ECE 579A Progress Report

Investigating Relationship between Covid-19 and Different Indexes for Happiness, Freedom and Population.

Derrell D'Souza
Department of Electrical and Computer Engineering
University of Victoria

EXECUTIVE SUMMARY

This report presents work related to the identification of the relationship between different factors affecting happiness, freedom and population and the Covid-19 situation in each country. COVID-19 pandemic has severely affected the health, safety and overall well being of an individual and community. It is therefore essential to investigate whether there is a specific relationship between the way COVID-19 is unfolding itself in different countries and their happiness, freedom and overall population-based features.

In our analysis, we specifically cluster 140 countries into two different groups and then try to investigate how these clusters behave according to different features. First, we calculate three different parameters i.e. average infection parameter, average mortality parameter and average recovery parameter that define the overall situation of the country during COVID-19. We perform PCA on these countries with six features i.e., the calculated parameters along with the total number of infections, deaths and recovered cases and use the three larget principal components for clustering using K-means algorithm. Then we visualize the clusters with different indexes and their features.

Our results indicate no clear relationship between the COVID-19 situation and the features. However, our analysis shows that countries with better life satisfaction, life expectancy and economy are affected most due to COVID-19.

Contents

List of Tables

List of Figures

List of Acronyms

AIP Average Infection Parameter

AMP Average Mortality Parameter

ARP Average Recovery Parameter

COVID-19 Corona Virus Disease 2019

EIU Economist Intelligence Unit

GWP Gallup World Poll

HFI Human Freedom Index

HLE Healthy Life Expectancy

JHU John Hopkins University

PCA Principal Component Analysis

PPP Purchasing Power Parity

SARS-CoV-2 Severe Acute Respiratory Syndrome Corona Virus 2

WHO World Health Organization

Chapter 1

Introduction

COVID-19 is an infectious disease caused by a new coronavirus first identified in December 2019 during a local breakout of pneumonia cases (with causes initially unknown) in Wuhan (Hubei, China). After identifying and isolating the pathogen, it was initially named as 2019 novel coronavirus (2019-nCoV). Later, it was officially called as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the WHO. Coronaviruses are a family of viruses known that cause respiratory infections. Compared with the SARS-CoV which resulted in an outbreak of SARS in 2003, it was found that SARS-CoV-2 had a stronger transmission capacity. On 30 January 2020, SARS-CoV-2 outbreak was declared as a public health emergency of international concern. Currently, human-to-human transmission of SARS-CoV-2 accounts for most infections worldwide. There is no vaccine yet to prevent COVID-19, and no specific treatment for it, other than managing the symptoms [1, 2].

1.1 Motivation

A public health emergency like SARS-CoV-2 has important effects on the health, safety and overall well-being of both individuals and communities. In an individual, emergencies may cause confusion, insecurity, emotional isolation and stigma whereas in communities it may cause economic loss, work and school closures, lack of resources for medical response, and scarce distribution of necessities. These effects may trigger a range of emotional reactions (such as distress or psychiatric conditions), morbid behaviours (such as excessive substance abuse), and non-compliance with public health directives (such as home confinement and vaccination) in people who get the disease

and in the general population. Extensive research in disaster mental health has recognized the ubiquity of emotional distress in the affected population, which will be also evident within populations affected by the Covid-19 pandemic [3].

The government determined relevant actions to flatten the curve, such as social distancing, restricting personal lives and private quarantining thousands of people, are a burden to the community. A recent work about the psychological impact of quarantine indicates the psychological strain on those who are not allowed or unable to participate in social life. Separation from loved ones, the loss of freedom, uncertainty over the status of the disease, and boredom can, on occasion, create dramatic effects [4, 5].

The COVID-19 outbreak has also aggravated threats to independent reporting. This comes during a time where trusted, accurate, impartial and timely information is key to fighting the pandemic. The pandemic has ravaged newsrooms, increased the pressure on media freedom, produced a lot of false information and has put the lives of journalists at risk [6].

1.2 Objective

With respect to the cause, scope and the severity, COVID-19 pandemic is very different from crisis in the past [7]. Therefore, observations relative to the health, safety, happiness and freedom motivates research into factors governing the response of different countries to the pandemic.

The objective of this analysis is to identify relationship between different factors influencing life of people i.e., happiness, freedom and population based parameters and the COVID-19 situation in different countries using different datasets relative to happiness and freedom. We first perform cluster analysis to group countries behaving similarly during the COVID-19 pandemic and then try to identify relationship between these clusters and their score for features in these datasets.

1.3 Report Organization

The rest of this report is organized as follows: Since we are using multiple datasets for our analysis, a description of the datasets used is provided in Chapter 2. We also describe the various pre-processing operations, calculations and normalization operations performed on these datasets. In Chapter 3, we describe our cluster analysis

approach where we use PCA for dimensionality reduction and K-means for clustering. Subsequently, we visualize the clusters with respect to features based on different indexes. In Chapter 4, we state our conclusions and try to infer our results.

Chapter 2

Data Collection and Preprocessing

In this chapter, we describe the different datasets used in our analysis as well as the preprocessing operations required to transform the available data into data useful for our analysis.

2.1 Datasets

We have used six different datasets to perform our analysis. Each of these datasets is explained below:

2.1.1 COVID-19 Data Repository (JHU CSSE)

The COVID-19 data repository contains data used by 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) [8]. We use the time series summary tables for confirmed, death and recovered cases. Each of these time series tables consists of latitude and longitude information and the cumulative reported cases for each country (for some countries data is reported at the province level) since January $22^{\text{nd}}, 2020$.

2.1.2 World Happiness Report, 2020

The World Happiness Report is a landmark survey of the state of global happiness that ranks 153 countries by how happy their citizens perceive themselves to be [9]. The National average life evaluations are carried out in terms of six key variables: GDP per capita, social support, healthy life expectancy, freedom to make life choices,

generosity and perception of corruption. Some of the important features of this dataset that we are interested in exploring include:

- The Happiness Score for each country based on answers to the main life evaluation question asked in the Gallup World Poll (GWP).
- Logged GDP per capita represents purchasing power parity (PPP) at constant 2011 international dollar prices.
- Social support (or having someone to count on in times of trouble) represents the national average of the binary responses (either 0 or 1) to the GWP question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
- Healthy Life Expectancy (HLE) at birth is based on the data extracted from the World Health Organization's (WHO) Global Health Observatory data repository.
- Freedom to make life choices is the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

2.1.3 Human Freedom Index, 2019

The Human Freedom Index (HFI) report presents the state of human freedom in the world based on a broad measure that encompasses personal, civil, and economic freedom. This report is claimed to be a resource that can help understand relationships between freedom and other social and economic phenomena, as well as the ways in which the various dimensions of freedom interact with one another. The report is co-published by the Cato Institute, the Fraser Institute, and the Liberales Institut at the Friedrich Naumann Foundation for Freedom and uses 76 different indicators in the following areas [10].

- Rule of Law
- Security and Safety
- Movement
- Religion

- Association, Assembly, and Civil Society
- Expression and Information
- Identity and Relationships
- Size of Government
- Legal System and Property Rights
- Access to Sound Money
- Freedom to Trade Internationally
- Regulation of Credit, Labor, and Business

These indicators are grouped into two sub-indexes personal freedom score and economic freedom score which are equally weighted and averaged to calculate the human freedom index.

2.1.4 World Press Freedom Index, 2020

The World Press Freedom Index every year since 2002 by Reporters Without Borders (RSF) that ranks 180 countries and regions as per the level of freedom available to journalists. It highlights the situation of medium based on evaluation of pluralism, independence of the media, quality of legislative framework and safety of journalists in each country and region. The degree of freedom available to journalists in 180 countries and regions is determined by pooling the responses of experts to a questionnaire devised by RSF. This qualitative analysis is combined with quantitative data on abuses and acts of violence against journalists during the period evaluated [11].

2.1.5 Index of Economic Freedom, 2020

The Index of Economic Freedom is an annual index and ranking created in 1995 by The Heritage Foundation and The Wall Street Journal to measure the degree of economic freedom in different countries. It focuses on four key aspects of the economic and entrepreneurial environment over which governments typically exercise policy control.

- Rule of Law (property rights, government integrity, judicial effectiveness)

- Government Size (government spending, tax burden, fiscal health)
- Regulatory Efficiency (business freedom, labour freedom, monetary freedom)
- Open Markets (trade freedom, investment freedom, financial freedom)

In assessing conditions in these four categories, the Index measures 12 specific components of economic freedom, each of which is graded on a scale from 0 to 100. Scores on these 12 components of economic freedom, which is calculated from several sub-variables, are equally weighted and averaged to produce an overall economic freedom score for each economy [12].

2.1.6 Democracy Index, 2019

The Democracy Index is an index compiled by the Economist Intelligence Unit (EIU) to measure the state of democracy in different countries [13]. The Democracy Index is based on five categories:

- Electoral process and pluralism
- Functioning of government
- Political participation
- Political culture
- Civil liberties

The index is based on a rating of 0 to 10 scale based on average of five category indexes with each category haviing a rating on a 0 to 10 scale. Based on its scores on a range of indicators within these categories, each country is then itself classified as one of four types of regime: “full democracy”, “flawed democracy”, “hybrid regime” or “authoritarian regime”.

2.1.7 World Population data, 2020

The World Population data, 2020 is the data provided by Worldometer [14] which complies the data from the United Nations Population Division estimates and contains data related to population, yearly change of population, net change in population, density, land area, migrants, fertility rate, median age, urban population and world share .

2.2 Data Pre-processing

Data pre-processing is necessary to convert the available data into a useful format for analysis. We perform the following pre-processing operations:

2.2.1 Data Cleaning

Before converting the data to a useable format, it is essential to clean the data to ensure correct, consistent and usable data. This is achieved by identifying inaccurate, incomplete, corruption and irrelevancy in the given data and either correcting or deleting them from the datasets.

