Exploring Factors Related to Global Rates of COVID-19

Project for Introduction to Programming Module by Amy Reidy

May 17th, 2021

---

1. Project Scope

The overall aim of this project is to utilize skills learned in the 'Introduction to Programming' module to explore factors that may explain differences in global rates of COVID-19. There are two main goals:

1. Test the claim that female-led countries have lower rates of COVID-19.

While researching factors related to COVID-19 rates, I was intrigued by studies and media reports claiming that countries with female heads of state were handling the pandemic better than male leaders, and that female-led countries have lower rates of COVID-19 deaths per capita (Garikipati & Kambhampati, 2020, Coscieme et al., 2020). However, a study by Winsor et al (2020) asserts that this perception has been caused by data selection bias and Western media bias that has amplified the successes of female leaders in OECD countries, and that there is not a significant difference globally in countries led by women versus men.

And so, this project tests the hypothesis that female-led countries have lower rates of COVID-19 than male-led countries, using all the countries in the dataset and for a separate sample of just OECD countries.

2. Create a linear regression model to predict rates of COVID-19 in countries around the world.

For the regression analysis section, multiple linear regression models were created using both Tensorflow and Scikit-learn. The target variable was 'Confirmed Cases of COVID-19 per 100,000 Population', and the following independent variables were used:

- Smoking – Smoking impairs lung functioning and COVID-19 is a virus which primarily attacks the lungs, and there is evidence that smokers may have a higher likelihood of developing more severe symptoms of the disease than non-smokers (Shastri et al., 2021, Lewis, 2020).

- Obesity – Studies around the world have identified obesity and severe obesity as risk factors for hospitalization and mechanical ventilation in relation to COVID-19 (Popkin et al., 2020, Gao et al., 2020, Dietz, W., Santos-Burgoa, C., 2020).

- Life Expectancy - Older people are more vulnerable to COVID-19, and they are more likely to have serious COVID-19 symptoms and require hospitalization or even succumb to the virus. And studies show that countries with high life expectancy, such as Italy, have had high case rates and death rates (Dowd et al., 2020).

- OECD Membership – this is a categorial variable for whether the country is part of the Organisation for Economic Co-operation and Development (OECD).

Most OECD countries are high-income, developed countries, which seem to generally have higher confirmed rates of COVID-19 compared to middle- and low-income countries. One of the many reasons for this could be due to higher testing rates and less deaths misattributed to other causes, as these countries may have more robust health systems than poorer countries. Unfortunately, there is no accurate data available to confirm this.

2. Description of the Data

COVID-19 Cases and Deaths:
The data for COVID-19 rates around the world was retrieved by sending an API request to 'https://api.covid19api.com/summary'. The API request was extracted in JSN format and restructured into a dataframe which has four columns ('Country': name of country, 'Code': ISO 3166-1 alpha-2 codes for each country, 'Cases': total confirmed cases, 'Deaths': total confirmed deaths). There were 190 countries in the dataset, with zero null entries.

Population:
The population CSV file was sourced from https://ourworldindata.org/grapher/population?time=2019..latest. The data is from 2019 and the table was restructured to contain three columns ('Country', 'Code': the alpha-2 country codes, 'Code3': the alpha-3 country codes which were used when creating the folium maps, and 'Population'). The population for each country was used to calculate the cases and deaths per 100,000 people.

Smoking:
The smoking CSV file was sourced from https://ourworldindata.org/smoking. The data is from 2017 and it shows the % share of deaths attributed to direct smoking for each country. The table was restructured to contain three columns ('Country', 'Code' and 'Smoking': the % share of smoking deaths).

Obesity:
The obesity CSV file was sourced from https://ourworldindata.org/obesity. The data is from 2017 and it shows the % share of deaths attributed to obesity (defined as having a BMI of over 30.0) for each country. The table was restructured to contain three columns ('Country', 'Code' and 'Obesity': the % share of obesity deaths).

Life Expectancy:
The CSV file sourced from https://ourworldindata.org/life-expectancy. The data is from 2019 and was restructured to contain three columns ('Country', 'Code' and 'Life_Expectancy': the life expectancy for each country measured in years).

Female Leaders:
To get a list of countries that have had a female leader for the COVID-19 outbreak, I used BeautifulSoup to web-scrape a table with female heads of state and government from this Wikipedia page: https://en.wikipedia.org/wiki/List_of_elected_and_appointed_female_heads_of_state_and_government
The data is up-to-date, and the table was restructured to contain three columns ('Country', 'Code', 'Female_Leader': this column was used to create a dummy variable in a new table). There was a total of 19 countries in the table.

