2D Task 1 Report

SC09 Group 8

Zhang Chunjie - 1005604 Seetoh Jian Qing -1005600 Mohammed Fauzaan - 1005404 Tan Kay Wee - 1005468

Data

Dataset Choice (refer to Annex 1.1 for code)

Our chosen dataset consists of Covid-19 related information from all the countries around the world from the beginning of the pandemic, 25 February 2020 to 8th November 2021 (Because we extracted this dataset on the 8th of November 2021). This dataset was chosen because of the large amount of data points, containing 135,577 rows and 67 columns of data. Furthermore, it contained the independent and dependent variables that we were interested in studying, and the data for our chosen variables was relatively complete.

The dataset contained enough variables for us to explore. 65 columns of candidate variables meant that we could try out numerous variations of relationships between data sets, and in doing so we were able to find relations that we wanted to focus on.

In addition, the data set also contained both categorical data (such as number of female smokers, male smokers, or the number of persons aged above 70) and continuous data (such as new cases, deaths, etc). Hence, we were able to play with different types of models.

Data Preparation (refer to Annex 1.2 for code)

We did the following to clean up and improve the large amount of data.

1. **Cleaning up "NaN" values (unavailable data):** The original data consisted of large slices of 'NaN' values indicating that the data was unavailable. This could be due to multiple factors-the dataset we used showed pandemic data before the creation of the vaccine. There were also countries like Bangladesh where there were no effective data collection method in place(hard to collect accurate data for entire countries due to urban vs rural areas). Therefore, we removed the rows where there were NaN values in any of the columns.

2. **Choosing a period for our dataset that has a higher accuracy:** In order to make the data more accurate we made sure to only consider the data from the 180 days preceding 8th November 2021. We choose this period as it was when the vaccine was well circulated to most of the countries in Asia. Prior to that period, there was low access to vaccines for all the countries around the world. By the last 180 days, most countries have already chosen a strategy to reduce the number of deaths. The selected period allowed us to calculate and establish a concrete relationship

between the variables under consideration. (Most of which involve vaccination as a major component).

3. **Sufficient datapoints for our model:** It is a general rule of thumb that more data results in greater accuracy of the resultant model. Therefore, we decided to take in all the countries in Asia. We could have gone on to include all the countries from more than one continent, or perhaps even the whole world. We decided to stick to only one continent due to homogeneity in terms of cultural landscape, and laws imposed. In addition, Asian countries constitute more than half the world's population.

The dataset we used, after cleaning up the data, consists of 4576 rows of data and 9 columns.

Choice of Variables

Our chosen metric to evaluate the accuracy of our models is the adjusted R-squared value.

Adjusted R-squared value is chosen since we are trying to build a multiple linear regression model to predict the number of deaths. Hence, metrics that measure how linearly related our values are would be the most suitable to determine the accuracy of our model. Furthermore, since we are dealing with multiple predictors, it is more appropraite to use adjusted R-squared value rather than R-squared value.

First Attempt

x variables: New Vaccinations/Individual, Median Age **y variable:** New Deaths **Adjusted R^2: 0.03822**

Our x variables were Median Age and New Vaccinations/Individual. Our hypothesis was that there would be a positive correlation between countries with a higher median age and the number of new deaths at any specific period of time. However, our adjusted R-Squared value was very poor. This means that there is little correlation between the chosen x variables and the number of new deaths in a country.

While discussing as a group, we realised that while age does make a difference in the fatality of the virus, median age was a mediocre variable. Countries with a lower median age tend to be less developed. This affects their access to vaccinations. Hence, while the population of less developed countries could be relatively young, the number of deaths could still be high.

Second Attempt

x variables: New Vaccinations, New Cases **y variable:** New Deaths **Adjusted R^2: 0.73923362**

We then considered other variables that might have a stronger correlation to the number of new deaths. The new x variables gave us a much higher adjusted R^2 value. This shows

us that the number of new cases affected the number of new deaths, as with an increased number of new vaccinations.

