Prediction of Death Toll during Covid-19 using ML

MACHINE LEARNING COURSE PROJECT WHICH USES DATA TO PREDICT DEATH TOLL AND AFFECTED PEOPLE.

SARFRAZ HUSSAIN; SYED ATEEB AHMED; SM RAZA NAQVI;

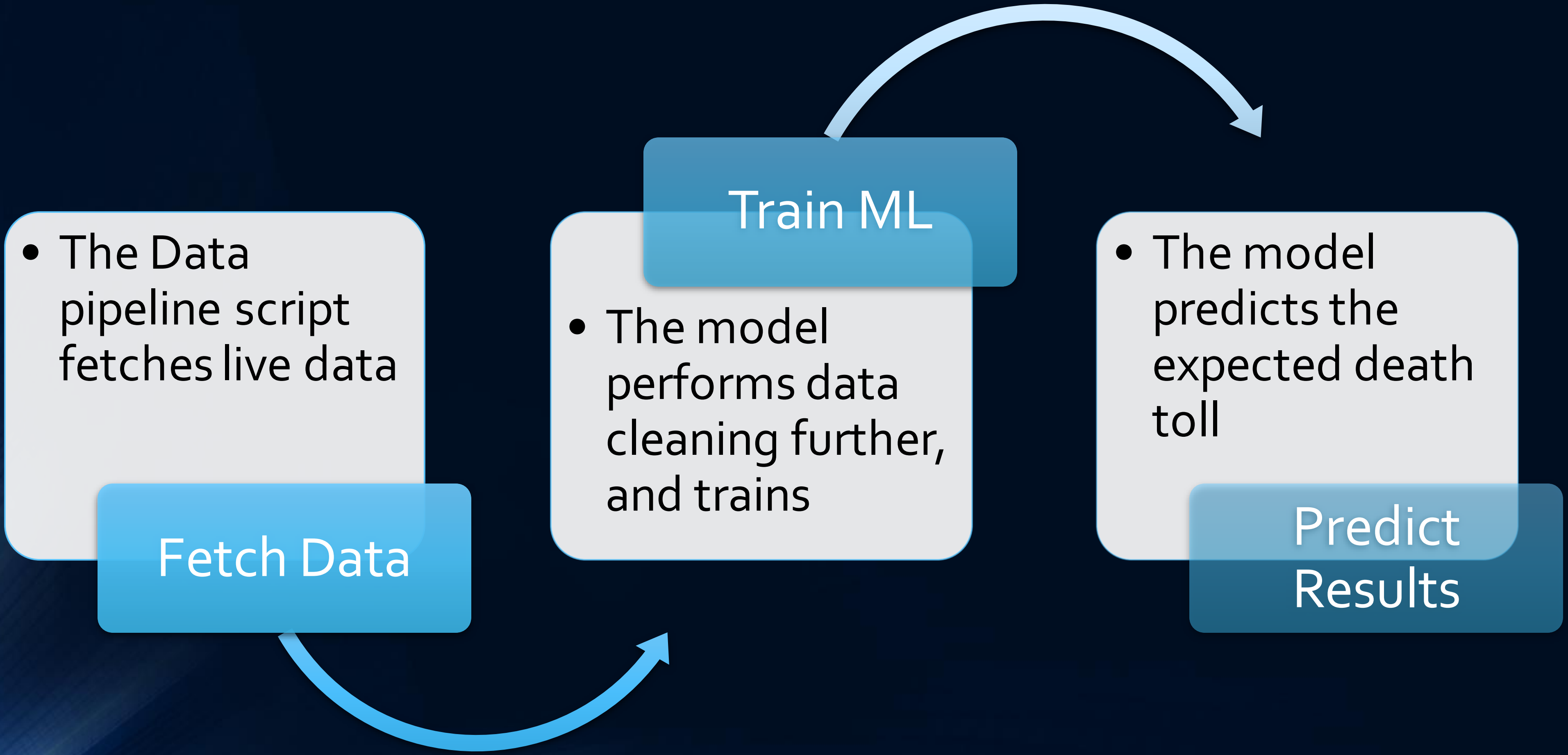
In this Project we have tried and implemented

- A data pipeline which fetches real life data from various sources [1][2][3]
- Neural Network to predict the total death toll under various circumstances, from when Covid-19 arrived in the country until now. [4]
- The project is uploaded and maintained at GITHUB_LINK [5]

Data Pipeline Methodology

- We have extracted information from two sources
  1. *Worldometers [3]*
  2. *Our World in Data [1]*
- To extract data from *Worldometers*, we used the *BeautifulSoup* Python library to scrape html data from Worldometers' website. After cleaning and processing, the scraped data is written on a csv file.
- *Our World in Data (OWID)* provides a downloadable csv file that contains all relevant data.

