 12; 170)
===============================================================================================
Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

(a) Growth of Confirmed Covid-19 Cases

```
Death DGR Models
===============================================================================================
                                                       Dependent variable:
                                        -------------------------------------------------------
                                                            Deaths_DGR
                                          Standard Death Model               Death K-model
                                                  (1)                             (2)
-----------------------------------------------------------------------------------------------
Population                                  0.0000000069**                 0.0000000047**
                                            (0.0000000028)                 (0.0000000020)

Population_Density                           -0.0002614772                  -0.0000293620
                                            (0.0002408615)                 (0.0001708506)

Urban_Population                            0.0931119900***                0.0495308900**
                                            (0.0268884700)                 (0.0193384200)

Life_Expectancy                              -0.0173503300                   0.0054305080
                                            (0.1530634000)                 (0.1079552000)

Median_Age                                   -0.0397163400                   0.1232952000
                                            (0.2031870000)                 (0.1452324000)

Smokers                                      -0.0370395800                  -0.0493281100
                                            (0.0614644300)                 (0.0434588200)

Population_over_60_years                    0.4794334000***                  0.1251850000
                                            (0.1812398000)                 (0.1313303000)

Hospital_Beds_per1000                       -0.5841894000**                 -0.2199058000
                                            (0.2780716000)                 (0.1980383000)

Healthcare_Expenditure_per_capita             0.0003793513                   0.0003272918
                                            (0.0004541490)                 (0.0003207347)

Access_to_Sanitation_Facilities              -0.0176057400                  -0.0140876200
                                            (0.0334181800)                 (0.0237520500)

Cluster1                                                                   9.3894460000***
                                                                           (0.7091579000)

Cluster2                                                                   5.3691490000***
                                                                           (0.9081242000)

Constant                                      2.5699860000                  -2.6526830000
                                            (9.0205640000)                 (6.3947640000)

-----------------------------------------------------------------------------------------------
Observations                                       183                            183
R2                                            0.3557179000                   0.6833165000
Adjusted R2                                   0.3182596000                   0.6609624000
Residual Std. Error                     5.6913360000 (df = 172)        4.0135500000 (df = 170)
F Statistic                     9.4963800000*** (df = 10; 172) 30.5678000000*** (df = 12; 170)
===============================================================================================
Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

(b) Growth of Death Rates

```
Mortality Rate Models
===============================================================================================
                                                       Dependent variable:
                                        -------------------------------------------------------
                                                            Mortality
                                       Standard Moratlity Model          Mortality K-model
                                                  (1)                            (2)
-----------------------------------------------------------------------------------------------
Population                                    0.0000000017                 0.0000000023**
                                            (0.0000000020)                 (0.0000000011)

Population_Density                           -0.0000316326                   0.0000864394
                                            (0.0001670891)                 (0.0000918940)

Urban_Population                             -0.0053364160                   0.0030415450
                                            (0.0186529200)                 (0.0104014000)

Life_Expectancy                               0.1070929000                 0.1168917000**
                                            (0.1061823000)                 (0.0580649800)

Median_Age                                  -0.3326280000**                 -0.0826713800
                                            (0.1409538000)                 (0.0781149700)

Smokers                                      -0.0054163730                   0.0223145600
                                            (0.0426387700)                 (0.0233748400)

Population_over_60_years                    0.3405642000***                  0.1048248000
                                            (0.1257287000)                 (0.0706375600)

Hospital_Beds_per1000                        -0.2965820000                  -0.1697430000
                                            (0.1929023000)                 (0.1065172000)

Healthcare_Expenditure_per_capita             0.0003073087                   0.0001206916
                                            (0.0003150498)                 (0.0001725110)

Access_to_Sanitation_Facilities               0.0115523000                  -0.0178722300
                                            (0.0231826800)                 (0.0127753200)

Cluster1                                                                   2.0011590000***
                                                                           (0.3814290000)

Cluster2                                                                   9.7344280000***
                                                                           (0.4884454000)

Constant                                      1.4940700000                  -5.2014830000
                                            (6.2576960000)                 (3.4395000000)

-----------------------------------------------------------------------------------------------
Observations                                       183                            183
R2                                            0.0914800800                   0.7315497000
Adjusted R2                                   0.0386591500                   0.7126003000
Residual Std. Error                     3.9481620000 (df = 172)        2.1587360000 (df = 170)
F Statistic                       1.7318910000* (df = 10; 172) 38.6053700000*** (df = 12; 170)
===============================================================================================
Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

(c) Mortality Rates

Figure 2: Comparison of Growth, Death, and Mortality Models with K-means classes

As we know the classes are from the K-means clustering based upon the similarity of countries based on the whole matrix of attributes. The effectiveness of including an attribute based upon this clustering can be seen in some of the models. In the cases DGR model, the class variable can help inform the growth level rate of a country. The cases DGR rates of countries in cluster 1 and 2 are significantly different from each other. In the death model, class 1 is significantly different from class 0. In the mortality model, class 2 is significantly different from class 0 at a 5% level. Altogether, clustering has produced interesting results in this multi-level regression and overall has increased the fit of these models. From the standard to k-model, the adjusted R-squared values goes from 14% to 51% in the Cases DGR models, 35% to 68% in the deaths DGR models, and 9% to 73% in the mortality models. The general morality model is the only model without a F-stat significant at a 1% level. Overall, the clusters are good predictors of morality, followed by death and growth.