Below, we show the data cleaning operations on our datasets:

- Removing the latitude and longitude information from the time series datasets of COVID-19 data repository, which are irrelevant for our analysis.
- Removing the state information and cumulating the data for countries from the time series data where cases have been reported on province/state level.
- Dropping the data corresponding to countries which have information present in some datasets and absent in other datasets before merging our datasets.
- Imputing missing values in feature of some datasets by either educated guessing or manually filling values for certain features that have data easily available through online means. For example, in the case of population data, if the fertility rate and median age information are missing for some countries, online sources can help us easily fill these missing values. In the case of datasets for indexes of happiness and freedom, the imputation method would depend on the methodology used for the construction of the dataset.
- Discarding features of datasets where imputation is not possible. For example, in cases of datasets for indexes and freedom, it may not be possible to fill missing values if values of the significant number of features are missing for a specific dataset.

2.2.2 Data Transformation

Before starting our analysis, we need to transform the data, into an appropriate form suitable for our analysis.

Below, we show the transformation operations that we perform on our datasets:

- Calculating average parameter values and the total number of cases for each country.

  Following operations are performed on time series data for confirmed cases, death cases and recovered cases to get their average infection parameter (AIP), average mortality parameter (AMP) and average recovery parameter (ARP), respectively. :

  1. A 7 day moving median for smoothing and removing noisy data is used to get a robust estimate of the trend for each country.
  2. 1-week differences are calculated on the smoothed time series data.
  3. The average of 1-week difference data for each country (from the day of its first report) gives the average rate for each time-series data.
  4. The average rate for each country is divided by its population to get the average parameter for each time-series data.

  The total number of cases for each country in each time series data is the more recent cumulated data from the table.

- Normalization

  Before performing analysis, it is essential to normalize the data at hand so that each and every feature in the dataset is represented on a common scale. The end goal of normalization is to spread the data over the normalized range to yield maximum information from the given data. For our analysis, we need to normalize the average infection parameter, the average mortality parameter, the average recovery parameter, total number of infections, total number of deaths and the total number of recoveries. The normalization technique for these features will take all of the data points in these features (i.e., the data for each country) and redistribute them as evenly as possible across the range [-1 1]. For each feature we perform the non-linear normalization as follows:

  1. Calculate the $\log_{10}$ of these features. The logarithm function squeezes together the larger values in our features and stretches out the smaller values.

2. Convert the logarithmic data x_i to normalized data $\hat{x}$ in the range [-1 1] using tanh normalization.

$$\hat{x_i} = \frac{1 - e^{-\left(\frac{x_i - \mu_i}{\sigma_i}\right)}}{1 + e^{-\left(\frac{x_i - \mu_i}{\sigma_i}\right)}} \tag{2.1}$$

where μ_i represents the mean of logarithmic data and σ_i represents the standard deviation.

This transformation is almost linear to the mean values and has smooth linearity at both extremities to ensure values remain in a limited range. As a result, the resolution is maintained for most values that are within the standard deviation of the mean [15]. Hence tanh normalization is a robust and efficient normalization technique which is not sensitive to outliers and converges faster than Z-score normalization.

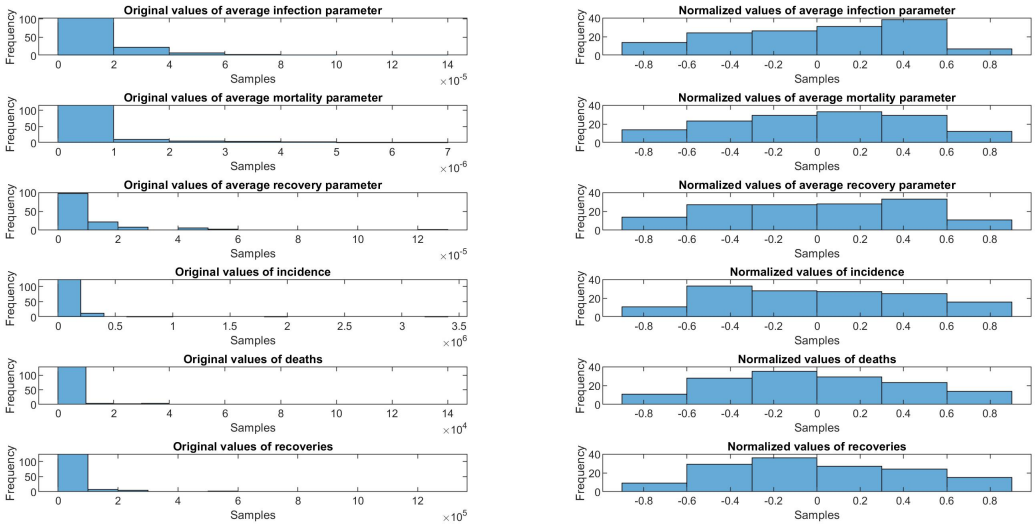


Figure 2.1: Histogram of original and normalized features.

In Figure 2.1, the histogram of each feature is plotted before and after normalization. As evident, our normalization method is successful in transforming the original data where most of the data points are concentrated in a specific region to normalized data where data points are spread over the [-1 1] range.

Chapter 3

Data Analysis and Visualization

In this chapter, we perform the analysis of our datasets to get actionable insights. We will visualize our results and highlight our observations and details gleaned from our analysis.

3.1 Exploratory Data Analysis

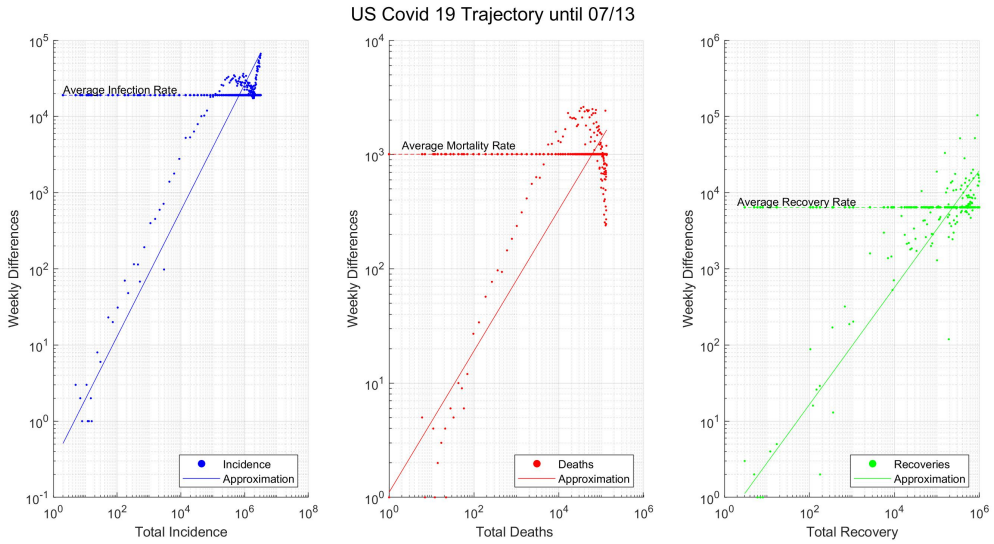
In this section, we explore the general relationship between different variables and plot calculated variables.

3.1.1 Trajectory Plot

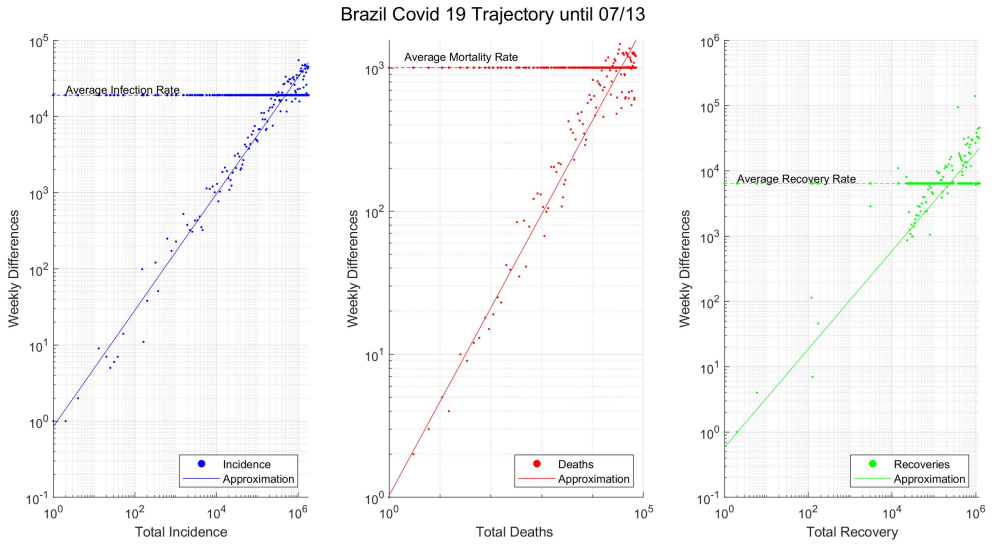
The trajectory plot is useful for understanding how a specific country is dealing with COVID-19. In the Figure 3.1 the trajectory plot is shown for the cases of infections, deaths and recovery for United States 3.1(a) and Brazil 3.1(b). On the Y-axis we have 7-day difference data and on the X-axis we have the moving median filtered data for number of cases . The average rate for each case i.e. infections, death and recovery is also plotted in the same plot. On each of the plots, we have also plotted the best fit linear approximation of the trajectory to show what would be the number of cases if the trajectory actually followed a linear path.

3.1.2 Top 5 countries affected due to COVID-19

In Figure 3.2, we plot the top 5 countries affected due to COVID-19 in terms of the total number of confirmed cases, deaths and recoveries. We also plot the top 5



(a)



(b)

Figure 3.1: Trajectory plots. (a) United States, (b) Brazil.

countries in terms of average infection parameter, average mortality parameter and average recovery parameter.