Complete Layout from Data pipeline to NN



- The Data pipeline script fetches live data

Fetch Data

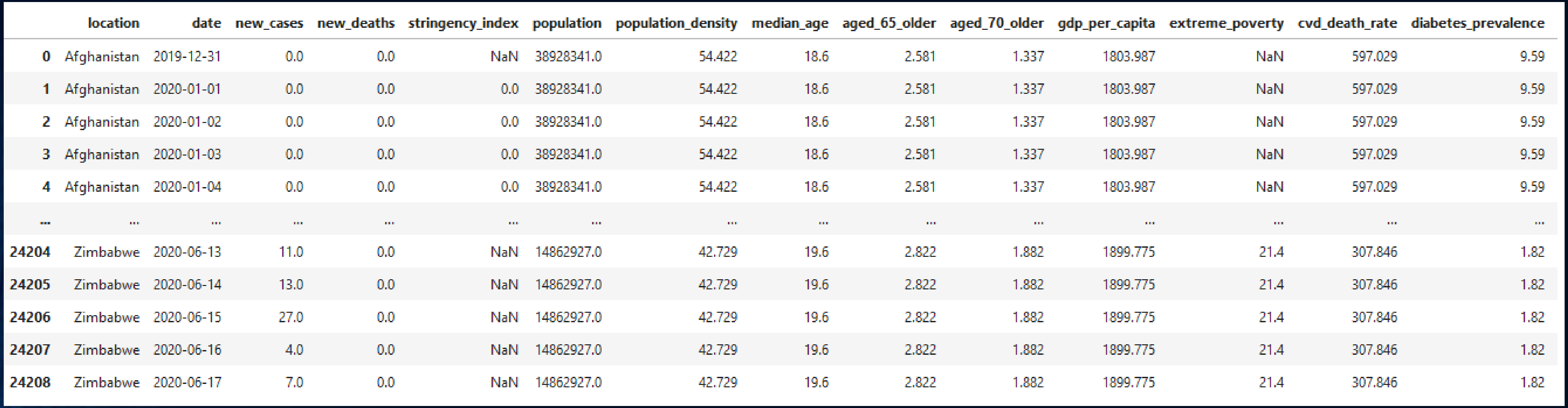
Train ML

- The model performs data cleaning further, and trains

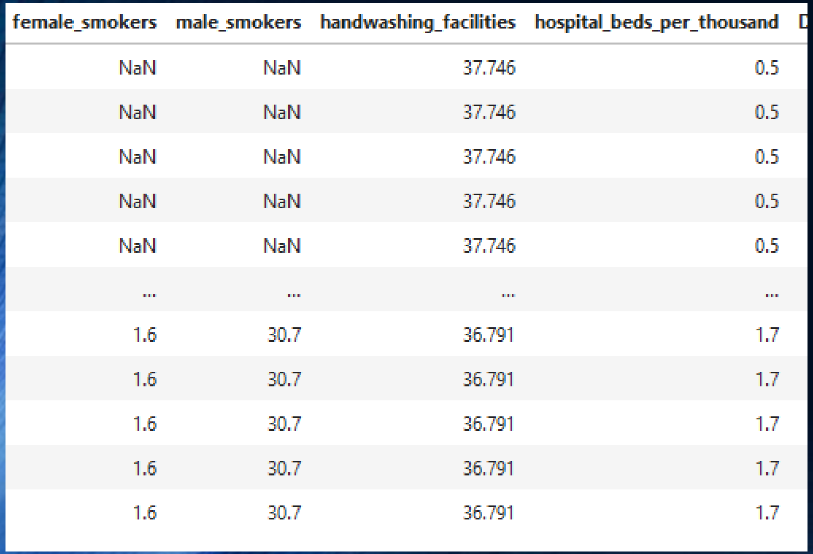
- The model predicts the expected death toll

