Angela DeLeo
Roman Saddi
Kinsey Vo
CPSC 375-02
Project Report

Project: Predicting COVID-19

Data Wrangling

In the first phase of data analysis, data wrangling must occur to properly group and form data that is “tidy.” Tidy data organizes data in a consistent and standardized manner, which simplifies data manipulation, analysis, and visualization tasks. Before wrangling the data, we had selected various variables when preparing our Population dataset. These variables are as follows: population of individuals 80 and up both male and female, total population, total urban and rural populations, population of males and females, and life expectancy at birth.

After reading our datasets in R, we began the data wrangling and data tidying steps. The steps we took are as follows:

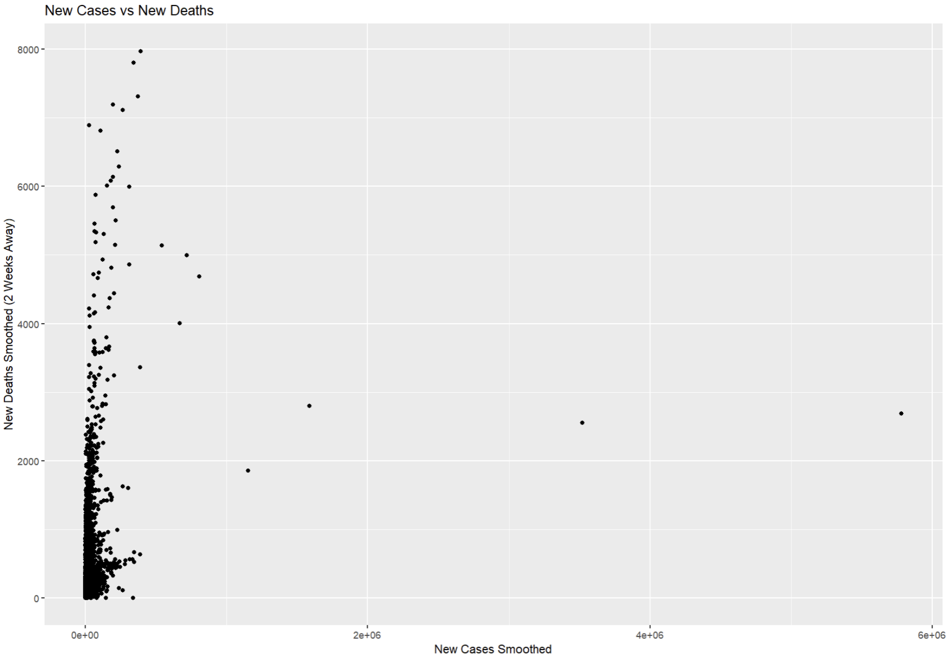
- Removed country-level data by removing rows in which country_code is not exactly three letters. We did this to both the Covid dataset and the Population dataset.
- Removed countries with low population (under 1 million). We did this to both the Covid dataset and the Population dataset.
- We then created our dependent variable, new_deaths_smoothed_2wk, by calculating the data that is two weeks ahead in new_deaths_smoothed. This dependent variable was created in our Covid dataset.
- Used the pivot_wider function R provides on our population data retrieved from The World Bank in order for each variable to be its own column. We achieved this by removing the column “Series Name” and pivoting the remaining data from long to wide format based on the values in the “Series Code” column, with the value from the “2023 [YR2023]” column filling in the new wide-format columns.
- Finally, we merged the two tables using the full_join function and merged them by iso_code and “Country Code.”

By doing these steps, we have successfully tidied our data and wrangled it to be more readable and simple for the next tasks. Next we will plot our data to visualize it.

Scatterplots

Scatterplot 1

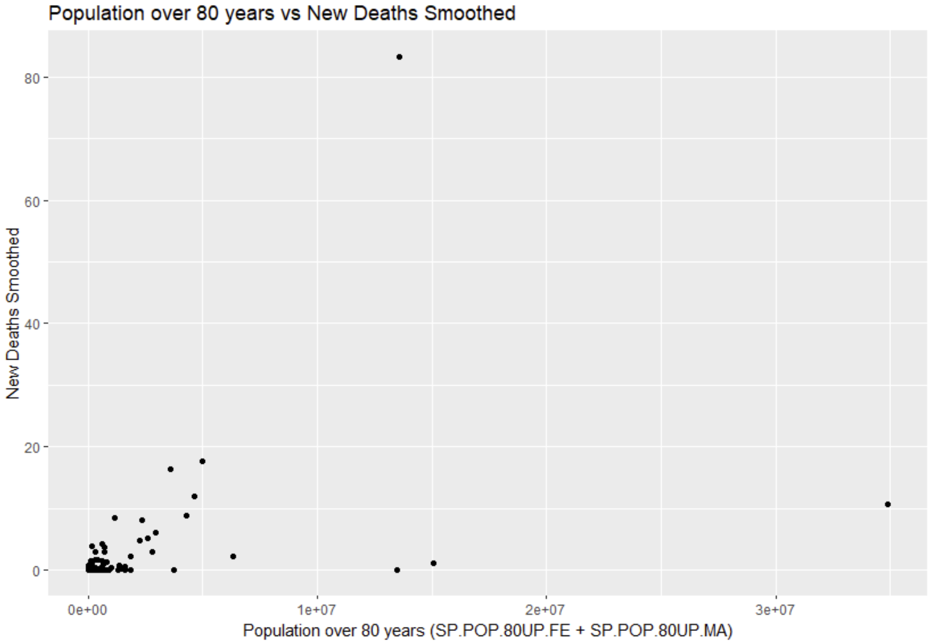
Most recent new deaths per day two weeks ahead and the corresponding new cases per day (new_cases_smoothed) for every country



