FAKULTI TEKNOLOGI MAKLUMAT DAN KOMUNIKASI
SEMESTER 2 2019/2020
BITI 2513

INTRODUCTION TO DATA SCIENCE

TASK 2

PREPARED BY:

BIL	STUDENT NAME	MATRIC NO
1	ABDUL HAZIQ BIN ABD KHALID	B031810256
2	MUHAMMAD NABIL IMRAN BIN SOLEHAN	B031810234
3	AHMAD NAUFAL BIN MOHD SALEH	B031810382
4	MEOR AMIRUL ASHRAF BIN JAMALULAIL	B031810468

TASK 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 handwashing facilities on premises, most recent year available
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010

The dataset are 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

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

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 shows 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

The goal of our project is to find which attributes affects 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	This attributes are originally used as the ID attribute (Primary Key). Our group decided to use the attribute “location” which hold 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 have different cultures,infrastructures,etc. Therefore we concluded 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 have more than 25% missing values in its data are to be removed.

$\frac{25}{100} \times 208 = 52$ Hence, any attributes with missing values that are higher than 52 will be remove.

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 have 14 attributes. All attributes now have reasonable amount of reliable data. However, we found out that certain country does not provide very much information. Therefore there are a lot of missing value in their row. We also concluded to removed 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 have 184 reliable data with 14 reliable attributes. The dataset is now ready to be studied and manipulated.