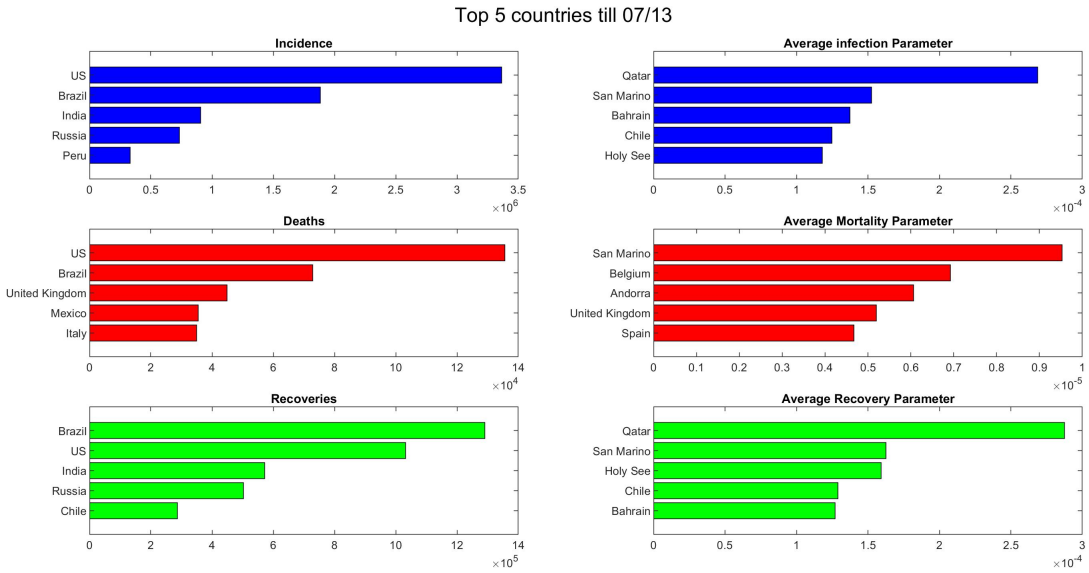


Figure 3.2: Top 5 countries with respect to infections, deaths and recoveries and their calculated parameter values.

3.1.3 Correlation between different parameter values.

The correlation plot between average infection parameter (AIP), average mortality parameter (AMP), average recovery parameter (ARP) is shown in Figure 3.3. The histogram of these parameters appear across the diagonal and the scatter plot between parameters appear in the off-diagonal elements. The slope of the least-square reference lines in the scatter plots is equal to the displayed correlation coefficients. The plot shows that a strong correlation exists between AIP and ARP and between AIP and AMP.

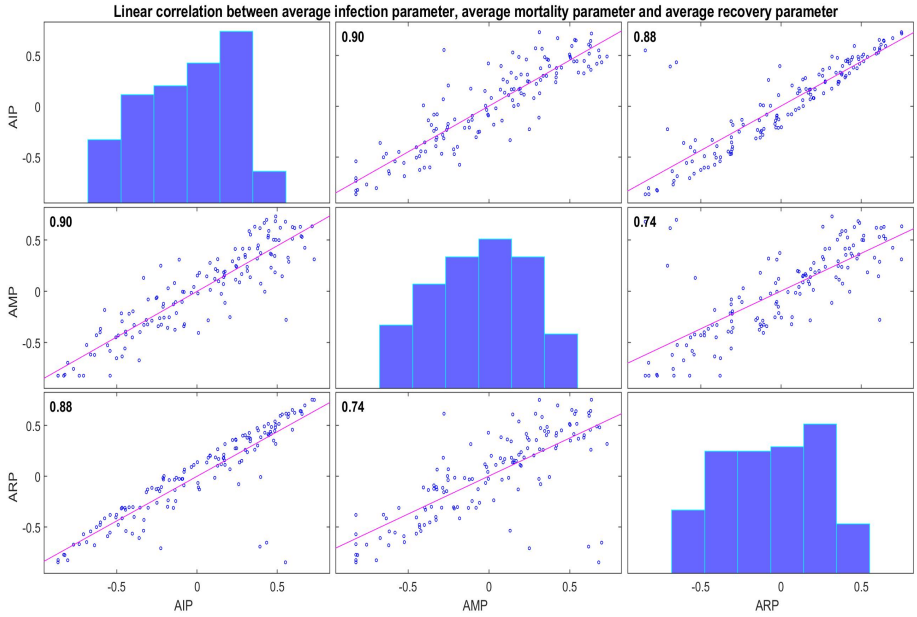


Figure 3.3: Correlation values between AIP, ADP and AMP.

3.2 Confirmatory Data Analysis

In this section, we will perform dimensionality reduction using PCA, cluster analysis using K-means clustering and visualize our clusters with respect to different indexes and their features.

3.3 Principal component analysis for dimensionality reduction

Principal component analysis (PCA) helps to reduce the dimensions and select factors that influence the variance of samples. We perform PCA before clustering since distance metrics run into a lot of problems due to the curse of dimensionality at higher dimensions.

We use the following six features for performing PCA.

- Average infection parameter (AIP)
- Average mortality parameter (AMP)

- Average recovery parameter (ARP)
- Total infections
- Total deaths
- Total recovery

The way we perform PCA is as follows:

1. Centralize the data. Our data consists of 6 features and each sample corresponds to a specific country.
2. Calculate the covariance of the data. This is done to identify biased and redundant information from the given data
3. Use the covariance to calculate the eigenvalues and the eigenvectors. We calculate eigenvectors and eigenvalues to identify where in the data, we have more variation since more variation indicates more information about the data.
4. Sort the eigenvectors in descending order. The highest eigenvector denotes the most significant variation in data and hence forms the first principal component. Because we have 6 features we have 6 eigenvectors. The percentage of contribution of each principal component is given by dividing the eigenvalue corresponding to the principal component by the sum of the eigenvalues. As shown in Table 3.1, most of the variation in the data is explained by the first three principal components.
5. The data is finally projected into a new subspace with only three dimensions corresponding to first three principal components by multiplying the data with the first three principal components.

The heatmap in Figure 3.4 explains the contribution of each feature (variable) to the Principal components and also explains the contribution of each principal component to the total variation.

Principal component	Percentage contribution (%)
PC1	78.40
PC2	12.71
PC3	6.82
PC4	1.75
PC5	0.30
PC6	0.02

Table 3.1: Percentage of variation explained by principal components

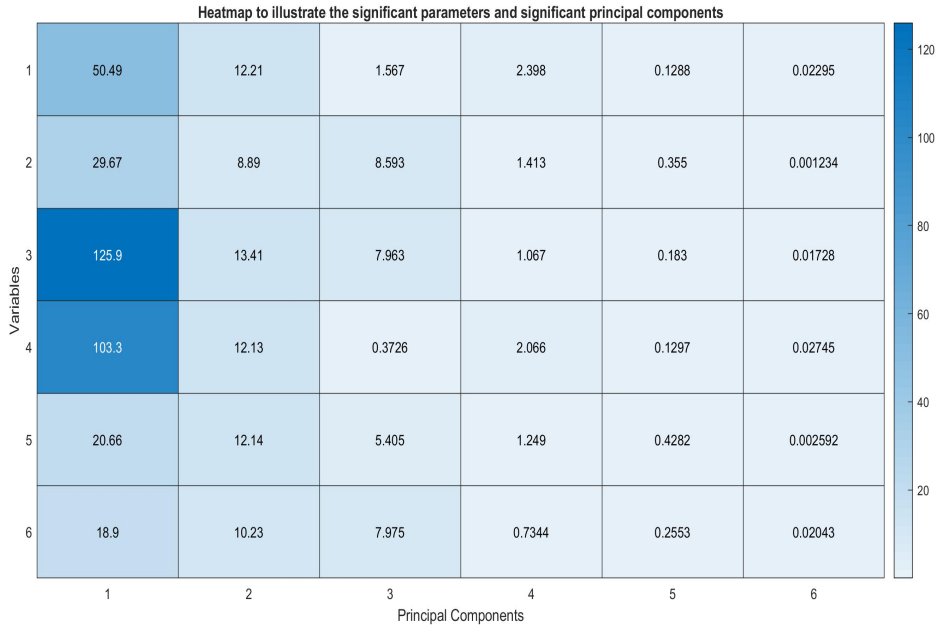


Figure 3.4: Heatmap showing percentage contribution of each variable and each principal component

The Table 3.2 shows the contribution of each variable to the principal components.

Feature	Percentage contribution (%)
Average infection parameter	14.51
Average mortality parameter	10.63
Average recovery parameter	32.27
Total infections	25.64
Total deaths	8.67
Total recovery	8.28

Table 3.2: Percentage contribution of each feature to the principal components

3.4 K-means clustering

K-means clustering is an unsupervised machine learning technique aimed at forming a cluster of data points that behave similarly. In k-means clustering, a set of k clusters is formed such that every data point is assigned to the closest centre, and the sum of the distances of all such assignments is minimized. In our analysis, we use the new data obtained using PCA to form two clusters from 140 countries that behave similarly during COVID-19.

The way we perform K-means clustering is as follows:

1. Randomly initialize two cluster centres.
2. Assign countries to the closest cluster centre based on the minimum Euclidean distance to the cluster centres.
3. Update cluster centres according to the mean of the assigned countries.
4. Repeat step 2 and 3 until convergence. We perform 10 iterations of clustering.

Table 3.3 shows how the 140 countries were assigned to cluster 1 and cluster 2. Figure 3.5 shows cluster assignments and principal components.

Cluster 1	Cluster 2
Albania	Algeria
Australia	Argentina
Benin	Armenia
Botswana	Austria
Burkina Faso	Azerbaijan
Burma	Bahrain
Burundi	Bangladesh
Cambodia	Belarus
Central African Republic	Belgium
Chad	Bolivia
Congo (Brazzaville)	Bosnia and Herzegovina
Congo (Kinshasa)	Brazil
Costa Rica	Bulgaria
Cote d'Ivoire	Cameroon
Croatia	Canada
Cyprus	Chile
Eswatini	China
Ethiopia	Colombia
Gambia	Czechia
Georgia	Denmark
Greece	Dominican Republic
Guinea	Ecuador
Haiti	Egypt
Jamaica	El Salvador
Japan	Estonia
Jordan	Finland
Kenya	France
Korea, South	Gabon
Laos	Germany
Latvia	Ghana
Lebanon	Guatemala
Lesotho	Honduras
Liberia	Hungary
Lithuania	Iceland
Madagascar	India
Malawi	Indonesia
Malaysia	Iran
Mali	Ireland
Malta	Israel
Mauritius	Italy
Mongolia	Kazakhstan
Montenegro	Kuwait
Mozambique	Kyrgyzstan
Namibia	Luxembourg
Nepal	Mauritania
New Zealand	Mexico
Nicaragua	Moldova
Niger	Morocco
Nigeria	Netherlands
Paraguay	North Macedonia
Rwanda	Norway
Senegal	Pakistan
Sierra Leone	Panama
Slovakia	Peru
Slovenia	Philippines
Sri Lanka	Poland
Taiwan*	Portugal
Tanzania	Romania
Thailand	Russia
Togo	Saudi Arabia
Trinidad and Tobago	Serbia
Tunisia	Singapore
Uganda	South Africa
Uruguay	Spain
Venezuela	Sweden
Vietnam	Switzerland
Zambia	Tajikistan
Zimbabwe	Turkey
	US
	Ukraine
	United Arab Emirates
	United Kingdom

