CPSC 375 Project Report

Group Members:

- Jacob Armstrong
- Dane Camacho
- David Harboyan

Important Data Wrangling Steps

To perform data wrangling, we begin by loading the “tidyverse” library. Then we read both datasets into variables (“covid_data” and “population_data”) using the “read_csv()” function.

After that, we removed all rows where the “country_code”/”iso_code” was not 3 letters long. We only needed to do this for “covid_data” since all instances of the variable in “population_data” were exactly 3 letters long.

Following this, we removed countries whose total population was less than 1 million by retrieving the countries with a population of 1 million or more.

To add a “new_deaths_smoothed_2wk”, which has the same values as “new_deaths_smoothed” but two weeks ahead, we grouped by “country_code”/“iso_code” and added a new column that is set to the result of: `lead(new_deaths_smoothed, n = 14))`

The “lead()” function above gets the value of “new_deaths_smoothed” that is new_deaths_smoothed 14 days (2 weeks) in the future. We store the result in “new_deaths_smoothed_2wk”. Note that this was done for “covid_data.”

Next, we needed to tidy “population_data” using the “pivot_wider()” function. We set the “names_from” parameter to “Series Code” and the “values_from” parameter to “2023 [YR2023]”. We also chose to remove the “Series Name” column to reduce redundancy.

The last step in our data wrangling was to merge “covid_data” and “population_data.” We used an inner join and joined by the “country_code”/“iso_code.” We stored the merged table in “merged_table.”

Variables Selected From Population Webpage

From the population estimates webpage, we selected 2 variables:

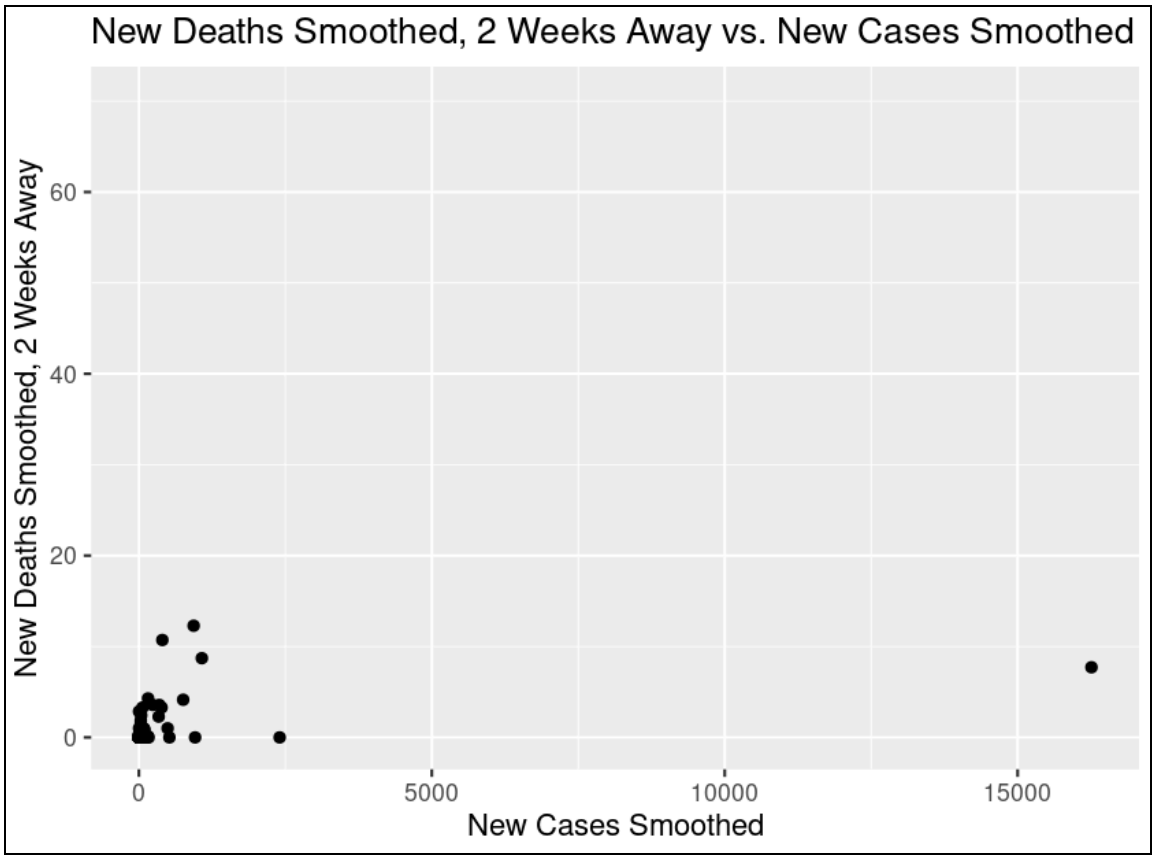
- Population ages 80 and above, female
- Population ages 80 and above, male

We selected these because it would be interesting to see how COVID-19 affects the older population since they are known to have weaker immune systems.

