FAKULTI TEKNOLOGI MAKLUMAT DAN KOMUNIKASI

SEMESTER 2 2019/2020

BITI 2513

INTRODUCTION TO DATA SCIENCE

FINAL REPORT

PREPARED BY:

BIL	**NAME**	**MATRIC NUMBER**
1	ABDUL HAZIQ BIN ABD KHALID	B031810256
2	AHMAD NAUFAL BIN MOHD SALEH	B031810382
3	MEOR AMIRUL ASHRAF BIN JAMALULAIL	B031810468
4	MUHAMMAD NABIL IMRAN BIN SOLEHAN	B031810234

Business Understanding

Objective

- To apply knowledge learned in Data Science for real life application.

Goal

- Predicting the death rate of COVID-19 virus of a country based on several factors.

Problem

- The appearance of a new virus called COVID-19 has been plaguing the whole world. Almost all countries in the world are affected by it.

Clear Question

- What factors affect the death rate of COVID-19 for a country?

Measurable Outcome

- Death per million of a country due to COVID-19 based on several factors.

Data Sources

- https://github.com/owid/covid-19-data/tree/master/public/data[2].

Data Management

The dataset our group is using is the Data on COVID-19 (coronavirus) provided by Our World in Data. There are exactly 24 234 data and 33 attributes in this dataset. Below are the details for each attribute :

iso_code	ISO 3166-1 alpha-3 – three-letter country codes
continent	Continent of the geographical location
location	Geographical location
date	Date of observation
total_cases	Total confirmed cases of COVID-19
new_cases	New confirmed cases of COVID-19
total_deaths	Total deaths attributed to COVID-19
new_deaths	New deaths attributed to COVID-19
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people

total_tests	Total tests for COVID-19
new_tests	New tests for COVID-19
new_tests_smoothed	New tests for COVID-19 (7-day smoothed).
total_tests_per_thousand	Total tests for COVID-19 per 1,000 people
new_tests_per_thousand	New tests for COVID-19 per 1,000 people
new_tests_smoothed_per_thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people
tests_units	Units used by the location to report its testing data
stringency_index	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
population	Population in 2020
population_density	Number of people divided by land area, measured in square kilometers, most recent year available
median_age	Median age of the population, UN projection for 2020
aged_65_older	Share of the population that is 65 years and older, most recent year available
aged_70_older	Share of the population that is 70 years and older in 2015
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available

extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
cvd_death_rate	Death rate from cardiovascular disease in 2017
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017
female_smokers	Share of women who smoke, most recent year available
male_smokers	Share of men who smoke, most recent year available
handwashing_facilities	Share of the population with basic hand washing facilities on premises, most recent year available
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010

Table 1: Attributes in this data set.

The dataset is updated daily as the pandemic outbreak is still ongoing. The structure of the dataset is very complicated. In short, each row represents each day and the data are sorted according to their country of origin.

A	B	C	D	E	F	G	H	I	J	K	L
iso_code	continent	location	date	total_cases	new_cases	total_death	new_deaths	total_cases	new_cases_	total_death	new_deaths
ALB	Europe	Albania	9/03/2020	2	2	0	0	0.695	0.695	0	0
ALB	Europe	Albania	10/03/2020	6	4	0	0	2.085	1.39	0	0
ALB	Europe	Albania	11/03/2020	10	4	0	0	3.475	1.39	0	0
ALB	Europe	Albania	12/03/2020	11	1	1	1	3.822	0.347	0.347	0.347
ALB	Europe	Albania	13/03/2020	23	12	1	0	7.992	4.17	0.347	0
ALB	Europe	Albania	14/03/2020	33	10	1	0	11.467	3.475	0.347	0
ALB	Europe	Albania	15/03/2020	38	5	1	0	13.205	1.737	0.347	0
ALB	Europe	Albania	16/03/2020	42	4	1	0	14.594	1.39	0.347	0
ALB	Europe	Albania	17/03/2020	51	9	1	0	17.722	3.127	0.347	0
ALB	Europe	Albania	18/03/2020	55	4	1	0	19.112	1.39	0.347	0
ALB	Europe	Albania	19/03/2020	59	4	2	1	20.502	1.39	0.695	0.347
ALB	Europe	Albania	20/03/2020	70	11	2	0	24.324	3.822	0.695	0
ALB	Europe	Albania	21/03/2020	70	0	2	0	24.324	0	0.695	0
ALB	Europe	Albania	22/03/2020	76	6	2	0	26.409	2.085	0.695	0
ALB	Europe	Albania	23/03/2020	89	13	2	0	30.926	4.517	0.695	0
ALB	Europe	Albania	24/03/2020	100	11	4	2	34.749	3.822	1.39	0.695
ALB	Europe	Albania	25/03/2020	123	23	5	1	42.741	7.992	1.737	0.347
ALB	Europe	Albania	26/03/2020	146	23	5	0	50.733	7.992	1.737	0
ALB	Europe	Albania	27/03/2020	174	28	6	1	60.463	9.73	2.085	0.347
ALB	Europe	Albania	28/03/2020	186	12	9	3	64.633	4.17	3.127	1.042
ALB	Europe	Albania	29/03/2020	197	11	10	1	68.455	3.822	3.475	0.347
ALB	Europe	Albania	30/03/2020	212	15	10	0	73.667	5.212	3.475	0
ALB	Europe	Albania	31/03/2020	223	11	12	2	77.49	3.822	4.17	0.695

Figure 1: Original dataset.

Our group decides to study the overall case for each country instead of each individual day. Therefore, we decided to only choose the latest date (16th of June 2020). It is possible to omit all the data from the previous date because each attribute "total_x" keeps the cumulative data from all the previous date.