Table 3.3: Cluster assignments

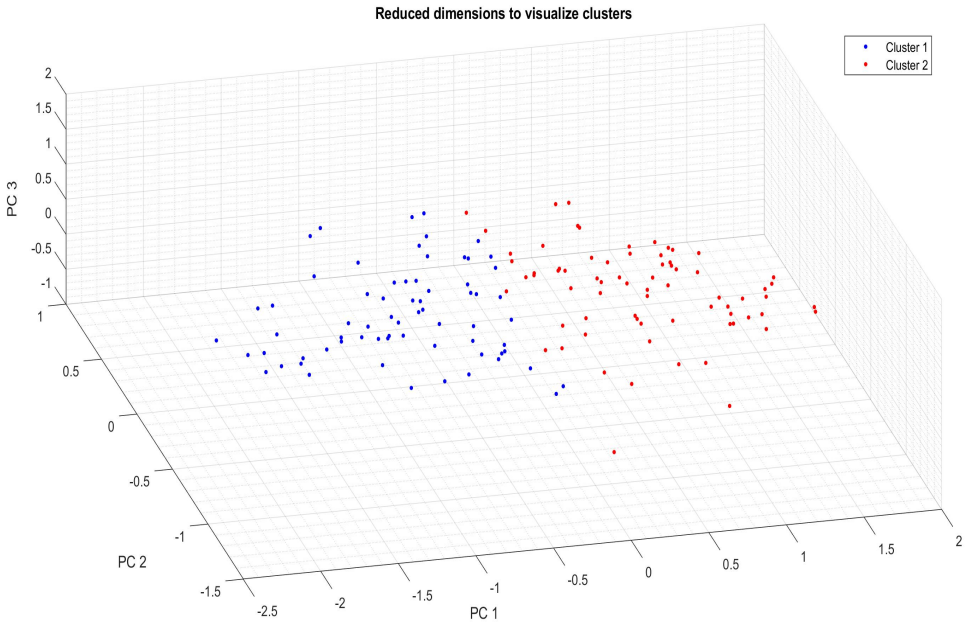


Figure 3.5: Visualization of Cluster assignments

3.5 Visualization of clusters versus different features

In this section, we will visualize our clusters with respect to different features. In the following scatter plots, various features are plots versus average infection parameter. The size of the marker for each country indicates the total number of infections in that particular country.

3.5.1 Happiness score

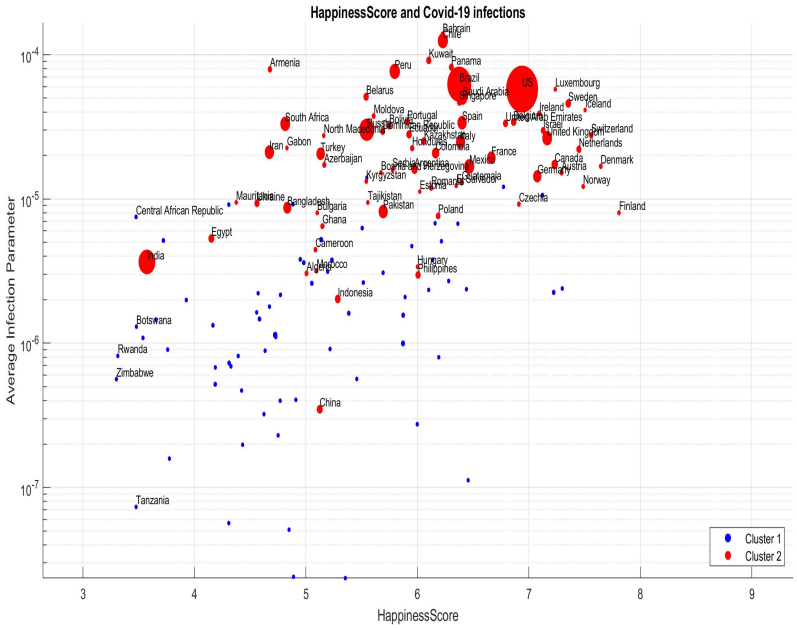


Figure 3.6: Happiness versus average infection parameter.

In Figure 3.6, we can see most of the countries in cluster 2 have higher values of average infection parameter compared to countries in the cluster. For most of the countries in cluster 2, the happiness index is greater than 5 and for most of the countries in cluster 1, the happiness index is lesser than 5.

3.5.2 Life expectancy

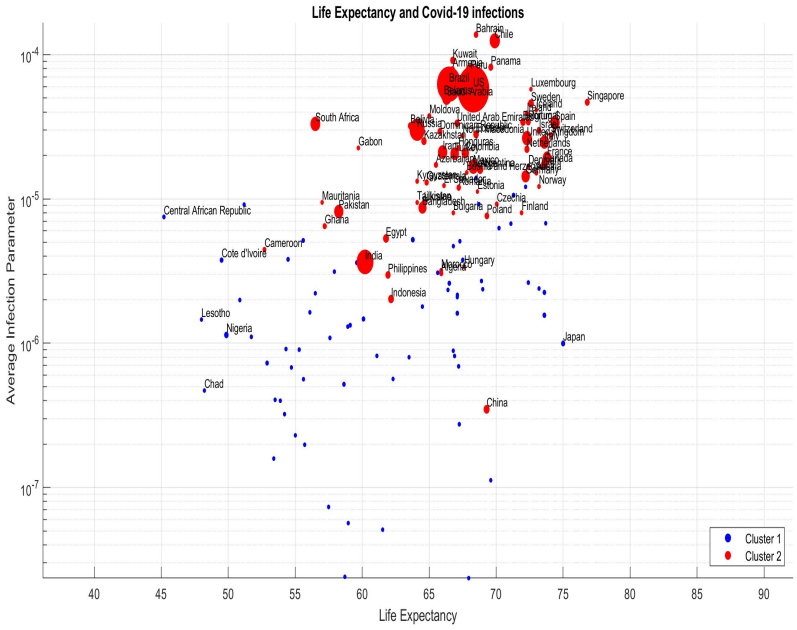


Figure 3.7: Life expectancy versus average infection parameter.

In Figure 3.7, we can see most of the countries in cluster 2, life expectancy is greater than 60 years. For countries in cluster 1, there is a lot of variation in life expectancy.

3.5.3 GDP per capita

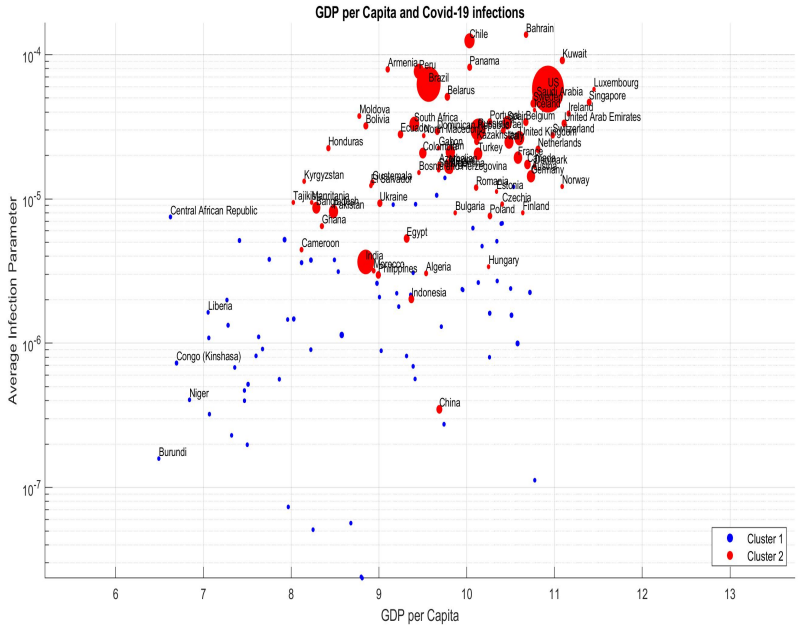


Figure 3.8: GDP per capita versus average infection parameter.

In Figure 3.8, we can see most of the countries in cluster 2, the GDP per capita is greater than 9. For countries in cluster 1, there is a lot of variation in GDP per capita.

3.5.4 Urban population

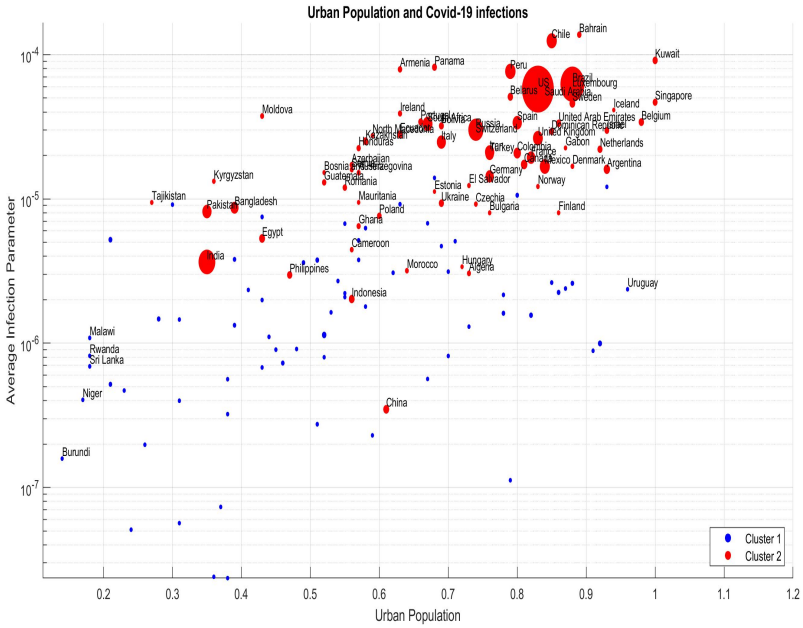


Figure 3.9: Urban population versus average infection parameter.