6.3 Dataset Clusters:

These are clusters formed using epidemiological factors.

1. Cluster 1: Afghanistan, Albania, Andorra, Argentina, Armenia, Austria, Bangladesh, Belarus, Bolivia, Bosnia and Herzegovina, Brazil, Bulgaria, Burkina Faso, Cameroon, Canada, Chile, Colombia, Congo (Kinshasa), Cote dIvoire, Croatia, Cuba, Cyprus, Czechia, Denmark, Djibouti, Dominican Republic, Ecuador, El Salvador, Estonia, Finland, Germany, Ghana, Greece, Guinea, India, Indonesia, Iran, Iraq, Ireland, Israel, Kenya, Kuwait, Kyrgyzstan, Latvia, Lithuania, Luxembourg, Mali, Mexico, Moldova, Morocco, Netherlands, New Zealand, Niger, Nigeria, North Macedonia, Norway, Pakistan, Panama, Peru, Poland, Portugal, Qatar, Romania, Russia, Saudi Arabia, Serbia, Slovenia, Somalia, South Africa, Spain, Switzerland, Tanzania, Tunisia, Turkey, US, Ukraine, Uruguay, Uzbekistan, Venezuela

2. Cluster 2: Australia, Azerbaijan, Bahrain, Benin, Bhutan, Botswana, Brunei, Burma, Cabo Verde, Cambodia, Central African Republic, Chad, China, Congo (Brazzaville), Costa Rica, Dominica, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Fiji, Gabon, Georgia, Grenada, Guatemala, Guinea-Bissau, Haiti, Holy See, Iceland, Jamaica, Japan, Jordan, Kazakhstan, Korea, South, Kosovo, Laos, Lebanon, Libya, Liechtenstein, Madagascar, Malaysia, Maldives, Malta, Mauritius, Monaco, Mongolia, Montenegro, Mozambique, Namibia, Nepal, Oman, Papua New Guinea, Paraguay, Rwanda, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Singapore, Slovakia, South Sudan, Sri Lanka, Taiwan*, Thailand, Timor-Leste, Uganda, United Arab Emirates, Vietnam, West Bank and Gaza, Western Sahara, Yemen, Zambia

3. Cluster 3: Algeria, Angola, Antigua and Barbuda, Bahamas, Barbados, Belgium, Belize, Burundi, Egypt, France, Gambia, Guyana, Honduras, Hungary, Italy, Liberia, Malawi, Mauritania, Nicaragua, Philippines, San Marino, Sudan, Suriname, Sweden, Syria, Togo, Trinidad and Tobago, United Kingdom, Zimbabwe

6.4 Best Safety Measures:

These are safety measures used frequently among countries that have maintained a daily growth rate of cases at less than 8%

1. Recommendation 1: Health screenings in airports and border crossings

2. Recommendation 2: Domestic travel restrictions

3. Recommendation 3: International flights suspension

4. Recommendation 4: Schools closure

5. Recommendation 5: Border closure
6. Recommendation 6: Awareness campaigns
7. Recommendation 7: Public services closure
8. Recommendation 8: Strengthening the public health system
9. Recommendation 9: Emergency administrative structures activated or established

7 Future Work

1. Research surrounding COVID-19 is constantly evolving, which could help us understand our findings more.
2. Datasets with more features would allow for more exploration of how variables interact with each other
3. It would be useful to know the attributes that determine the vulnerability of a country to COVID-19

We hoped to find a linear combination of attribute values which determined the vulnerability of certain countries to the disease.

The higher the score, the greater the vulnerability. The lower the score, the lesser the vulnerability.

