# CPSC 375 - Introduction to Data Science and Big Data

# Final Project

Bryce Lin and Serop Kelkelian

May 1, 2023

# 1. Data Preprocessing

## 1.1 Data Wrangling

The first step of the data wrangling/preparation was to read the two data files needed for this project using `read_csv()`. Most data manipulation was performed on the OWID dataset. Non-country-level data was removed using `nchar(string)` to determine whether or not the country code was for a continent. Following that, `filter()` was used to remove countries with a total population under one million people. Additionally, columns that were not used for linear modeling such as “deaths” columns excluding `new_deaths_smoothed` were removed using `select()`. With unnecessary columns now removed, an additional column was created that was used for linear modeling. The column is `new_deaths_smoothed_2wk`, and it contains the same values as `new_deaths_smoothed` just two weeks ahead of time. This was done by making a copy of the entire table, subtracting 14 from the `date` column, renaming the new column, and joining both tables together using `inner_join()`**.** To make the `demographic` dataset tidy, `pivot_wider()` was used.

## 1.2 Variable Transformations

Prior to variable transformation, the chosen variables are very important. First and foremost, the variables were chosen based on the percentage of NA values in the column. This was done using the formula `sum(is.na(dataset$columnName))/ nrow(dataset)`. Using RMSE, the chance of getting NaN is greatly minimized. Variables that were somewhat related to one another were put together. This was done to see if there was any correlation between the variables that were similar.

