Factors that affects Covid Spread

Group 38

Introduction

The purpose of this report is to find out about the factors which affect the spread of covid virus in every country or continent using the datasets regarding the covid virus. There would be three sections in this report where the first section would describe about the first dataset used such as its summary statistics, explains removing of the missing values and outliers and describing the problem statement. The second section would include another dataset and explains how the second dataset would be joined with the first dataset. Lastly, the last section would interpret the visualisation plots to further understand the problem statement and devise methods into overcoming the problem statement.

Section A

```
> summary(covid_data)
   iso_code          continent           location            date            total_cases          new_cases
 Length:166326      Length:166326      Length:166326      Length:166326      Min.   :        1   Min.   :      0
 Class :character   Class :character   Class :character   Class :character   1st Qu.:     2001   1st Qu.:      1
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :    26117   Median :     79
                                                                            Mean   :  2536044   Mean   :  11571
                                                                            3rd Qu.:   298702   3rd Qu.:   1063
                                                                            Max.   :445129499   Max.   :4206334
                                                                            NA's   :3033        NA's   :3193
 new_cases_smoothed  total_deaths       new_deaths      new_deaths_smoothed total_cases_per_million
 Min.   :      0    Min.   :      1   Min.   :    0.0   Min.   :    0.000   Min.   :     0.0
 1st Qu.:      7    1st Qu.:     79   1st Qu.:    0.0   1st Qu.:    0.143   1st Qu.:   623.6
 Median :    107    Median :    783   Median :    2.0   Median :    2.429   Median :  4731.5
 Mean   :  11566    Mean   :  57664   Mean   :  171.1   Mean   :  172.673   Mean   : 29447.8
 3rd Qu.:   1146    3rd Qu.:   7307   3rd Qu.:   20.0   3rd Qu.:   21.286   3rd Qu.: 37724.5
 Max.   :3444237    Max.   :5995245   Max.   :18020.0   Max.   :14689.143   Max.   :706541.9
 NA's   :5176       NA's   :20875     NA's   :20839     NA's   :22936       NA's   :3791
 new_cases_per_million new_cases_smoothed_per_million total_deaths_per_million new_deaths_per_million
 Min.   :    0.00      Min.   :    0.00               Min.   :   0.00          Min.   :  0.000
 1st Qu.:    0.04      1st Qu.:    1.63               1st Qu.:  18.58          1st Qu.:  0.000
 Median :   11.44      Median :   18.83               Median : 127.74          Median :  0.127
 Mean   :  166.43      Mean   :  165.51               Mean   : 509.38          Mean   :  1.687
 3rd Qu.:  101.29      3rd Qu.:  120.86               3rd Qu.: 711.96          3rd Qu.:  1.369
 Max.   :51427.49      Max.   :16052.61               Max.   :6322.26          Max.   :453.772
 NA's   :3951          NA's   :5928                   NA's   :21620           NA's   :21584
 new_deaths_smoothed_per_million reproduction_rate  icu_patients     icu_patients_per_million hosp_patients
 Min.   :  0.000                 Min.   :-0.08     Min.   :    0.0   Min.   :  0.00           Min.   :     0
 1st Qu.:  0.018                 1st Qu.: 0.81     1st Qu.:   27.0   1st Qu.:  3.90           1st Qu.:   129
 Median :  0.292                 Median : 1.00     Median :  146.0   Median : 13.38           Median :   702
 Mean   :  1.688                 Mean   : 1.00     Mean   :  903.8   Mean   : 23.98           Mean   :  4192
 3rd Qu.:  1.765                 3rd Qu.: 1.18     3rd Qu.:  594.0   3rd Qu.: 34.84           3rd Qu.:  2759
 Max.   :144.167                 Max.   : 6.14     Max.   :28891.0   Max.   :177.28           Max.   :154536
 NA's   :23675                   NA's   :40506     NA's   :142863    NA's   :142863           NA's   :141709
 hosp_patients_per_million weekly_icu_admissions weekly_icu_admissions_per_million weekly_hosp_admissions
 Min.   :   0.00           Min.   :   0.0        Min.   :  0.00                    Min.   :     0
 1st Qu.:  26.39           1st Qu.:  47.0        1st Qu.:  3.88                    1st Qu.:   320
 Median :  84.22           Median : 219.0        Median : 11.09                    Median :  1330
 Mean   : 167.23           Mean   : 469.8        Mean   : 15.47                    Mean   :  6058
 3rd Qu.: 227.84           3rd Qu.: 671.0        3rd Qu.: 20.45                    3rd Qu.:  5184
 Max.   :1544.08           Max.   :4838.0        Max.   :221.21                    Max.   :154696
 NA's   :141709            NA's   :160893        NA's   :160893                    NA's   :155403
 weekly_hosp_admissions_per_million   new_tests        total_tests        total_tests_per_thousand
 Min.   :  0.00                     Min.   :      1   Min.   :        0   Min.   :    0.00
 1st Qu.: 23.51                     1st Qu.:   2478   1st Qu.:   372813   1st Qu.:   34.72
 Median : 72.66                     Median :   9809   Median :  1951879   Median :  179.34
 Mean   :103.63                     Mean   :  68308   Mean   : 17325202   Mean   :  740.15
 3rd Qu.:142.32                     3rd Qu.:  38926   3rd Qu.:  9148832   3rd Qu.:  703.46
 Max.   :839.13                     Max.   :3740296   Max.   :820619379   Max.   :29001.02
 NA's   :155403                     NA's   :99009     NA's   :97071       NA's   :97071
```

```
new_tests_per_thousand new_tests_smoothed new_tests_smoothed_per_thousand positive_rate   tests_per_case
Min.   :  0.00         Min.   :      0    Min.   :  0.00                  Min.   :0.00    Min.   :     1.0
1st Qu.:  0.26         1st Qu.:   2153    1st Qu.:  0.23                  1st Qu.:0.02    1st Qu.:     7.2
Median :  0.94         Median :   8683    Median :  0.90                  Median :0.06    Median :    17.3
Mean   :  3.23         Mean   :  61026    Mean   :  2.91                  Mean   :0.10    Mean   :   198.9
3rd Qu.:  2.89         3rd Qu.:  36079    3rd Qu.:  2.70                  3rd Qu.:0.14    3rd Qu.:    52.3
Max.   :534.01         Max.   :3080396    Max.   :147.60                  Max.   :0.99    Max.   :422065.6
NA's   :99009          NA's   :82291      NA's   :82291                   NA's   :87671   NA's   :88242
tests_units        total_vaccinations  people_vaccinated   people_fully_vaccinated total_boosters
Length:166326      Min.   :0.000e+00   Min.   :0.000e+00   Min.   :1.000e+00       Min.   :1.000e+00
Class :character   1st Qu.:5.989e+05   1st Qu.:3.924e+05   1st Qu.:2.745e+05       1st Qu.:2.091e+03
Mode  :character   Median :4.754e+06   Median :2.920e+06   Median :2.230e+06       Median :4.366e+05
                   Mean   :1.720e+08   Mean   :8.814e+07   Mean   :6.945e+07       Mean   :1.906e+07
                   3rd Qu.:2.944e+07   3rd Qu.:1.713e+07   3rd Qu.:1.371e+07       3rd Qu.:3.984e+06
                   Max.   :1.085e+10   Max.   :4.976e+09   Max.   :4.401e+09       Max.   :1.424e+09
                   NA's   :121132      NA's   :123339      NA's   :126085          NA's   :148787
new_vaccinations   new_vaccinations_smoothed total_vaccinations_per_hundred people_vaccinated_per_hundred
Min.   :       0   Min.   :       0          Min.   :  0.00                  Min.   :  0.00
1st Qu.:    6356   1st Qu.:    1058          1st Qu.: 12.12                  1st Qu.:  8.62
Median :   41207   Median :    9335          Median : 58.50                  Median : 36.23
Mean   : 1170261   Mean   :  523267          Mean   : 72.79                  Mean   : 37.85
3rd Qu.:  275256   3rd Qu.:   66413          3rd Qu.:123.45                  3rd Qu.: 64.56
Max.   :54905988   Max.   :43536956          Max.   :335.81                  Max.   :124.57
NA's   :128879     NA's   :81928             NA's   :121132                  NA's   :123339
people_fully_vaccinated_per_hundred total_boosters_per_hundred new_vaccinations_smoothed_per_million
Min.   :  0.00                      Min.   : 0.00              Min.   :     0
1st Qu.:  5.11                      1st Qu.: 0.01              1st Qu.:   686
Median : 27.08                      Median : 3.02              Median :  2203
Mean   : 32.48                      Mean   :12.20              Mean   :  3317
3rd Qu.: 57.90                      3rd Qu.:19.90              3rd Qu.:  4763
Max.   :121.45                      Max.   :89.80              Max.   :117497
NA's   :126085                      NA's   :148787             NA's   :81928
new_people_vaccinated_smoothed new_people_vaccinated_smoothed_per_hundred stringency_index   population
Min.   :       0               Min.   : 0.00                              Min.   :  0.00   Min.   :4.700e+01
1st Qu.:     428               1st Qu.: 0.02                              1st Qu.: 40.74   1st Qu.:1.172e+06
Median :    4026               Median : 0.07                              Median : 54.63   Median :8.478e+06
Mean   :  216100               Mean   : 0.15                              Mean   : 54.59   Mean   :1.474e+08
3rd Qu.:   27327               3rd Qu.: 0.19                              3rd Qu.: 70.37   3rd Qu.:3.393e+07
Max.   :21419033               Max.   :11.75                              Max.   :100.00   Max.   :7.875e+09
NA's   :83238                  NA's   :83238                              NA's   :36254    NA's   :1075
population_density    median_age    aged_65_older    aged_70_older    gdp_per_capita     extreme_poverty
Min.   :    0.137   Min.   :15.10   Min.   : 1.144   Min.   : 0.526   Min.   :   661.2   Min.   : 0.10
1st Qu.:   36.253   1st Qu.:22.20   1st Qu.: 3.507   1st Qu.: 2.063   1st Qu.:  4449.9   1st Qu.: 0.60
Median :   85.129   Median :29.90   Median : 6.614   Median : 3.915   Median : 12951.8   Median : 2.20
Mean   :  464.327   Mean   :30.57   Mean   : 8.763   Mean   : 5.534   Mean   : 19644.3   Mean   :13.58
3rd Qu.:  212.865   3rd Qu.:39.10   3rd Qu.:14.178   3rd Qu.: 8.678   3rd Qu.: 27936.9   3rd Qu.:21.20
Max.   :20546.766   Max.   :48.20   Max.   :27.049   Max.   :18.493   Max.   :116935.6   Max.   :77.60
NA's   :18398       NA's   :28495   NA's   :29989    NA's   :29234    NA's   :27822      NA's   :75111

cardiovasc_death_rate diabetes_prevalence female_smokers   male_smokers   handwashing_facilities
Min.   : 79.37        Min.   : 0.990      Min.   : 0.10   Min.   : 7.70   Min.   :  1.19
1st Qu.:168.71        1st Qu.: 5.310      1st Qu.: 1.90   1st Qu.:21.60   1st Qu.: 19.35
Median :243.81        Median : 7.170      Median : 6.30   Median :31.40   Median : 49.84
Mean   :260.21        Mean   : 8.211      Mean   :10.63   Mean   :32.78   Mean   : 50.79
3rd Qu.:329.94        3rd Qu.:10.430      3rd Qu.:19.30   3rd Qu.:41.30   3rd Qu.: 83.24
Max.   :724.42        Max.   :30.530      Max.   :44.00   Max.   :78.10   Max.   :100.00
NA's   :29548         NA's   :22377       NA's   :60276   NA's   :61731   NA's   :97757
hospital_beds_per_thousand life_expectancy human_development_index excess_mortality_cumulative_absolute
Min.   : 0.10              Min.   :53.28   Min.   :0.394            Min.   : -37726.1
1st Qu.: 1.30              1st Qu.:69.50   1st Qu.:0.602            1st Qu.:    -75.2
Median : 2.40              Median :75.05   Median :0.743            Median :   3424.6
Mean   : 3.03              Mean   :73.58   Mean   :0.726            Mean   :  37613.0
3rd Qu.: 4.00              3rd Qu.:78.93   3rd Qu.:0.845            3rd Qu.:  24784.6
Max.   :13.80              Max.   :86.75   Max.   :0.957            Max.   :1080748.1
NA's   :42662              NA's   :11058   NA's   :30073            NA's   :160630
excess_mortality_cumulative excess_mortality excess_mortality_cumulative_per_million
Min.   :-28.45              Min.   :-95.92   Min.   :-1826.60
1st Qu.: -0.72              1st Qu.: -0.75   1st Qu.:  -29.79
Median :  6.07              Median :  7.20   Median :  473.39
Mean   :  9.40              Mean   : 15.97   Mean   :  972.20
3rd Qu.: 14.52              3rd Qu.: 23.00   3rd Qu.: 1656.36
Max.   :111.01              Max.   :374.93   Max.   : 9153.06
NA's   :160630              NA's   :160630   NA's   :160630
```