In Figure 3.9, we can see most of the countries in cluster 2, most of the countries in cluster 2 have urban population more than 60% of the total population. For countries in cluster 1, there is a lot of variation in the urban population.

3.5.5 Press Freedom

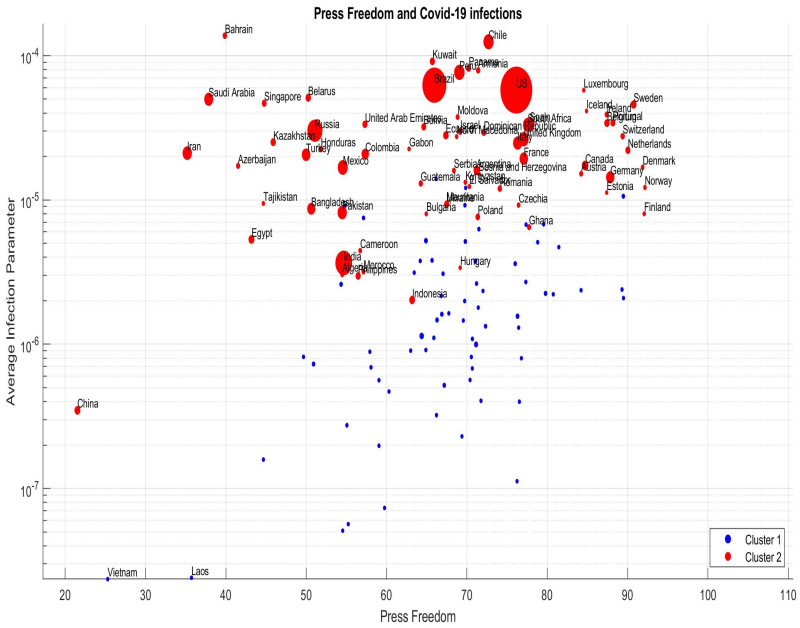


Figure 3.10: Press Freedom versus average infection parameter.

In Figure 3.9, it can be observed that there is no clearly defined trend between press freedom and average infection. However, we can see specific subclusters of cluster 2 formed at the boundaries of press freedom index values.

3.5.6 Index of Economic Freedom

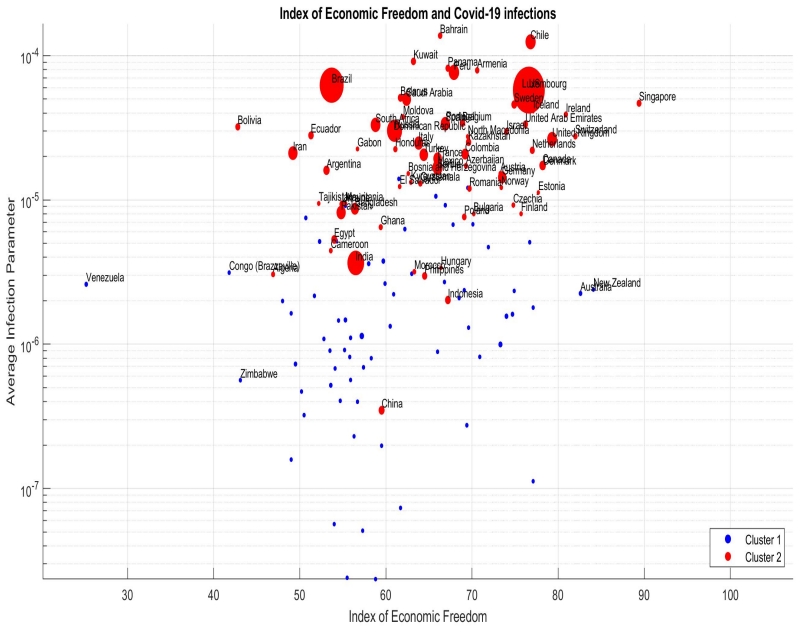


Figure 3.11: Economic Freedom versus average infection parameter.

In Figure 3.11, it can be observed that there is no clearly defined trend between economic freedom and average infection. However, we can see that most of the countries in cluster 2 have higher values of economic freedom and most of the countries in cluster 1 have a relatively low value of economic freedom.

Chapter 4

Conclusion

In this analysis, we try to examine the relationship between different happiness, freedom and population-based features and COVID-19 situation in different countries. Specifically, we calculate three different parameters average infection parameter, average mortality parameter and average recovery parameter, divide the countries into two separate clusters for analyses and try to visualize how these clusters perform with respect to the above features.

While we found no clear relationship between clusters of countries, we can conclude several things that might show how COVID-19 is affecting different countries.

- Countries will life satisfaction higher than average i.e above-average happiness have been affected the most due to COVID-19 pandemic.
- Most of the countries with high average life expectancy have been affected more as compared to countries with low average life expectancy.
- Most of the countries with above-average GDP per capita have been affected more as compared to countries with low GDP per capita.
- Most of the countries with above-average urban population have been affected more as compared to countries with a low urban population.
- Countries which are in cluster 1 and have low press freedom index might not be sharing accurate information.
- Most of the countries with high economic freedom are affected more as compared to countries with low economic freedom.

Prior research has shown that people who are 60 and older appear to be especially vulnerable to the COVID-19, while children appear to be less susceptible to it [16]. In countries with higher life expectancy, happiness, GDP per capita and happiness tend to have higher older populations and tend to be more economically developed, resulting in better nutrition, health care services and government support in retirement. In these countries, people tend to have fewer children and to live longer. In poorer and less happy countries, fertility is higher and people more often die in young adulthood as a consequence of war or from infectious diseases and complications related to childbearing.

COVID-19 has mostly affected urban population and infections tend to be higher because of a complex combination of factors, including population density, national and international connectivity and public health response.

Our analysis might involve several confounders that might create a spurious association. For example, testing rates and climate in different countries might significantly influence our results if taken into consideration.