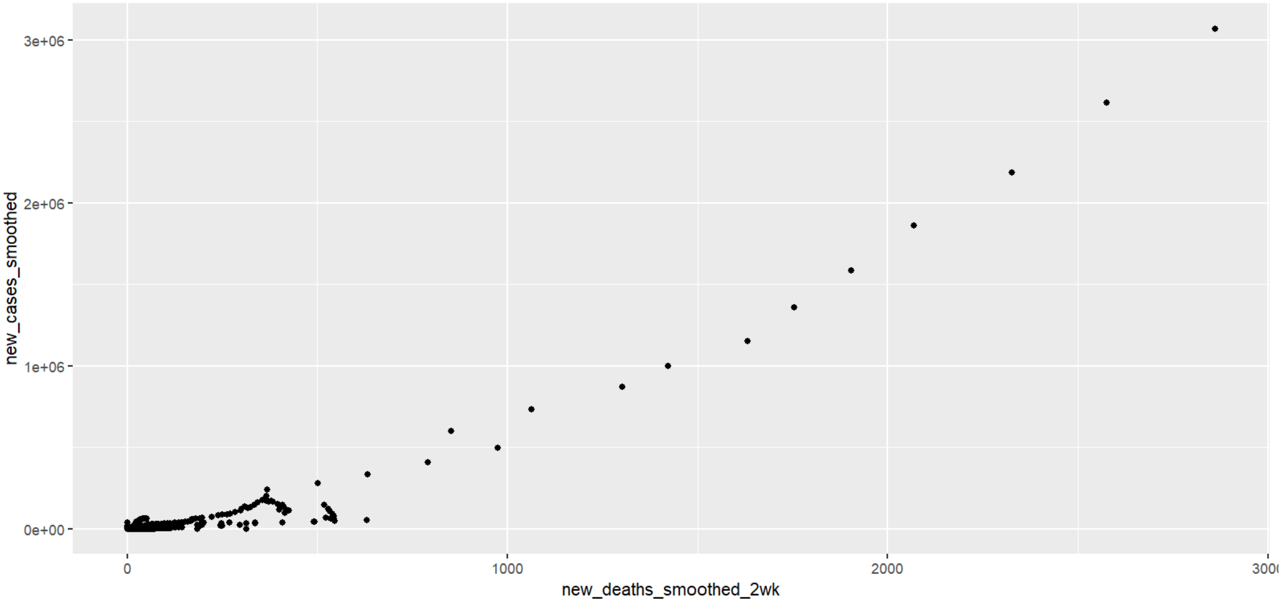
# 2. Training the Model

## 2.1 Data Modeling

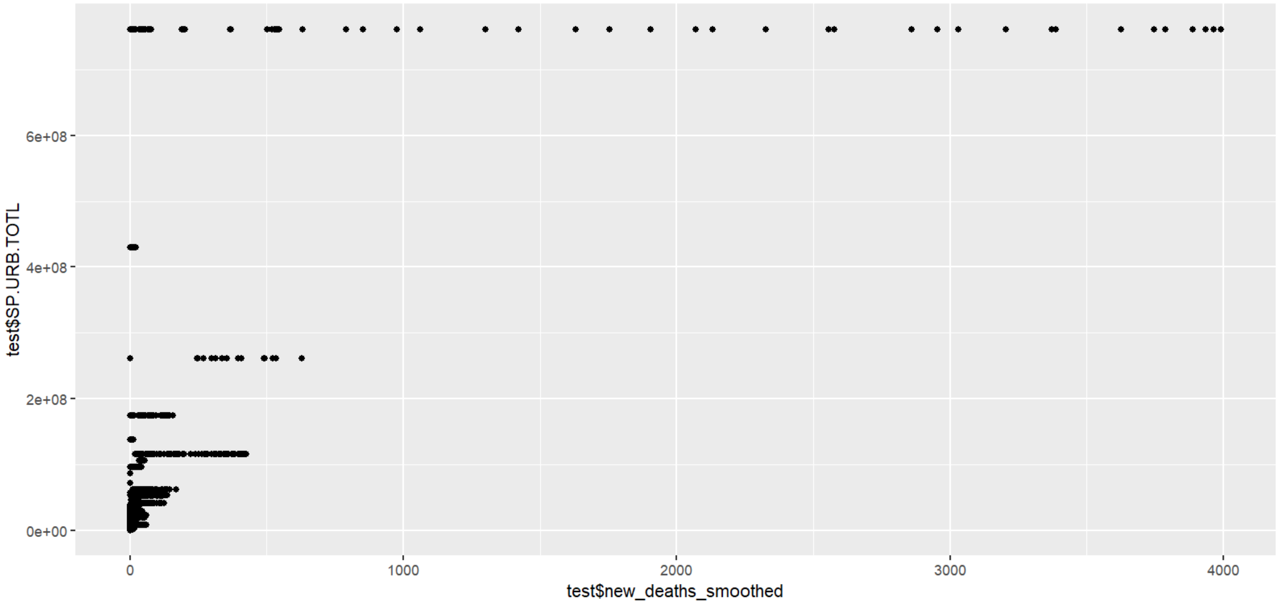
When performing data modeling, the variables were not chosen based on any particular traits such as percentage of NA values or correlation to one another. The deciding factor in regards to the variables chosen was variable importance. There is a pool of important variables which were randomly split into five different equations. The adjusted r-squared was kept track of to determine which combination of variables resulted in a higher adjusted r-squared.

# 3. Data Evaluation

## 3.1 Scatterplots



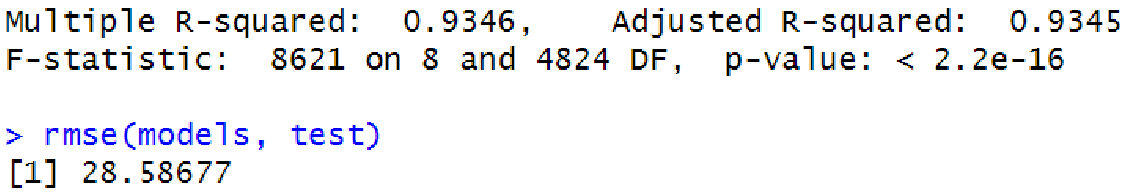
new_deaths_smoothed_2wk vs. new_cases_smoothed for every country



new_deaths_smoothed vs. urban population

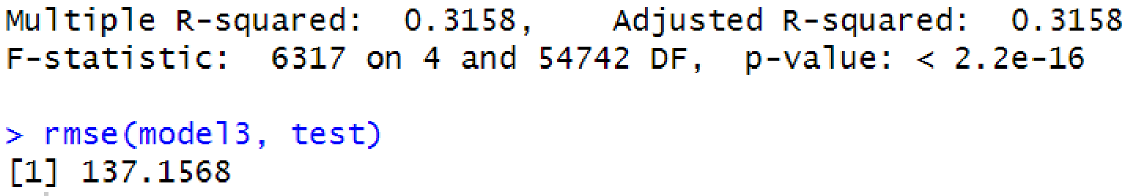
## 3.2 R2 and Root Mean Squared Error

**Model S:**



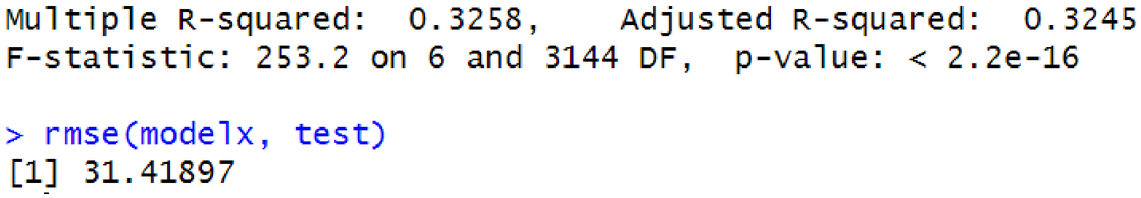
```
Multiple R-squared:  0.9346,	Adjusted R-squared:  0.9345 
F-statistic:  8621 on 8 and 4824 DF,  p-value: < 2.2e-16

> rmse(models, test)
[1] 28.58677
```