A	B	C	D	E	F	G	H	I	J	K	L
iso_code	continent	location	date	total_case	new_case:	total_deat	new_deatl	total_cases_	new_case:	total_death	new_deatl
AFG	Asia	Afghanista	16/06/2020	25527	761	478	7	655.743	19.549	12.279	0.18
ALB	Europe	Albania	16/06/2020	1590	69	36	0	552.505	23.977	12.51	0
DZA	Africa	Algeria	16/06/2020	11031	112	777	10	251.556	2.554	17.719	0.228
AND	Europe	Andorra	16/06/2020	853	0	51	0	11039.928	0	660.066	0
AGO	Africa	Angola	16/06/2020	142	2	6	0	4.321	0.061	0.183	0
AIA	North Ame	Anguilla	16/06/2020	3	0	0	0	199.973	0	0	0
ATG	North Ame	Antigua an	16/06/2020	26	0	3	0	265.501	0	30.635	0
ARG	South Ame	Argentina	16/06/2020	32772	1208	854	21	725.112	26.728	18.896	0.465
ARM	Asia	Armenia	16/06/2020	17064	397	285	16	5758.573	133.975	96.179	5.4
ABW	North Ame	Aruba	16/06/2020	101	0	3	0	945.994	0	28.099	0
AUS	Oceania	Australia	16/06/2020	7335	15	102	0	287.648	0.588	4	0
AUT	Europe	Austria	16/06/2020	17065	27	678	1	1894.764	2.998	75.28	0.111
AZE	Asia	Azerbaijan	16/06/2020	10324	367	122	3	1018.229	36.196	12.033	0.296
BHS	North Ame	Bahamas	16/06/2020	104	1	11	0	264.464	2.543	27.972	0
BHR	Asia	Bahrain	16/06/2020	19013	786	46	3	11173.713	461.923	27.034	1.763
BGD	Asia	Banglades	16/06/2020	90619	3099	1209	38	550.242	18.817	7.341	0.231
BRB	North Ame	Barbados	16/06/2020	97	1	7	0	337.543	3.48	24.359	0
BLR	Europe	Belarus	16/06/2020	54680	707	312	4	5786.659	74.82	33.018	0.423
BEL	Europe	Belgium	16/06/2020	60100	71	9661	6	5185.677	6.126	833.591	0.518
BLZ	North Ame	Belize	16/06/2020	21	0	2	0	52.814	0	5.03	0
BEN	Africa	Benin	16/06/2020	483	13	9	2	39.841	1.072	0.742	0.165

Figure 2: Dataset after removing previous date

Each row now represents each country instead of each day. There is 208 data and 33 attributes now.

Since our group decided to study the overall case, all attributes “new_x” were removed due to they only shows each data for each individual date. Only attributes that show the overall data for each country were kept. The current attributes are now 26.

The removed attributes are as follow :

new_cases	New confirmed cases of COVID-19
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people
new_tests	New tests for COVID-19
new_tests_smoothed	New tests for COVID-19 (7-day smoothed).
new_tests_per_thousand	New tests for COVID-19 per 1,000 people
new_tests_smoothed_per_thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people

The goal of our project is to find which attributes affect the death rate of COVID-19. Therefore, any attributes that is not relevant in finding the death rate of COVID-19 were removed. The current attributes are now 23.

The removed attributes are as follow :

Attributes	**Reasons for removal**
iso_code	These attributes are originally used as the ID attribute (Primary Key). Our group decided to use the attribute “location” which holds the name of the countries as the ID instead.

continent	Continent is very general as each continent represents so many countries. Each country in the same continent has different cultures,infrastructures,etc. Therefore we concluded the continent does not play a huge role in this study.
date	As we no longer study each date individually, this attribute was removed.

For further cleaning of our dataset, we decide to remove any attributes with too much missing values in it. This is due to the fact that lots of missing value in an attribute may lead to biased conclusions. We decided that any attribute that has more than 25% missing values in its data are to be removed.

Hence, any attributes with missing values that are higher than 52 will be removed.

Attributes	No. Of Missing Values
total_tests	208
total_tests_per_thousand	208
strigency_index	203
handwashing_facilities	117
extreme_poverty	88
male_smokers	71
female_smokers	69

The dataset now only has 14 attributes. All attributes now have reasonable amounts of reliable data. However, we found out that certain countries do not provide very much information. Therefore there are a lot of missing values in their row. We also concluded to remove any country with too much missing data.

The countries that were removed are as follows :

- Andorra
- Anguilla
- Bermuda
- Bonaire Sint Eustatius and Saba
- British Virgin Islands
- Cayman Islands
- Faeroe Islands
- Falkland Islands
- Gibraltar
- Greenland
- Guernsey
- Isle of Man
- Jersey
- Kosovo
- Montserrat
- Northern Mariana Islands
- Sint Maarten (Dutch part)
- Turks and Caicos Islands
- Vatican
- Liechtenstein
- Monaco
- Saint Kitts and Nevis
- San Marino
- Western Sahara

The dataset now has 184 reliable data with 14 reliable attributes. The dataset is now ready to be studied and manipulated.

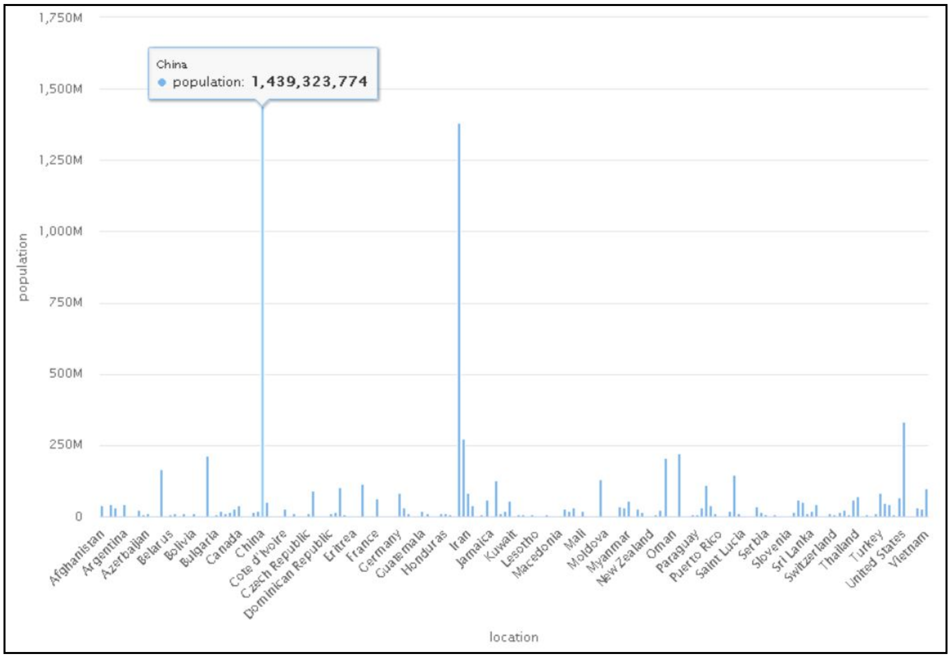


Figure 3: Bar graph of population versus location.