Predict Results

Pre-Processing – Initial Form

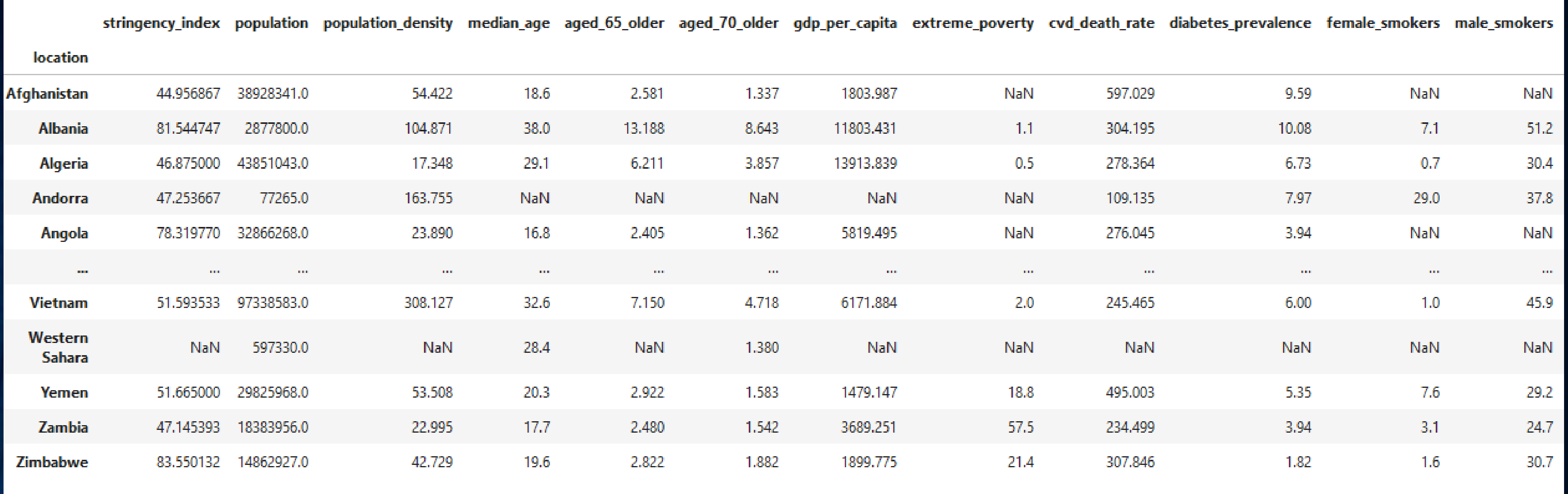


	location	date	new_cases	new_deaths	stringency_index	population	population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita	extreme_poverty	cvd_death_rate	diabetes_prevalence
0	Afghanistan	2019-12-31	0.0	0.0	NaN	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59
1	Afghanistan	2020-01-01	0.0	0.0	0.0	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59
2	Afghanistan	2020-01-02	0.0	0.0	0.0	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59
3	Afghanistan	2020-01-03	0.0	0.0	0.0	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59
4	Afghanistan	2020-01-04	0.0	0.0	0.0	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24204	Zimbabwe	2020-06-13	11.0	0.0	NaN	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82
24205	Zimbabwe	2020-06-14	13.0	0.0	NaN	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82
24206	Zimbabwe	2020-06-15	27.0	0.0	NaN	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82
24207	Zimbabwe	2020-06-16	4.0	0.0	NaN	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82
24208	Zimbabwe	2020-06-17	7.0	0.0	NaN	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82

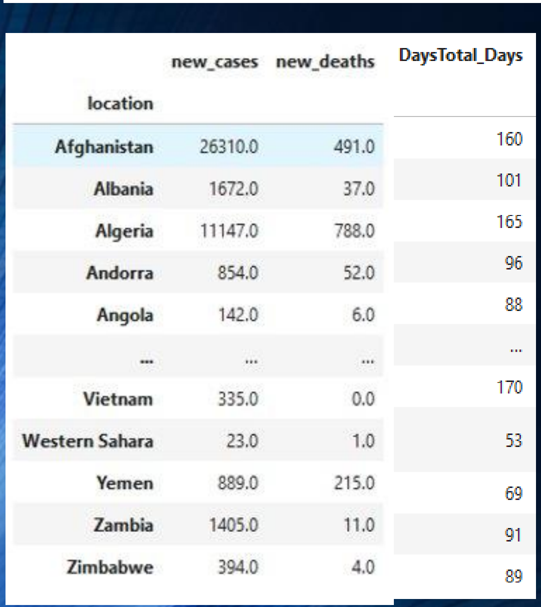


female_smokers	male_smokers	handwashing_facilities	hospital_beds_per_thousand	D
NaN	NaN	37.746	0.5	
NaN	NaN	37.746	0.5	
NaN	NaN	37.746	0.5	
NaN	NaN	37.746	0.5	
NaN	NaN	37.746	0.5	
...	...	...	...	
1.6	30.7	36.791	1.7	
1.6	30.7	36.791	1.7	
1.6	30.7	36.791	1.7	
1.6	30.7	36.791	1.7	
1.6	30.7	36.791	1.7	