Scatterplot: New Deaths Smoothed, 2 Weeks Away vs. New Cases Smoothed

R code for scatterplot:

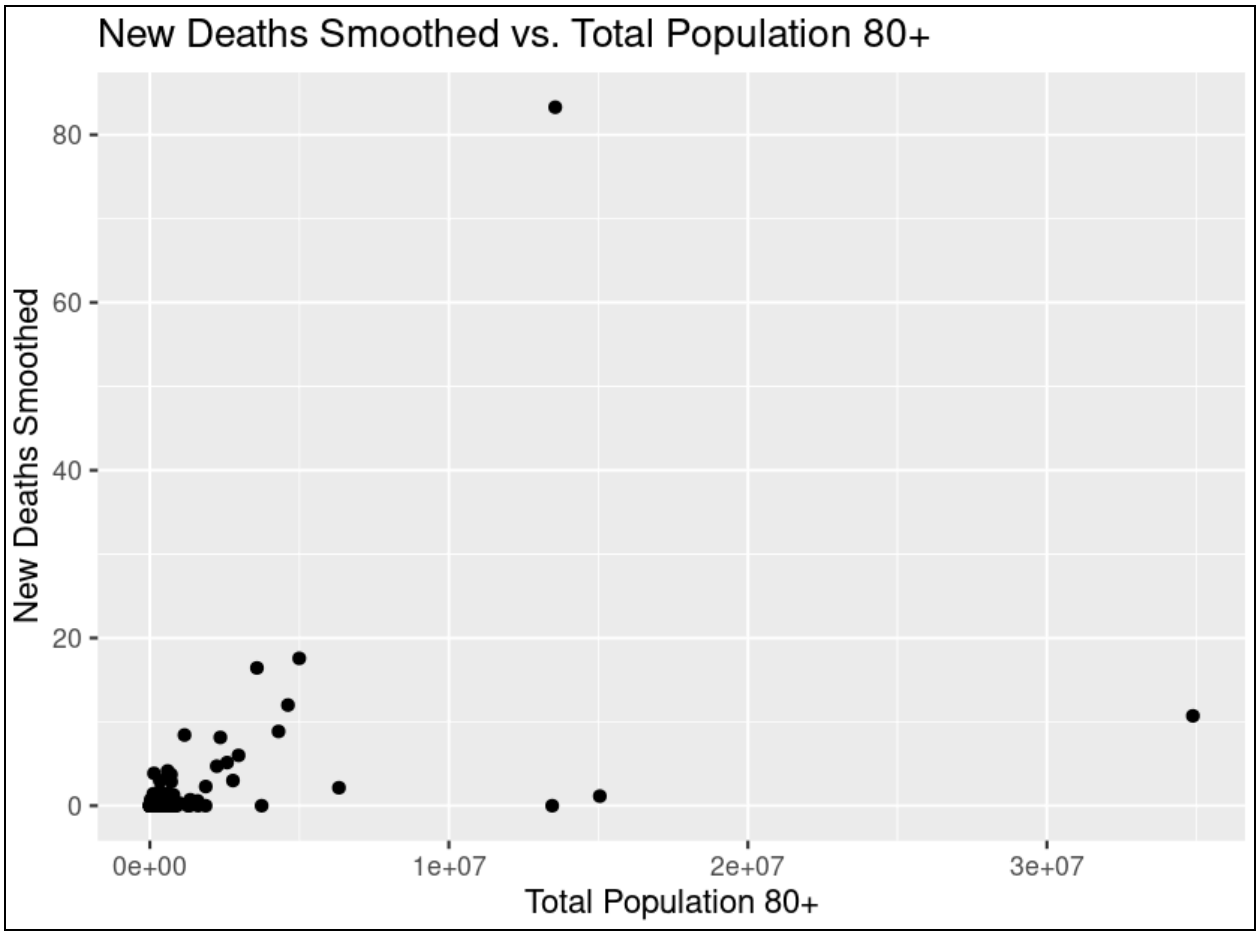
```
> plot_data <- merged_table %>% filter(date == “2023-06-30”)
> ggplot(data=plot_data) + geom_point(aes(x =
new_cases_smoothed, y = new_deaths_smoothed_2wk)) + labs(x =
"New Cases Smoothed", y = "New Deaths Smoothed, 2 Weeks Away",
title = "New Deaths Smoothed, 2 Weeks Away vs. New Cases
Smoothed")
```



Scatterplot: New Deaths Smoothed vs. Total Population 80+

R code for scatterplot:

```
> plot_data <- merged_table %>% filter(date == "2023-06-30")
> ggplot(data=plot_data) + geom_point(aes(x =
(as.numeric(SP.POP.80UP.FE) + as.numeric(SP.POP.80UP.MA)), y =
new_deaths_smoothed)) + labs(x = "Total Population 80+", y =
"New Deaths Smoothed", title = "New Deaths Smoothed vs. Total
Population 80+")
```



Variable Transforms

We generated 3 transformed variables:

- diabetic_population
- total_smokers
- SP.POP.80UP.TOT

Description for diabetic_population:

diabetic_population = (diabetes_prevalence * 0.01) * population

This transformed variable represents the number of people in the population that have diabetes. We take diabetes_prevalence (which is a percentage), convert it to a decimal, and multiply it by the population.