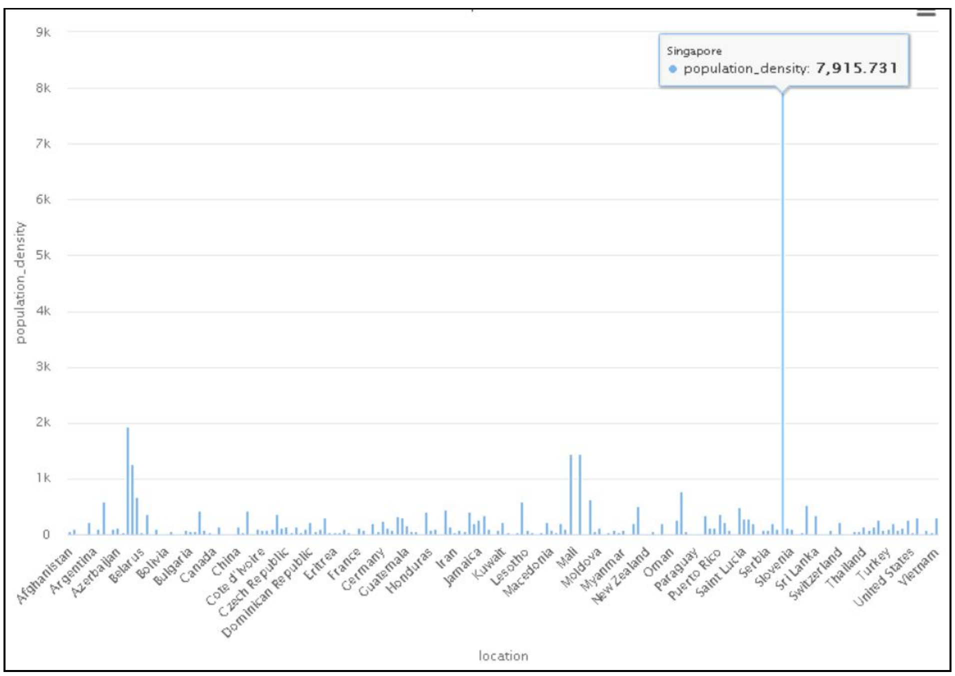


Figure 4: Bar graph of population density versus location.

Figure x shows that the location with the highest population according to the data is China with the population of 1439 million. The attribute population could be affecting the total death rate of the COVID-19 virus. From our assumption, we decide that the higher the population of a country is, the higher chances for the inhabitant of the country to be infected with this virus.

In some other cases the population is not completely affecting the total rate of death per million from the COVID-19 virus when it comes to land masses and other factors. As shown in figure y, Singapore has the highest value of population density which is 7915.731 people per square kilometer. We assume that as the population density goes higher, the chances of getting infected will be higher, thus leading to increase in the total death rate per million of the country.

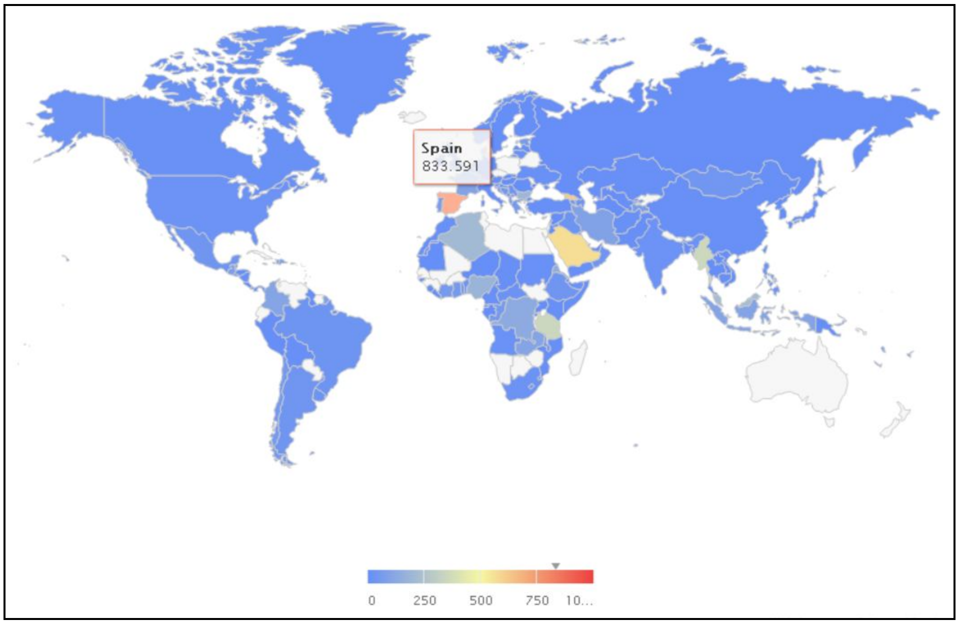


Figure 5: Map chart of total death per million.

The attribute total death rate per million represents the total death rate of a country over 1 million people. Figure A shows that Spain had the highest total death rate per million on 16th June 2020. During this time, Spain has a surge of COVID-19 cases because of the ignorance of the people about the severity of the pandemic happening at that time[1]. Thus causing the sudden spike of new cases of COVID-19.

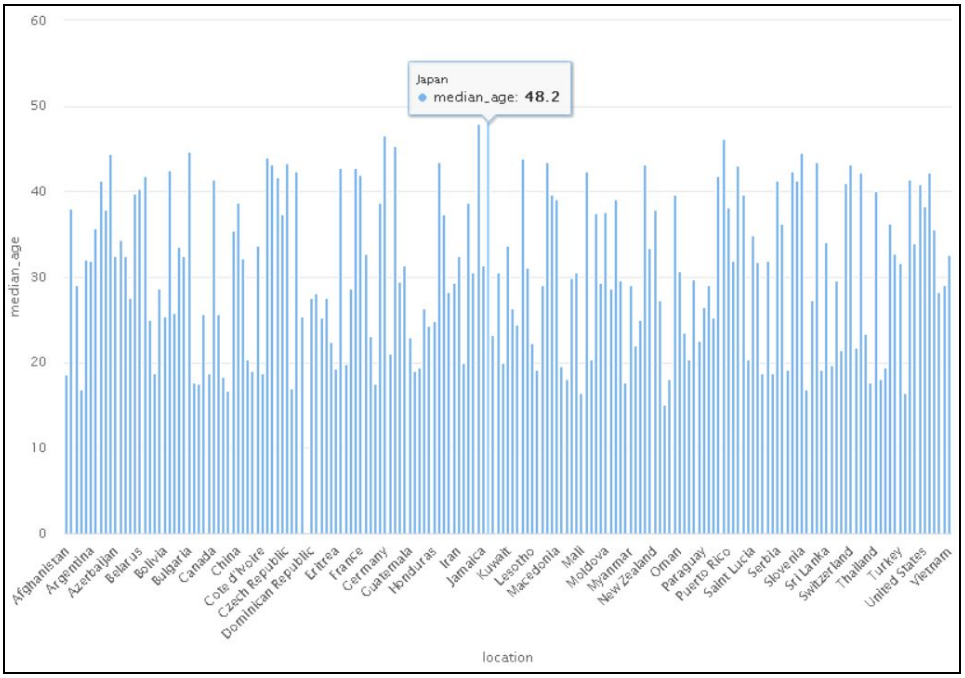


Figure 6: Bar graph of median age versus location.

