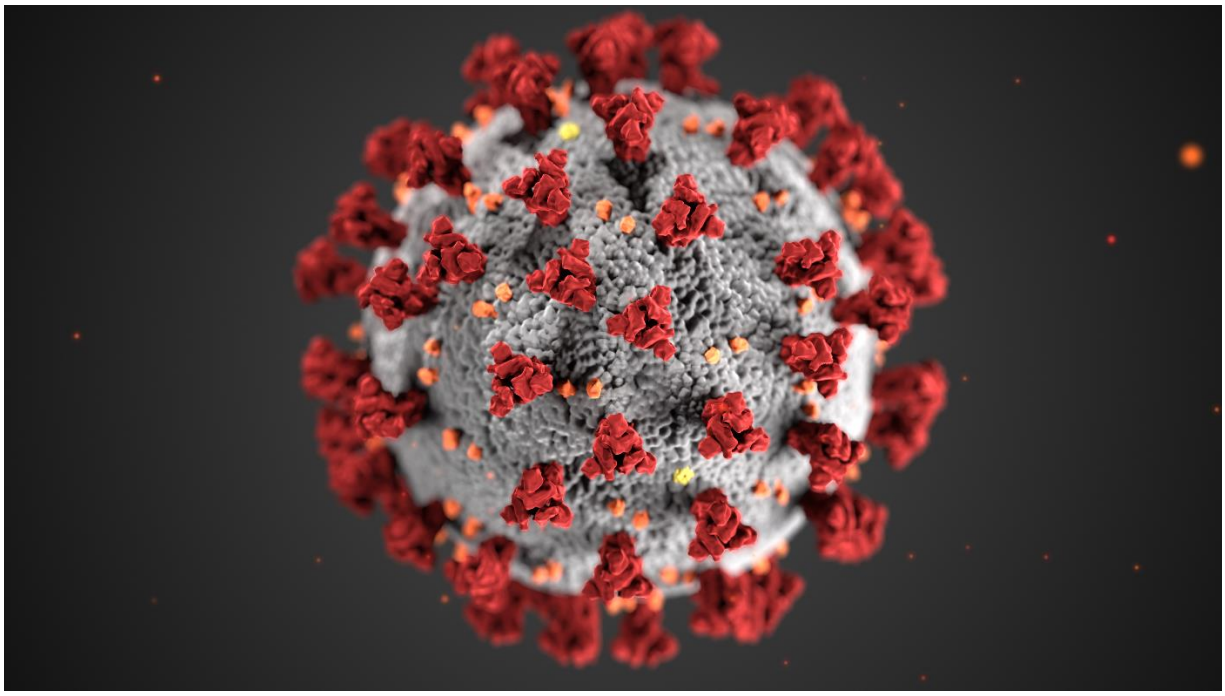


Covid-19 and Economic Status of the Globe

A STUDY ON COVID 19 AND ITS CORRELATIONS WITH GDP, HDI AND POVERTY RATE ACROSS COUNTRIES

Teoman Berkay Ayaz 1800004169
Mehmet Salih Çifci 1800004594 | Arman Sami Darı 1700002889
CSE0448 Big Data Analytics

Table of Contents

Introduction, Project Goals and Scope

The following work was assigned to us as the term project for CSE0448 Big Data Analytics at IKU for the 2021-2022 Fall semester. The project group (G2) conducting the following study consists of three members:

- Teoman Berkay Ayaz 1800004169
- Mehmet Salih Çifci 1800004594
- Arman Sami Darı 1700002889

The topic of the project was, as written in the project description:

- The Impact of Covid-19 Pandemic on the Global Economy: Emphasis on Poverty Alleviation and Economic Growth

The project, in accordance with the project description seeks to evaluate the correlations between human development index (HDI), GDP of a given country and the poverty rate with the mortality rate of Covid-19 in the given country.

Throughout this project our aim is to:

- Evaluate the correlations of Covid-19 mortality rate with economic factors
- Evaluate the correlations of Covid-19 mortality rate with HDI
- Unravel any unexpected correlations
- Extract information from the data at hand
- Visualize the extracted information
- Build a categorical prediction model using machine learning

Data Collection & Sources of Collected Data

Although we were handed a dataset initially, after running some correlation tests we realized it posed two main issues for us:

- Dataset was missing quite a lot for us to be able to conduct a meaningful study.
- It had too little features for the correlations that we initially aimed to study.