**Model 3:**



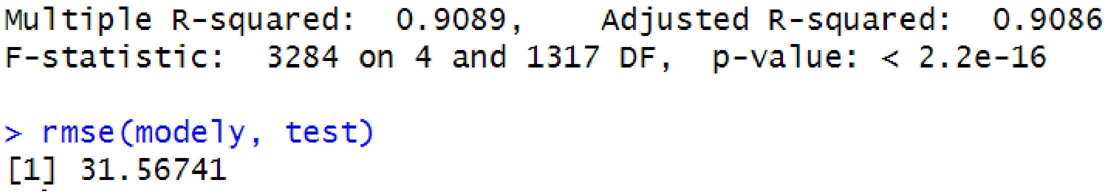
```
Multiple R-squared:  0.3158,	Adjusted R-squared:  0.3158 
F-statistic:  6317 on 4 and 54742 DF,  p-value: < 2.2e-16

> rmse(model3, test)
[1] 137.1568
```

**Model X:**



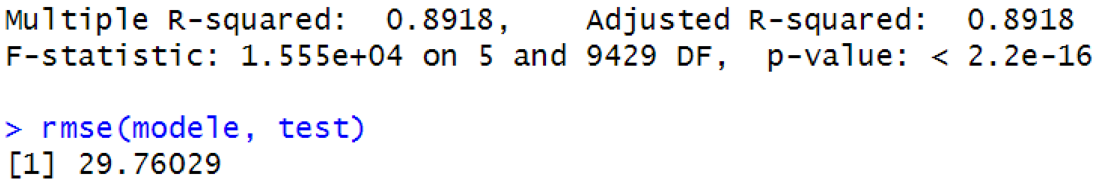
```
Multiple R-squared:  0.3258,	Adjusted R-squared:  0.3245 
F-statistic: 253.2 on 6 and 3144 DF,  p-value: < 2.2e-16

> rmse(modelx, test)
[1] 31.41897
```

**Model Y:**



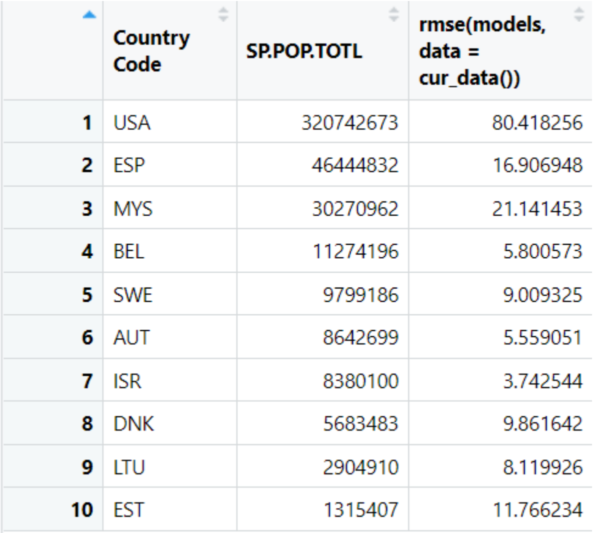
```
Multiple R-squared:  0.9089,	Adjusted R-squared:  0.9086 
F-statistic:  3284 on 4 and 1317 DF,  p-value: < 2.2e-16

> rmse(modely, test)
[1] 31.56741
```