OECD Countries:
The last dataset is a list of all countries part of the Organisation for Economic Co-operation and Development (OECD). This list was obtained from https://www.oecd.org/newsroom/global-oecd-welcomes-colombia-as-its-37th-member.htm#:~:text=The%20OECD's%2037%20members%20are,Poland%2C%20Portugal%2C%20Slovak%20Republic%2C. I pasted the string into the notebook and converted it to a list, then created a dataframe with the following columns: 'Country', 'Code' and 'OECD': a column of 1's to be used to obtain a dummy variable.

Data Cleaning and Preprocessing:

- All datasets were restructured to extract only the relevant columns with the most recent values
- A column was added with the ISO 3166-1 alpha-2 codes for each country to ensure consistency for joining the tables for analysis.
- Dataframes were checked for null values. Rows with regional data were dropped and missing country codes for other rows were entered manually.
- Dataframes were also checked for significant outliers, and any outliers which were more than 3 standard deviations away from the mean of the sample were omitted.

3. Database

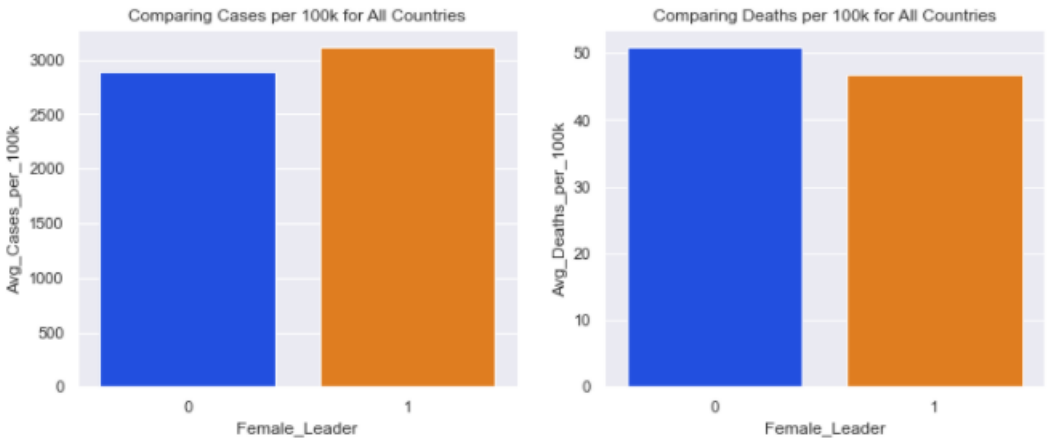
The original database I created contained seven tables ('leaders', 'covid', 'lifeexp', 'oecd', 'population', 'smoking', 'obesity'), and I merged these tables to create three more tables to aid the analysis ('covid_rates', 'femaleleaders', 'covid_factors'). All tables were joined using the alpha-2 country codes.

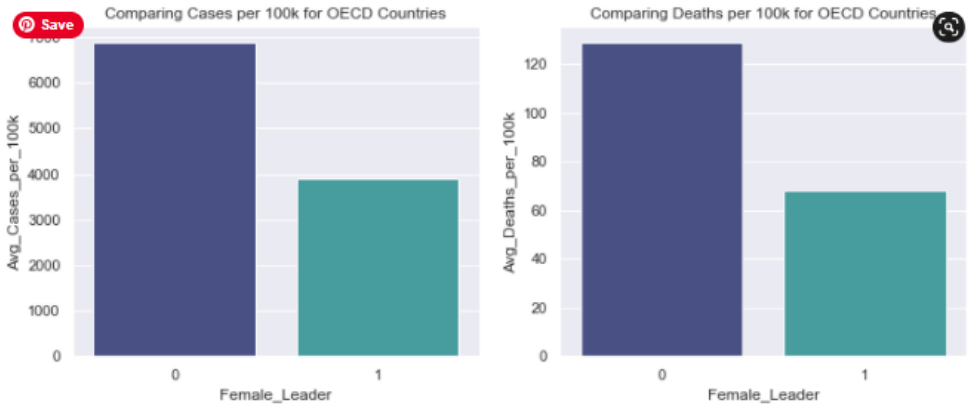
4. End Analysis/Visualisation

a) Testing the claim that female-led countries have lower rates of COVID-19.

Null Hypothesis:
Female-led countries have the same rates of COVID-19 as male-led countries.

Alternative Hypothesis:
Female-led countries have lower rates of COVID-19 than male-led countries.





We can see from the bar charts that there is not much difference between female-led and male-led countries in a global context, however for OECD countries, female-led countries do appear to have lower rates of COVID-19 for both confirmed cases and deaths.

The samples were not normally distributed (even after transforming the data) and there were lots of outliers. Therefore, the Mann-Whitney U test, which is suitable for non-parametric samples, was used to test the following hypotheses.