Figure 1: The summary of covid data

The summary and descriptive statistics of each variable in the raw dataset are shown in figure 1 above.

```
> str(covid_data)
'data.frame':   166326 obs. of  67 variables:
 $ iso_code                                  : chr  "AFG" "AFG" "AFG" "AFG" ...
 $ continent                                 : chr  "Asia" "Asia" "Asia" "Asia" ...
 $ location                                  : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ date                                      : chr  "2020-02-24" "2020-02-25" "2020-02-26" "2020-02-27" ...
 $ total_cases                               : num  5 5 5 5 5 5 5 5 5 5 ...
 $ new_cases                                 : num  5 0 0 0 0 0 0 0 0 0 ...
 $ new_cases_smoothed                        : num  NA NA NA NA NA NA 0.714 0 0 0 ...
 $ total_deaths                              : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_deaths                                : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_deaths_smoothed                       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ total_cases_per_million                   : num  0.126 0.126 0.126 0.126 0.126 0.126 0.126 0.126 0.126 0.126 ...
 $ new_cases_per_million                     : num  0.126 0 0 0 0 0 0 0 0 0 ...
 $ new_cases_smoothed_per_million            : num  NA NA NA NA NA NA 0.018 0 0 0 ...
 $ total_deaths_per_million                  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_deaths_per_million                    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_deaths_smoothed_per_million           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ reproduction_rate                         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ icu_patients                              : num  NA NA NA NA NA NA NA NA NA NA ...
 $ icu_patients_per_million                  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ hosp_patients                             : num  NA NA NA NA NA NA NA NA NA NA ...
 $ hosp_patients_per_million                 : num  NA NA NA NA NA NA NA NA NA NA ...
 $ weekly_icu_admissions                     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ weekly_icu_admissions_per_million         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ weekly_hosp_admissions                    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ weekly_hosp_admissions_per_million        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_tests                                 : num  NA NA NA NA NA NA NA NA NA NA ...
 $ total_tests                               : num  NA NA NA NA NA NA NA NA NA NA ...
 $ total_tests_per_thousand                  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_tests_per_thousand                    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_tests_smoothed                        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_tests_smoothed_per_thousand           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ positive_rate                             : num  NA NA NA NA NA NA NA NA NA NA ...
 $ tests_per_case                            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ tests_units                               : chr  "" "" "" "" ...
 $ total_vaccinations                        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ people_vaccinated                         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ people_fully_vaccinated                   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ total_boosters                            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_vaccinations                          : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_vaccinations_smoothed                 : num  NA NA NA NA NA NA NA NA NA NA ...
 $ total_vaccinations_per_hundred            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ people_vaccinated_per_hundred             : num  NA NA NA NA NA NA NA NA NA NA ...
 $ people_fully_vaccinated_per_hundred       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ total_boosters_per_hundred                : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_vaccinations_smoothed_per_million     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_people_vaccinated_smoothed            : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_people_vaccinated_smoothed_per_hundred: num  NA NA NA NA NA NA NA NA NA NA ...

$ stringency_index                          : num  8.33 8.33 8.33 8.33 8.33 ...
$ population                                : num  39835428 39835428 39835428 39835428 39835428 ...
$ population_density                        : num  54.4 54.4 54.4 54.4 54.4 ...
$ median_age                                : num  18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 ...
$ aged_65_older                             : num  2.58 2.58 2.58 2.58 2.58 ...
$ aged_70_older                             : num  1.34 1.34 1.34 1.34 1.34 ...
$ gdp_per_capita                            : num  1804 1804 1804 1804 1804 ...
$ extreme_poverty                           : num  NA NA NA NA NA NA NA NA NA NA ...
$ cardiovasc_death_rate                     : num  597 597 597 597 597 ...
$ diabetes_prevalence                       : num  9.59 9.59 9.59 9.59 9.59 9.59 9.59 9.59 9.59 9.59 ...
$ female_smokers                            : num  NA NA NA NA NA NA NA NA NA NA ...
$ male_smokers                              : num  NA NA NA NA NA NA NA NA NA NA ...
$ handwashing_facilities                    : num  37.7 37.7 37.7 37.7 37.7 ...
$ hospital_beds_per_thousand                : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
$ life_expectancy                           : num  64.8 64.8 64.8 64.8 64.8 ...
$ human_development_index                   : num  0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 ...
$ excess_mortality_cumulative_absolute      : num  NA NA NA NA NA NA NA NA NA NA ...
$ excess_mortality_cumulative               : num  NA NA NA NA NA NA NA NA NA NA ...
$ excess_mortality                          : num  NA NA NA NA NA NA NA NA NA NA ...
$ excess_mortality_cumulative_per_million   : num  NA NA NA NA NA NA NA NA NA NA ...
```