Data Collection & Sources of Collected Data (Cont'd)

After this realization we started to look through data, sources ranging from World Bank to United Nations Development Program until we settled on certain sets for certain features we wished to have. Throughout the process of collecting more data, we started to hypothesize about the correlations and collected more and more data based on the hypotheses we developed. By the end of the data collection process, we ended up with two separate datasets:

- main: Our main dataset for evaluate correlations, extract information from alongside enriching mldata
- mldata: Our dataset for the classified prediction model we will be developing

Following is a list of the sources of the data we have used to create our datasets, alongside some description on the data:

- Initial Data: The dataset provided to us by the instructor and the dataset we have used as a base for mldata.
    - Source: https://data.mendeley.com/datasets/b2wvnbnpj9/1
- Covid: The data we used as a basis for our main dataset which contains covid data on countries. (Data taken live at 16/12/2021 5:30AM)
    - Source: https://www.worldometers.info/coronavirus/
- Poverty Rate: The poverty rate data across countries around the globe.
    - Source: https://worldpopulationreview.com/country-rankings/poverty-rate-by-country
- Age: The mean age data on countries of the world.
    - Source: https://worldpopulationreview.com/country-rankings/poverty-rate-by-country
- HDI: The human development index for each individual country.
    - Source: https://worldpopulationreview.com/country-rankings/hdi-by-country
- GDP per capita: GDP data of individual nations across the globe.
    - Source: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
- Population: Total population of countries.
    - Source: https://worldpopulationreview.com/
- Travel: Arrival and departure data on a given country.
    - Source: https://data.worldbank.org/indicator/ST.INT.ARVL
    - Source: https://data.worldbank.org/indicator/ST.INT.DPRT

Features of the Dataset (Collected, Engineered, Categorical)

Our main dataset contains 17 features (10 collected, 2 categorical, 5 engineered) and our mldata dataset contains 21 features (13 collected, 3 categorical, 5 engineered) 17 of them found in both datasets. The following table contains descriptions of each feature:

Feature	Description	Type	Found in
COUNTRY	The name of the country in the row	Collected	both
AGE	The mean age of the given country	Collected	both
POR	Poverty rate of a given country	Collected	both
HDI	Human Development Index of a country	Collected	both
GDPCAP	GDP per capita of countries	Collected	both
ARRVL	Arrivals at that country	Collected	both
DPRT	Departures from that country	Collected	both
TD	Total Deaths in a country	Collected	both
TC	Total Cases in a country	Collected	both
POP	Population of a given country	Collected	both
MR	Mortality rate of covid in each country	Engineered	both
SPI	Spread Index (TC/POP)	Engineered	both
HDS	Human Development Status (Developed or not)	Categorical	both
WLHS	Wealth Status (High Income, Low Income)	Categorical	both
ARI	Arrival Index (ARRVL/POP)	Engineered	both
DPI	Departure Index (DPRT/POP)	Engineered	both
TRI	Travel Index ((ARI+DPI)/2)	Engineered	both
CODE	Three letter ISO Code of countries	Collected	mldata
DATE	Date info of a given row	Collected	mldata
STI	Stringency Index	Collected	mldata
MRG	Mortality rate-based group (1 = low, 3 = high)	Categorical	mldata

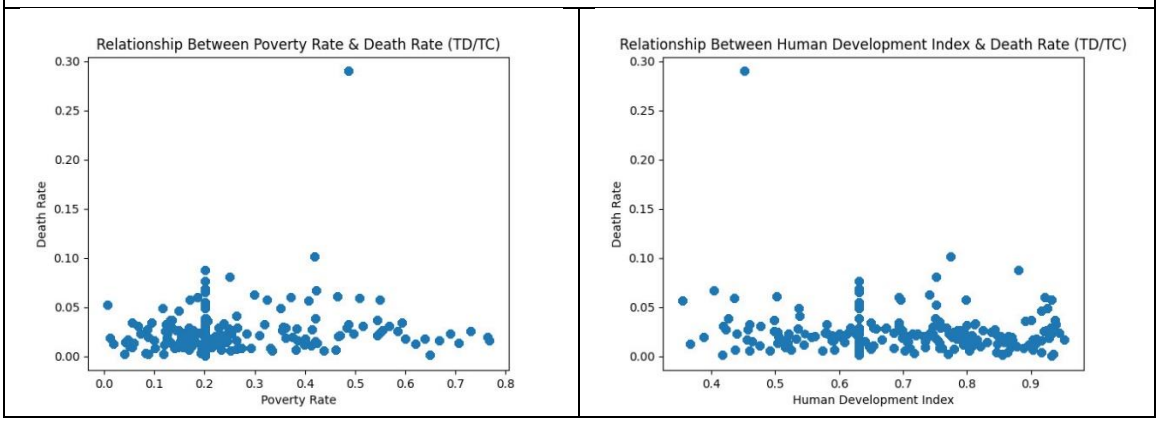
Additional Info:

- main dataset was used to enrich mldata (hence everything in main is also in mldata).
- Intervals of WLHS (based on GDPCAP) are:
    - [Extreme Low<2000], [2000 <= Low < 6000], [6000<=Moderate<24000], [24000 <= Moderate High < 44000], [44000<=Extreme High]
- Interval of HDS (based on HDI) are:
    - [Underdeveloped < 0.5], [0.5 <= Developing < 8], [8 <= Developed]

Data Cleaning and Missing Values

When we initially plotted a correlation heatmap we were unable to find any significant correlations. When we tried by filling the missing values by the means the correlations were stronger albeit not by much. When we plotted scatterplots to inspect the correlations, we saw that means were building up and affecting correlations drastically.

Scatter Plots of Initial Dataset



Since data was rather scarce especially on columns like ARRVL and DPRT, our hands were tied when it came to data cleaning. And since filling by means was weakening the correlations, we had to fill the missing values using a more logical approach which required different methods for different columns.

For ARRVL and DPRT columns, which was the scarcest data we engineered ARI and DPI (ratio of arrivals and departures to and from a country with its total population) and we filled each the ARI and DPI country using the mean of the wealth status group they belong.

Process of Thought on Collecting Data & Filling Missing Values

When we first looked at the initial correlation heatmap, and categorized the distribution into groups we asked some questions:

- Is there any correlation between poverty rate and mortality rate?
- Is there any correlation between age and mortality rate?
- Why does the virus spread more in wealthier & developed countries?
- Is it possible that virus spreads more in wealthier & developed countries because people living there tend to travel more often?

