Project Final Report: Jerry He (jgh39), Andrew Wu (aw339), Alex Whang (amw166), Lucy Gu (wg74), Steven Bochner (sb487)
4/20/22

Part 1: Introduction and Research Questions

COVID has made a significant impact on our world over the course of the past few years - but for many countries, that is not the leading cause of death. Whether it's through viral diseases such as COVID or deaths caused by heart disease, cancer, violence, and suicide, our team was interested in studying general trends between a country's standing in the world and the reason their citizens are dying. Furthermore, we would like to explore if there is a way to identify key factors that influence a country's total deaths as well as COVID-19 deaths.
Our research questions are:

- How do different types of economic, sociological, and demographical data correlate with various causes of death?
- What factors identify a country's total deaths?
- Can we identify the best indicators for deaths due to COVID-19?

Our motivation behind these research questions ultimately stems from wanting to understand how mortality rates of countries correlate with different factors. Ultimately, we want to discover ways to lower mortality rates by identifying these factors. Then, we can use these results to see what factors should be focused on by a country and allot more resources.

Part 2: Data Sources

To investigate our research questions, we used the "Death Cause by Country" dataset from Kaggle (which contains information regarding the causes of deaths across all genders and age groups) as well as multiple datasets found at http://data.un.org/**.** We are looking at 10 datasets from the UN website with information about country population growth rates, imports & exports, consumer price indices, land, population densities, employment, GDP, expenditure on health, and migration. These datasets are very relevant to our research questions as they provide the statistics needed to investigate a relationship between a country's causes of deaths and its other characteristics (demographics, economic indicators, social conditions, etc.).

Part 3: Results and Methods

GitHub repo: https://github.com/awu339/cs216_finalproject
LGBook.ipynb- includes a cleaning method, several sample visualizations, and work to demonstrate a potential/straightforward merge method.
AWBook.ipynb- includes the actual application of the cleaning method onto the csv files, as well as the method which merges all the cleaned data frames from each csv files together and extracts relevant rows for the actual data visualization performed later. Also includes various filters, ranging from standardization, to summing, etc.

Visualizations.ipynb - includes scatterplots of 5 different economic, sociological, and demographic factors in relation to 5 different causes of death in a country
Visualizations Results.ipynb - includes scatterplots of 5 different economic, sociological, and demographic factors in relation to the number of COVID-19 deaths in a country; also includes the linear regression estimates used to produce r-squared and MSE values from each of the scatterplots (10 total across Visualizations.ipynb and Visualizations Results.ipynb).

Cleaning

```
def clean(dataset):
    codes = dataset.groupby("Area Code").count().index
    i = 0
    data1 = [pd.DataFrame()] * len(codes)
    curName = ""
    for ind in codes:
        data1[i] = dataset[dataset["Area Code"] == ind]
        curName = data1[i]["Area Name"].reset_index(drop=True)[0]
        data1[i] = data1[i].pivot_table("Value",index="Year",columns="Series",aggfunc='mean')
        data1[i]["Area Code"] = [ind] * len(data1[i])
        data1[i]["Area Name"] = [curName] * len(data1[i])
        data1[i]=data1[i].reset_index()
        i = i+1

    data_final = data1[0]
    for ind in range(len(data1)):
        if(ind!=0):
            data_final = pd.concat([data_final, data1[ind]])
    data_final = data_final.reset_index(drop=True)
    return data_final
```

The clean method is tailored to deal with the particular type of csv tables we found on data.un.org. The dataset is formulated such that the "Series" column contains the variable name that corresponds to the values in the values column, and ideally for visualization purposes we would want these variable names to be standalone column headers. The best way to do this is naturally via a pivot table - however, there's a catch, because of how each dataset contains data about every country and the rows are only differentiated by country names in the Area Name and Area Code columns, doing a direct pivot table would mush them all together and that is not something we want.

The clean function is designed to solve this problem. It first grabs all the possible unique Area Codes that exist in the dataframe, then creates an array of empty data frames that is the same length. For each entry of the array then, the code first filters out only the rows with the corresponding Area Code, then uses the pivot table function to put the filtered dataset into Year vs [Different Variable Names in Series column] format (ie, index is Year, and column is Series). The Year index column is then manually added back as a column, and the actual index of the data frame is resetted to avoid potential duplicate index problems when the data frames are merged back together and used for visualization (the index was in fact resetted in a later stage too

so potentially this is duplicate functionality, but it's never bad to be careful). The Area Code and the Area Name are also manually re-added to this new data frame to distinguish between different sub-data frames once they're re-merged together.