Figure 2: The structure of covid data

The total observations in the raw dataset is 166326 and the number of variables is 67. This raw dataset contains 4 categorical variables and 63 numerical variables based on the data types shown in Figure 2.

```
> covid_data_filter = covid_data[,c("continent", "location", "date", "total_cases", "new_cases",
+                                   "total_deaths", "new_deaths", "population", "population_density",
+                                   "median_age", "aged_65_older", "aged_70_older", "gdp_per_capita",
+                                   "life_expectancy")]
> summary(covid_data_filter)
  continent           location             date           total_cases          new_cases        total_deaths       new_deaths
 Length:166326      Length:166326      Length:166326      Min.   :       1   Min.   :     0   Min.   :      1   Min.   :    0.0
 Class :character   Class :character   Class :character   1st Qu.:    2001   1st Qu.:     1   1st Qu.:     79   1st Qu.:    0.0
 Mode  :character   Mode  :character   Mode  :character   Median :   26117   Median :    79   Median :    783   Median :    2.0
                                                          Mean   :  2536044   Mean   :  11571   Mean   :  57664   Mean   :  171.1
                                                          3rd Qu.:   298702   3rd Qu.:   1063   3rd Qu.:   7307   3rd Qu.:   20.0
                                                          Max.   :445129499   Max.   :4206334   Max.   :5995245   Max.   :18020.0
                                                          NA's   :3033        NA's   :3193      NA's   :20875     NA's   :20839
   population        population_density    median_age    aged_65_older    aged_70_older    gdp_per_capita     life_expectancy
 Min.   :4.700e+01   Min.   :    0.137   Min.   :15.10   Min.   : 1.144   Min.   : 0.526   Min.   :   661.2   Min.   :53.28
 1st Qu.:1.172e+06   1st Qu.:   36.253   1st Qu.:22.20   1st Qu.: 3.507   1st Qu.: 2.063   1st Qu.:  4449.9   1st Qu.:69.50
 Median :8.478e+06   Median :   85.129   Median :29.90   Median : 6.614   Median : 3.915   Median : 12951.8   Median :75.05
 Mean   :1.474e+08   Mean   :  464.327   Mean   :30.57   Mean   : 8.763   Mean   : 5.534   Mean   : 19644.3   Mean   :73.58
 3rd Qu.:3.393e+07   3rd Qu.:  212.865   3rd Qu.:39.10   3rd Qu.:14.178   3rd Qu.: 8.678   3rd Qu.: 27936.9   3rd Qu.:78.93
 Max.   :7.875e+09   Max.   :20546.766   Max.   :48.20   Max.   :27.049   Max.   :18.493   Max.   :116935.6   Max.   :86.75
 NA's   :1075        NA's   :18398       NA's   :28495   NA's   :29989    NA's   :29234    NA's   :27822      NA's   :11058
```

Figure 3: Summary of the filtered dataset

Since we are only focusing on a few variables, we filtered unused variables. Figure 3 shows the filtered dataset.

```
> #Convert chr to date type
> str(covid_data_filter)
'data.frame':   166326 obs. of  14 variables:
 $ continent          : chr  "Asia" "Asia" "Asia" "Asia" ...
 $ location           : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ date               : chr  "2020-02-24" "2020-02-25" "2020-02-26" "2020-02-27" ...
 $ total_cases        : num  5 5 5 5 5 5 5 5 5 5 ...
 $ new_cases          : num  5 0 0 0 0 0 0 0 0 0 ...
 $ total_deaths       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ new_deaths         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ population         : num  39835428 39835428 39835428 39835428 39835428 ...
 $ population_density: num  54.4 54.4 54.4 54.4 54.4 ...
 $ median_age         : num  18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 ...
 $ aged_65_older      : num  2.58 2.58 2.58 2.58 2.58 ...
 $ aged_70_older      : num  1.34 1.34 1.34 1.34 1.34 ...
 $ gdp_per_capita     : num  1804 1804 1804 1804 1804 ...
 $ life_expectancy    : num  64.8 64.8 64.8 64.8 64.8 ...
```

Figure 4: Structure of filtered dataset

Figure 4 shows the structure of the filtered dataset. After looking at the variable types, we found out that the date variable should be in date type instead of character type so we converted it into date type.

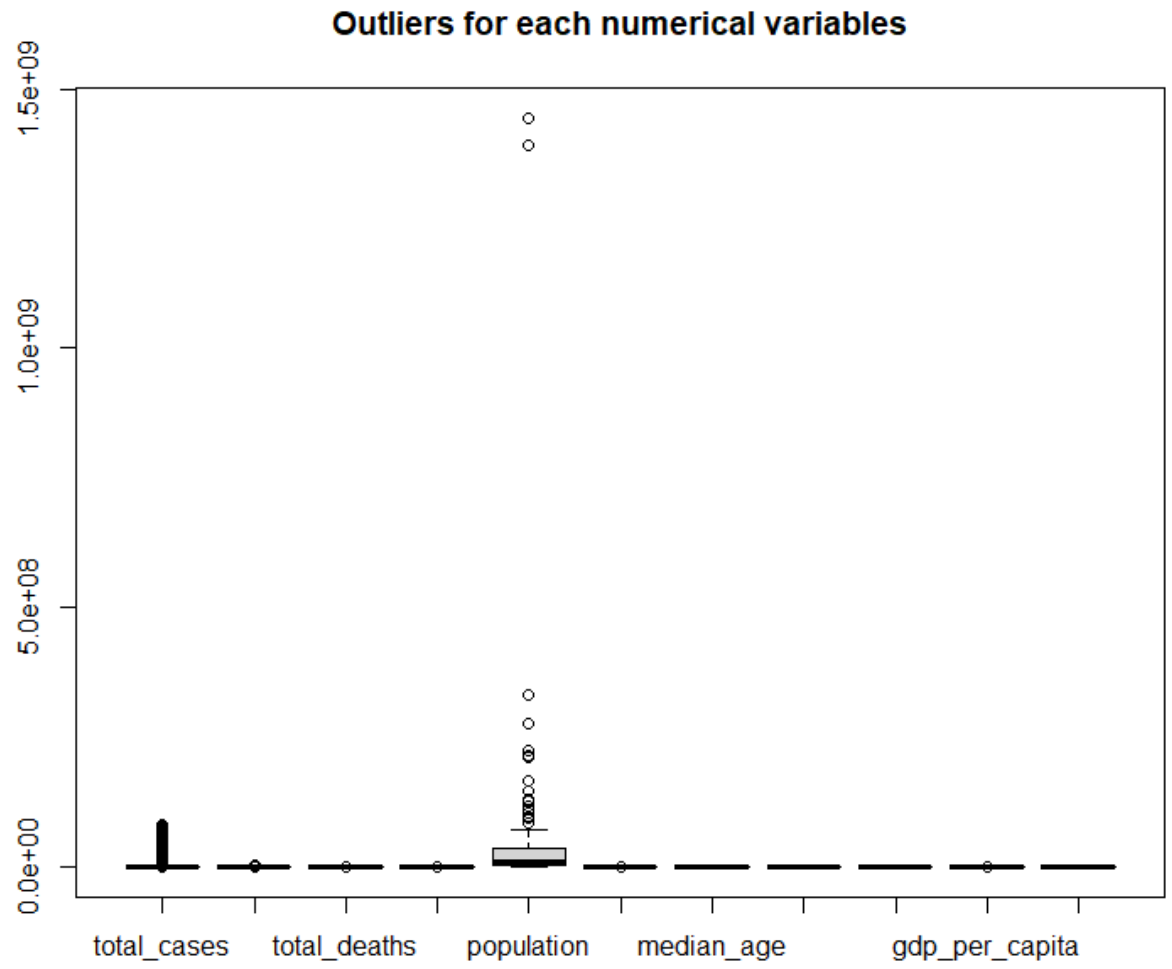
```
> #Empty string in continent
> unique(covid_data_filter$continent)
[1] "Asia"          ""              "Europe"        "Africa"        "North America" "South America" "Oceania"
> covid_data_filter = covid_data_filter[!(covid_data_filter$continent==""), ]
```

Figure 5: Empty string present in continent

Figure 5 shows that an empty string is present in the continent variable. The rows containing empty strings in the continent variable were removed.