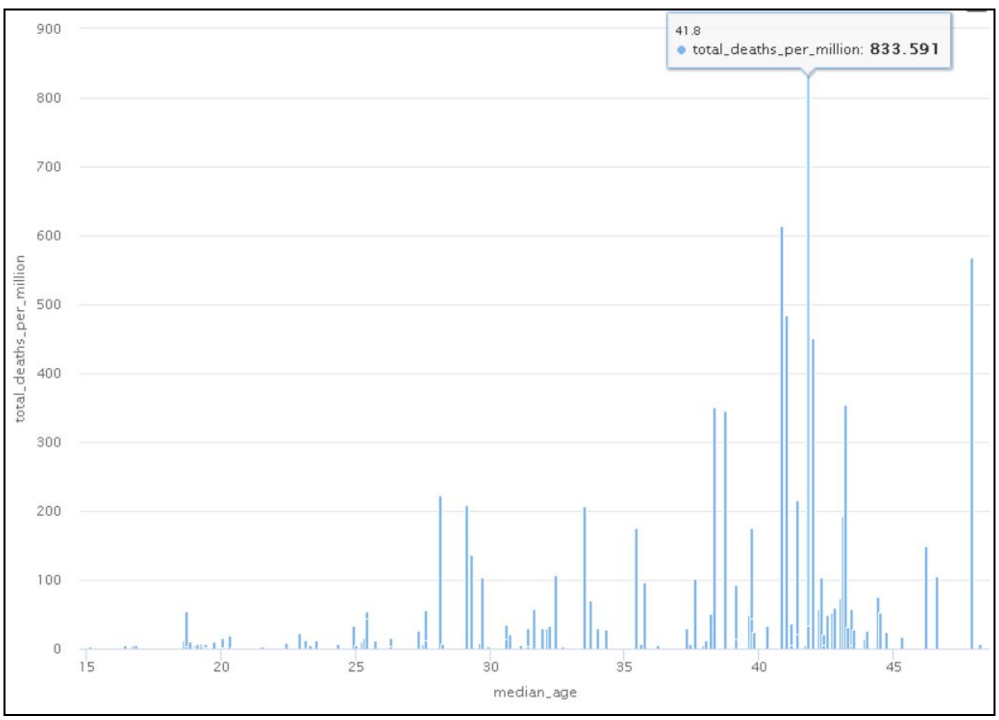


Figure 7: Bar graph of total death per million versus median age.

Figure 6 shows the bar graph of median age of a country as Japan with the highest median age compared to the others. Figure 7 shows that the highest total death rate per million of 833.591 is related to the median age of 41.8 years old. From this statement we can assume that the older people become, the probability of getting infected thus causing the increases of the total death rate per million of a country.

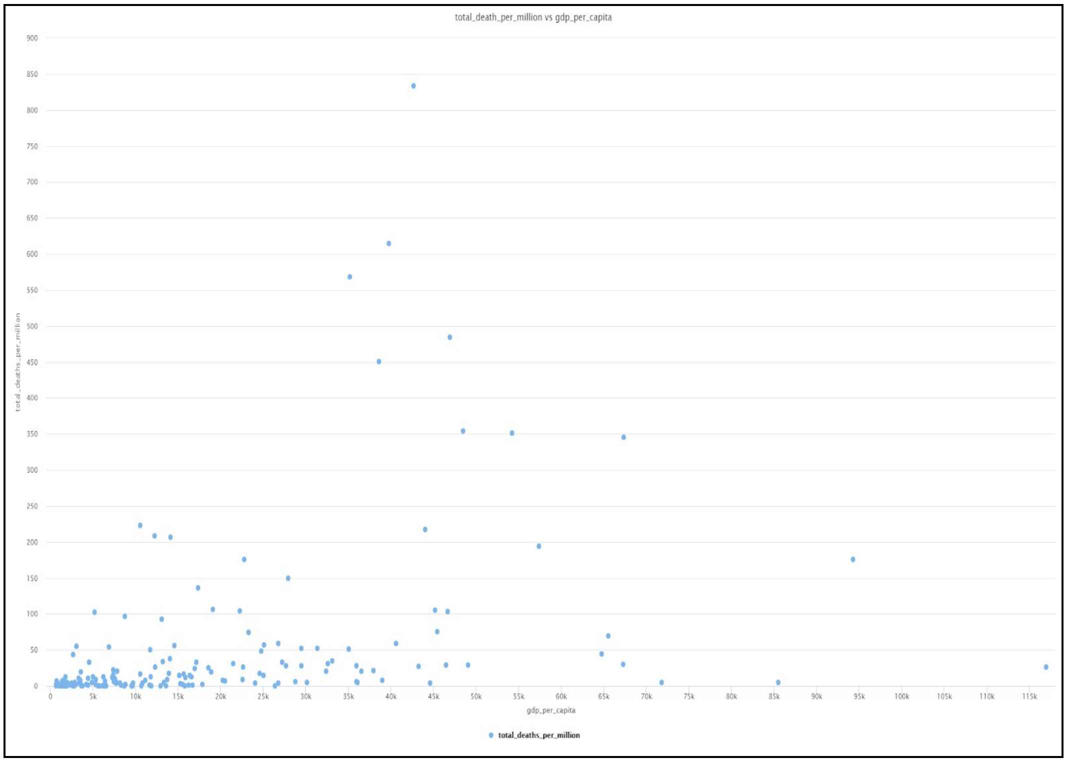


Figure 8: Scattered graph of total_deaths_per_million vs gdp_per_capita

The graph shows the data gdp_per_capita. Gdp stands for gross domestic products which means the average total income of each country. We can see that most of the countries which have below 35k gdp have less number of total deaths per million compared to countries which have more gdp. This could also mean that, in my assumption, countries which have more than 35k gdp have a stable financial status of citizens are not aware of the virus. They could hang around their cities without protection and lack of awareness about the global virus. From the data, we can also know that there are no countries who have gdp between 95k until 110k. Most of the countries which are affected by the global virus is country who have below than 70k gdp.