Pre-Processing – Components-wise Aggregation

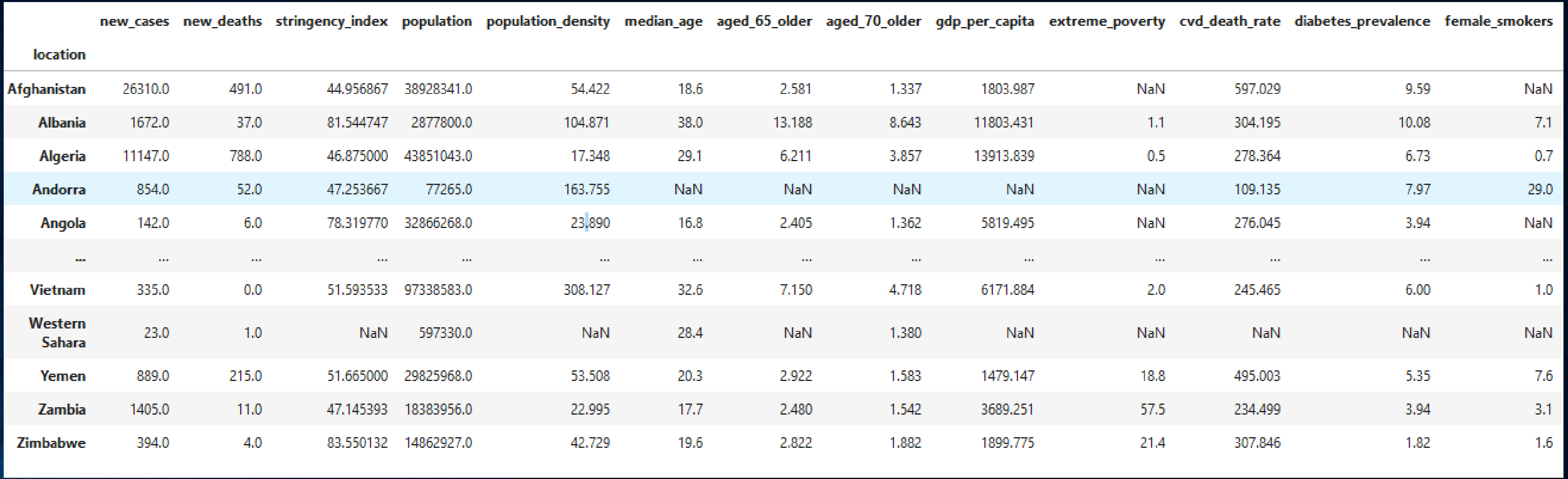


	stringency_index	population	population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita	extreme_poverty	cvd_death_rate	diabetes_prevalence	female_smokers	male_smokers
location												
Afghanistan	44.956867	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59	NaN	NaN
Albania	81.544747	2877800.0	104.871	38.0	13.188	8.643	11803.431	1.1	304.195	10.08	7.1	51.2
Algeria	46.875000	43851043.0	17.348	29.1	6.211	3.857	13913.839	0.5	278.364	6.73	0.7	30.4
Andorra	47.253667	77265.0	163.755	NaN	NaN	NaN	NaN	NaN	109.135	7.97	29.0	37.8
Angola	78.319770	32866268.0	23.890	16.8	2.405	1.362	5819.495	NaN	276.045	3.94	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
Vietnam	51.593533	97338583.0	308.127	32.6	7.150	4.718	6171.884	2.0	245.465	6.00	1.0	45.9
Western Sahara	NaN	597330.0	NaN	28.4	NaN	1.380	NaN	NaN	NaN	NaN	NaN	NaN
Yemen	51.665000	29825968.0	53.508	20.3	2.922	1.583	1479.147	18.8	495.003	5.35	7.6	29.2
Zambia	47.145393	18383956.0	22.995	17.7	2.480	1.542	3689.251	57.5	234.499	3.94	3.1	24.7
Zimbabwe	83.550132	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82	1.6	30.7