Outliers

The outliers are shown after omitting the NA's and filtering the variables from the raw data




```
> covid_clean$median_age[covid_clean$median_age %in% boxplot.stats(covid_clean$median_age) $ out]
numeric(0)
> covid_clean$aged_65_older[covid_clean$aged_65_older %in% boxplot.stats(covid_clean$aged_65_older) $ out]
numeric(0)
> covid_clean$aged_70_older[covid_clean$aged_70_older %in% boxplot.stats(covid_clean$aged_70_older) $ out]
numeric(0)
> covid_clean$life_expectancy[covid_clean$life_expectancy %in% boxplot.stats(covid_clean$life_expectancy) $ out]
numeric(0)
```

Figure 6: Boxplot of cleaned data

The figure 6 shows the outliers of numerical variables after removing the NA's and unused variables from the raw data. The circle above the variable name shows the outliers. There are outliers present in variables such as total cases and total deaths. Some variables such as median age and life expectancy have no outliers.

Explain the methods used for data cleaning

```
> summary(covid_clean_final)
  continent           location              date             total_cases       new_cases       total_deaths      new_deaths
 Length:52823       Length:52823       Min.   :2020-03-01   Min.   :     1   Min.   :   0.0   Min.   :   1.0   Min.   :0.00
 Class :character   Class :character   1st Qu.:2020-08-23   1st Qu.:  2366   1st Qu.:   0.0   1st Qu.:  38.0   1st Qu.:0.00
 Mode  :character   Mode  :character   Median :2021-02-15   Median :  9049   Median :  19.0   Median : 149.0   Median :0.00
                                       Mean   :2021-02-21   Mean   : 35562   Mean   : 112.6   Mean   : 556.0   Mean   :1.06
                                       3rd Qu.:2021-08-17   3rd Qu.: 34281   3rd Qu.: 107.0   3rd Qu.: 615.5   3rd Qu.:1.00
                                       Max.   :2022-03-05   Max.   :864564   Max.   :1582.0   Max.   :4653.0   Max.   :7.00
   population        population_density   median_age     aged_65_older     aged_70_older     gdp_per_capita    life_expectancy
 Min.   :   98728   Min.   :  1.98      Min.   :15.10   Min.   : 2.339   Min.   : 1.285   Min.   :  661.2   Min.   :53.28
 1st Qu.: 1449891   1st Qu.: 21.19      1st Qu.:19.20   1st Qu.: 3.002   1st Qu.: 1.756   1st Qu.: 2014.3   1st Qu.:63.71
 Median : 5548361   Median : 51.76      Median :24.40   Median : 4.213   Median : 2.380   Median : 6222.6   Median :69.82
 Mean   :10363804   Mean   : 97.82      Mean   :26.45   Mean   : 6.617   Mean   : 4.167   Mean   :11358.6   Mean   :69.12
 3rd Qu.:16914985   3rd Qu.:124.03      3rd Qu.:32.10   3rd Qu.: 8.273   3rd Qu.: 5.200   3rd Qu.:16409.3   3rd Qu.:74.62
 Max.   :44909351   Max.   :494.87      Max.   :46.20   Max.   :21.502   Max.   :14.924   Max.   :49045.4   Max.   :83.44
```

Figure 7: The summary of the cleaned data

The summary and descriptive statistics of each variable in the cleaned dataset are shown in figure 7 above.

The first step was to filter the variables that we considered using in our visualization. After that, empty string was found inside the continent so we removed rows containing them. The next step was using na.omit function to remove rows containing NA. The last step was to remove numerical variables containing outliers. Initially, there were 166326 samples but after data cleaning, there are 52823 samples remaining.

- **Design statement**

The problem statement revolves around analyzing the amount of covid cases and deaths from the virus in every country or continent. The objective is to identify the main factors which affect the number of total covid cases and deaths in every country or continent and develop strategies to reduce the amount of covid cases and deaths in every country or continent effectively. In identifying the main factors which have an effect on the spread of the covid virus, every country or continent would be able to have an easier time in developing precautionary measures to reduce the spread of the virus and improve their economy to the time before the virus was discovered.

Total cases is one of the important factors to compare what contributes to fast covid spread between different countries. Total cases are calculated by adding today's new cases with yesterday's total cases. The spike of total cases means that there were many new covid cases. To determine what are the variables that contribute to fast covid spread, we can compare total cases with other variables that might have an effect on covid spread.

Population density is another important factor that might contribute to covid spread. Covid19 is known to transmit the virus through air and the spread should be faster if everyone is closer. By comparing population density with total cases, we might see if population density has an effect on covid spread through the visualization.

Instead of using countries, we could also use continents to see if different continents have higher covid cases compared to others. Since it is easier to travel within the same continent, the spread can be much higher if the covid cases are higher. The culture of different continents can also affect the spread of covid. For example, the majority of people in Asia started wearing masks after the covid breakout while only a minority of people in Europe wore them.

Population of each continent can be linked with the total cases of each continent to compare if higher populations have higher cases. If continents with low populations have unreasonably high covid cases, we can further investigate on that continent to see what effects the covid cases and deaths.

Upon finding out the main factors which have an effect on the spread of the covid virus in every country or continent by using data analysis such as population and population density. Every country or continent would have a better understanding on how to control the spread of the covid virus or even prevent the covid virus from causing another outbreak such as implementing an effective lockdown system, improving their health regulations and also improving their health infrastructure.

Reducing the amount of covid cases and deaths in every country or continent would be beneficial for the economy as the virus has caused complications for people to carry out their daily lives which negatively impacts the economy. Therefore, by implementing the suggested suggestions or recommendations, every country or continent would be able to overcome the virus and maintain the same productivity level as before the virus started resulting in the improvement of the economy.

Section B

```
> #Description of raw data
> summary(covid_medical)
    Entity           Continent            Latitude        Longitude       Average.temperature.per.year Hospital.beds.per.1000.people
 Length:38472       Length:38472       Min.   :-40.90   Min.   :-106.35   Min.   :-2.00                Min.   : 0.200
 Class :character   Class :character   1st Qu.:  8.62   1st Qu.:  -3.44   1st Qu.:11.00                1st Qu.: 1.400
 Mode  :character   Mode  :character   Median : 27.51   Median :  21.82   Median :20.00                Median : 2.500
                                       Mean   : 23.74   Mean   :  20.21   Mean   :17.72                Mean   : 3.165
                                       3rd Qu.: 45.94   3rd Qu.:  47.48   3rd Qu.:25.00                3rd Qu.: 4.490
                                       Max.   : 64.96   Max.   : 179.41   Max.   :29.00                Max.   :13.050

 Medical.doctors.per.1000.people   GDP.Capita         Population          Median.age    Population.aged.65.and.over....      Date
 Min.   :0.020                   Min.   :   411.6   Min.   :3.413e+05   Min.   :16.00   Min.   : 1.00                   Length:38472
 1st Qu.:0.820                   1st Qu.:  3659.0   1st Qu.:4.794e+06   1st Qu.:27.00   1st Qu.: 5.00                   Class :character
 Median :1.890                   Median :  8821.8   Median :1.148e+07   Median :32.00   Median : 8.00                   Mode  :character
 Mean   :2.086                   Mean   : 19002.3   Mean   :4.897e+07   Mean   :32.75   Mean   :10.66
 3rd Qu.:3.210                   3rd Qu.: 25946.2   3rd Qu.:4.286e+07   3rd Qu.:41.00   3rd Qu.:16.00
 Max.   :7.520                   Max.   :114704.6   Max.   :1.339e+09   Max.   :48.00   Max.   :28.00

  Daily.tests          Cases              Deaths
 Min.   :-239172   Min.   :      1   Min.   :    1
 1st Qu.:   1505   1st Qu.:   2074   1st Qu.:   77
 Median :   5520   Median :  21431   Median :  527
 Mean   :  39441   Mean   : 287903   Mean   : 8090
 3rd Qu.:  20382   3rd Qu.: 137377   3rd Qu.: 3480
 Max.   :2945871   Max.   :28605669   Max.   :513091
 NA's   :7895      NA's   :254        NA's   :3610
```