Null Hypotheses	Statistic	p-value	Reject H0 at α = 0.050?
1. H0: Average number of cases per 100,000 population are the same globally for female-led and male-led countries.	1345.0	0.166	Fail to Reject H0
2. H0: Average number of deaths per 100,000 population are the same globally for female-led and male-led countries.	1492.0	0.382	Fail to Reject H0
3. H0: Average number of cases per 100,000 population are the same for female-led and male-led OECD countries.	58.0	0.014	Reject H0
4. H0: Average number of deaths per 100,000 population are the same for female-led and male-led OECD countries.	59.0	0.015	Reject H0

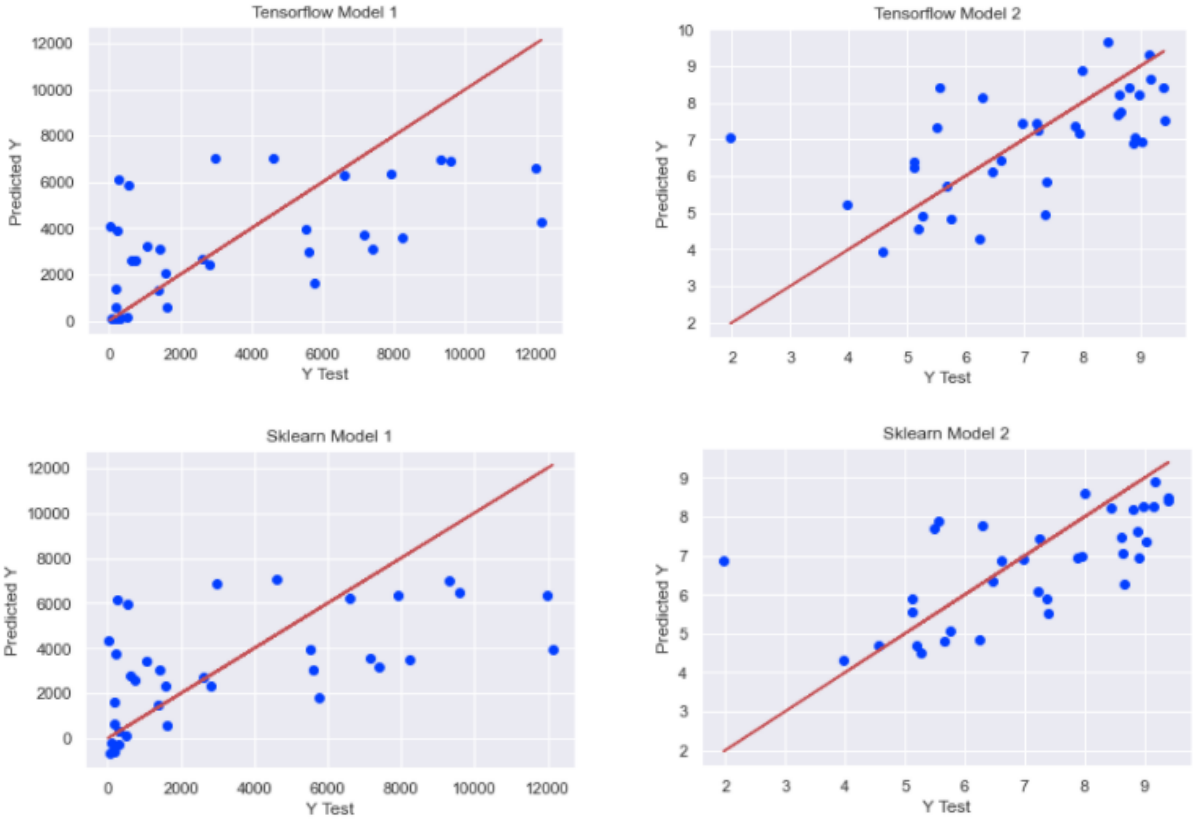
The table shows that there is not a significant difference between female-led and male-led countries worldwide as the p-values are greater than 0.05. However, there is indeed a significant difference between the female-led and male-led OECD countries, with p= 0.014 for differences in case rates and p= 0.015 for differences in deaths.

b) Linear regression model to predict rates of COVID-19 in countries across the world.

While visualizing the relationships between the variables on pairplots, I noticed that the distribution of the dependent variable seemed to be very skewed. And I carried out a log transformation on the target variable to check if this would increase the linearity between the target and independent variables.

However, although there seemed to be a bit more linearity with the log transformed target, there were still lots of outliers. So, I decided to create different multiple linear regression models (using Tensorflow and then scikit-learn) with the untransformed target and the log transformed target to compare the results.