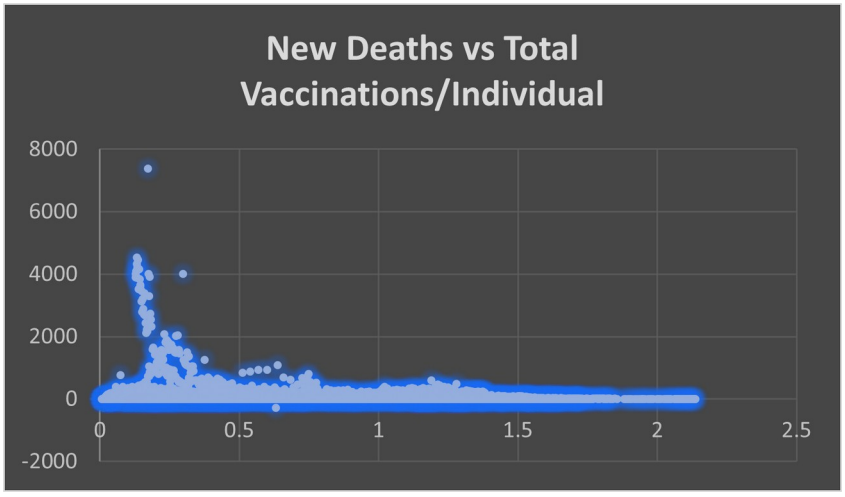
Third Attempt

x variables: Total Vaccinations/Individual, New Cases **y variable:** New Deaths **New Adjusted R^2: 0.742771149**

We improved on our model by changing New Vaccinations/Individual to Total Vaccinations/Individual. The number of new vaccinations are likely to go down in the long run once a country has a large population of their citizens fully vaccinated. Total Vaccinations/Individual is a variable that would increase and stay constant in cases of countries with a high access to vaccines. This should allow our output of new deaths to be more stable. In short, as long as Total Vaccinations/Individual > 1, that should represent a population that has at least 1 jab per person.

Relationship between Total Vaccinations/Individual and New Deaths

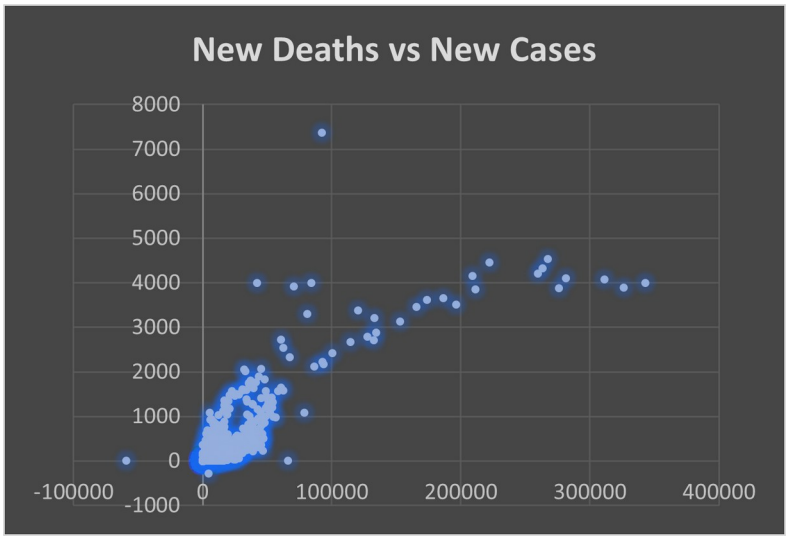
We can see that New Deaths were extremely high when the Total Vaccinations/Individual was under 0.5. As more people got vaccinated, the curve started to drop. When Total Vaccinations/Individual increased to above 1, the curve started to flatten. This showed the efficacy of the vaccine once everyone in the country has had at least 1 dosage of the vaccine.



Relationship between New Cases and New Deaths

We see a relationship that resembles a logarithimic curve. As New Cases increase, New Deaths increase. However, at a certain point, the number of New Deaths started to plateau even while New Cases started to increase.

We think this is because of the effect of vaccination taking place. There will be more and more asymptomatic Covid cases that does not suffer any serious symptoms that would otherwise lead to death.