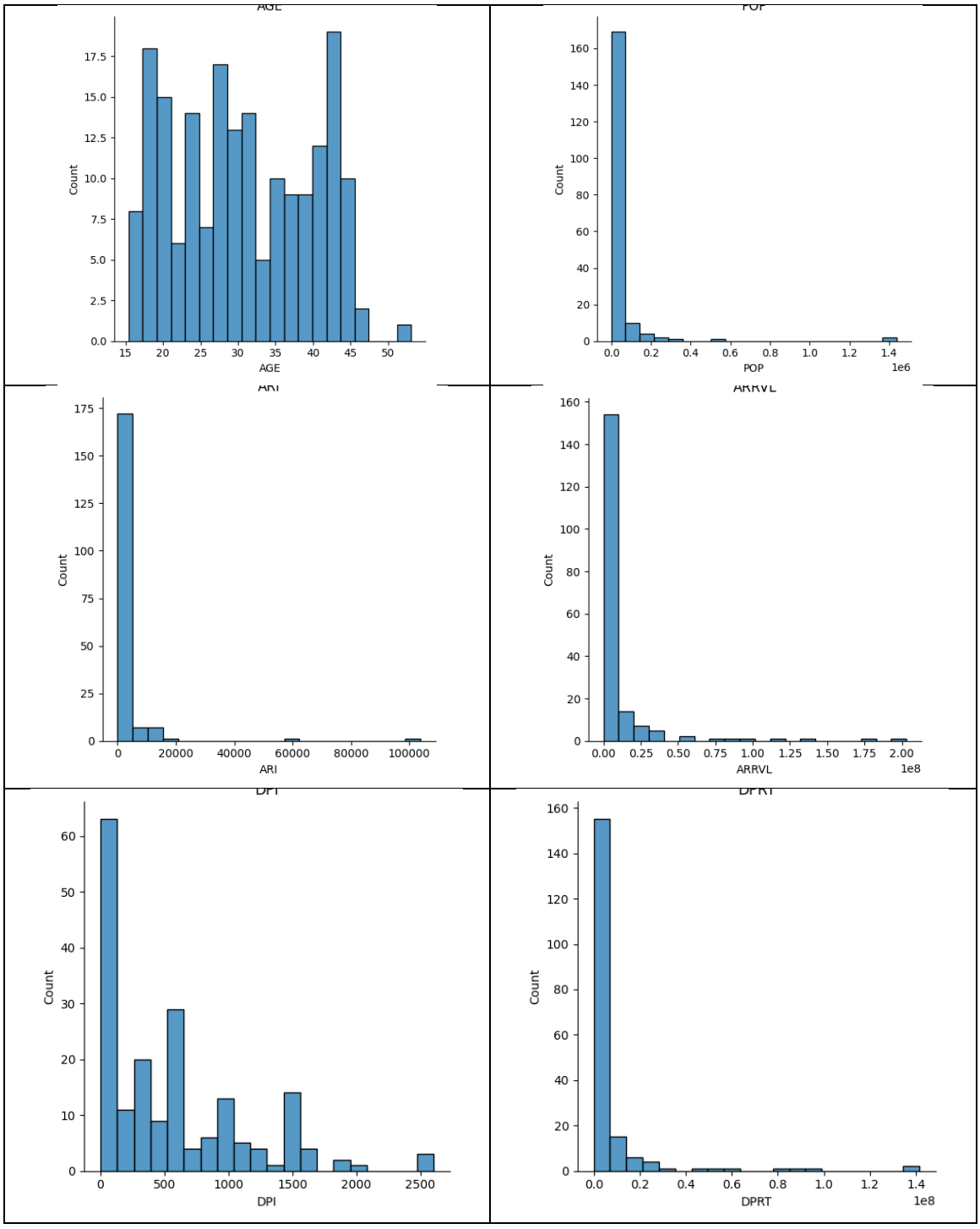
To answer these questions and test our hypotheses we decided to search for data which could help us answer them, hence we decided to collect data from the sources cited earlier in the section "Data Collection & Sources of Collected Data". Midway through our collection, we realized that the unique values of the columns that we are looking for the correlations or lack thereof, are about 200 unique values each. So, we decided to create a new dataset (main dataset or newData) with about 200 countries and their full data.

With regards to missing values, since filling empty cells by the mean of the entire column was causing means to pile up and weaken the correlations, we had to prevent means to pile up. For that we created two categorical features (WLHS & HDS) which would enable us to fill the missing values by the means of the groups they belong. About the travel data however, since the means of ARRVL and DPRT would be more than the population of some countries we had to come up with a different approach. Hence, we engineered ARI and DPI. We used ARI and DPI to help with the correlations of travel and spread of the virus alongside to generate TRI (travel index, avg. of ARI and DPI in a row).