Scatterplot 2

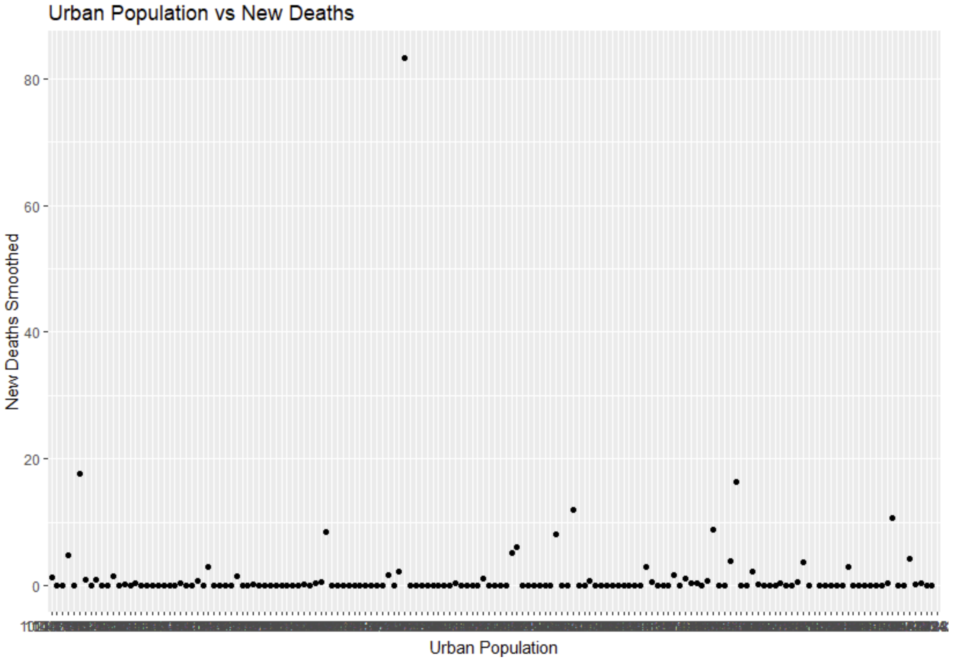
Most recent new deaths and the total (female + male) population over 80 for every country with date ‘2023-06-30’

- Convert SP.POP.80UP.FE & SP.POP.80UP.MA to numeric before plotting



Scatterplot 3

Most recent new deaths per day and the urban population



Linear Modeling

In the second phase of data analysis, linear modeling must occur to begin to visualize the data at hand. We first listed out all the predictor variables available in our datasets: total_vaccinations, new_cases_smoothed, total_cases, SP.POP.TOTL, stringency_index, median_age, gdp_per_capita, life_expectancy, population_density, diabetes_prevalence, and cardiovase_death_rate.

Next, we generated a variety of transformed variables in order to not only normalize the variables but also to ensure variables are prepared for linear modeling, even if they are nonlinear variables. We generated the following four transformed variables: cardiovasc_deaths, total_vaccinations_per_capita, economic_vulnerability_index, and healthcare_capacity_index. We chose to generate these variables in particular to improve the predictive performance of our model by accounting for additional factors.

Below is a snippet of our code generating those transformed variables.

```
merged_data <- merged_data %>%
   mutate(cardiovasc_deaths =
     cardiovasc_death_rate * population) %>%
   mutate(total_vaccinations_per_capita =
     total_vaccinations / as.numeric(SP.POP.TOTL)) %>%
   mutate(economic_vulnerability_index =
     (gdp_per_capita * population) / extreme_poverty) %>%
   mutate(healthcare_capacity_index =
     (icu_patients_per_million + hosp_patients_per_million) / 2)
```

Next, we split our dataset in order to train and test our model. We used data from 2022 in order to build and train our linear models using R’s lm() function. We used data from 2023 in order to test, and later on, evaluate our data.

Finally, we ran eight linear regression models with each containing at least five different combinations of predictor variables in the formula. Within these eight linear regression models, we combined a variety of predictor variables in the formulas. We decided on which variables to combine based on our prior knowledge of COVID-19 related deaths in addition to understanding the age groups, economic groups, and a variety of other discerning factors that we know COVID-19 had impacts on. In each model, we had a few common variables that we included in our formulas: new_cases_smoothed, total_vaccinations_per_capita, median_age, and population_density.

We included these four variables in each of our linear model's formulas as we know that those variables are the most relevant when discussing COVID-19. The rest of the variables we utilized in our linear models varied, we focused on economics, health conditions, and health care to determine which variables to include in our formulas. We wanted to ensure we covered as much of the variables that we believed were relevant with COVID-19 as possible.