Conclusion

Our third model is the best at predicting Covid-19 deaths in various countries as it has the highest adjusted R-squared value among our three models, which corresponds to having the strongest linear correlation. From building the model, we obtained the following coefficient values and hence, final equation:

Let y be the number of new daily deaths per day, x_1 be the number of total vaccinations per individual, and x_2 be the number of new daily cases.

$$y = -42.585\,x_1 + 0.01752\,x_2 + 31.614$$

Annex

1.1: Pulling Data(Selecting Columns)

```
import pandas as pd
df=pd.read_csv('https://covid.ourworldindata.org/data/owid-covid-data.csv',dtype='unicode')
#these are the columns under consideration :
columns_from_main=['location','date','total_cases','new_cases','new_deaths','new_vaccinations','median_age','population','total_vaccinations']
print(df)
```

```
       iso_code continent     location        date total_cases
new_cases  \
0           AFG      Asia  Afghanistan  2020-02-24         5.0
5.0
1           AFG      Asia  Afghanistan  2020-02-25         5.0
0.0
```

```
2           AFG      Asia  Afghanistan  2020-02-26         5.0
0.0
3           AFG      Asia  Afghanistan  2020-02-27         5.0
0.0
4           AFG      Asia  Afghanistan  2020-02-28         5.0
0.0
...         ...       ...          ...         ...         ...       .
..
135576      ZWE    Africa     Zimbabwe  2021-11-20    133615.0
22.0
135577      ZWE    Africa     Zimbabwe  2021-11-21    133647.0
32.0
135578      ZWE    Africa     Zimbabwe  2021-11-22    133674.0
27.0
135579      ZWE    Africa     Zimbabwe  2021-11-23    133674.0
0.0
135580      ZWE    Africa     Zimbabwe  2021-11-24    133747.0
73.0

       new_cases_smoothed total_deaths new_deaths new_deaths_smoothed
...  \
0                     NaN          NaN        NaN                  NaN
...
1                     NaN          NaN        NaN                  NaN
...
2                     NaN          NaN        NaN                  NaN
...
3                     NaN          NaN        NaN                  NaN
...
4                     NaN          NaN        NaN                  NaN
...
...                   ...          ...        ...                  ...
...
135576             31.714       4699.0        0.0                0.429
...
135577             31.286       4699.0        0.0                0.429
...
135578             33.714       4699.0        0.0                0.286
...
135579             24.143       4699.0        0.0                0.143
...
135580             27.143       4703.0        4.0                0.571
...

       female_smokers male_smokers handwashing_facilities  \
0                 NaN          NaN                  37.746
1                 NaN          NaN                  37.746
2                 NaN          NaN                  37.746
3                 NaN          NaN                  37.746
4                 NaN          NaN                  37.746
```

```
...                 ...          ...                    ...
135576              1.6         30.7                 36.791
135577              1.6         30.7                 36.791
135578              1.6         30.7                 36.791
135579              1.6         30.7                 36.791
135580              1.6         30.7                 36.791

       hospital_beds_per_thousand life_expectancy
human_development_index  \
0                             0.5           64.83
0.511
1                             0.5           64.83
0.511
2                             0.5           64.83
0.511
3                             0.5           64.83
0.511
4                             0.5           64.83
0.511
...                           ...             ...
...
135576                        1.7           61.49
0.571
135577                        1.7           61.49
0.571
135578                        1.7           61.49
0.571
135579                        1.7           61.49
0.571
135580                        1.7           61.49
0.571

       excess_mortality_cumulative_absolute
excess_mortality_cumulative  \
0                                       NaN
NaN
1                                       NaN
NaN
2                                       NaN
NaN
3                                       NaN
NaN
4                                       NaN
NaN
...                                     ...                        ..
.
135576                                  NaN
NaN
135577                                  NaN
NaN
```

```
135578                                    NaN
NaN
135579                                    NaN
NaN
135580                                    NaN
NaN

       excess_mortality excess_mortality_cumulative_per_million
0                   NaN                                    NaN
1                   NaN                                    NaN
2                   NaN                                    NaN
3                   NaN                                    NaN
4                   NaN                                    NaN
...                 ...                                    ...
135576              NaN                                    NaN
135577              NaN                                    NaN
135578              NaN                                    NaN
135579              NaN                                    NaN
135580              NaN                                    NaN

[135581 rows x 67 columns]
```