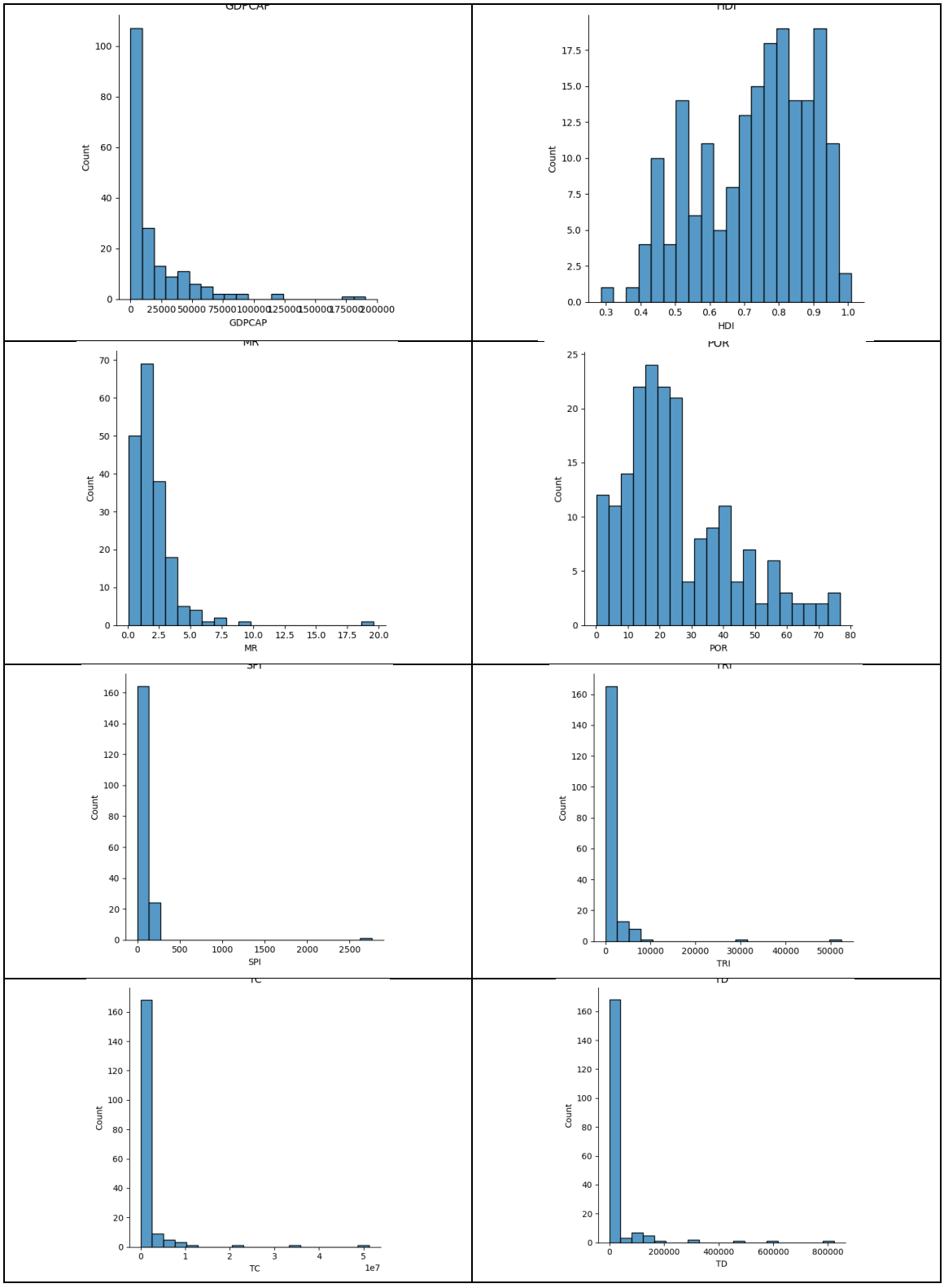
When it came to filling certain columns however, since we did not want to weaken the correlation once more, we had to manually add data to the columns for they were missing from the sources we have acquired data from.

Numerical Distribution of the Dataset Population

The following collection of distribution plots show the numerical distribution of our data in each respective column:



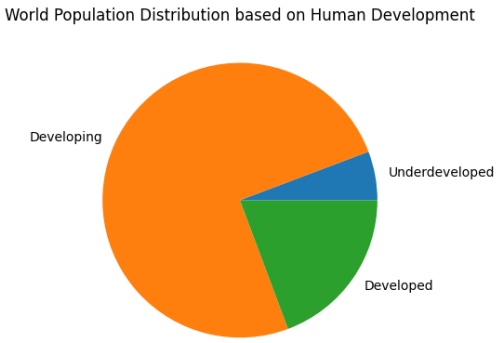
Numerical Distribution of the Dataset Population (Cont'd)

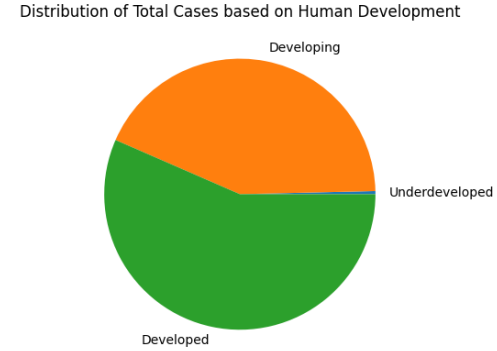


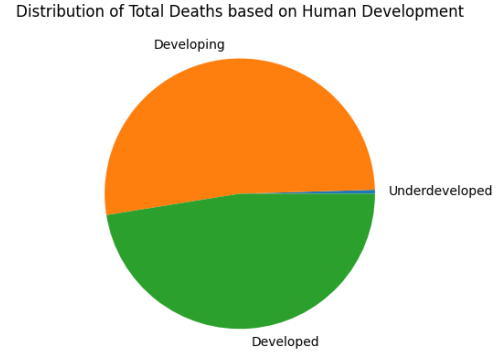
Categorical Distribution of The Dataset Population