Figure 8: The raw covid medical data

Figure 8 shows the summary of the raw covid medical data. The data contains variables such as entity(country name), continent, latitude…etc. Summary function also shows the summary statistics of the numerical variables (like minimum, maximum average). There is NA's present in daily tests(7895 NA's), cases(254), and deaths(3610).

```
> str(covid_medical)
'data.frame':   38472 obs. of  15 variables:
 $ Entity                          : chr  "Albania" "Albania" "Albania" "Albania" ...
 $ Continent                       : chr  "Europe" "Europe" "Europe" "Europe" ...
 $ Latitude                        : num  41.1 41.1 41.1 41.1 41.1 ...
 $ Longitude                       : num  20.2 20.2 20.2 20.2 20.2 ...
 $ Average.temperature.per.year    : int  14 14 14 14 14 14 14 14 14 14 ...
 $ Hospital.beds.per.1000.people   : num  2.89 2.89 2.89 2.89 2.89 2.89 2.89 2.89 2.89 2.89 ...
 $ Medical.doctors.per.1000.people: num  1.29 1.29 1.29 1.29 1.29 1.29 1.29 1.29 1.29 1.29 ...
 $ GDP.Capita                      : num  5353 5353 5353 5353 5353 ...
 $ Population                      : int  2873457 2873457 2873457 2873457 2873457 2873457 2873457 2873457 2873457 2873457 ...
 $ Median.age                      : int  38 38 38 38 38 38 38 38 38 38 ...
 $ Population.aged.65.and.over....: int  14 14 14 14 14 14 14 14 14 14 ...
 $ Date                            : chr  "2020-02-25" "2020-02-26" "2020-02-27" "2020-02-28" ...
 $ Daily.tests                     : num  8 5 4 1 8 3 2 5 6 8 ...
 $ Cases                           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Deaths                          : num  NA NA NA NA NA NA NA NA NA NA ...
```

Figure 9: The structure of covid medical data

Figure 9 shows the structure of covid medical data. The data has 38472 observations and 15 variables. The variable type for entity, continent, and date are in characters. Variable types for latitude, hospital beds per 1000 people, and GDP capita are in number. Variable types for population, median age… is in integers.

```
#NOTE: We only want to use average temp, hospital beds, and medical doctors to merge with out first data.
#NOTE: So we filter them and remove duplicate rows.
#Filter variable not used
covid_medical_filter = covid_medical[,c("Entity", "Continent", "Average.temperature.per.year",
                                        "Hospital.beds.per.1000.people", "Medical.doctors.per.1000.people")]

#Remove duplicate rows
covid_medical_final <- covid_medical_filter[!duplicated(covid_medical_filter),]
```

Figure 10: Data process for covid medical data

Since we only need to use average temperature per year, hospital beds per 1000 people, and medical doctors per 1000 people for visualization, we filtered out other variables not used. We kept the Entity(country) and continent because these must be used to join with the first data.

Data Joining

There are 3 different join functions used to perform data joining between the covid_clean_final dataset and covid_medical_final data set which are inner join, anti join and full join shown in the figures below.

```
> #Data Joining
> #Inner join
> covid_inner_join = covid_clean_final %>%
+   inner_join(covid_medical_final)
Joining with `by = join_by(location)`
> summary(covid_inner_join)
  continent           location             date           total_cases       new_cases      total_deaths      new_deaths      population
 Length:22497       Length:22497       Length:22497       Min.   :     1   Min.   :   0   Min.   :   1.0   Min.   :0.000   Min.   :  368792
 Class :character   Class :character   Class :character   1st Qu.:  2507   1st Qu.:   8   1st Qu.:  31.0   1st Qu.:0.000   1st Qu.: 2689862
 Mode  :character   Mode  :character   Mode  :character   Median : 15613   Median :  51   Median : 246.0   Median :0.000   Median : 5813302
                                                          Mean   : 55792   Mean   : 176   Mean   : 814.2   Mean   :1.471   Mean   :10757205
                                                          3rd Qu.: 70697   3rd Qu.: 208   3rd Qu.:1075.0   3rd Qu.:2.000   3rd Qu.:17196308
                                                          Max.   :875440   Max.   :1602   Max.   :4691.0   Max.   :7.000   Max.   :44616626
 population_density   median_age    aged_65_older    aged_70_older    gdp_per_capita  life_expectancy  Continent
 Min.   :  1.98     Min.   :17.7   Min.   : 2.355   Min.   : 1.525   Min.   : 1095   Min.   :60.85   Length:22497
 1st Qu.: 19.75     1st Qu.:21.1   1st Qu.: 3.158   1st Qu.: 1.882   1st Qu.: 4228   1st Qu.:67.44   Class :character
 Median : 74.23     Median :30.6   Median : 6.991   Median : 4.625   Median :11803   Median :74.47   Mode  :character
 Mean   :101.52     Mean   :30.7   Mean   : 9.425   Mean   : 6.071   Mean   :17415   Mean   :73.43
 3rd Qu.:135.58     3rd Qu.:37.9   3rd Qu.:14.799   3rd Qu.: 9.720   3rd Qu.:29481   3rd Qu.:78.57
 Max.   :494.87     Max.   :46.2   Max.   :21.502   Max.   :14.924   Max.   :49045   Max.   :83.44
 Average.temperature.per.year Hospital.beds.per.1000.people Medical.doctors.per.1000.people
 Min.   :-1.00                Min.   : 0.200                Min.   :0.020
 1st Qu.:11.00                1st Qu.: 1.300                1st Qu.:0.140
 Median :21.00                Median : 2.300                Median :1.490
 Mean   :18.69                Mean   : 2.667                Mean   :1.734
 3rd Qu.:26.00                3rd Qu.: 3.400                3rd Qu.:3.200
 Max.   :29.00                Max.   :11.000                Max.   :6.260
```

Figure 11 : Inner join between covid_clean_final and covid_medical_final datasets

The inner join function is used to match observations in covid_clean_final and covid_medical_final datasets using the key variable location. Therefore, observations in the location variable of covid_clean_final that have a match with observations in the location variable of covid_medical_final would be joined together as shown in figure 11 above.

```
> #Anti join
> covid_anti_join = anti_join(covid_clean_final, covid_medical_final)
Joining with `by = join_by(location)`
> summary(covid_anti_join)
  continent           location            date           total_cases       new_cases        total_deaths      new_deaths
 Length:30378       Length:30378       Length:30378       Min.   :     1   Min.   :   0.00   Min.   :   1.0   Min.   :0.0000
 Class :character   Class :character   Class :character   1st Qu.:  2317   1st Qu.:   0.00   1st Qu.:  43.0   1st Qu.:0.0000
 Mode  :character   Mode  :character   Mode  :character   Median :  7248   Median :   6.00   Median : 125.0   Median :0.0000
                                                          Mean   : 20838   Mean   :  66.14   Mean   : 370.6   Mean   :0.7578
                                                          3rd Qu.: 19790   3rd Qu.:  53.00   3rd Qu.: 347.0   3rd Qu.:1.0000
                                                          Max.   :384668   Max.   :1595.00   Max.   :4686.0   Max.   :7.0000
   population       population_density   median_age    aged_65_older    aged_70_older    gdp_per_capita    life_expectancy
 Min.   :   98728   Min.   :  3.612    Min.   :15.10   Min.   : 2.339   Min.   : 1.285   Min.   :  661.2   Min.   :53.28
 1st Qu.: 1002197   1st Qu.: 22.662    1st Qu.:18.80   1st Qu.: 2.922   1st Qu.: 1.583   1st Qu.: 1703.1   1st Qu.:61.58
 Median : 5180208   Median : 51.667    Median :21.50   Median : 3.556   Median : 2.155   Median : 3393.5   Median :65.31
 Mean   :10073627   Mean   : 95.023    Mean   :23.31   Mean   : 4.538   Mean   : 2.757   Mean   : 6870.4   Mean   :65.93
 3rd Qu.:13497237   3rd Qu.: 99.110    3rd Qu.:26.30   3rd Qu.: 4.800   3rd Qu.: 2.954   3rd Qu.: 7824.4   3rd Qu.:71.68
 Max.   :44909351   Max.   :437.352    Max.   :43.30   Max.   :19.027   Max.   :11.580   Max.   :32605.9   Max.   :79.38
```