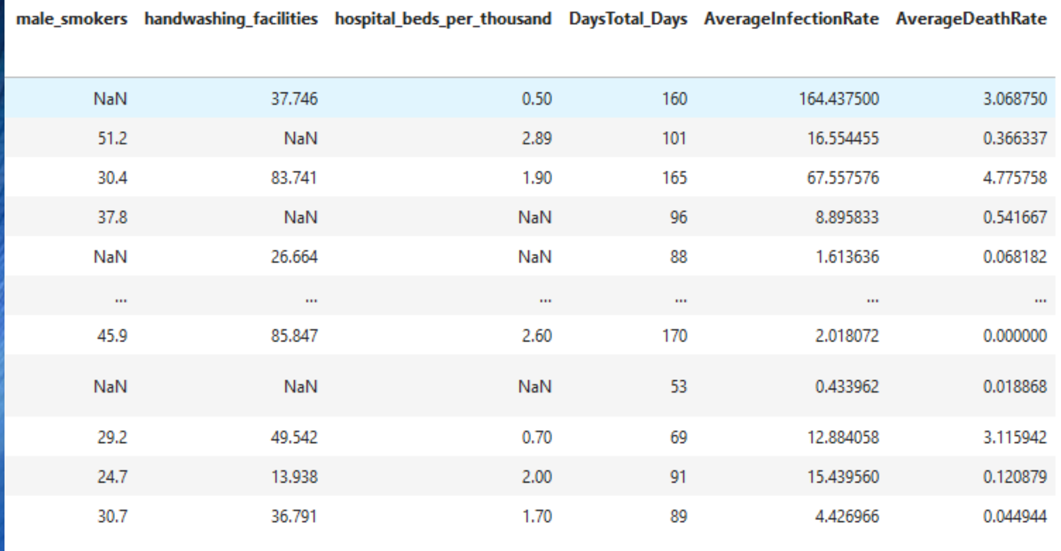


	new_cases	new_deaths	DaysTotal_Days
location			
Afghanistan	26310.0	491.0	160
Albania	1672.0	37.0	101
Algeria	11147.0	788.0	165
Andorra	854.0	52.0	96
Angola	142.0	6.0	88
...	...	...	...
Vietnam	335.0	0.0	170
Western Sahara	23.0	1.0	53
Yemen	889.0	215.0	69
Zambia	1405.0	11.0	91
Zimbabwe	394.0	4.0	89