The following collection of distribution plots show the categorical distribution of our data:

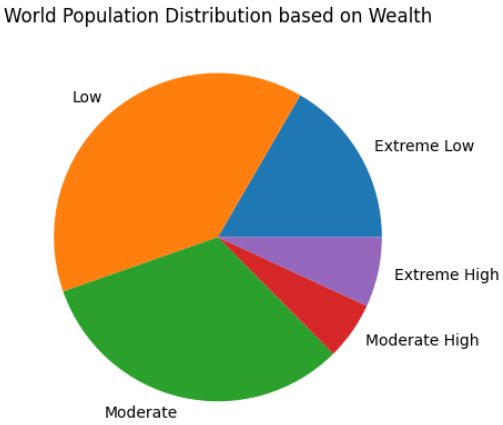
HDI Based Distribution

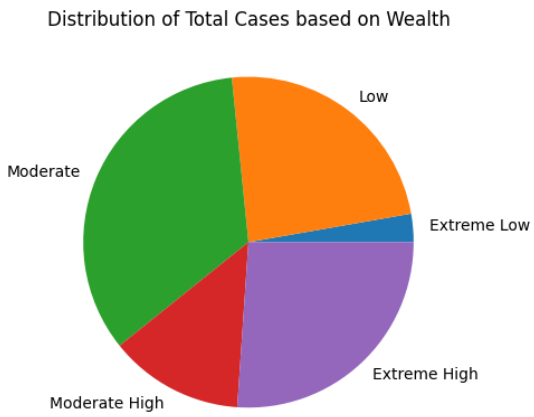


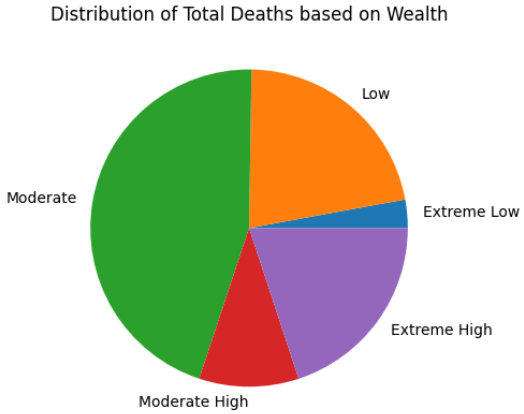




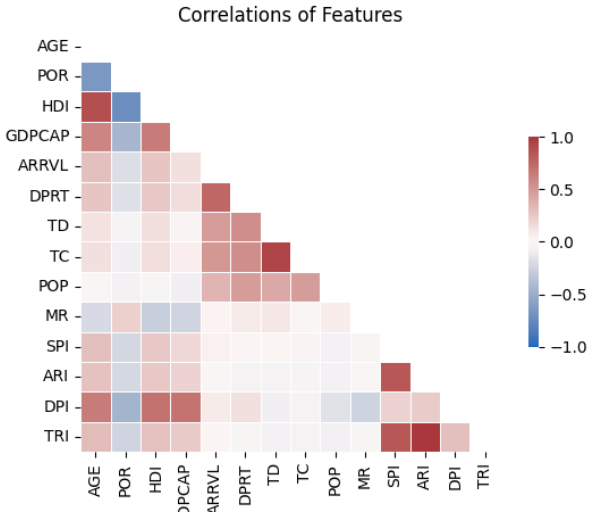
Wealth (GDP per Capita) Based Distribution







Correlation Analysis



When we look at the final correlation graph, we can see that mortality rate is positively correlated with POR (0.22), and negatively correlated with AGE (-0.20), HDI (-0.28) and GDPCAP (-0.25).

Aside from mortality rates, HDI is positively correlated with SPI (0.26), TRI (0.29) and GDPCAP (0.64).