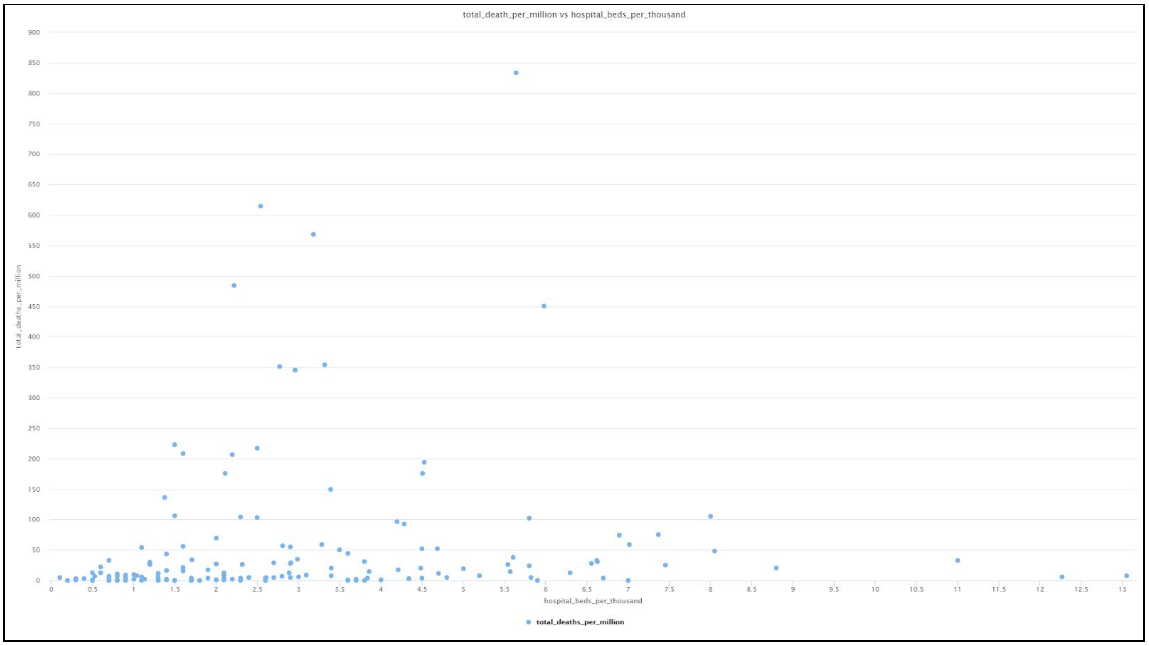


Figure abc: Scattered graph of total_deaths_per_million vs hospital_beds_per_thousand

From the factor of a thousand hospital beds of thousand, we can see that most countries in the world have less than 7 thousand hospital beds in each country. Countries that have less hospital beds might not be able to contain the number of patients infected with the global virus. That is why we can see in the figure above, a lot of countries have almost equal number of hospital beds and its affecting the total number of deaths per million in most of the country.

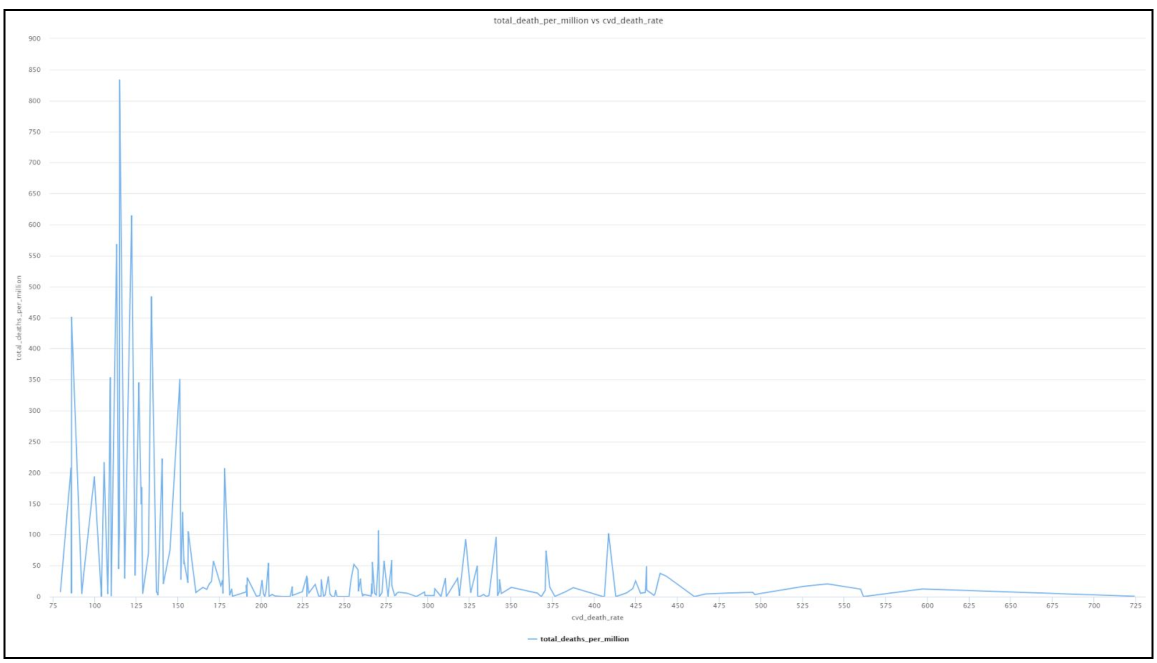


Figure abcde : Line graph of total_death_per_million vs cvd_death_rate

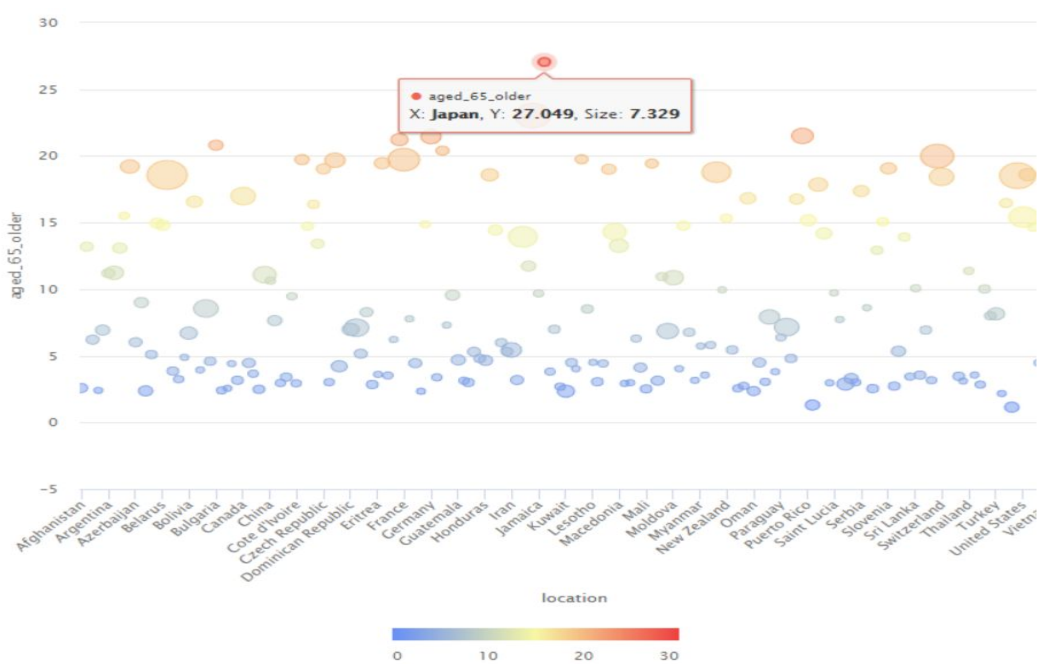
Factors that could lead affect the person to the COVID-19 is if the person has low immunity of the body and has some other chronic disease. For example, the data showed that cardiovascular disease also known as heart attack also affects the total number of deaths per million. Generally, we are known that COVID-19 affects the person's hard to breathe. While cardiovascular disease involves the heart which is related to the human respiratory system that helps them breathe. That is why this factor is also counted in collecting the data.