After we ran the eight linear regression models, we determined the adjusted R-squared for each of them using R's built-in function summary(). By doing this, we could predict which model would be the best.

Evaluating the Linear Models

In the third and final phase of data analysis, we evaluate all the data after linear modeling was completed. In this phase of data analysis, we determine the best model to use for predicting COVID-19 related deaths by determining and comparing R-squared values in addition to Root Mean Squared Error of each model. The best model is the one with the highest R-squared value and the lowest Root Mean Squared Error, meaning it has the most accuracy rate and the lowest error rate. To achieve this, we used the rmse() function in library(modelr) to calculate the Root Mean Squared Error (RMSE) on all of our models and we used summary() to determine the R-squared (R2) value of our models.

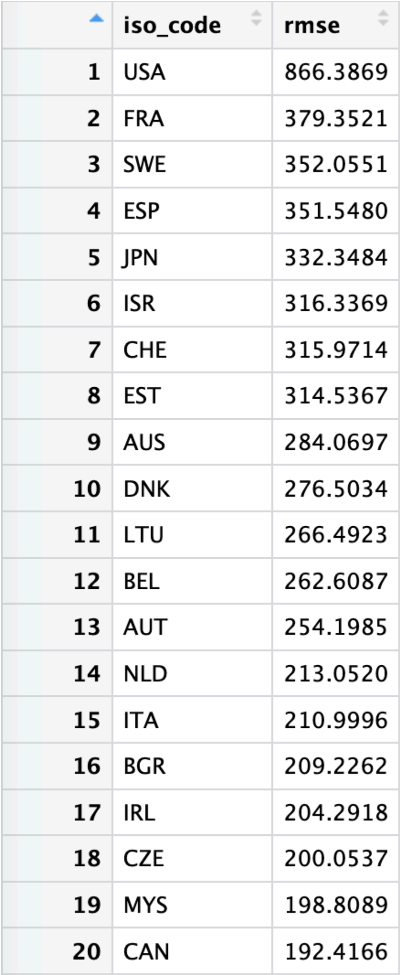
Below is a table listing the R2 and RMSE values we calculated for all eight of our models.

All Models	R2	RMSE
lm1	0.2466	NaN
lm2	0.2116	355.942
lm3	0.6379	520.0336
lm4	0.2107	356.2785
lm5	0.9032	484.939
lm6	0.9093	563.9375
lm7	0.5834	NaN
lm8	0.7622	585.2031

Our R-squared was the most accurate for our fifth and sixth models, lm5 and lm6. However, lm5 has a lower RMSE than lm6 does, thus making lm5 the best model out of the eight models we ran. Although lm6 has a slightly higher R2 value, its higher RMSE suggests that its predictions might be less accurate compared to lm5. Thus, we believe model lm5 is the best model as it has a high R2 value and a lower RMSE compared to the other seven models.

For our best model, lm5, we then calculated the Root Mean Squared Error for each country in our dataset. We first calculated the RMSE for each country using the best model, lm5. Then, after filtering out all of the NAs from the table using na.omit(), the data went down from 159 rows to 20 rows. Finally, we arranged the data in decreasing order of population.

Below is a table that displays the 20 most populous countries arranged in decreasing order of population.



	iso_code	rmse
1	USA	866.3869
2	FRA	379.3521
3	SWE	352.0551
4	ESP	351.5480
5	JPN	332.3484
6	ISR	316.3369
7	CHE	315.9714
8	EST	314.5367
9	AUS	284.0697
10	DNK	276.5034
11	LTU	266.4923
12	BEL	262.6087
13	AUT	254.1985
14	NLD	213.0520
15	ITA	210.9996
16	BGR	209.2262
17	IRL	204.2918
18	CZE	200.0537
19	MYS	198.8089
20	CAN	192.4166

Conclusion

In conclusion, our most accurate linear regression model, lm5, focuses on several key implications regarding COVID-19 death rates in addition to the factors influencing them. As our model lm5 has a high adjusted R-squared value and a relatively low Root Mean Squared Error, we know that this is our best model out of the eight models we developed, thus making it the most reliable model for predicting COVID-19 related deaths. The variables utilized in the making of this model were as follows: new_cases_smoothed, total_vaccinations_per_capita, median_age, population_density, cardiovasc_deaths, diabetes_prevalence, icu_patients, and healthcare_capacity_index.

These variables align with the knowledge we already have regarding COVID-19's impact. It is known that those with lower immune systems, such as individuals with cardiovascular diseases or diabetes, are more prone to contracting COVID-19 as well as dying from it. We also know that higher transmission rates, lower vaccination rates, age, and population density correlate with higher death rates with regards to COVID-19.

Our best model, lm5, shows that COVID-19 mortality is influenced by many factors. It goes beyond just the number of new cases and vaccination rates, highlighting the importance of demographic and health-related factors. Things like age distribution, existing health conditions, and healthcare capacity all affect death rates.

Datasets and References

1. Data on COVID-19 from *Our World in Data*
2. Population estimates from The World Bank's DataBank