After the array is populated thoroughly, where each entry has a "cleaned" data table from a particular country, we loop through the array again to concat them all back together. Although our final analysis and visualization might be drawn on individual countries and we might need to separate them by filtering again, we think that it would be "cleaner" and easier to manage if all data are accessible from one big dataframe instead of having to go through various small data frames for each area in order to find them. The end result of this method is then a "cleaned" dataset, including all rows from the original dataset, where the columns are reformatted to have only Year, Area Code, Area Name, Value, and [List of variable names appearing in original Series column].

The merge method in this example book was a simple one, whose sole purpose is to demonstrate a possible merge strategy on the cleaned data. Different datasets are merged with the outer method (to keep all data even if one table might not have data in certain years matching the other), and on columns "Area Code", "Area Name", and "Year". We ended up using a more complicated merging method in our prep to actual visualization, since we decided we only care about the most recent data for each country from each table. The method and rationale will be described below.

Merging and Filtering (AWBook.ipynb)

After cleaning up the data, we decided on what we wanted to do with the cleaned data before merging it. Because the UN data that we were using were spread across the different dates, the first thing we did was only identify the most recent data for each country, which is what the "recent" function does, which only takes the most recent year of each data set for each country. One concern is that the last year of data is sometimes different from year to year, so for the purpose of this prototype, we removed the year by setting it all the same and then removing it on the merge. Next step is just merging all the data sets together based on 'Area Code' and 'Area Name' (which is the country name). This is done by a for loop that iterates through all the datasets and merges them for 192 countries. Finally, after creating the "merged" df with all the UN data, we do an outer join with the death data in order to create our final data table with comparison. We do left on 'Area Name' and right on "Country Name" in order to create our final data table for comparison. Contains all dataframe work including standardization of deaths (dividing by total population) as well as various summations for overall data.

Visualizations & Results (Visualizations.ipynb)

For visualizing the data, we decided to create multiple scatterplots to investigate the relationship between economic, sociological, and demographic factors of a country and its number of deaths due to various causes. We also performed linear regression estimates on several

sets of variables (found in Visualization Results.ipynb). Our observations are listed below with the most important scatterplots shown:

- **1st scatterplot (on left)**: By plotting the current health expenditure of a country with its deaths due to COVID-19, we can see a slight negative correlation between these two variables for countries that have a significant number of COVID-19 deaths (5000+). However, there are also many countries without a significant number of COVID-19, so this slight negative correlation may change as we further investigate the relationship between health expenditure and COVID-19 deaths in countries with fewer than 5000 COVID-19 deaths. After performing a linear regression estimate on these two variables, we found an r-squared value of 0.016 and an MSE value of about 627,817,861.
- **2nd scatterplot (found in Visualizations.ipynb):** By plotting the urban population of a country with its deaths due to interpersonal violence we can see that there appears to be very little correlation between these variables. After performing a linear regression estimate on these two variables, we found an r-squared value of 0.136 and an MSE value of about 49,580.
- **3rd scatterplot (found in Visualizations.ipynb):** By plotting the population annual rate of increase of a country with its deaths due to interpersonal violence we can see that there appears to be very little correlation between these variables. After performing a linear regression estimate on these two variables, we found an r-squared value of 0.0006 and an MSE value of about 50,300,828.
- **4th scatterplot (found in Visualizations.ipynb):** By plotting a country's life expectancy at birth for both sexes with its deaths due to COVID-19, we can see that there appears to be a slight positive correlation between these variables for countries that have a significant number of COVID-19 deaths (5000+). However, there are also many countries without a significant number of COVID-19, so this slight positive correlation may change as we further investigate the relationship between life expectancy at birth for both sexes and COVID-19 deaths in countries with fewer than 5000 COVID-19 deaths. After performing a linear regression estimate on these two variables, we found an r-squared value of 0.026 and an MSE value of about 665,704,642.
- **5th scatterplot (found in Visualizations.ipynb):** By plotting the general consumer price index of a country with its deaths due to self-harm, we can see that there appears to be very little correlation between the variables. After performing a linear regression estimate on these two variables, we found an r-squared value of 0.002 and an MSE value of about 343,919,454.