The data collected here is to see the total death of people who are infected with COVID-19. The countries that have lower cvd death rate have higher total deaths per million compared to countries that have higher cvd death rate. This means the country with high cvd death rate not affecting the total number of deaths because they might die to the disease instead.

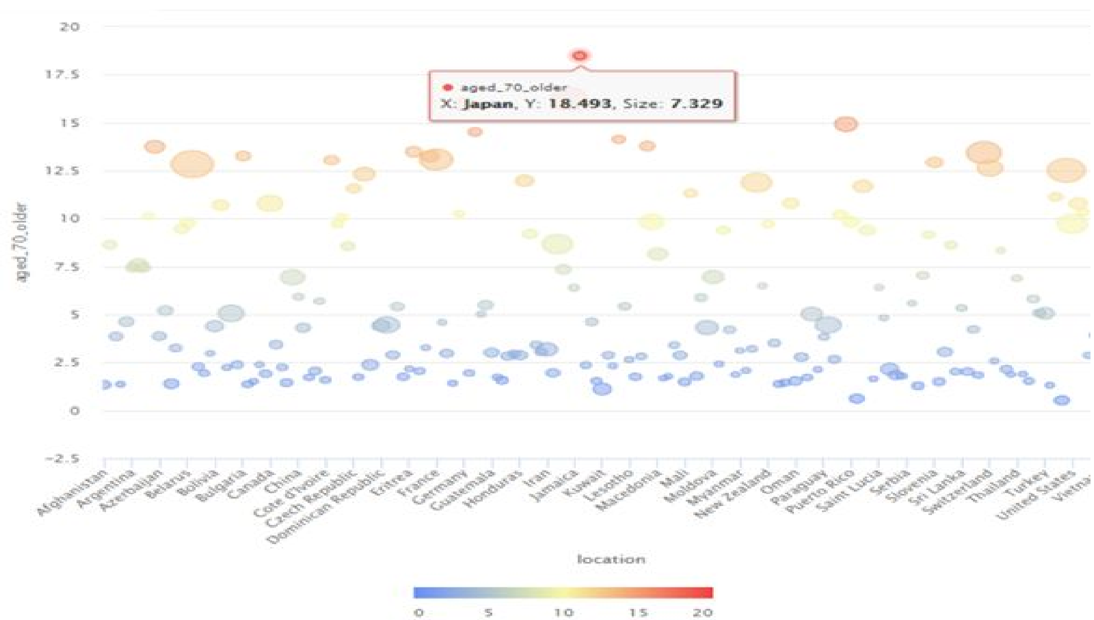
location	total_deaths...	population	aged_65... ↓	aged_70_ol...
Japan	7.329	126476458	27.049	18.493
Italy	568.474	60461828	23.021	16.240
Portugal	149.068	10196707	21.502	14.924
Germany	105.032	83783945	21.453	15.957
Finland	58.837	5540718	21.228	13.264
Bulgaria	25.329	6948445	20.801	13.272
Greece	17.653	10423056	20.396	14.524
Sweden	484.292	10099270	19.985	13.433
Latvia	14.845	1886202	19.754	14.136

Table

The table below shows the percentage of aged 65 and age 70 for each country based on the population. Based on the table we can find the country with the highest percentage of both age and the relationship between total death per million and both age percentages using graphs.



Graph



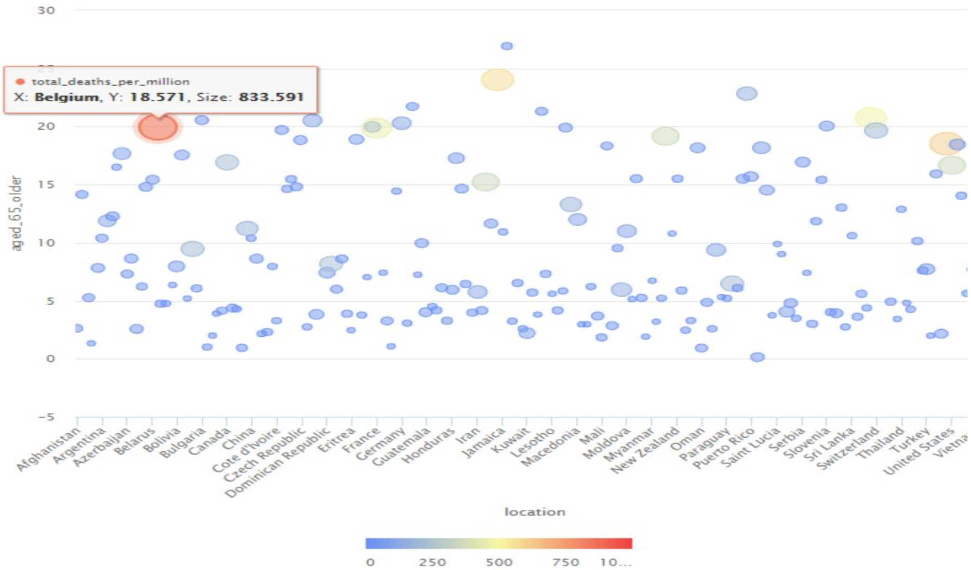
Graph

As we can see with both scatter graphs, Japan has the highest percentage than the other country. Many countries have below 10 of percentage age 65 citizens, while below 5 of percentage for age 70 citizens. In the scatter plot x-axis is set as the location and y-axis for the percentage of both ages. Next, the color of each plot represents the amount of percentage for both age based on the color indicator.

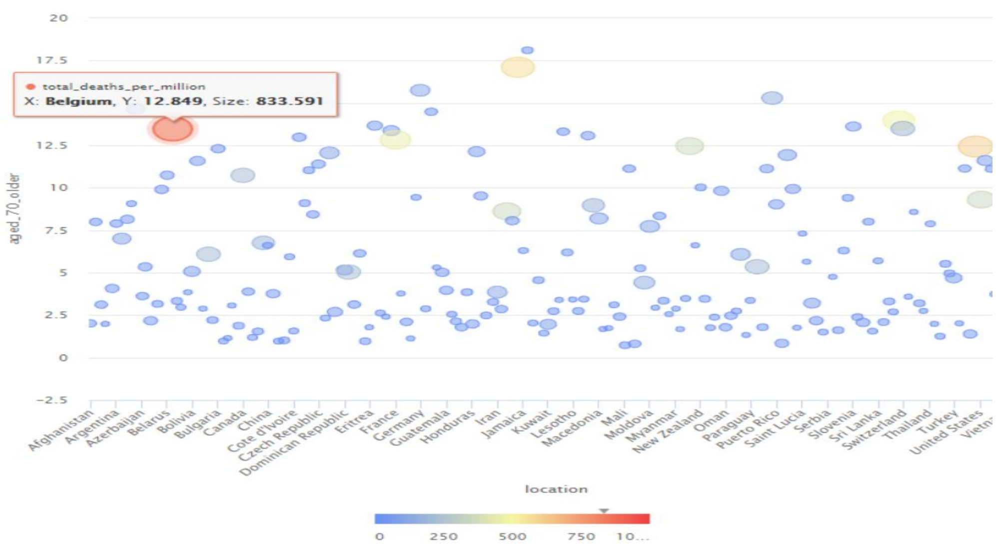
location	total_de...	median_age	aged_65_ol...	aged_70_ol...
Belgium	833.591	41.800	18.571	12.849
United Kingd...	614.795	40.800	18.517	12.527
Italy	568.474	47.900	23.021	16.240
Sweden	484.292	41	19.985	13.433
France	450.964	42	19.718	13.079
Netherlands	353.606	43.200	18.779	11.881
United States	350.834	38.300	15.413	9.732
Ireland	345.498	38.700	13.928	8.678
Ecuador	222.694	28.100	7.104	4.458
Canada	216.601	41.400	16.984	10.797