Appendix A

MATLAB code

```
% Processing and Cleaning
% *Step I: Initial loading and processing of Covid 19 data*
%
% We use the data <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series
% from Johns Hopkins University>. This database contains COVID case numbers by
% country.
%
% The next task is to load the data <https://www.mathworks.com/help/matlab/ref/websave.html?s_tid=srchtitle
% from URL> into MATLAB. Then data preprocessing steps are taken (column name
% and type, specifying variable property...) to prepare the data for calculation.

% Do some housekeeping
clc
clear
close all
%%
% *a) Processing the confirmed time series data*


fileName = [tempdir 'time_series_covid19_confirmed_global.csv'];
%url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master" + ...
%   "/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv";
% fileName = websave(fileName, url);
opts = detectImportOptions(fileName); % Detect import parameters

% Fix range and delimiter
opts.DataLines = [2, Inf];
opts.Delimiter = ",";

% Fix column names and types for first columns
opts.VariableNames(1:4) = [{'ProvinceState'}, {'CountryRegion'}, {'Lat'}, {'Long'}];
opts.VariableTypes(1:4) = [{'string'}, {'string'}, {'double'}, {'double'}];

% Fix file level properties
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";

% Fix variable properties
opts = setvaropts(opts, "ProvinceState", "WhitespaceRule", "preserve");
opts = setvaropts(opts, ["ProvinceState", "CountryRegion"], "EmptyFieldRule", "auto");

% Import the data
data_confirmed = readtable(fileName, opts);

% Clear temp variable, opts
clear opts url fileName;
%%
% *b) Processing the death time series data*
```

```
fileName = [tempdir 'time_series_covid19_deaths_global.csv'];
%url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master" + ...
%   "/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv";
% fileName = websave(fileName, url);
opts = detectImportOptions(fileName); % Detect import parameters

% Fix range and delimiter
opts.DataLines = [2, Inf];
opts.Delimiter = ",";

% Fix column names and types for first columns
opts.VariableNames(1:4) = [{'ProvinceState'}, {'CountryRegion'}, {'Lat'}, {'Long'}];
opts.VariableTypes(1:4) = [{'string'}, {'string'}, {'double'}, {'double'}];

% Fix file level properties
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";

% Fix variable properties
opts = setvaropts(opts, "ProvinceState", "WhitespaceRule", "preserve");
opts = setvaropts(opts, ["ProvinceState", "CountryRegion"], "EmptyFieldRule", "auto");

% Import the data
data_deaths = readtable(fileName, opts);

% Clear temp variable, opts
clear opts url fileName;
%%
% *c) Processing the recovered time series data*

fileName = [tempdir 'time_series_covid19_recovered_global.csv'];
%url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master" + ...
%    "/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv";
% fileName = websave(fileName, url);
opts = detectImportOptions(fileName); % Detect import parameters

% Fix range and delimiter
opts.DataLines = [2, Inf];
opts.Delimiter = ",";

% Fix column names and types for first columns
opts.VariableNames(1:4) = [{'ProvinceState'}, {'CountryRegion'}, {'Lat'}, {'Long'}];
opts.VariableTypes(1:4) = [{'string'}, {'string'}, {'double'}, {'double'}];

% Fix file level properties
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";

% Fix variable properties
opts = setvaropts(opts, "ProvinceState", "WhitespaceRule", "preserve");
opts = setvaropts(opts, ["ProvinceState", "CountryRegion"], "EmptyFieldRule", "auto");

% Import the data
data_recovered = readtable(fileName, opts);

% Clear temp variable, opts
clear opts url fileName;
%
% b) Perform moving average and calculate various parameters
% We cumulated the country data and calculated the moving median<https://www.mathworks.com/help/matlab/ref/movmean.html
% >(7 days).
%
% *i) Calculation of average infection rate*

avg = 7; % Days to average
c = size(data_confirmed, 2); % Number of columns
ds = c-4; % Number of days with data (first four columns give different information)
t0 = datetime('22/1/2020'); % Date of first data
lastDate = t0 + days(ds-1); % Date of last data

% Prepare variables for cumulate country data
```

```
cats = unique(data_confirmed.CountryRegion); % Countries
lc = length(cats); % Number of countries
data2_confirmed = data_confirmed; % Copy data to force same structure for both dataset
data2_confirmed(lc+1:end, :) = []; % Remove unnecessary rows

% Cumulate country data
for i = 1:lc
    tmp1 = mean(data_confirmed{data_confirmed.CountryRegion == cats(i), 3:4}, 1); % take mean of the coordinates
    tmp2 = sum(data_confirmed{data_confirmed.CountryRegion == cats(i), 5:end}, 1); % take sum of the data
    data2_confirmed(i, 1:2) = table("", cats(i)); % Assign country to first columns
    data2_confirmed(i, 3:end) = array2table([tmp1 tmp2]); % Assign mean of coordiantes and sum of cases
end
clear tmp1 tmp2

data_covid = data2_confirmed(:,2);
%detect the first infection for each country
% Smoothing and daily differences data (7-day moving median)
data  = data2_confirmed{:,5:c};
data = movmedian(data,avg,2,'Endpoints','discard');
data_diff = diff(data,[],2);
data_confirmed_series = [array2table(data) array2table(data_diff)];

[m,n] = find(data_diff);
firstIndex = accumarray(m,n,[size(data_diff,1),1],@min,size(data_diff,2));
for i = 1:length(firstIndex)
    data_covid.Average_Infection_Rate(i) = mean(data_diff(i,firstIndex(i):end));
end
%%
%
%
% *ii) Calculate of average mortality rate*

c = size(data_deaths, 2); % Number of columns
ds = c-4; % Number of days with data (first four columns give different information)
t0 = datetime('22/1/2020'); % Date of first data
lastDate = t0 + days(ds-1); % Date of last data

% Prepare variables for cumulate country data
cats = unique(data_deaths.CountryRegion); % Countries
lc = length(cats); % Number of countries
data2_deaths = data_deaths; % Copy data to force same structure for both dataset
data2_deaths(lc+1:end, :) = []; % Remove unnecessary rows

% Cumulate country data
for i = 1:lc
    tmp1 = mean(data_deaths{data_deaths.CountryRegion == cats(i), 3:4}, 1); % take mean of the coordinates
    tmp2 = sum(data_deaths{data_deaths.CountryRegion == cats(i), 5:end}, 1); % take sum of the data
    data2_deaths(i, 1:2) = table("", cats(i)); % Assign country to first columns
    data2_deaths(i, 3:end) = array2table([tmp1 tmp2]); % Assign mean of coordiantes and sum of cases
end
clear tmp1 tmp2

%detect the first infection for each country
% Smoothing and daily differences data (7-day moving median)
data  = data2_deaths{:,5:c};
data = movmedian(data,avg,2,'Endpoints','discard');
data_diff = diff(data,[],2);
data_deaths_series = [array2table(data) array2table(data_diff)];
[m,n] = find(data_diff);
firstIndex = accumarray(m,n,[size(data_diff,1),1],@min,size(data_diff,2));
for i = 1:length(firstIndex)
    data_covid.Average_Mortality_Rate(i) = mean(data_diff(i,firstIndex(i):end));
end
%%
% *iii) Calculate of average recovery rate*

c = size(data_recovered, 2); % Number of columns
ds = c-4; % Number of days with data (first four columns give different information)
t0 = datetime('22/1/2020'); % Date of first data
lastDate = t0 + days(ds-1); % Date of last data
```

```
% Prepare variables for cumulate country data
cats = unique(data_recovered.CountryRegion); % Countries
lc = length(cats); % Number of countries
data2_recovered = data_recovered; % Copy data to force same structure for both dataset
data2_recovered(lc+1:end, :) = []; % Remove unnecessary rows

% Cumulate country data
for i = 1:lc
    tmp1 = mean(data_recovered{data_recovered.CountryRegion ...
        == cats(i), 3:4}, 1); % take mean of the coordinates
    tmp2 = sum(data_recovered{data_recovered.CountryRegion ...
        == cats(i), 5:end}, 1); % take sum of the data
    data2_recovered(i, 1:2) = table("", cats(i)); % Assign country to first columns
    data2_recovered(i, 3:end) = array2table([tmp1 tmp2]); % Assign mean of coordiantes and sum of cases
end
clear tmp1 tmp2

%detect the first infection for each country
% Smoothing and daily differences data (7-day moving median)

data  = data2_recovered{:,5:c};
data = movmedian(data,avg,2,'Endpoints','discard');
data_diff = diff(data,[],2);
data_recovered_series = [array2table(data) array2table(data_diff)];
[m,n] = find(data_diff);
firstIndex = accumarray(m,n,[size(data_diff,1),1] ...
    ,@min,size(data_diff,2));
for i = 1:length(firstIndex)
    data_covid.Average_Recovery_Rate(i) = mean(data_diff ...
        (i,firstIndex(i):end));
end

%%
% *iv) Calculate of average prevalence rate*

% % Smoothing and daily differences data (7-day moving median)
%
% data  = data2_confirmed{:,5:c} - (data2_deaths{:,5:c} + data2_recovered{:,5:c});
% data = movmedian(data,avg,2,'Endpoints','discard');
% data_diff = diff(data,[],2);
% data_prevalence_series = [array2table(data) array2table(data_diff)];
% [m,n] = find(data_diff);
% firstIndex = accumarray(m,n,[size(data_diff,1),1],@min,size(data_diff,2));
% for i = 1:length(firstIndex)
%     data_covid.Average_Prevalence_Rate(i) = mean(data_diff(i,firstIndex(i):end));
% end
%%
% *Plotting the trajectory*

total_days = ds-avg+1;
series_type = {data_confirmed_series data_deaths_series data_recovered_series};
display_name = {'Incidence','Deaths','Recoveries'};
display_name1 = {'Average Infection Rate','Average Mortality Rate','Average Recovery Rate'};
color = {'blue','red','green','cyan'};
xlabels = {'Total Incidence','Total Deaths ','Total Recovery'};
Brazil_rate = data_covid{24,{'Average_Infection_Rate','Average_Mortality_Rate','Average_Recovery_Rate'}};
Brazil_rate = repmat(Brazil_rate,[total_days,1]);
US_rate = data_covid{175,{'Average_Infection_Rate','Average_Mortality_Rate','Average_Recovery_Rate'}};
US_rate = repmat(US_rate,[total_days,1]);
plot_ind = [175,24];
country_name = {'US','Brazil'}
for j = 1:2
    figure(j)
    for i = 1:3
        subplot(1,3,i)
        X = series_type{i}{plot_ind(j),1:total_days};
        Y = [0 series_type{i}{plot_ind(j),total_days+1:end}];
        R = US_rate(:,i);
        h = scatter(X','Y',3,color{i},'filled','DisplayName', ...
            display_name{i}); % Create plot with logarithmic scale
        hold on;
```

```
        plot(X',R,'--','Color',color{i},'Marker','.');
        logX = log10(X);
        logX(logX == -inf) = 0;
        logY = log10(Y);
        logY(logY == -inf) = 0;
        P = polyfit(logX' ,logY',1);
        yfit = (X).^P(1).*(10^(P(2)));
        g = plot(X',yfit','Color',color{i}, 'DisplayName', 'Approximation');
        text(2,R(1,1)+0.2*R(1,1),display_name1{i},'FontSize',9)
        set(gca,'Xscale','log')
        set(gca,'Yscale','log')
        xlabel(xlabels{i});
        ylabel('Weekly Differences');
        legend([h g], 'Location', 'southeast');
        grid on
        hold off
    end

formatOut = 'mm/dd';
lastDate = datestr(lastDate, formatOut);
sgtitle( sprintf('%s Covid 19 Trajectory until %s ', ...
    country_name{j},char(lastDate)));
end



clear data data_diff cats cc ds avg lastDate lc t0 n firstIndex
clear xlabels X logX Y logY color
%%
% *iv) Add total infected, total deaths and total recovered and join population
% data*

data_covid.Total_Infections = data2_confirmed{:,c};
data_covid.Total_Deaths = data2_deaths{:,c};
data_covid.Total_Recovery = data2_recovered{:,c};
data_covid.Total_Prevalence = data2_confirmed{:,c} - (data2_deaths{:,c} + data2_recovered{:,c});
filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\Population2020.csv';
data_population = readtable(filename);
data_population.Properties.VariableNames{1} = 'CountryRegion';
data_population.CountryRegion = string(data_population.CountryRegion);
[data_join, ileft, iright] = innerjoin(data_covid,data_population(:,[1:2]),'keys','CountryRegion');
indx = find(ismember(1:188,ileft)==0);
data_covid = data_join;

data_covid.percentage_of_infections = data_covid.Total_Infections./data_join.Population;
data_covid.percentage_of_deaths = data_covid.Total_Deaths./data_covid.Population;
data_covid.percentage_of_recovered = data_covid.Total_Recovery./data_covid.Population;
%data_covid.percentage_of_prevalence = data_covid.Total_Prevalence./data_covid.Population;


%normalize by population to get the average parameter
data_covid{:,{'Average_Infection_Rate','Average_Mortality_Rate','Average_Recovery_Rate'}} = ...
    data_covid{:,{'Average_Infection_Rate','Average_Mortality_Rate','Average_Recovery_Rate'}}./data_covid.Population;
head(data_covid,5);
clear c data_confirmed data2_confirmed data_deaths data2_deaths data_recovered data2_recovered data_join data_population ileft iright
%%
% *c) Visualize corona virus data.*

%sort_coronavirus_data
covid_confirmed = sortrows(data_covid(:,{'CountryRegion','Total_Infections'}),'Total_Infections','Ascend');
covid_confirmed_rate = sortrows(data_covid(:,{'CountryRegion','Average_Infection_Rate'}),'Average_Infection_Rate','Ascend');
covid_deaths = sortrows(data_covid(:,{'CountryRegion','Total_Deaths'}),'Total_Deaths','Ascend');
covid_deaths_rate = sortrows(data_covid(:,{'CountryRegion','Average_Mortality_Rate'}),'Average_Mortality_Rate','Ascend');
covid_recovery = sortrows(data_covid(:,{'CountryRegion','Total_Recovery'}),'Total_Recovery','Ascend');
covid_recovery_rate = sortrows(data_covid(:,{'CountryRegion','Average_Recovery_Rate'}),'Average_Recovery_Rate','Ascend');
%covid_prevalence = sortrows(data_covid(:,{'CountryRegion','Total_Prevalence'}),'Total_Prevalence','Ascend');
%covid_prevalence_rate = sortrows(data_covid(:,{'CountryRegion','Average_Prevalence_Rate'}),'Average_Prevalence_Rate','Ascend');
covid_rate = data_covid(:,{'CountryRegion','Average_Infection_Rate','Average_Mortality_Rate','Average_Recovery_Rate'});
total_data = {covid_confirmed covid_confirmed_rate covid_deaths covid_deaths_rate covid_recovery covid_recovery_rate };
titles_for_plots = {'Incidence','Average infection Parameter','Deaths', 'Average Mortality Parameter', 'Recoveries','Average Recovery Parameter'};
```

```
j = 1;
color = {'b','r','g','c'};
%get top 20 countries
figure(3)
for i = 1:3
    subplot(3,2,j)
    barh(total_data{j}{end-4:end,2},color{i})
    set(gca,'YTick',[1:5],'yticklabel',total_data{j}(end-4:end,:).CountryRegion);
    title(titles_for_plots{j})

    subplot(3,2,j+1)
    barh(total_data{j+1}{end-4:end,2},color{i})
    set(gca,'YTick',[1:5],'yticklabel',total_data{j+1}(end-4:end,:).CountryRegion);
    title(titles_for_plots{j+1})
    j = j+2;
end
    sgtitle('Top 5 countries')
clear color total_data titles_for_plots covid_confirmed_rate covid_deaths_rate covid_recovery_rate
