Alex Jun (885103481)

CPSC 375

Due May 5, 2024

Final Report

1.Data Wrangling Steps

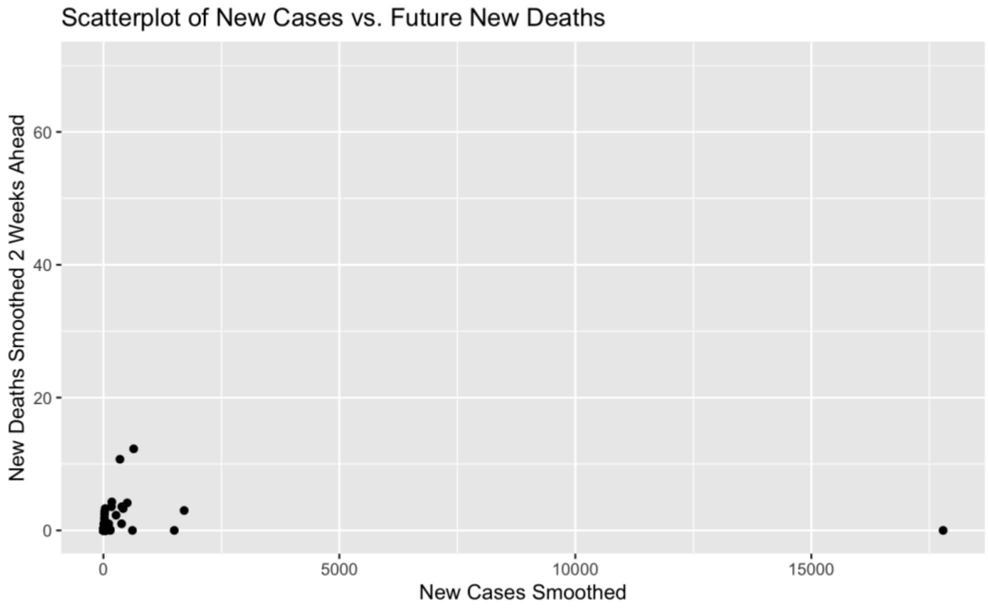
- Imported covid data which is “owid-covid-data.csv” two excel data from databank which are 2022 and 2023 data.
- Had two data with joining the table with 2022 to training data and 2023 to testing data. (Approved by Professor)
- Filtering the COVID dataset for valid ISO codes
- Selecting and removing certain columns from population data.
- Pivoting wider for both population training and validation datasets to organize data by 'Series Code'.
- Coercing population counts to numeric and handling NAs.
- Filtering for populations greater than 1 million.

2. Variables chose from webpages

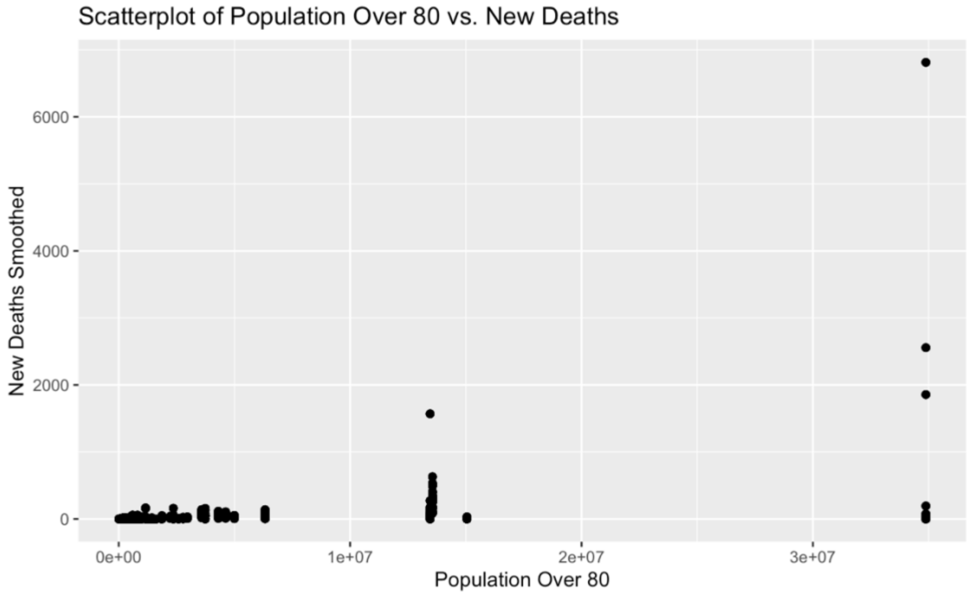
I chose the following 6 variables, because there is so many NA values for other variables so that I would not really use those data.