Figure 12 : Anti join between covid_clean_final and covid_medical_final datasets

The anti join function is used to diagnose join mismatches between datasets covid_clean_final and covid_medical_final. Therefore, observations which do not match in the location variable of covid_clean_final with observations in the location variable of covid_medical_final are shown in figure 12 above.

```
> #Full join
> covid_full_join = full_join(covid_clean_final, covid_medical_final)
Joining with `by = join_by(location)`
> summary(covid_full_join)
  continent           location            date           total_cases       new_cases        total_deaths      new_deaths
 Length:52911       Length:52911       Length:52911       Min.   :     1   Min.   :    0.0   Min.   :    1.0   Min.   :0.000
 Class :character   Class :character   Class :character   1st Qu.:  2369   1st Qu.:    0.0   1st Qu.:   38.0   1st Qu.:0.000
 Mode  :character   Mode  :character   Mode  :character   Median :  9066   Median :   19.0   Median :  149.0   Median :0.000
                                                          Mean   : 35710   Mean   :  112.9   Mean   :  559.3   Mean   :1.061
                                                          3rd Qu.: 34404   3rd Qu.:  107.0   3rd Qu.:  617.0   3rd Qu.:1.000
                                                          Max.   :875440   Max.   :1602.0    Max.   :4691.0    Max.   :7.000
                                                          NA's   :36       NA's   :36        NA's   :36        NA's   :36
   population        population_density   median_age    aged_65_older    aged_70_older    gdp_per_capita    life_expectancy
 Min.   :    98728   Min.   :  1.98     Min.   :15.10   Min.   : 2.339   Min.   : 1.285   Min.   :  661.2   Min.   :53.28
 1st Qu.:  1449891   1st Qu.: 21.19     1st Qu.:19.20   1st Qu.: 3.002   1st Qu.: 1.756   1st Qu.: 2014.3   1st Qu.:63.71
 Median :  5548361   Median : 51.76     Median :24.40   Median : 4.213   Median : 2.380   Median : 6222.6   Median :69.82
 Mean   : 10364472   Mean   : 97.79     Mean   :26.45   Mean   : 6.617   Mean   : 4.167   Mean   :11356.9   Mean   :69.12
 3rd Qu.: 16914985   3rd Qu.:124.03     3rd Qu.:32.10   3rd Qu.: 8.273   3rd Qu.: 5.200   3rd Qu.:16409.3   3rd Qu.:74.62
 Max.   :44909351   Max.   :494.87     Max.   :46.20   Max.   :21.502   Max.   :14.924   Max.   :49045.4   Max.   :83.44
 NA's   :36         NA's   :36         NA's   :36      NA's   :36       NA's   :36       NA's   :36        NA's   :36
   Continent         Average.temperature.per.year Hospital.beds.per.1000.people Medical.doctors.per.1000.people
 Length:52911       Min.   :-2.00                Min.   : 0.200                 Min.   :0.020
 Class :character   1st Qu.:11.00                1st Qu.: 1.300                 1st Qu.:0.140
 Mode  :character   Median :21.00                Median : 2.300                 Median :1.490
                    Mean   :18.69                Mean   : 2.668                 Mean   :1.734
                    3rd Qu.:26.00                3rd Qu.: 3.400                 3rd Qu.:3.200
                    Max.   :29.00                Max.   :13.050                 Max.   :7.520
                    NA's   :30378                NA's   :30378                  NA's   :30378
```

Figure 13 : Full join between covid_clean_final and covid_medical_final datasets

Full join function is used to join all the observations in covid_clean_final and covid_medical_final regardless of whether there is a match or mismatch. Therefore, the observations in both covid_clean_final and covid_medical_final would be joined together forming a new dataset called covid_full_join as shown in figure 13 above.

Section C

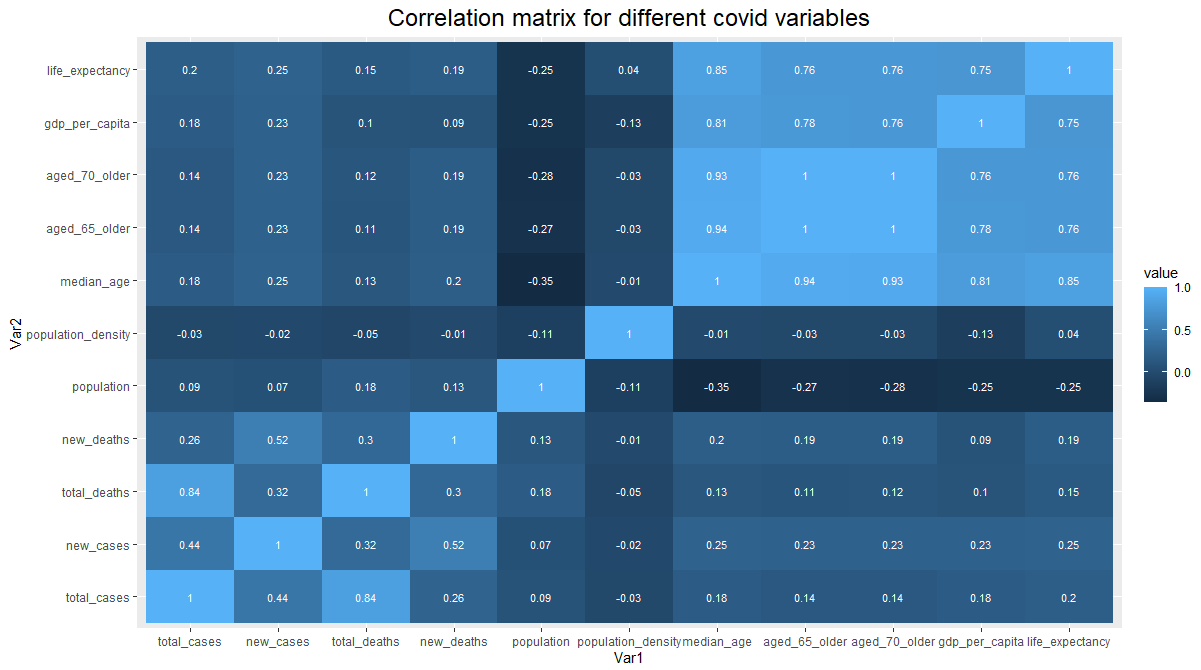


Figure 14: Correlation matrix for different covid variables

The correlation matrix in figure 14 shows the correlation between different covid variables. The first dataset was used in this correlation matrix after data cleaning and removing outliers. The correlation of positive 1 means it has the highest positive correlation while negative 1 means it has the highest negative correlation. As the colour moves from dark blue to light blue, the correlation will lean towards positive 1. Moving on to the variables, there is a strong positive correlation between life expectancy and gdp per capita, meaning that better developed countries live longer. There is a weak positive correlation between total cases and life expectancy or gdp per capita meaning that more developed countries have a higher total cases, but this could be due to them having more populations. There is no correlation between total cases and population density and this could be due to the fact that populations are different between countries. There is also the possibility that population density has little to no effect on total cases.