Pre-Processing – Final Form

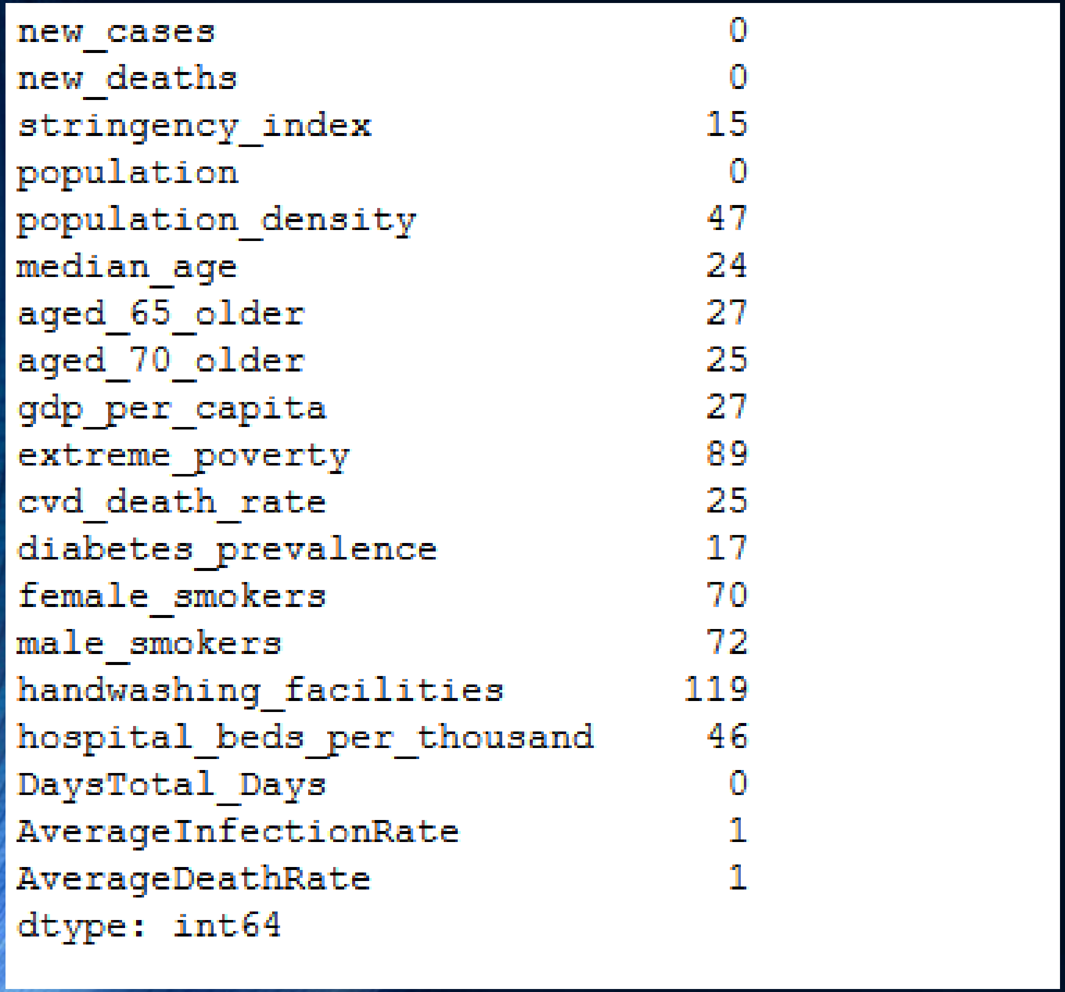


	new_cases	new_deaths	stringency_index	population	population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita	extreme_poverty	cvd_death_rate	diabetes_prevalence	female_smokers
location													
Afghanistan	26310.0	491.0	44.956867	38928341.0	54.422	18.6	2.581	1.337	1803.987	NaN	597.029	9.59	NaN
Albania	1672.0	37.0	81.544747	2877800.0	104.871	38.0	13.188	8.643	11803.431	1.1	304.195	10.08	7.1
Algeria	11147.0	788.0	46.875000	43851043.0	17.348	29.1	6.211	3.857	13913.839	0.5	278.364	6.73	0.7
Andorra	854.0	52.0	47.253667	77265.0	163.755	NaN	NaN	NaN	NaN	NaN	109.135	7.97	29.0
Angola	142.0	6.0	78.319770	32866268.0	23.890	16.8	2.405	1.362	5819.495	NaN	276.045	3.94	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Vietnam	335.0	0.0	51.593533	97338583.0	308.127	32.6	7.150	4.718	6171.884	2.0	245.465	6.00	1.0
Western Sahara	23.0	1.0	NaN	597330.0	NaN	28.4	NaN	1.380	NaN	NaN	NaN	NaN	NaN
Yemen	889.0	215.0	51.665000	29825968.0	53.508	20.3	2.922	1.583	1479.147	18.8	495.003	5.35	7.6
Zambia	1405.0	11.0	47.145393	18383956.0	22.995	17.7	2.480	1.542	3689.251	57.5	234.499	3.94	3.1
Zimbabwe	394.0	4.0	83.550132	14862927.0	42.729	19.6	2.822	1.882	1899.775	21.4	307.846	1.82	1.6



male_smokers	handwashing_facilities	hospital_beds_per_thousand	DaysTotal_Days	AverageInfectionRate	AverageDeathRate
NaN	37.746	0.50	160	164.437500	3.068750
51.2	NaN	2.89	101	16.554455	0.366337
30.4	83.741	1.90	165	67.557576	4.775758
37.8	NaN	NaN	96	8.895833	0.541667
NaN	26.664	NaN	88	1.613636	0.068182
...	...	...	...	...	...
45.9	85.847	2.60	170	2.018072	0.000000
NaN	NaN	NaN	53	0.433962	0.018868
29.2	49.542	0.70	69	12.884058	3.115942
24.7	13.938	2.00	91	15.439560	0.120879
30.7	36.791	1.70	89	4.426966	0.044944