- Urban population
- Population, total
- Population, male
- Population, female
- Population ages 80 and above, female
- Population ages 80 and above, male

3. Scatter Plot (News cases vs Future New Deaths)



4. Scatter Plot (Population Over 80 vs New Deaths)



5. Variable Transformations

- Cardiovasc deaths = (cardiovasc_death_rate * population)
- Population percentage elder than 80s = ((SP.POP.80UP.FE + SP.POP.80UP.MA) / SP.POP.TOTL) * 100)
- Urban Population Percentage = (SP.URB.TOTL / SP.POP.TOTL) * 100)

6. Different 4 Models and Why I chose

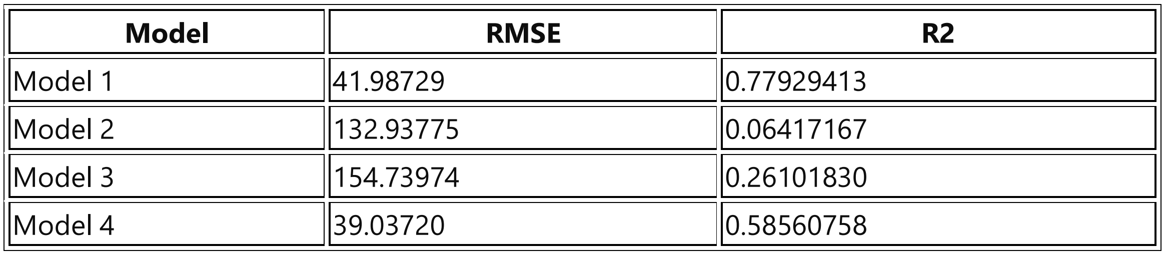
- **model_1** : new_cases_smoothed ,total_cases ,icu_patients , total_vaccinations ,people_fully_vaccinated , gdp_per_capita , urban_population_percentage, life_expectancy, elderly_population_percentage
  Reason: This is Comprehensive model, I was trying to put all the predictors which logically think that it is relevant to covid and these are the most make sense variables as a general.

- model_2 : gdp_per_capita , extreme_poverty , population_density , urban_population_percentage ,human_development_index
  Reason: This is the model with social and economic variables that I chose. All the variables are relating with economics.

- model_3: total_vaccinations, people_vaccinated, people_fully_vaccinated, total_boosters, new_vaccinations_smoothed
  Reason: I thought that putting the predictors regarding the vaccination is make sense. These are the all variables regarding to vaccination and boosters.

- model_4 <- lm(new_deaths_smoothed_2wk ~ population + hospital_beds_per_thousand + icu_patients + hosp_patients + handwashing_facilities, data = final_train)
  Reason: This is about the hosipital infrastructure. Preventing and fighting back to covid might be based on the infrastructure.

7. RMSE of the Best Model for 20 Most Populous Countries



Model	RMSE	R2
Model 1	41.98729	0.77929413
Model 2	132.93775	0.06417167
Model 3	154.73974	0.26101830
Model 4	39.03720	0.58560758

Based on the table, Model 1 has the highest the R squared values and the relatively the lowest RMSE, little bit larger than RMSE. Therefore, we choose the Model 1 as the best Model.

8. Top 20 Countries with RMSE

iso_code	location	population	RMSE
CHN	**China**	**1425887360**	**NAN**
IND	**India**	**1417173120**	**NAN**
USA	**United States**	**338289856**	**81.77700**
IDN	**Indonesia**	**275501344**	**NAN**
PAK	**Pakistan**	**235824864**	**NAN**
NG	**Nigeria**	**218541216**	**NAN**
BRA	**Brazil**	**215313504**	**NAN**
BGD	**Bangladesh**	**171186368**	**NAN**
RUS	**Russia**	**144713312**	**NAN**
MEX	**Mexico**	**127504120**	**NAN**
JPN	**Japan**	**123951696**	**200.55290**
ETH	**Ethiopia**	**123379928**	**NAN**
PHL	**Philippines**	**115559008**	**NAN**
EGY	**Egypt**	**110990096**	**NAN**
COD	**Democratic Republic of Congo**	**99010216**	**NAN**
VNM	**Vietnam**	**98186856**	**NAN**
IRN	**Iran**	**88550568**	**NAN**
TUR	**Turkey**	**85341248**	**NAN**
DEU	**Germany**	**83369840**	**25.23073**
THA	**Thailand**	**71697024**	**NAN**

9. Conclusion

The model 1 identifies ICU patient counts, vaccination rates, economic factors (GDP per capita), urbanization, and life expectancy as significant determinants of COVID-19 death rates. Total Cases and Elderly Population Percentage showed no significant impact on new deaths smoothed over 2 weeks in this model. Additionally, the strategy to prevent could be focus on enhancing critical care capacity, accelerating vaccination efforts, and leveraging economic and urban planning to mitigate the pandemic's impact, especially in vulnerable populations.