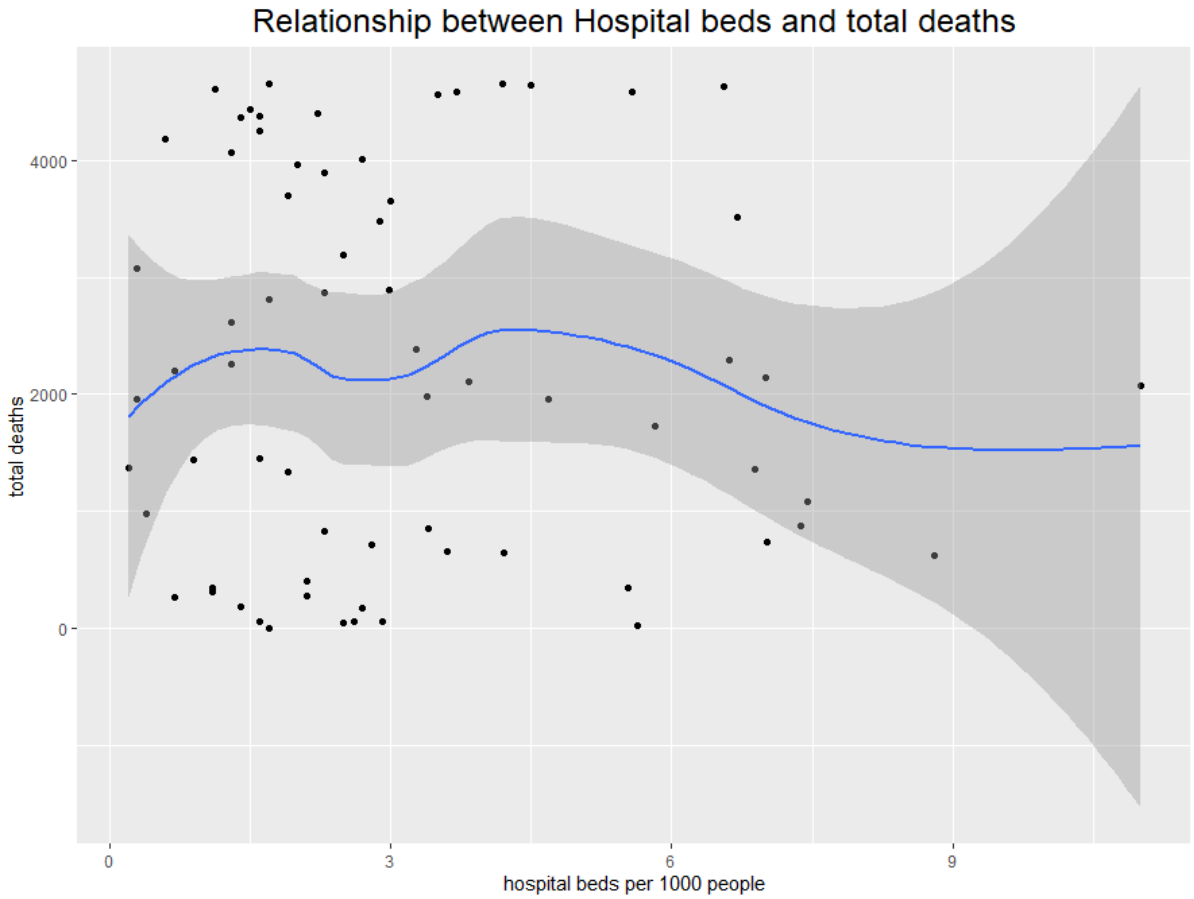


Figure 15 : Scatterplot of total deaths against hospital beds per 1000 people for every country

This scatter plot shows the relationship between total deaths and hospital beds per 1000 people of every country where each country would represent a dot in the scatter plot. The data set used to produce this scatter plot was made by performing an inner join function between covid_cases_final and covid_medical_final datasets. A smooth line was added into the scatter plot to show a better visualisation of the distribution of data. It was expected to show a negative correlation between hospital beds per 1000 people and total deaths. However, it can be seen that there is no correlation between the two variables. Thus, it can be concluded that the amount of hospital beds per 1000 people in every country would have little to no effect on the total deaths of each country from the Covid-19 virus.

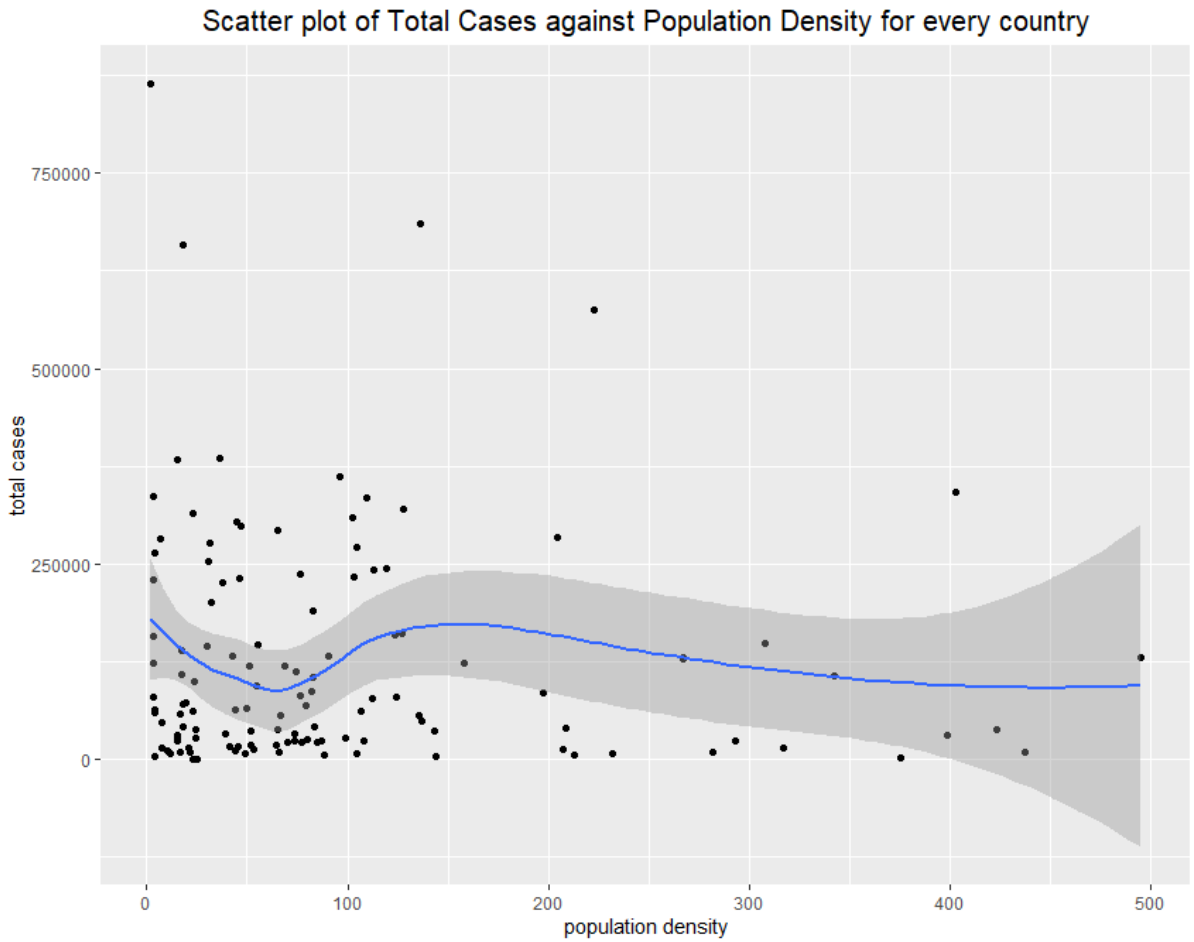


Figure 16 : Scatter plot of total cases against population density for every country

The scatter plot in figure 16 shows the latest total number of covid cases against population density for each country. The first dataset was used for this plot after data cleaning and removing outliers. A smooth line was added into the scatter plot to show a better visualisation of the distribution of data. There seems to be slight negative correlation when the population density is between 0 to 60 but it changes to positive correlation after that. We expected the total cases to increase when population density increases because covid transmits through air and it is easier to transmit covid when they are closer to each other. However, it can be seen that there is no correlation between them.