1.2: Preparing Data (Selecting date, Changing to Integer, Removing Null)

```
#this is to extract all the that fall within the date range of 13th
May 2021 to 8th November 2021, i.e., 180 days
df['date']=pd.to_datetime(df['date'])
mask=(df['date']>='5/13/2021')&(df['date']<='11/8/2021')
df_location=df.loc[(mask)&(df['continent']=='Asia'),columns_from_main]

#convert the string value to a number:
df_location["new_deaths"]=pd.to_numeric(df_location["new_deaths"],
downcast='integer')
df_location["new_cases"]=pd.to_numeric(df_location["new_cases"],
downcast='integer')
df_location['new_vaccinations']=pd.to_numeric(df_location['new_vaccina
tions'], downcast='integer')
df_location['population']=pd.to_numeric(df_location['population'],
downcast='integer')
df_location['median_age']=pd.to_numeric(df_location['median_age'],
downcast='integer')
df_location['total_vaccinations']=pd.to_numeric(df_location['total_vac
cinations'], downcast='integer')

#This block removes the NaN slices from the excel sheet
df_location['total_vaccinations/population']=df_location['total_vaccin
ations']/df_location['population']
df_location.dropna(subset=['new_vaccinations'],inplace=True)
df_location.dropna(subset=['total_cases'],inplace=True)
df_location.dropna(subset=['new_deaths'],inplace=True)
df_location.dropna(subset=['new_cases'],inplace=True)
```

```
print(len(df_location))#number of columns
print(df_location.size)#number of columns * number of rows (8929*9)
print("The data under observation are:\n",df_location)

df_location.to_csv('new data.csv',index=False)#this line will save the
new data into your current folder. This was the final excel sheet that
we worked with.
#Further visualizations are done on this file
```

```
4576
45760
The data under observation are:
            location       date total_cases  new_cases  new_deaths  \
458     Afghanistan 2021-05-27     68366.0      623.0        14.0
465     Afghanistan 2021-06-03     75119.0     1093.0        27.0
8715     Azerbaijan 2021-05-13    328668.0      509.0        16.0
8716     Azerbaijan 2021-05-14    328994.0      326.0        12.0
8717     Azerbaijan 2021-05-15    329371.0      377.0        14.0
...             ...        ...         ...        ...         ...
132808      Vietnam 2021-10-31    921122.0     5519.0        53.0
132809      Vietnam 2021-11-01    926720.0     5598.0        48.0
132810      Vietnam 2021-11-02    932357.0     5637.0        74.0
132811      Vietnam 2021-11-03    939463.0     7106.0        78.0
132812      Vietnam 2021-11-04    946043.0     6580.0        59.0

        new_vaccinations  median_age  population
total_vaccinations  \
458               2859.0        18.6  39835428.0           593313.0

465               4015.0        18.6  39835428.0           630305.0

8715             21015.0        32.4  10223344.0          1748525.0

8716             13107.0        32.4  10223344.0          1761632.0

8717             15794.0        32.4  10223344.0          1777426.0

...                  ...         ...         ...                ...

132808          554499.0        32.6  98168829.0         81929875.0

132809         1201589.0        32.6  98168829.0         83131464.0

132810          959435.0        32.6  98168829.0         84090899.0

132811          793175.0        32.6  98168829.0         84884074.0
```

```
132812         1435734.0        32.6  98168829.0          86319808.0


        total_vaccinations/population
458                          0.014894
465                          0.015823
8715                         0.171033
8716                         0.172315
8717                         0.173860
...                               ...
132808                       0.834581
132809                       0.846821
132810                       0.856595
132811                       0.864674
132812                       0.879300

[4576 rows x 10 columns]
```