Lastly, GDPCAP is positively correlated with SPI (0.18), TRI (0.25), ARI (0.21) and DPI (0.69)

Full Correlation Table (Note: ARRVL and DPRT were removed to make the table fit since they were not relevant after ARI & DPI)

	AGE	POR	HDI	GDPCAP	TD	TC	POP	MR	SPI	ARI	DPI	TRI
AGE	1.000	-0.656	0.877	0.599	0.126	0.136	-0.002	-0.203	0.301	0.282	0.640	0.317
POR	-0.656	1.000	-0.721	-0.440	-0.054	-0.081	-0.064	0.215	-0.228	-0.221	-0.442	-0.244
HDI	0.877	-0.721	1.000	0.647	0.138	0.148	-0.039	-0.286	0.256	0.249	0.691	0.287
GDPCAP	0.599	-0.440	0.647	1.000	0.030	0.070	-0.079	-0.245	0.180	0.206	0.682	0.244
TD	0.126	-0.054	0.138	0.030	1.000	0.936	0.411	0.108	0.022	-0.058	-0.078	-0.062
TC	0.136	-0.081	0.148	0.070	0.936	1.000	0.471	-0.011	0.029	-0.053	-0.060	-0.056
POP	-0.002	-0.064	-0.039	-0.079	0.411	0.471	1.000	0.076	-0.063	-0.067	-0.165	-0.076
MR	-0.203	0.215	-0.286	-0.245	0.108	-0.011	0.076	1.000	-0.042	-0.028	-0.256	-0.043
SPI	0.301	-0.228	0.256	0.180	0.022	0.029	-0.063	-0.042	1.000	0.843	0.212	0.842
ARI	0.282	-0.221	0.249	0.206	-0.058	-0.053	-0.067	-0.028	0.843	1.000	0.233	0.998
DPI	0.640	-0.442	0.691	0.682	-0.078	-0.060	-0.165	-0.256	0.212	0.233	1.000	0.290
TRI	0.317	-0.244	0.287	0.244	-0.062	-0.056	-0.076	-0.043	0.842	0.998	0.290	1.000

Correlation Tests

To detect how significant a correlation is we conducted Pearson's correlation test on significant correlations in our dataset alongside Spearman's correlation test to detect any non-linear correlation in our dataset.

Results are in the following table:

	Correlation Test Results			
	Pearson	p value %	Spearman	p value %
MR AGE	-0.2033	0.005	-0.2027	0.005
MR POR	0.2151	0.003	0.2602	0.0002
MR HDI	-0.2861	6.577	-0.3354	2.39
MR GDPCAP	-0.2448	0.0007	-0.3622	3.033
HDI SPI	0.2563	0.0003	0.7136	1.003
HDI TRI	0.2866	6.379	0.7749	4.1623
HDI GDPCAP	0.6468	8.831	0.9459	2.195
GDPCAP TRI	0.2444	0.0007	0.8293	3.792
GDPCAP SPI	0.1797	0.0133	0.6869	9.912

Note: p value < 0.05 (5%) is regarded statistically significant

According to the tests we conducted, amongst statistically significant correlations ($p < 0.05$ or $< 5\%$), HDI and SPI have a much stronger correlation (0.7136) than seen on the heatmap alongside MR and HDI which is also significantly stronger (-0.3354).

Other significant changes in correlation from the table seem to be HDI&TRI (0.7749), GDPCAP&TRI (0.8293) and lastly GDPCAP&SPI (0.6869).

Our Hypotheses on the Correlations of Spread and HDI, GDP and Mortality Rate

Our first hypothesis was that, when it came to mortality rate, the biggest factor was GDP per capita for it is the main deciding factor on how good healthcare a country can afford.

While the hypothesis has truth to it, the correlation tests show us that poverty caused by income inequality has a rather significant effect on the mortality rate.

One notable detail about the correlations is that, even though that it is a well-known fact that covid is more fatal for people of older age, countries that have higher mean age tend to have less mortality due to the correlation of age, HDI and GDP per capita.

Our second hypothesis was that the virus spreads farther and to more people in more developed and wealthier countries because people who live in these countries tend to travel more frequently.

The correlations between GDPCAP&TRI (0.8293) and HDI&TRI (0.7749) proves this hypothesis.

Aside from our hypotheses, the negative correlation between HDI and poverty rate (-0.721) shows us that the more a country is developed, income inequality becomes that less of a problem.