%%
% *Step 2: Processing happiness index data*

filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\happiness_index_2020.csv'
data_happiness = readtable(filename);
data_happiness.Properties.VariableNames{1} = 'CountryRegion';
data_happiness.Properties.VariableNames{2} = 'HappinessScore';
data_happiness.CountryRegion = string(data_happiness.CountryRegion);
%joining covid 19 and happiness data
[data_join, ileft, iright] = innerjoin(data_covid,data_happiness(:,1:end),'keys','CountryRegion');
indx = find(ismember(1:153,iright)==0);
head(data_happiness,5);
%%
% *Step 3: Process press freedom index data*

filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\world_press_freedom_2020.csv';
data_pfi = readtable(filename);
data_pfi = data_pfi(:,2:end);
data_pfi.Properties.VariableNames{1} = 'PFRank';
data_pfi.Properties.VariableNames{2} = 'CountryRegion';
data_pfi.Properties.VariableNames{5} = 'PressScore';
data_pfi{:,3:end} = 100 - data_pfi{:,3:end}; %reversing the trend
data_pfi.CountryRegion = string(data_pfi.CountryRegion);
%joining covid 19 and happiness data
[data_join2, ileft, iright] = innerjoin(data_join,data_pfi(:,2:end),'keys','CountryRegion');
indx = find(ismember(1:149,ileft)==0);
%data_join.CountryRegion{indx};
head(data_pfi,5);
%%
% *Step 4: Process democracy index data*

filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\democracy_index.csv';
data_di = readtable(filename);
data_di.Properties.VariableNames{1} = 'CountryRegion';
data_di.Properties.VariableNames{2} = 'DemocracyScore';
data_di.CountryRegion = string(data_di.CountryRegion);
[data_join3, ileft, iright] = innerjoin(data_join2,data_di,'keys','CountryRegion');
indx = find(ismember(1:149,ileft)==0);
head(data_di,5);
%%
% *Step 5: Process economic freedom index data*

filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\economic_freedom_index.csv';
data_ei = readtable(filename);
data_ei.Properties.VariableNames{1} = 'CountryRegion';
data_ei.WorldRank = [];
data_ei.Properties.VariableNames{2} = 'EconomicScore';
data_ei.CountryRegion = string(data_ei.CountryRegion);
[data_join4, ileft, iright] = innerjoin(data_join3,data_ei,'keys','CountryRegion');
indx = find(ismember(1:147,ileft)==0);
head(data_ei,5);
%%
% *Step 6: Human Freedom Index data*
```

```
filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\human_freedom_index.csv';
data_hfi = readtable(filename);
data_hfi.Properties.VariableNames{1} = 'CountryRegion';
data_hfi.Properties.VariableNames{3} ='HumanFreedomScore';
data_hfi(:,[2 4]) = [];
data_hfi.CountryRegion = string(data_hfi.CountryRegion);
head(data_hfi,5);
[data_join5, ileft, iright] = innerjoin(data_join4,data_hfi,'keys','CountryRegion');
indx = find(ismember(1:146,ileft)==0);
%%
% *Step 7: World Population data*

filename = 'C:\Users\User\Desktop\Summer 2020\Data Analysis\ECE 579A\covid19_analysis\datasets\Population2020.csv';
data_pop = readtable(filename);
data_pop.Properties.VariableNames{1} = 'CountryRegion';
data_pop.CountryRegion = string(data_pop.CountryRegion);
head(data_pop,5);
[data_set, ileft, iright] = innerjoin(data_join5,data_pop(:,[1 3:end]),'keys','CountryRegion');
indx = find(ismember(1:143,ileft)==0);
data_set.Migrants_net_ = data_set.Migrants_net_ ./data_set.Population;
data_set.Migrants_net_ = data_set.Migrants_net_ + repmat(abs(min(data_set.Migrants_net_)) + 0.00001,[height(data_set),1]);
%%
% *Step 8: Create final data, address missing values*

data_set = rmmissing(data_set); %remove the missing values
data_set.Properties.VariableNames{1} = 'Country';

data_set.Properties.RowNames = data_set.Country;
data_set = movevars(data_set,{'PressScore','DemocracyScore','EconomicScore','HumanFreedomScore'},'Before','GDPPerCapita');
covid_rate.Properties.RowNames = covid_rate.CountryRegion;
covid_rate = covid_rate(data_set.Country,:);

%normalize rates by Population and density
%%
% *Step 9: Normalization*

data_set_log  = data_set;
replicate = repmat(min(data_set{:,2:8},[],1),[size(data_set{:,2:8},1) 1]);
data_set_log{:,2:8} = data_set_log{:,2:8} + abs(replicate);
data_set_log(:,2:8) = varfun(@log10,data_set_log(:,2:8));
%data_set_log{:,51} = log(data_set{:,51})/log(8);
data = data_set_log{:,2:8};
for i  = 1:size(data,2)
    data_column = data(:,i);
    sorted =  sort(data_column);
    sorted(sorted==-inf) = [];
    data_column(data_column == -inf) = sorted(1) ;
    data(:,i) = data_column;
end
data_set_log{:,2:8} = data;
data_set_normalized = data_set;
data_set_normalized{:,2:8} = (1 - exp(-normalize(data_set_log{:,2:8})))./(1 + exp(-normalize(data_set_log{:,2:8}))); %tanh normalization


clear data_happiness data_di data_ei data_hfi data_pfi data_join data_join1 data_join2 data_join3 data_join4 data_covid
head(data_set,5);
head(data_set_normalized,5);
titles_for_plots = {'average infection parameter', 'average mortality parameter' ,'average recovery parameter','incidence','deaths','recoveries'};
figure(4)
j = 1;
for i = 1:6
    subplot(6,2,j)
    histogram(data_set{:,i+1})
    xlabel('Samples')
    ylabel('Frequency')
    title(sprintf('Original values of %s',titles_for_plots{i}))
    subplot(6,2,j+1)
    histogram(data_set_normalized{:,1+i})
    xlabel('Samples')
    ylabel('Frequency')
```

```
    title(sprintf('Normalized values of %s',titles_for_plots{i}))
    j = j+2;
end
%%
% *Correlation between Average Infection Parameter, Average Mortality Parameter
% and Average Recovery Parameter.*

figure(5)
varname = {'AIP','AMP', 'ARP'};
corrplot(data_set_normalized(:,["Average_Infection_Rate","Average_Mortality_Rate", "Average_Recovery_Rate"]),'varnames',varname)
title('Linear correlation between average infection parameter, average mortality parameter and average recovery parameter')
clear indx ileft iright filename
%%
% *Visualization*
%
% *a) Perform PCA for dimension reduction*

%Create a scatter plot to find the relationship between infection rate and
%how to people recoover in happier counters and counteries with good gdp
%per capita?
%group the data according to infection rate, recovery rate and death rate
%find threshold for grouping
%Perform K means clustering and use percentage of recovered, percentage of
%deaths as the predictors

%create a group for clustering


%"AIR_population","ADR_population","ARR_population",
rand('state',47);
Xc = data_set_normalized{:,["Average_Infection_Rate","Average_Mortality_Rate","Average_Recovery_Rate","Total_Infections","Total_Deaths","Total_Recovery"]};
%Xr = data_set_by_region_normalized{:,["AIR_population","ADR_population","ARR_population","Total_Infections","Total_Deaths","Total_Recovery"]};

% Apply PCA for dimension reduction
[Uqc,score_c,latent,tsquared,explainedc,mu] = pca(Xc);
projc = Xc*Uqc;
% [Uqr,score_r,latent,tsquared,explainedr,mu] = pca(Xr);
%
% projr = Xr*Uqr;
value = abs(normalize(Xc'*projc));
figure(5)
heatmap(value.*explainedc','XLabel','Principal Components','YLabel','Variables');
title('Heatmap to illustrate the significant parameters and significant principal components')
rank_variables = sum(value.*explainedc',2);
rank_variables = rank_variables./sum(rank_variables)
%%
% *b) Perform cluster analysis using K-means algorithm*



%group 142 countries into 3 clusters using K means clustering

K  = 2;
[idxc,Cc] = kmeans(projc(:,1:3),K,'MaxIter',10);
% [idxr, Cr] = kmeans(projr,K,'MaxIter',100);
% subplot(1,2,1)
figure(6)
grpstats(data_set{:,["Average_Infection_Rate","Average_Mortality_Rate", "Average_Recovery_Rate"]},idxc,0.05);
xlabel('Covid19 groups')
ylabel('Mean per group');
set(gca,'xtick',1:length(unique(idxc)),'xticklabel',{'Cluster 1', 'Cluster 2'});
title('Mean and 95% confidence intervals (Country)')
legend("Average Infection Parameter","Average Mortality Parameter", "Average Recovery Parameter", "Average Prevalence Parameter",'location','northwest')

figure(7)
%number of dimensions to reduce
% proj = (proj-min(proj)./(max(proj)-min(proj))); % 0 1 range
proj1 = projc(idxc==1,:);
proj2 = projc(idxc==2,:);
proj3 = projc(idxc==3,:);
plot3(proj1(:,1),proj1(:,2), proj1(:,3),'.b','linew',1.5,'MarkerSize',10)
```

```
hold on
plot3(proj2(:,1),proj2(:,2), proj2(:,3),'.r', 'linew',1.5,'MarkerSize',10)
plot3(proj3(:,1),proj3(:,2), proj3(:,3),'.g', 'linew',1.5,'MarkerSize',10)
grid on
grid minor
xlabel('PC 1')
ylabel('PC 2')
zlabel('PC 3')
title('Reduced dimensions to visualize clusters')
legend('Cluster 1', 'Cluster 2','location','northeast');
cluster_1 = data_set{:,1}(idxc == 1)
%%
% From inspection, cluster 1 represents countries that are highly affected by
% covid, cluster 2 represents moderately affected countries and cluster 3 represents
% low affected countries.

cluster_2 = data_set{:,1}(idxc == 2)
cluster_3 = data_set{:,1}(idxc == 3)

%text(proj1(:,1),proj1(:,2),poorcovid,'FontSize',7,'HorizontalAlignment', 'left','VerticalAlignment', 'bottom');
% set(gca,'xscale','log');
hold off;
%%
% *c) How is the performance of Covid 19 with respect to Happiness in various
% countries?*

%clf([6 7 8])
visualize_plots(data_set,covid_rate, 'HappinessScore',idxc,K,0,'HappinessScore',1,1,7);
%%
% *c) How is the performance of Covid 19 with respect to Life Expectancy in
% various countries?*

%clf([9 10 11])
visualize_plots(data_set, covid_rate, 'HealthyLifeExpectancy',idxc,K,0,'Life Expectancy',1,1,10);
%%
% *d) How is the performance of Covid 19 with respect to gdp per capita in various
% countries?*

%clf([9 10 11])
visualize_plots(data_set, covid_rate, 'GDPPerCapita',idxc,K,0,'GDP per Capita',1,1,13);
%%
visualize_plots(data_set, covid_rate, 'PressScore',idxc,K,0,'Press Freedom',1,1,16);
%%
%clf([9 10 11])
visualize_plots(data_set, covid_rate, 'HumanFreedomScore',idxc,K,0,'Human Freedom Index',1,1,19);
%%
visualize_plots(data_set, covid_rate, 'EconomicScore',idxc,K,0,'Index of Economic Freedom',1,1,22);
%%
visualize_plots(data_set, covid_rate, 'UrbanPop_',idxc,K,0,'Urban Population',1,1,25);
%%
visualize_plots(data_set, covid_rate, 'Med_Age',idxc,K,0,'Median Age',1,1,28);

%%

% figure(6)
% plot(data_set_normalized{:,'GDPPerCapita'})
% hold on
% plot(data_set_normalized{:,'Total_Infections'})
% plot(data_set_normalized{:,'Total_Deaths'})
% plot(data_set_normalized{:,'Total_Recovery'})
% hold off
% legend('total infections','total deaths','total recovery','location','northeast')
% legend boxoff
%%
% *Visualizations*

% %how does GDP per capita affect the covid spread?
% data_set_new = sortrows(data_set_normalized,'FreedomToMakeLifeChoices','descend');
% head(data_set_new,15);
% figure(7)
% stackedplot([data_set_new(1:20,4:6) data_set_new(1:20, ...
```

```
% 'FreedomToMakeLifeChoices') data_set_new(1:20,'pf_score') data_set_new(1:20,'ef_score') data_set_new(1:20,'HumanFreedomScore')])
% cell2table(data_set_new(1:20,:).Properties.RowNames,'VariableNames',{'Country'})
%%
% figure(8)
% ax = axes; % create axes
% data_set_new = [data_set_new(:,1:3) data_set_new(:,'GDPPerCapita')];
% plot(ax,table2array(data_set_new(1:20,1:4))); % plot data
% ax.XTick = 1:20; % limit X-axis ticks no. to columns
% ax.XTickLabel = data_set_new(1:20,1:3).Properties.RowNames; % get columns names
% xtickangle(ax,45);
% legend('Average Infection Rate', 'Average Mortality Rate', 'Average Recovery Rate', 'GDPPerCapita'); % get algorithm names




%
% Step 3: Visualize the results
% Let's answer our questions about case numbers vs specific countries.
%
% In the figure below, you can see the most significant worldwide cases, or
% you can filter to the territory of Gamax Laboratory Solutions and Europe.
%
% It was easy to add country-specific filtering with Live Editor. You can add
% many<https://www.mathworks.com/help/matlab/matlab_prog/add-live-editor-tasks-to-a-live-script.html?s_tid=srchtitle
% interactive tasks> to your Live Script.% % Filter by region

%functions
%%
function  visualize_plots(dcountry,covid_rate, param,idxc,K,xscale,xlabelv,yscale,textc,lastfigno)
    color = {'b','r','g'};
    C = {'Total_Infections','Total_Deaths','Total_Recovery'};
    D = {"Average_Infection_Rate","Average_Mortality_Rate", "Average_Recovery_Rate"};
    E = string({sprintf('%s and Covid-19 infections',xlabelv), sprintf('%s and Covid-19 deaths',xlabelv), sprintf('%s and Covid-19 recovery',xlabelv)});
    covid_confirmed_rate = sortrows(covid_rate(:,{'CountryRegion','Average_Infection_Rate'}),'Average_Infection_Rate','Ascend');
    covid_deaths_rate = sortrows(covid_rate(:,{'CountryRegion','Average_Mortality_Rate'}),'Average_Mortality_Rate','Ascend');
    covid_recovery_rate = sortrows(covid_rate(:,{'CountryRegion','Average_Recovery_Rate'}),'Average_Recovery_Rate','Ascend');
    %covid_prevalence_rate = sortrows(covid_rate(:,{'CountryRegion','Average_Prevalence_Rate'}),'Average_Prevalence_Rate','Ascend');
    F = {covid_confirmed_rate, covid_deaths_rate, covid_recovery_rate};
    G = {'Average Infection Parameter', 'Average Mortality Parameter', 'Average Recovery Parameter'};
    countries = sortrows(dcountry(:,{'Country',param}),param,'descend');
    poorcovid = dcountry{:,1}(idxc == 2);
    best_worse = countries{:,1}([1:5 end-4:end]);

    for i = 1:length(C)
        figure(lastfigno +i);
        markerSizes = normalize(dcountry{:,C{i}},'range',[10 1000]);
%         markerSizes = markerSizes + abs(min(markerSizes));
%         markerSizes = round(markerSizes + 1);
        for j = 1:K
            scatter(dcountry{:,param}(idxc == j),dcountry{:,D{i}}(idxc == j),markerSizes(idxc == j),color{j},'filled');
            hold on;
            grid on;

        end
        countries_to_plot = F{i}{:,'CountryRegion'}(end-19:end);
        countries_to_plot = unique([countries_to_plot ; best_worse ; poorcovid]);
        legend('Cluster 1', 'Cluster 2', 'Cluster 3','location','southeast');
        xminvalc = min(dcountry{:,param});
        xmaxvalc = max(dcountry{:,param});
        yminvalc = min(dcountry{:,D{i}});
        ymaxvalc = max(dcountry{:,D{i}});

        if textc
            text(dcountry{countries_to_plot,param},(dcountry{countries_to_plot,D{i}}),countries_to_plot, ...
                'FontSize',7,'HorizontalAlignment', 'left','VerticalAlignment', 'bottom');
        end
        hold off;
        if xscale
           set(gca,'xscale','log');
        end
```

```
        if yscale
           set(gca,'yscale','log');
           axis([xminvalc - 0.2*xminvalc xmaxvalc + 0.2*xmaxvalc  0 ymaxvalc + 0.2*ymaxvalc])
        else
            axis([xminvalc - 0.2*xminvalc xmaxvalc + 0.2*xmaxvalc  yminvalc - 0.2*yminvalc ymaxvalc + 0.2*ymaxvalc])
        end

        title(E{i});
        ylabel(G{i});
        xlabel(xlabelv);
    end
end
```