We also created 5 scatterplots to investigate the relationship between 5 different economic, sociological, or demographic factors and the number of COVID-19 deaths of a country **(all found in Visualization Results.ipynb - cells #7 - 11)**. We also conducted linear regression estimates on these sets of variables; however, we did not find any significant results that would lead us to believe that any one of these 5 factors is a good indicator of a country's COVID-19 deaths. The highest r-squared value from the linear regression estimates was about 0.107.

- 1st scatterplot (Population density vs. Covid-19 deaths): We found an r-squared value of about 0.003 and an MSE value of about 152,255,276.
- 2nd scatterplot (Population annual rate of increase (percent) vs. Covid-19 deaths): We found an r-squared value of about 0.016 and an MSE value of about 634,869,527
- 3rd scatterplot (Capital city population (thousands) vs. Covid-19 deaths): We found an r-squared value of about 0.107 and an MSE value of about 1,688,970.
- 4th scatterplot (Total fertility rate (children per women) vs. Covid-19 deaths): We found an r-squared value of about 0.039 and an MSE value of about 657,278,226.
- 5th scatterplot (Employment by industry: Services (%) Male and Female vs. Covid-19 deaths): We found an r-squared value of about 0.012 and an MSE value of about 644,983,188.

- **SB_Book.ipynb:** This notebook contains more visualizations that are meant to explore the data.Combined with AWBook - these books also include correlation analysis between standardized data and all factors. This is sorted and shown in AWBook which shows which factors have a significant correlation in comparison to other correlations and returns us some decent correlations that also makes sense logically - such as countries with a generally older population tend to suffer more from COVID and that countries with less unemployment tend to have lower death rates. This book also plotted these correlations via scatterplots to showcase said correlations visually.

Part 4: Limitations and Future Work

- **Limitations**
    - Lack of consistent data was our biggest problem. The UN data had inequitable reporting since many countries only had data for certain years - which means that it's hard to track from a year to year basis and doing it from only the most recent standpoint forces us to have to absorb the error of change in time. Many countries also just lack accountable reporting which makes it hard to generalize across from country to country.
    - Confounding variables exist aplenty partly due to just the lack of power of our engine. We can't account for certain factors that maybe more qualitative in scale vs quantity (such as political factors or various other globalization trends).
- **Future work**
    - Through this project we focused on using linear regressions logistic regressions - but these are sometimes not the best models for prediction. Future work can be used to work on models that are more accurate models as well as classification models that can hopefully help us make predictions rather than identify trends.
    - As mentioned before - we want to be able to make predictions. Given certain factors of a country, we would like to know if we can predict the amount of deaths that the country should have for a certain factor, then we can compare it to real

numbers and see which countries are doing better than expected or the opposite and take a deep dive. This is something we did not get to in the span of this project and hope to explore in the future.
    - We'd also like to explore the possibility of combining different factors in our model. Most of our regressions are based on a 1 to 1 analysis, but hopefully in the future, we are able to combine multiple different factors based off of some sort of connector and be able to compare that to deaths to obtain a more accurate model.

Part 5: Conclusion

- The general conclusion we actually came to was that it's very hard to predict causes of death solely based off of country data. Even after standardization of our data - our correlations were not very strong in relation to any form of death due to various amounts of confounding factors. However, the general trends that we do believe do exist - as shown in some of our Jupyter notebooks - things such as older age or more employment or general trade balance can correlate to amount of deaths or reasons of death. Our models confirmed certain hypothesis such that countries with older population tend to suffer from COVID deaths,
- We also concluded that it's near impossible to find a reason of causation because of so many factors that can be quantified or not analyzed within our models. We can only find general correlations and make educated guesses about which factors affect which parts of death.
- One thing we also noted was that covid deaths is an anomaly compared to total death trends. Countries with higher imports and exports (positive trade balance) suffered more from covid even though they were developed countries, same thing with GDP calculations. For example countries with high health expenditure also tended to have more people die from covid.
- We cannot predict causes of death in a 1 to 1 scenario because each factor is so highly correlated with other factors within a country and thus the best way to approach this problem is by trying to blend together and then comparing it to death causes in order to make the most accurate predictions.

Part 6: Appendix of additional figures and tables

(All figures and tables are connected within our github and our notebooks, since we did try to work out charts on multitudes of factors and exchanging x and y axis - too many figures to attach to an appendix.)