Table

Based on the clean data as shown on the table above, the next objective is to identify the relationship for both ages with the total death per million whether if the percentage of the age is higher might affect the number of total deaths per million.



Graph



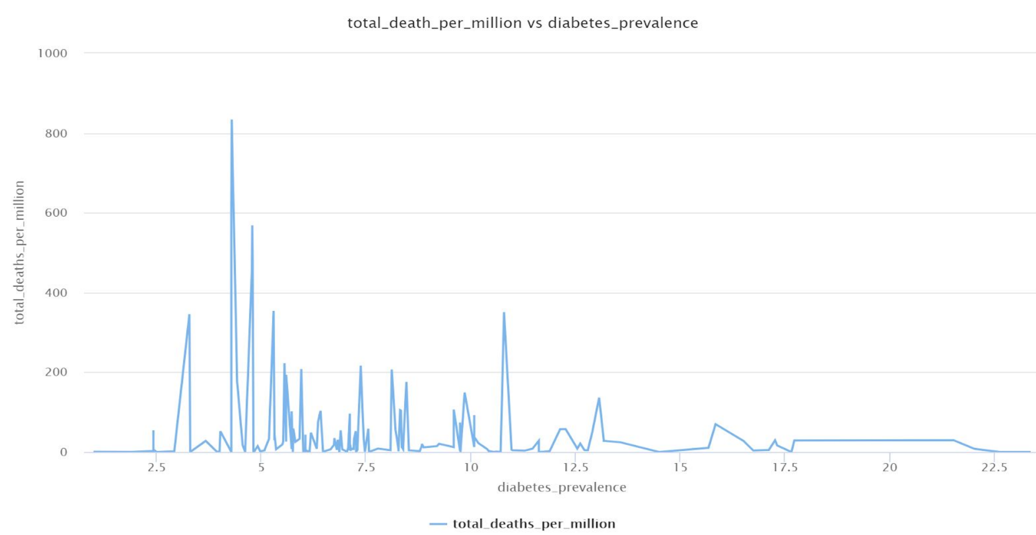
Graph

The size of each scatter plot is based on the total death per million and the color is based on the indicator. Thus, we can see that some of the countries with the amount of percentage above 15 for the percentage age 65 and 7.5 for the percentage age 70, show the increasing total death per million and the highest total death per million is Belgium with 18.571 percent of age 65 and 12.849 percent of age 70 . However, it will not be the main cause of the total death per million because the highest percentage of both ages, which is Japan, has a small total death per million. As an assumption, the percentage of both age can be one of the causes of death per million but it is related to medical facilities and the awareness for each country.

location	total_deaths_per_million	diabetes_prevalence
Afghanistan	12.279	9.590
Albania	12.510	10.080
Algeria	17.719	6.730
Angola	0.183	3.940
Antigua and ...	30.635	13.170
Argentina	18.896	5.500
Armenia	96.179	7.110
Aruba	28.099	11.620
Australia	4	5.070
Austria	75.280	6.350
Azerbaijan	12.033	7.110

Table

The table above shows the attribute of location , total death per million and diabetes prevalence based on the clean data. The objective is to find the relationship between total death per million and diabetes prevalence percentage in each country.



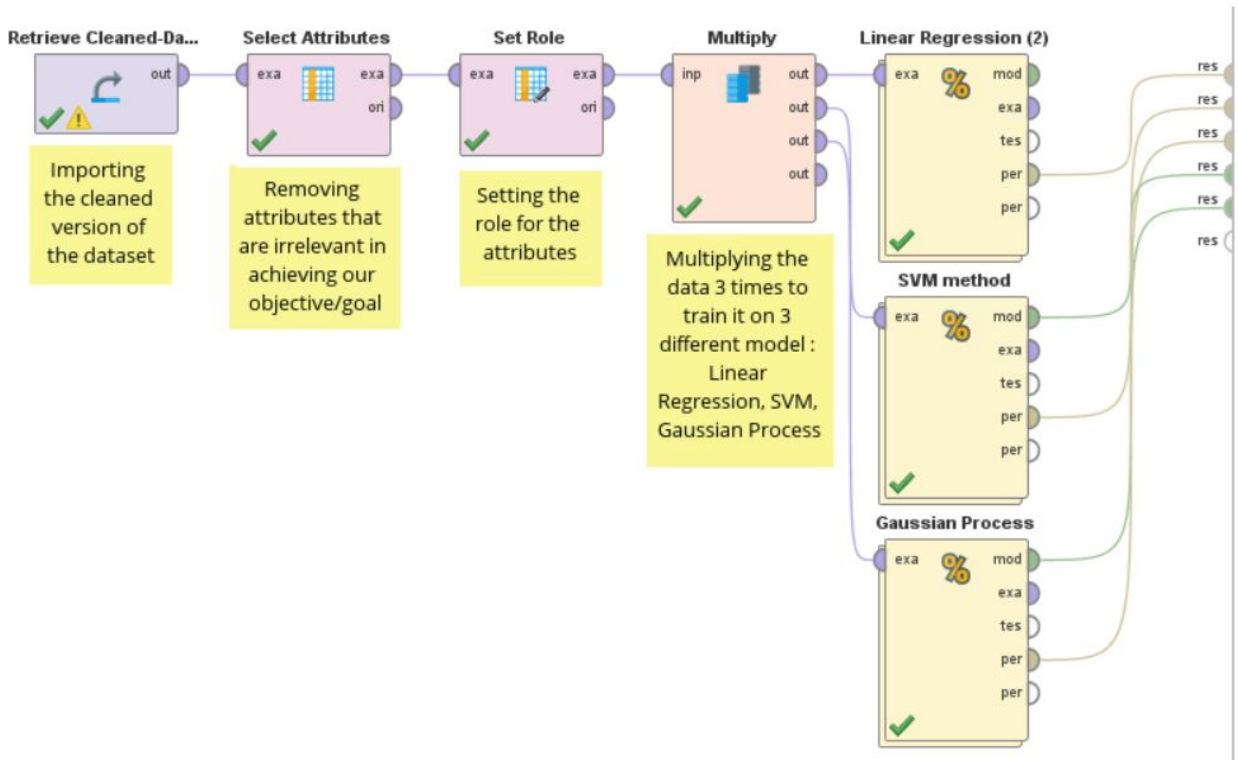
Graph

As the result based on the line graph, we can see that if the lower percentage of diabetes prevalence in certain countries the higher total of death in million has been recorded. Therefore, for the assumption the higher percentage of diabetes prevalence may come from a country with better facilities and higher income that takes precautions to prevent the diabetes prevalence to be infected by the disease.