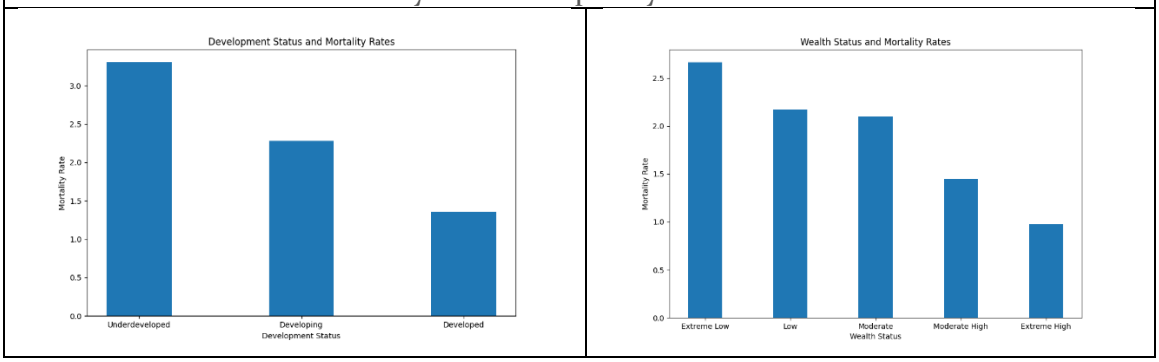
Summation of Extracted Information

Regarding how far the virus spreads, it seems that the more a country is developed and wealthier, the farther the virus seems to spread. Considering the relationship between HDI and GDP per capita, it is easier for people living in wealthy and developed countries to travel around the globe and carry around the virus.

However, despite being able to spread farther, the virus has less mortality in developed and wealthy countries due to the fact that they have better healthcare access compared to people living in poor/underdeveloped countries.

One interesting find is that, even though it is a known fact that the virus is deadlier for people of older age, it is more preferable to be in a country with higher avg. age due to the relationships between age, HDI and GDP per capita.

Mortality Rates Grouped by HDI & Wealth



Aside from that, income inequality and lack thereof of GDP seems to have a significant relationship with mortality rates. What this tells us is that even though the virus is less likely to spread as far as it would in a developed country, it will be significantly more lethal in underdeveloped countries.

Machine Learning: Classified Prediction

The goal of the machine learning part was to develop a machine learning model to predict which country will fall into which mortality rate group. (less than 1.5% is low, 1.5 to 2.5 is moderate, above 2.5 is high).

MRG as our target variable we started testing multiple classified prediction algorithms, with the intent to build a stacked classifier for increased accuracy, which yielded the following results:

Results		
Logistic Regression	0.5285172778561354	0.5368901229670766
Gaussian Naïve Bayes	0.4187015162200282	0.42106307021023404
K Neighbors Classifier	0.9990964386459803	0.9980166600555335

After seeing that KNC yielding an accuracy of 99% we thought that it could be overfitting, so we tried cross validating it, yielding:

[0.99933892, 0.99845747, 0.99669458, 0.99779639, 0.99779639, 0.99779639, 0.99713467, 0.99889795, 0.99845713, 0.99911836]

[0.96633663, 0.96831683, 0.97619048, 0.95436508, 0.9702381, 0.96825397, 0.96230159, 0.96428571, 0.9702381, 0.97619048]

Looking at the cross validation results we can tell that it is not overfitting. So why such an unusual accuracy rate? The answer to that is that, since differences between each class is so significant, alongside the similarities within a group which are rather significant as well. Considering these, it is possible for the machine learning model reach an unusual accuracy of 99%.

References

- Cover Image: https://www.cdc.gov/media/subtopic/images.htm //#23311
- Age Data: https://worldpopulationreview.com/country-rankings/median-age
- Poverty Rate: https://worldpopulationreview.com/country-rankings/poverty-rate-by-country
- Travel Data: https://data.worldbank.org/indicator/ST.INT.ARVL , https://data.worldbank.org/indicator/ST.INT.DPRT
- HDI Data: https://worldpopulationreview.com/country-rankings/hdi-by-country
- GDP per capita: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
- Population: https://worldpopulationreview.com/
- Covid Data: Taken live at 16/12/2021 5:30AM from: https://www.worldometers.info/coronavirus/