Pre-Processing – Adjusting Missing Values



```
new_cases                        0
new_deaths                       0
stringency_index                15
population                       0
population_density              47
median_age                      24
aged_65_older                   27
aged_70_older                   25
gdp_per_capita                  27
extreme_poverty                 89
cvd_death_rate                  25
diabetes_prevalence             17
female_smokers                  70
male_smokers                    72
handwashing_facilities         119
hospital_beds_per_thousand      46
DaysTotal_Days                   0
AverageInfectionRate             1
AverageDeathRate                 1
dtype: int64
```

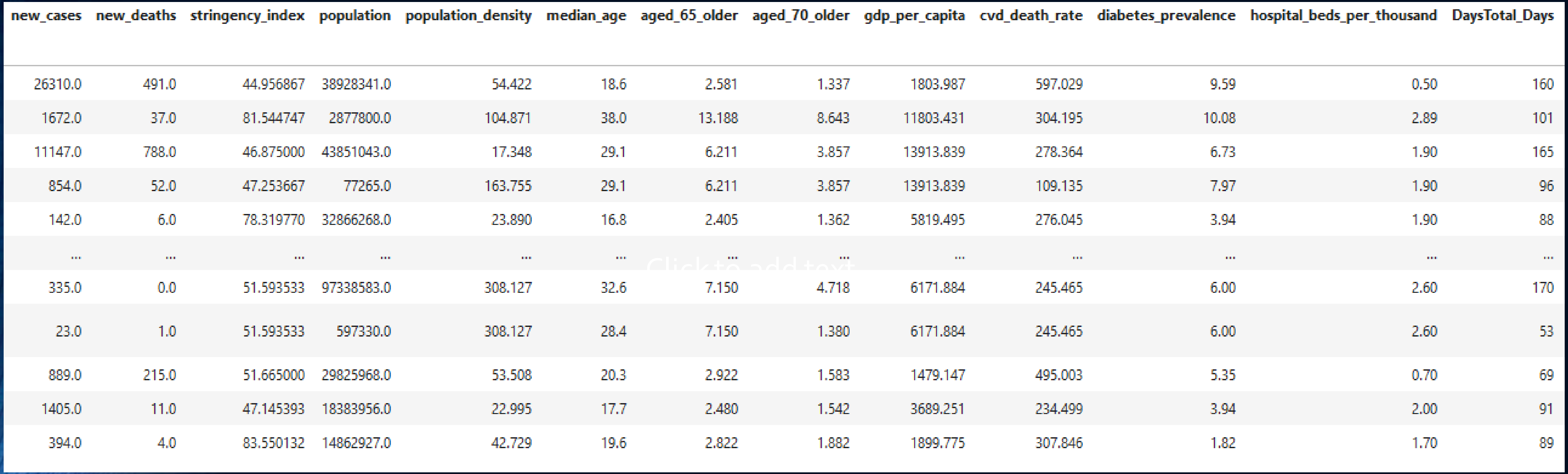
Dropped columns which are mostly empty:

1. handwashing_facilities
2. female_smokers
3. male_smokers
4. extreme_poverty

Rest of the Columns:

- Hot Deck Imputation
  - A randomly chosen value from an individual in the sample who has similar values on other variables.
  - In other words, find all the sample projects who are similar on other variables, then randomly choose one of their values on the missing variables.

Pre-Processing – Feature Scaling



new_cases	new_deaths	stringency_index	population	population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita	cvd_death_rate	diabetes_prevalence	hospital_beds_per_thousand	DaysTotal_Days
26310.0	491.0	44.956867	38928341.0	54.422	18.6	2.581	1.337	1803.987	597.029	9.59	0.50	160
1672.0	37.0	81.544747	2877800.0	104.871	38.0	13.188	8.643	11803.431	304.195	10.08	2.89	101
11147.0	788.0	46.875000	43851043.0	17.348	29.1	6.211	3.857	13913.839	278.364	6.73	1.90	165
854.0	52.0	47.253667	77265.0	163.755	29.1	6.211	3.857	13913.839	109.135	7.97	1.90	96
142.0	6.0	78.319770	32866268.0	23.890	16.8	2.405	1.362	5819.495	276.045	3.94	1.90	88
...	...	...	...	...	...	...	...	...	...	...	...	...
335.0	0.0	51.593533	97338583.0	308.127	32.6	7.150	4.718	6171.884	245.465	6.00	2.60	170
23.0	1.0	51.593533	597330.0	308.127	28.4	7.150	1.380	6171.884	245.465	6.00	2.60	53
889.0	215.0	51.665000	29825968.0	53.508	20.3	2.922	1.583	1479.147	495.003	5.35	0.70	69
1405.0	11.0	47.145393	18383956.0	22.995	17.7	2.480	1.542	3689.251	234.499	3.94	2.00	91
394.0	4.0	83.550132	14862927.0	42.729	19.6	2.822	1.882	1899.775	307.846	1.82	1.70	89