Bibliography

[1] "Who statement regarding cluster of pneumonia cases in wuhan, china." [Online]. Available: https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china

[2] "Coronavirus." [Online]. Available: https://www.who.int/health-topics/coronavirus

[3] B. Pfefferbaum and C. S. North, "Mental health and the covid-19 pandemic," *New England Journal of Medicine*, 2020.

[4] A. Bäuerle, E.-M. Skoda, N. Dörrie, J. Böttcher, and M. Teufel, "Psychological support in times of covid-19: the essen community-based cope concept," *Journal of Public Health*, 2020.

[5] S. K. Brooks, R. K. Webster, L. E. Smith, L. Woodland, S. Wessely, N. Greenberg, and G. J. Rubin, "The psychological impact of quarantine and how to reduce it: rapid review of the evidence," *The Lancet*, 2020.

[6] L. Pullicino, "Covid-19: The impact on news media," *Available at SSRN 3640199*, 2020.

[7] W. Ding, R. Levine, C. Lin, and W. Xie, "Corporate immunity to the covid-19 pandemic," National Bureau of Economic Research, Tech. Rep., 2020.

[8] C. COVID, "Global cases by johns hopkins csse https://github.com/cssegisanddata," *COVID-19/blob/master/csse covid*, vol. 19, 19.

[9] J. F. Helliwell, R. Layard, J.-E. De Neve, and J. Sachs, "World happiness report 2020. new york: Sustainable development solutions network," 2020.

[10] I. Vasquez and T. Porcnik, *Human Freedom Index 2019*, 2019.

[11] RBF, "World press freedom index," 2020.

[12] T. Miller, A. B. Kim, and J. M. Roberts, "Index of economic freedom (washington, dc: The heritage foundation, 2019)," 2019.

[13] E. I. Unit, "Democracy index 2019: A year of democratic setbacks and popular protest."," 2020.

[14] "Countries in the world by population (2020)." [Online]. Available: https://www.worldometers.info/world-population/population-by-country/

[15] K. L. Priddy and P. E. Keller, *Artificial neural networks: an introduction.*, SPIE press, 2005, vol. 68.

[16] "Statement – older people are at highest risk from covid-19, but all must act to prevent community spread," Jul 2020. [Online]. Available: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/statements/statement-older-people-are-at-highest-risk-from-covid-19,-but-all-must-act-to-prevent-community-spread