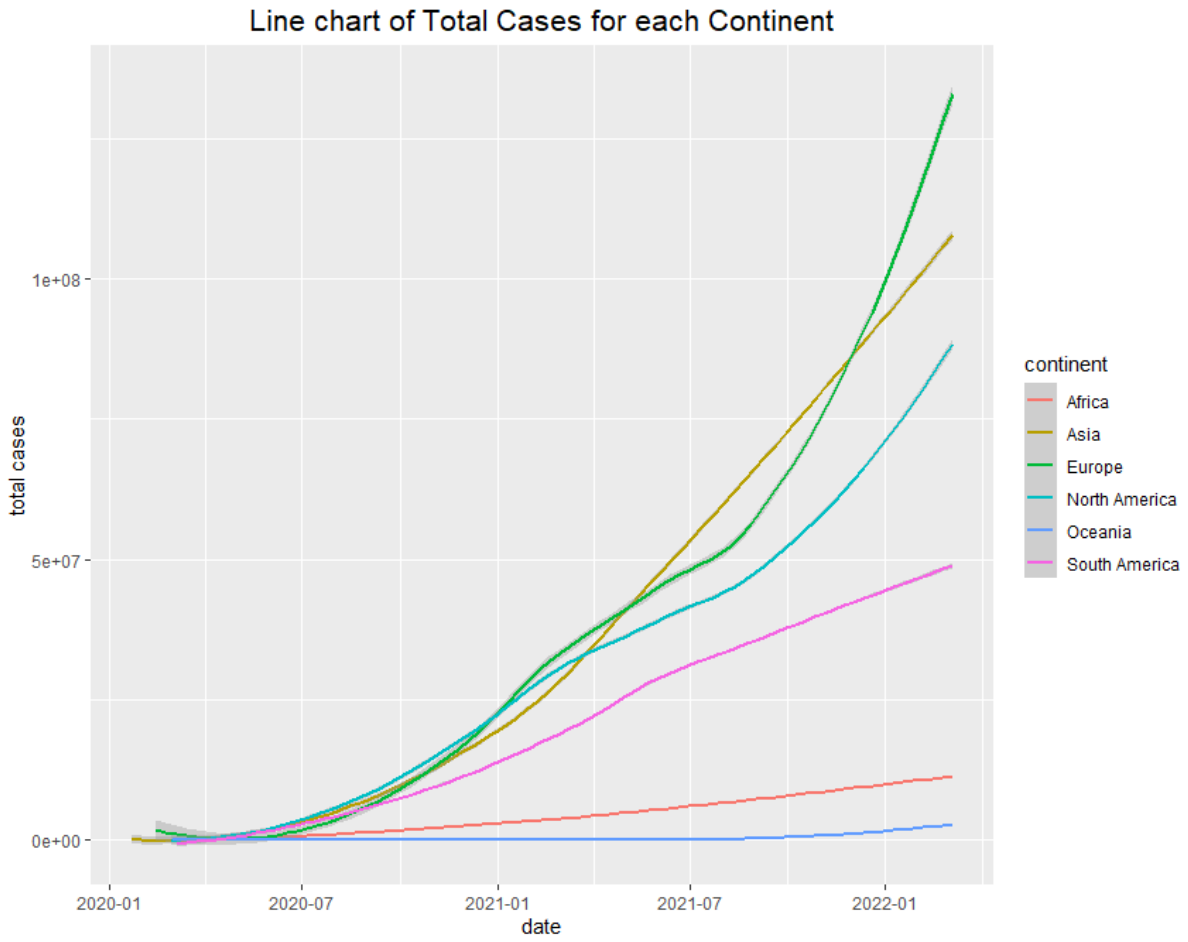


Figure 17 : Line Chart of total cases in each continent over a period of time

The line chart in figure 17 shows the total number of covid cases in each continent over a period of time which start from the year 2020 to year 2022. The dataset used for this line chart was cleaned by removing the missing values and empty strings but the outliers were not removed as removing the outliers from the dataset would have shown an incorrect distribution of the line chart. From this line chart we can see which continents would have the highest and lowest total cases. Starting from the continent from the highest total cases would be Europe, followed by Asia, North America, South America, Africa and lastly Ocenia. The number of total cases in each continent could have been affected by a few possible factors such as the population where a higher number of population would suggest that the virus might spread more easily, average temperature where colder continents might lower the amount of total cases as the virus is less active than in warmer continents and GDP per capita where a continent with a higher GDP per capita would generally have better health infrastructure and assess to healthcare resulting in the lower number of total cases.

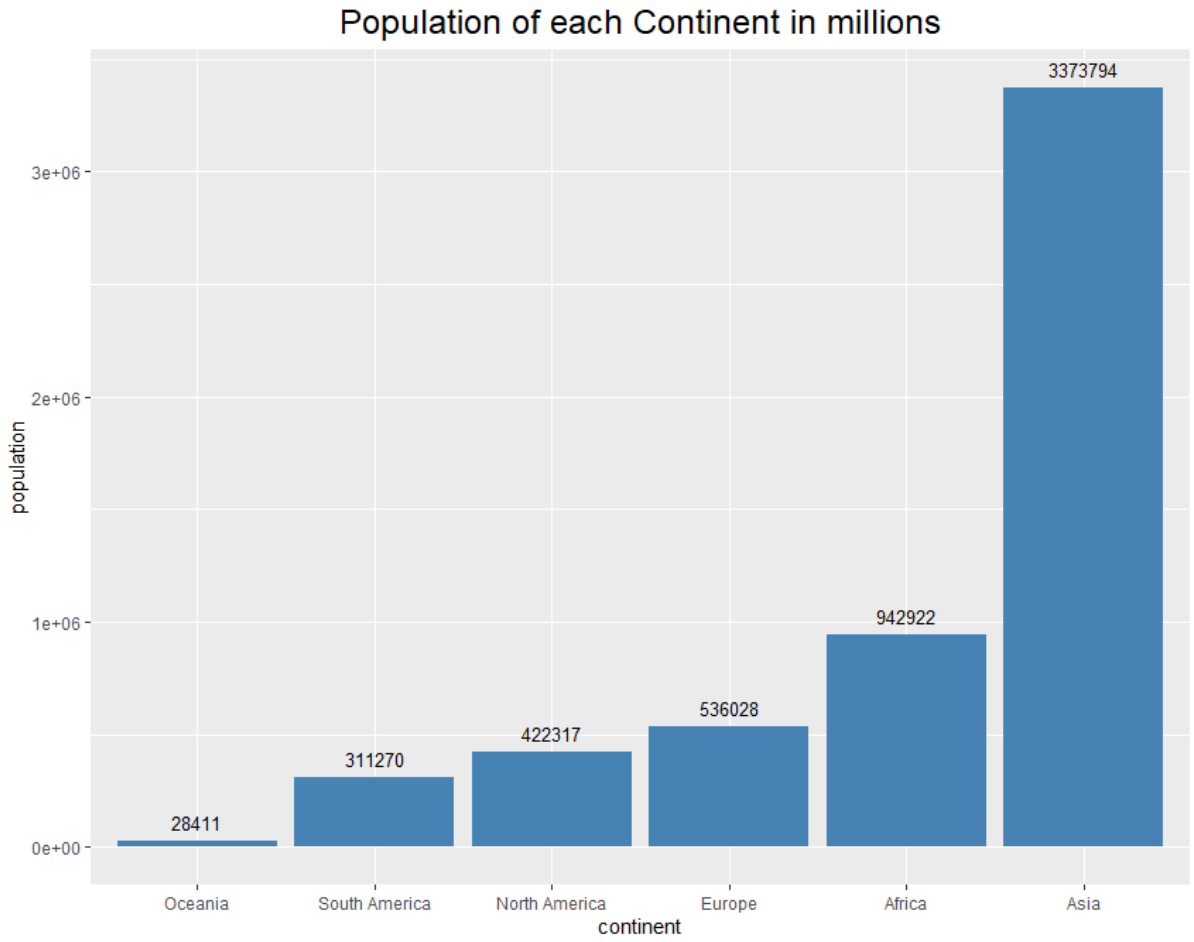


Figure 18 : Barchart of population of each continent in millions

The barchart in figure 18 shows the population of each continent in millions and the bar chart was reorganised to show the continent with the lowest population starting on the left side of the line chart to the continent with the highest population moving along the right side of the bar chart where is sequence would start with Oceania with a population of 28,411 million, followed by South America with a population of 311,270 million, North America with a population of 422,317, Europe with a population of 536,028 million, Africa with a population of 942,922 million and lastly Asia with a population of 3,373,794 million which is more than three times of the continent with the second highest population. It is expected that continents with a higher population would have a higher number of total cases however based on figure 17 shows that Asia which has the highest population did not have the highest number of total cases but the second highest. Furthermore, Europe, which is the continent with the third highest population, had the highest number of total cases. This could possibly be due to other factors such as health regulation process of each country where Asia have stricter health regulations such as having to wear masks out in public and having longer duration of lockdown whereas Europe may have less strict health regulations allowing the Europians to wear masks as they please and have short duration of lockdowns.

Conclusion

After analyzing the data using multiple variables and plot, we found that factors like number of hospital beds and population density has little effect on covid cases and deaths which is

not expected. When comparing total cases of each continent and their population, we found out that some continents with lower populations have higher total cases. This signifies that there might be other factors that could contribute to speed of covid spread. Investigation on each continent can be done in the future to identify the cause of covid spread.

Resources

First dataset:
https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset
Second dataset:
https://www.kaggle.com/datasets/sambelkacem/covid19-algeria-and-world-dataset