Description for total_smokers:

total_smokers = female_smokers + male_smokers

This transformed variable represents the total number of smokers in the population. We simply add female_smokers and male_smokers together.

Description for SP.POP.80UP.TOT:

SP.POP.80UP.TOT = as.numeric(SP.POP.80UP.FE) + as.numeric(SP.POP.80UP.MA))

This transformed variable represents the total number of people in the population that are 80 years old and up. We add SP.POP.80UP.FE and SP.POP.80UP.MA together (Note that “as.numeric()” was needed).

Combinations of Predictor Variables In Our Models

Model 1:

```
lm(new_deaths_smoothed_2wk~new_cases_smoothed+diabetic_populatio
n, data=merged_2022)
```

The predictor variables for this model are:

- new_cases_smoothed
- diabetic_population

We chose these predictors because people who are diabetic and have COVID-19 are more likely to die than those without diabetes. This is because individuals with diabetes usually have weaker immune systems (due to inflammation). So if they have COVID-19, they would have a harder time fighting it off.

Model 2:

```
lm(new_deaths_smoothed_2wk~new_cases_smoothed+total_smokers+card
iovasc_death_rate+life_expectancy, data=merged_2022)
```

The predictor variables for this model are:

- new_cases_smoothed
- total_smokers
- cardiovasc_death_rate
- life_expectancy

We chose these predictors because people who smoke typically develop cardiovascular diseases, lowering their life expectancy. Combining this with

COVID-19 would further affect smokers as they are more prone to serious respiratory infections such as pneumonia since smoking weakens the immune system.

Model 3:

```
lm(new_deaths_smoothed_2wk~new_cases_smoothed+people_fully_vaccinated+total_cases, data=merged_2022)
```

The predictor variables for this model are:

- new_cases_smoothed
- people_fully_vaccinated
- total_cases

We chose these predictors because as time goes on it is possible for the new deaths to decrease. This is because more people will get vaccinated and the general population should have been exposed to it, which both increase the odds of survival.

Model 4:

```
lm(new_deaths_smoothed_2wk~icu_patients+diabetic_population+total_smokers+SP.POP.80UP.TOT+people_fully_vaccinated, data=merged_2022)
```

The predictor variables for this model are:

- Icu_patients
- diabetic_population
- total_smokers

- SP.POP.80UP.TOT
- people_fully_vaccinated

We chose these predictors because those who are ICU patients are in critical condition. With this in mind, if they also have diabetes, are smokers, are older (80+), they have a higher chance of dying. But it would also be people who are fully vaccinated, that should help them fight off the infection.

Model 5:

```
lm(new_deaths_smoothed_2wk~life_expectancy+icu_patients+total_vaccinations, data=merged_2022)
```

The predictor variables for this model are:

- life_expectancy
- Icu_patients
- total_vaccinations

We chose these predictors because if overall life expectancy is high and total vaccinations are high, there will likely be less patients ending up in the ICU due to covid complications. Lower life expectancy and less vaccinations, combined with a higher number of ICU patients could lead to more deaths.

R2 and RMSE of All Our Models

	R2	RMSE
Model 1	0.3041	133.985
Model 2	0.2833	134.9722
Model 3	0.6391	131.9727
Model 4	0.8681	50.73344
Model 5	0.8871	48.23147

Based on the results above, it appears that Model 5 is the best (highest R2 / lowest RMSE values).

RMSE Of The 20 Most Populous Countries In Our Best Model (Model 5)

	RMSE
United States	71.1
Japan	348
Germany	67.6
France	80.9
Italy	13.6
Spain	15.9
Canada	15.0
Malaysia	12.0
Australia	62.9
Chile	5.31
Netherlands	9.48
Belgium	6.13
Sweden	8.10
Czechia	3.03
Israel	3.86
Austria	3.50
Switzerland	4.92
Bulgaria	2.95
Denmark	2.67
Ireland	1.58

Conclusion

Our most accurate model, Model 5, consisted of predictor variables containing information on life expectancy, the number of vaccinations, and the number of ICU patients. We chose these predictor variables as we believed higher life expectancy in general and more vaccinations would lead to less ICU patients dying from COVID-19 complications. The reverse would also be true – lower life expectancy and less vaccinations, with a higher number of ICU patients would lead to more deaths from COVID-19. The R2 value for Model 5 was the highest among all the models we tested, and the RMSE value was lower than the others, meaning this model would be able to more accurately predict the number of deaths considering the predictor variables we chose. It is likely that the variables do correlate with the number of COVID-19 deaths in a given location. Considering the results from all models, significant factors when predicting death rates include diabetes, smoking, old age, being in the ICU, life expectancy, and vaccinations. As for non-significant factors, new cases of COVID-19 do not actually seem to be a good predictor of death rates. This is most likely because the significant factors mentioned play a much bigger role in determining the odds of someone dying from COVID-19.