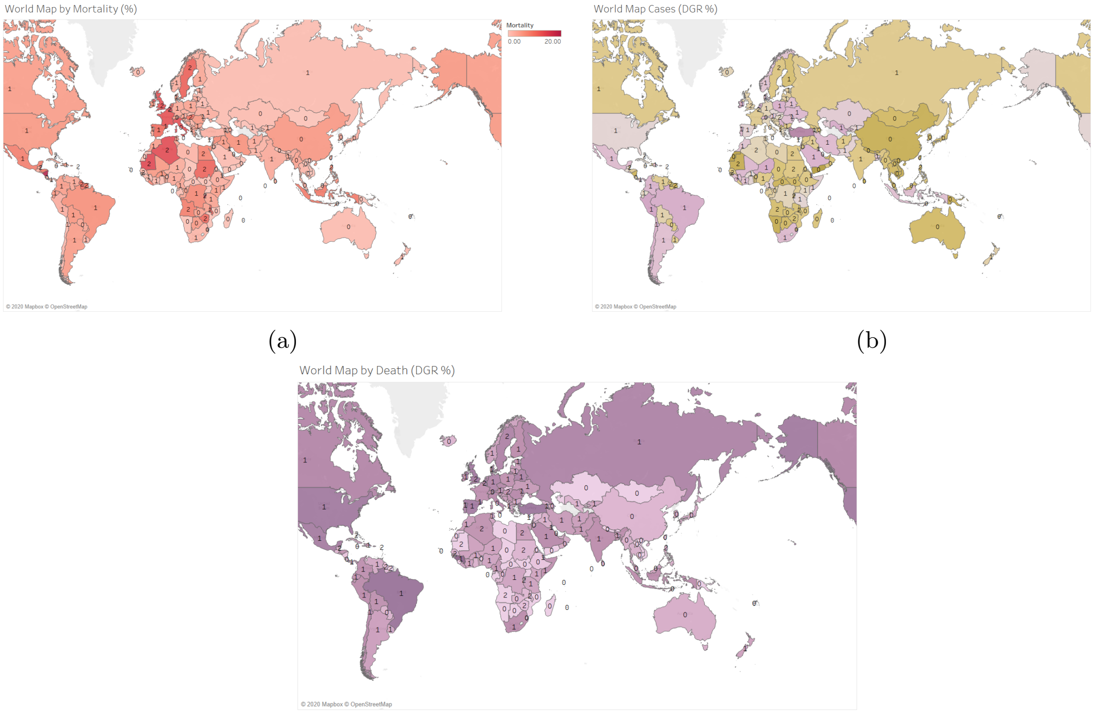
The challenge was to find the coefficients to each attribute with attributes that caused faster spread having positive coefficients and attributes that caused slower spread having negative coefficients.

While finding the sign of each coefficient is possible intuitively, the magnitude of the coefficients is harder to derive.

8 Appendix

8.1 Figures

Maps:


World Map by Mortality (%)
Mortality
0.00
20.00
© 2020 Mapbox © OpenStreetMap
World Map Cases (DGR %)
Cases Daily Growt..
0.00
31.40
© 2020 Mapbox © OpenStreetMap
World Map by Death (DGR %)
© 2020 Mapbox © OpenStreetMap

(a)

(b)

(c)

(d) World Maps for Epidemiological Rates

Plots:

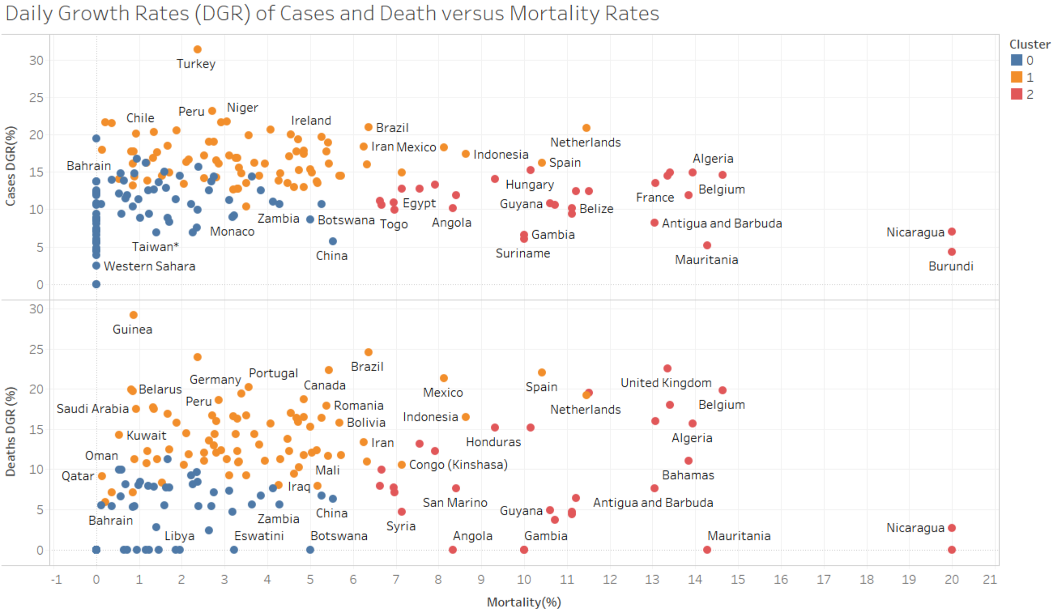


Figure 4

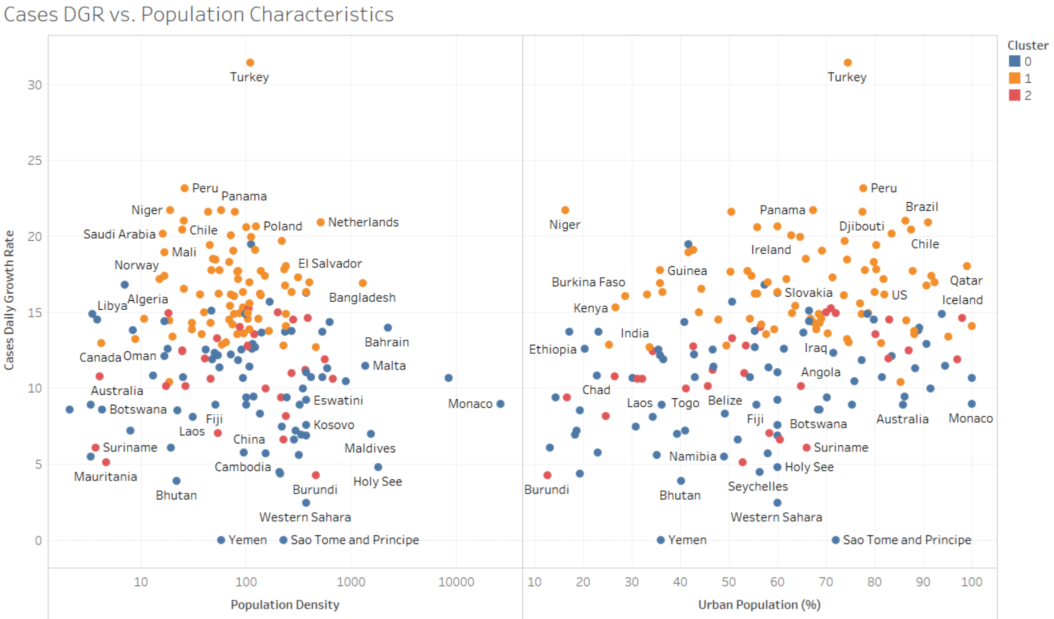


Figure 5

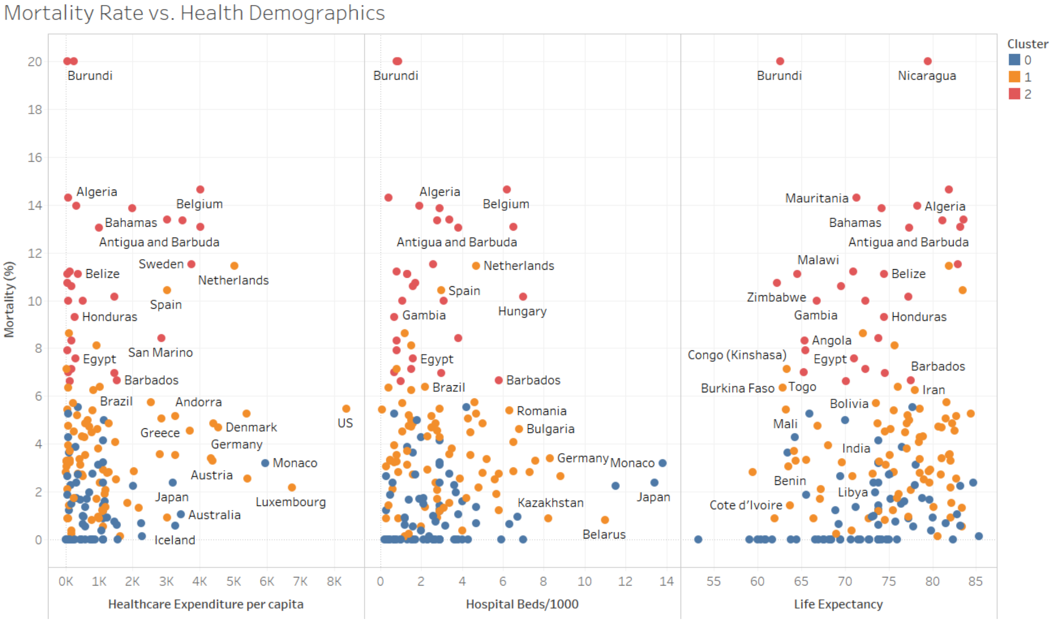


Figure 6