	Mean Absolute Error	Root Square Mean Error	Explained Variance Score
TF 1 - Cases Untransformed	2182.873490	2971.264505	0.355178
SK 1 - Cases Untransformed	2304.412780	3041.817929	0.326387
TF 2 - Log Transformed Cases	1.147537	1.505135	0.296414
SK 2 - Log Transformed Cases	1.079075	1.412329	0.407280



From the table and charts above, we can see that the best performing linear regression model was built using sklearn and the log transformed target, as it had the lowest scores for mean absolute error and root mean square, and the highest explained variance score at 41%.

5. Conclusion

Overall, I think this project was successful as it produced some interesting findings. While the hypothesis testing showed evidence of a significant difference between female-led OECD countries and male-led OECD countries, there was no significant difference when considering all countries in the dataset. This seems to reinforce Winsor et al.'s critique of the media and other research studies for focusing too much on OECD countries when reporting on lower rates of COVID-19 for female-led countries compared to countries led by men.

I also feel that the regression analysis was successful to a certain extent, as the strongest regression model has an explained variance score of 41%, and although this might be considered low, it is higher than I expected to achieve due to the complex nature of the virus.

However, there are still lots of ways in which the regression analysis could be improved, for example:

- More independent variables could be added, such as air pollution rates, the population density of a country, if the country is an island, etc.
- The moderate to high correlation between the independent variables suggests that there may be an issue with multicollinearity. This should be investigated further as multicollinearity undermines the statistical significance of an independent variable.
- More outliers could be omitted, such as small island nations with very low populations, to make the data less noisy, or the significant outliers which were dropped could have been kept in the dataset, such as Czech Republic and Hungary, to make the data less biased.
- The data for obesity and smoking is for 2017, but trends show that obesity is increasing globally every year, while smoking deaths rates are generally decreasing in most countries (but increasing in some low-to-middle income countries), so more current data would give more accurate results.

How could the data be used in the future?

It is important for us to try to understand the factors that are related to high COVID-19 cases and deaths around the world, for us to try to minimize the spread and severity of this virus and other viruses in the future. For example, if findings show that the elderly population are more at risk from these types of viruses, governments should try to introduce more policies that protect older citizens.

Also, if further studies show more evidence of strong correlation between factors such as smoking or obesity and a higher risk of hospitalization or death due to viruses like COVID-19, then there are more incentives for individuals to make better lifestyle choices.

Lastly, it is encouraging to see that in developed countries, there is evidence to suggest female leaders have been more effective at handling this pandemic. These findings may boost the public's perceptions of women leaders, encourage more women to enter politics and leadership positions, and it may also inspire male heads of state to reflect upon what they can learn from the leadership and communication styles of leaders such as Jacinda Ahern (New Zealand), Angela Merkel (Germany) and Tsai Ing-wen (Taiwan).

6. References

- Coscieme, L., Fioramonti, L., Mortensen, L. F., Pickett, K. E., Kubiszewski, I., Lovins, H., ... & Wilkinson, R. (2020). Women in power: female leadership and public health outcomes during the COVID-19 pandemic. *MedRxiv*.
- Dietz, W., & Santos-Burgoa, C. (2020). Obesity and its implications for COVID-19 mortality. *Obesity*, *28*(6), 1005-1005.
- Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., ... & Mills, M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences*, *117*(18), 9696-9698.
- Gao, F., Zheng, K. I., Wang, X. B., Sun, Q. F., Pan, K. H., Wang, T. Y., ... & Zheng, M. H. (2020). Obesity is a risk factor for greater COVID-19 severity. *Diabetes care*, *43*(7), e72-e74. Garikipati, S., & Kambhampati, U. (2021). Leading the Fight against the Pandemic: Does Gender really matter?. *Feminist Economics*, *27*(1-2), 401-418.

- Lewis, T. (2020). Smoking or vaping may increase the risk of a severe coronavirus infection. *Scientific American*, *17*.
- Popkin, B. M., Du, S., Green, W. D., Beck, M. A., Algaith, T., Herbst, C. H., ... & Shekar, M. (2020). Individuals with obesity and COVID-19: A global perspective on the epidemiology and biological relationships. *Obesity Reviews*, *21*(11), e13128.
- Shastri, M. D., Shukla, S. D., Chong, W. C., Kc, R., Dua, K., Patel, R. P., Peterson, G. M., & O'Toole, R. F. (2021). Smoking and COVID-19: What we know so far. Respiratory medicine, 176, 106237. https://doi.org/10.1016/j.rmed.2020.106237
- Windsor, L. C., Yannitell Reinhardt, G., Windsor, A. J., Ostergard, R., Allen, S., Burns, C., ... & Wood, R. (2020). Gender in the time of COVID-19: Evaluating national leadership and COVID-19 fatalities. *PloS one*, *15*(12), e0244531.