**Model E:**



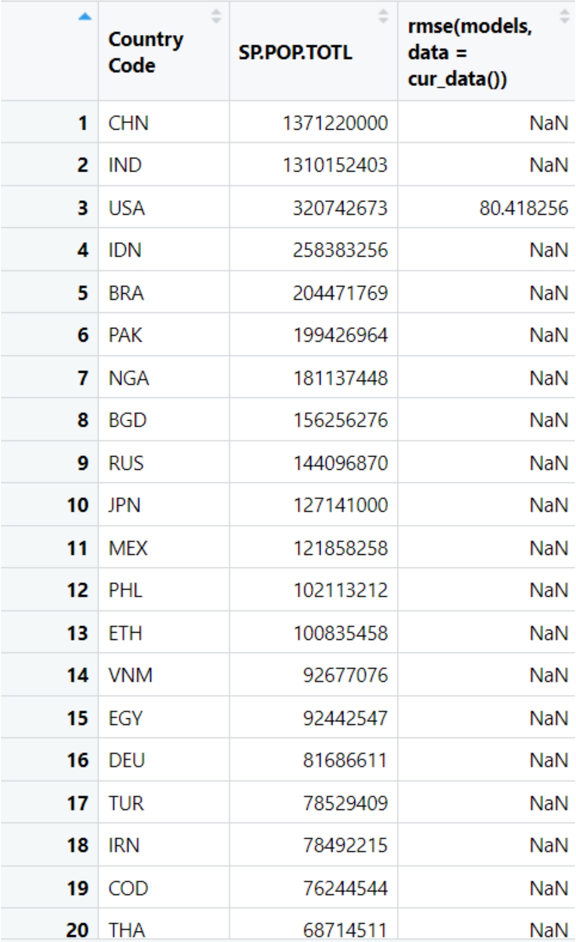
```
Multiple R-squared:  0.8918,	Adjusted R-squared:  0.8918 
F-statistic: 1.555e+04 on 5 and 9429 DF,  p-value: < 2.2e-16

> rmse(modele, test)
[1] 29.76029
```

## 3.3 Tables



| | Country Code | SP.POP.TOTL | rmse(models, data = cur_data()) |
|---|---|---|---|
| 1 | USA | 320742673 | 80.418256 |
| 2 | ESP | 46444832 | 16.906948 |
| 3 | MYS | 30270962 | 21.141453 |
| 4 | BEL | 11274196 | 5.800573 |
| 5 | SWE | 9799186 | 9.009325 |
| 6 | AUT | 8642699 | 5.559051 |
| 7 | ISR | 8380100 | 3.742544 |
| 8 | DNK | 5683483 | 9.861642 |
| 9 | LTU | 2904910 | 8.119926 |
| 10 | EST | 1315407 | 11.766234 |

RMSE of the Best Model for 20 Most Populous Countries



| | Country Code | SP.POP.TOTL | rmse(models, data = cur_data()) |
|---|---|---|---|
| 1 | CHN | 1371220000 | NaN |
| 2 | IND | 1310152403 | NaN |
| 3 | USA | 320742673 | 80.418256 |
| 4 | IDN | 258383256 | NaN |
| 5 | BRA | 204471769 | NaN |
| 6 | PAK | 199426964 | NaN |
| 7 | NGA | 181137448 | NaN |
| 8 | BGD | 156256276 | NaN |
| 9 | RUS | 144096870 | NaN |
| 10 | JPN | 127141000 | NaN |
| 11 | MEX | 121858258 | NaN |
| 12 | PHL | 102113212 | NaN |
| 13 | ETH | 100835458 | NaN |
| 14 | VNM | 92677076 | NaN |
| 15 | EGY | 92442547 | NaN |
| 16 | DEU | 81686611 | NaN |
| 17 | TUR | 78529409 | NaN |
| 18 | IRN | 78492215 | NaN |
| 19 | COD | 76244544 | NaN |
| 20 | THA | 68714511 | NaN |

R2 and RMSE of the Different Models

## 4. Conclusion

The best model created in our project was model S and the worst model was by far model 3. Model S had an RMSE value of 28.58677. This was the lowest of all the models meaning the predictions from model S were much closer to their actual values. Model 3 had an RMSE value of 137.1568. This was by far the highest of all the models tested which means the predictions from the models were nowhere near the actual values. When taking a closer look at both models, multiple observations were made. First, model S has two specific variables, `new_vaccinations` and `extreme_poverty` which have proven to be significant factors when determining accuracy. These variables are significant because when removed, the accuracy of the model tanks drastically. To that note, model 3 included variables such as `gdp_per_capita` which were proven to be quite insignificant in the grand scheme of things. Furthermore, model 3 included `population_density` and `urban_population`. While both these variables seem relatively important in relation to COVID data, the two variables have a decent bit of overlap as they are similar in nature. This means that these two variables are not as significant as variables like `new_vaccinations` or `extreme_poverty`. With these variables being less significant, removing them from the model training did not cause a noticeable drop in accuracy. In conclusion, the evaluation of the models proved that test related data such as new_vaccinations and `weekly_icu_admissions` are much more significant than general economic data such as `gdp_per_capita`.