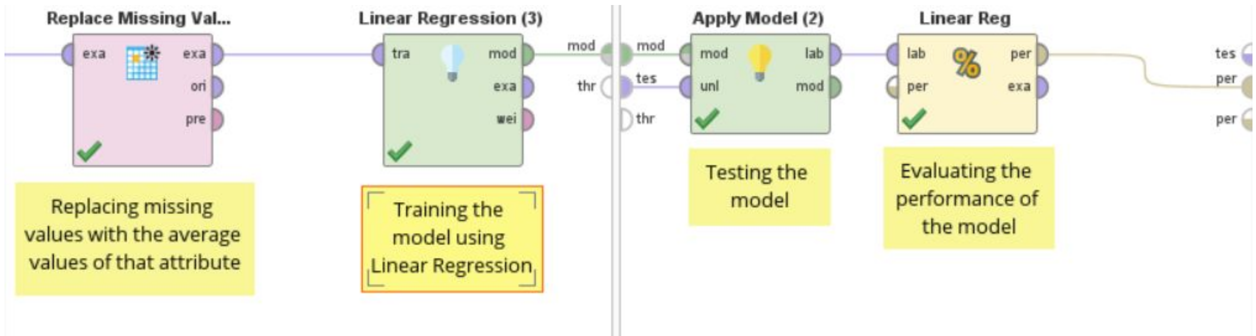
Modelling & Evaluation

After we have finished exploring and understanding the data, we can now create a model to predict the death rate of COVID-19 virus of a country based on several factors. With this we can also see which factors affect the death rate of COVID-19 the most.

Our group decided to use RapidMiner to create three different models that use different methods which are : Support Vector Machine (SVM), Gaussian Process and Linear Regression. Below are how we use RapidMiner to create the model and evaluate them.



Linear Regression



Regression is a technique used to calculate the strength of the relationship between one dependent variable(Output) and other different independent variables(Input). With this it can then predict the dependent variable(Output) based on the independent variables(Input).

Evaluation of Linear Regression with different Feature Selection Method :

M5 Prime :

```
root_mean_squared_error: 93.899 +/- 46.888
```

Greedy :

```
root_mean_squared_error: 93.441 +/- 46.988
```

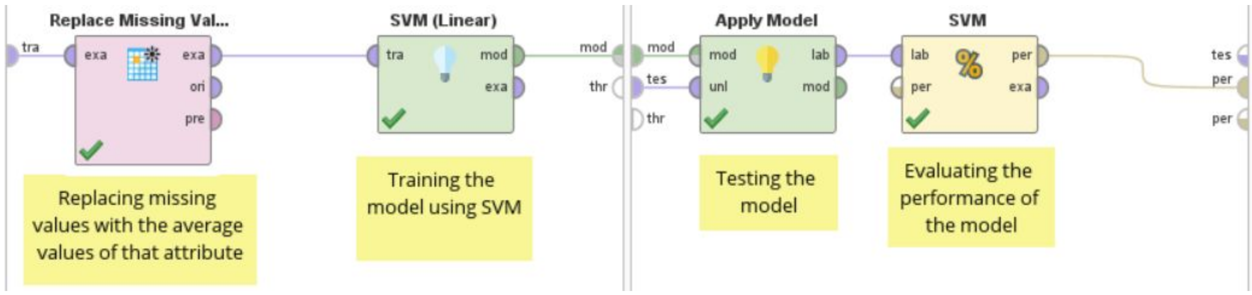
T-Test :

```
root_mean_squared_error: 94.420 +/- 47.258
```

It is noted here that the Greedy Feature Selection provide a better model compared to the other two. This means that the feature that Greedy method chose were the determining factors that affect the death rate of COVID-19. These factors are :

POPULATION DENSITY, AGED 70 OLDER, GDP PER CAPITA, DIABETES PREVALENCE, HOSPITAL BEDS PER THOUSAND

SVM

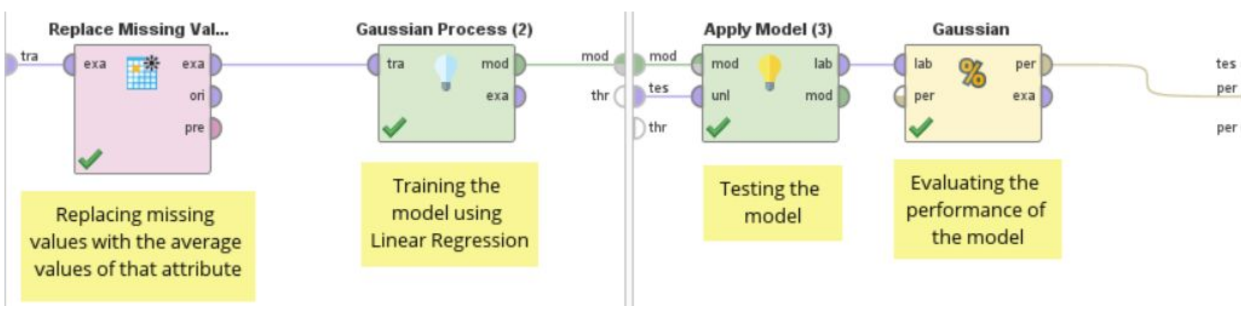


SVM uses the Java implementation of the support vector machine *mySVM* by Stefan Rüping. It is restricted to the dot (linear) kernel, but outputs a high performance model that only contains the linear coefficient for faster model application.

Evaluation of Support Vector Machine :

```
root_mean_squared_error: 97.311 +/- 76.225
```

Gaussian Process



A Gaussian process is a stochastic process whose realizations consist of random values associated with every point in a range of times (or of space) such that each such random variable has a normal distribution.

Evaluation of Gaussian Process :

```
root_mean_squared_error: 106.190 +/- 77.857
```

We can conclude that the Linear Regression Model (Greedy) have the best performance compared to others.

Reference

[1]Spain: Measures tightened as COVID-19 cases surge. (n.d.). Retrieved July 15, 2020, from https://www.aa.com.tr/en/europe/spain-measures-tightened-as-covid-19-cases-surge/1909229

[2](2020).*owid/covid-19-data*.GitHub.
https://github.com/owid/covid-19-data/tree/master/public/data