- To Avoid Exploding Gradients

Neural Networks – Regression Problem

- Neural Network V1
  - Input Layer
  - Hidden Layer – H1
    - 4 Neurons – RELU Activation
  - Output Layer – O1
    - 1 Neurons – Linear

- Neural Network V2
  - Input Layer
  - Hidden Layer – H1
    - 4 Neurons – RELU Activation
  - Hidden Layer – H1
    - 4 Neurons – RELU Activation
  - Output Layer – O1
    - 1 Neurons – Linear

- Neural Network V2
  - Input Layer
  - Hidden Layer – H1
    - 4 Neurons – RELU Activation
  - Hidden Layer – H1
    - 4 Neurons – RELU Activation
  - Hidden Layer – H1
    - 4 Neurons – RELU Activation
  - Output Layer – O1
    - 1 Neurons – Linear

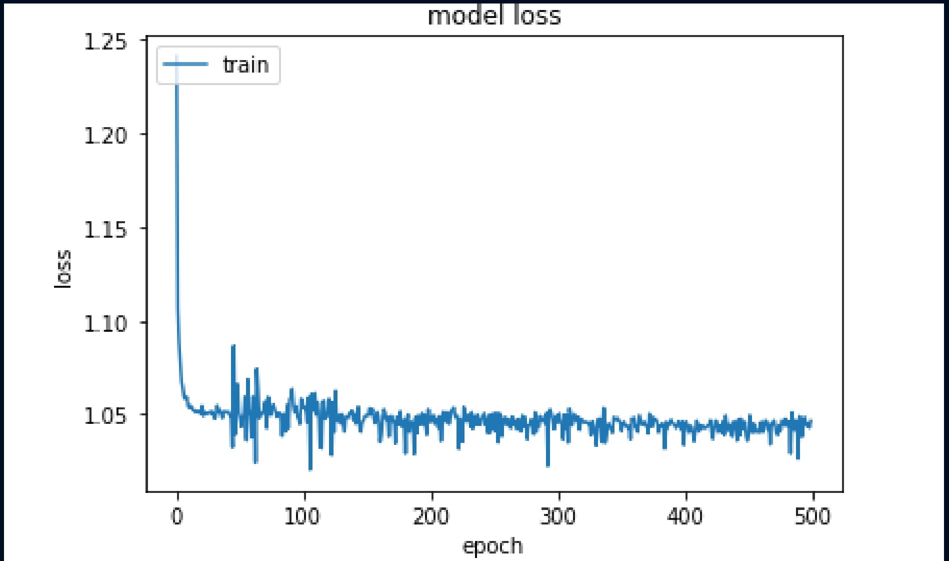
Neural Networks – Regression Problem

MSE	Hidden Layers	Neurons – H Layer
0.87005	1	4
0.95637	2	4|4
0.96911	3	4|4|4

MSE	Hidden Layers	Neurons – H Layer
0.87005	1	8
0.84937	2	4|8
1.83121	2	8|8
1.07042	3	4|8|4

Neural Networks – Regression Problem

MSE	Hidden Layers	Neurons – H Layer	Epochs
0.94759	1	4	200
0.89544	2	4	500
0.96911	3	4	1000



MSE	Hidden Layers	Neurons – H Layer	Learning Rate
0.98835	2	4|4	0.1
0.98943	2	4|4	0.3
0.99134	2	4|4	0.5
1.14691	2	4|4	0.7

Results and future work

- We would work on the model to generate more accurate prediction (HOW?)
  - Improve upon the Data
    - Many of the features in the data are not necessarily 2019 stats because WHO adds the most recent data, which could be from 2015, 2012 etc.
    - We would develop a front-end application.
- We would generate a Flask API for others to use

References:

1. https://ourworldindata.org/coronavirus-source-data
2. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
3. https://www.worldometers.info/coronavirus
4. https://scikit-learn.org/

Thank you!

Open to further questions