**Project 1**

**CPSC 375 - 01**

Group Members:

Parthiv Desai – Sec 02

Sarthak Gajjar – Sec 01

Siddharth Chauhan- Sec 02

Title: Predicting Daily COVID-19 Deaths with Linear Modeling

1. **Introduction**

This project aims to predict daily COVID-19 deaths in different countries using linear modeling. We consider various factors, such as the number of cases, vaccination rates, development level, age demographics, and pre-existing medical conditions like diabetes. The main dataset used is the Our World in Data COVID-19 dataset, which we combine with demographic data.

1. **Data Wrangling**

We began by loading and wrangling the OWID COVID-19 data and demographics data using the read\_csv() function. We filtered the data to only include rows with a valid 3-letter ISO code and countries with a population of at least 1 million. We removed irrelevant columns, such as excess death rates, and added a new column new\_deaths\_smoothed\_2wk using lead function for a 14-day lead for making a new column “new\_deaths\_smoothed\_2wk”.We combined Series Name and Series Code using the unite function to unite both the columns to tidy the demographics dataset but also separated the values with a “\_” to separate both the columns for the next part. The reason to combine both the columns is that It will form a wide data frame with one row per country which is the efficient order for a better regression model.

Renaming the column “Country\_code” to “iso\_code” in the demographics for merging purposes as “iso\_code” can be used for performing joins between the two datasets.

1. **Variable Selection and Transformation**

We carefully selected predictor variables based on prior knowledge, domain expertise, and trends observed in plots. For instance, the iso\_code variable was used for performing joins between datasets since it serves as a unique identifier. In addition, we created three transformed variables to better capture the underlying relationships in the data. These transformed variables include cardiovasc\_deaths, diabetes\_cases, and all\_vaccinations.

To compute cardiovasc\_deaths, we multiplied cardiovasc\_death\_rate by total\_cases. This new variable allowed us to account for the impact of cardiovascular disease on COVID-19 deaths. Similarly, diabetes\_cases was calculated by multiplying the diabetes\_prevalence variable by total\_cases, capturing the potential effect of diabetes on COVID-19 mortality rates. Lastly, we derived the all\_vaccinations variable by multiplying total\_vaccinations by total\_cases, highlighting the role of vaccination efforts in mitigating the severity of the pandemic.

By generating these transformed variables, we aimed to create a more robust and accurate linear model that would better account for the complexities of predicting daily COVID-19 deaths.

1. **Linear Modeling and Evaluation**

We divided our dataset into training (data from 2022) and testing (data from 2023) subsets. We built five different linear models using various combinations of predictor variables, including the transformed variables.

We calculated the RMSE for all models on the testing dataset and identified the best model based on the lowest RMSE. For the best model, we also calculated the RMSE for each country.

1. **Results**

The results of the RMSE values for the five linear models are as follows:

Model 1 RMSE: 52.72166

Model 2 RMSE: 54.17523

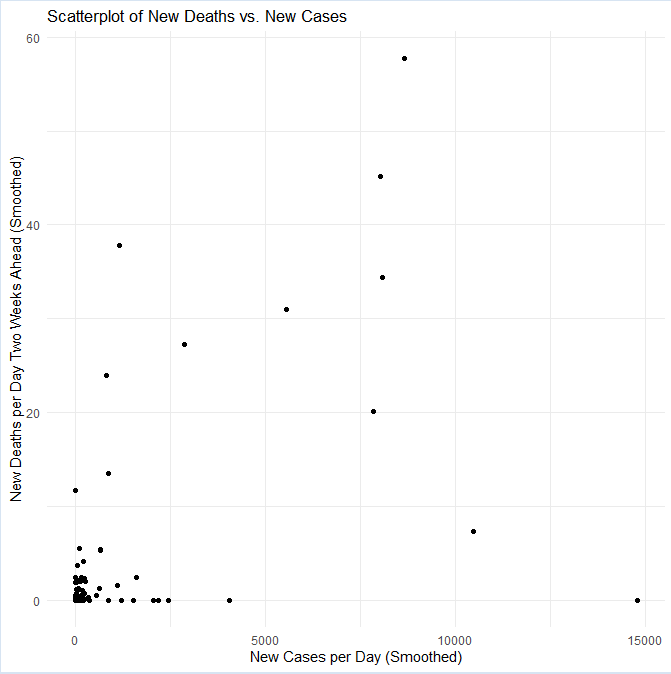
Model 3 RMSE: 45.59158

Model 4 RMSE: 54.16819

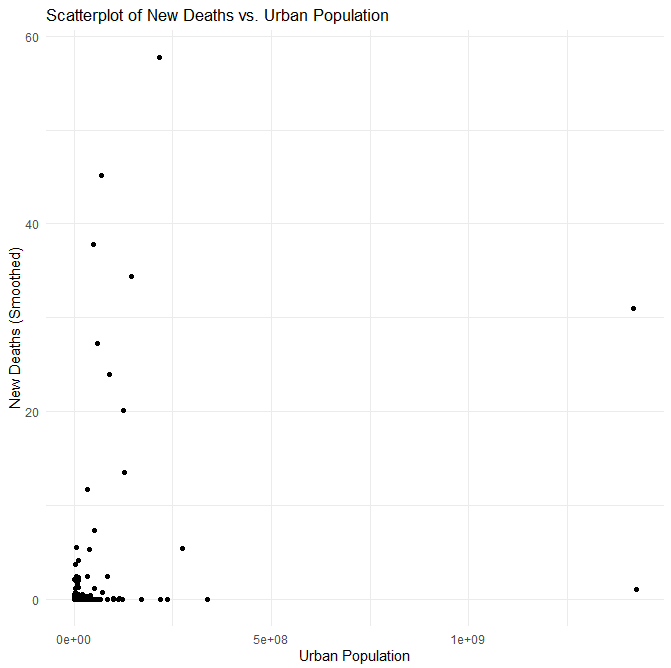
Model 5 RMSE: 52.67958

Based on the lowest RMSE value, Model 3 is the best model.

**The scatterplot of the most recent new\_deaths\_smoothed\_2wk and new\_cases\_smoothed for every country:**



**The scatterplot of the most recent new deaths per day and the urban population:**



The table below shows the R² and RMSE values for the different models:

**Model R² RMSE**

1 0.9056 52.72166

2 0.9054 54.17523

3 0.9082 45.59158

4 0.9054 54.16819

5 0.9056 52.67958

The table below shows the RMSE of the best model (Model X) for the 20 most populous countries (excluding NaN):

**Country RMSE (Model 3)**

United States 97.579

Japan 81.029

Germany 78.101

France 83.515

Italy 11.176

Spain 16.994

Canada 3.168

Malaysia 5.053

Australia 16.182

Chile 5.511

Belgium 5.748

Sweden 9.934

Czechia 2.077

Israel 1.962

Austria 1.194

Switzerland 0.749

Bulgaria 4.874

Denmark 5.378

Ireland 4.561

Estonia 1.415

1. **Conclusion**

Our most accurate model (Model 3) indicates that cardiovasc\_death\_rate, population, icu\_patients, and new\_cases\_smoothed are the primary factors affecting daily COVID-19 deaths. These findings suggest that in addition to new cases which is the intuitive factor, there are other factors as well such as the population, ICU patients, etc. In Scenarios like the recent pandemic, this type of data and the information that we were able to extract from it can prove to be really valuable for governments and health organizations to make informed decisions regarding public health measures and resource allocation.

1. **Code**

Our final R code is well-structured, using the tidyverse library for readability and efficient data manipulation. The code includes adequate comments to guide readers through the steps taken in the project.