UNIVERSITY OF CANBERRA
AUSTRALIA'S CAPITAL UNIVERSITY

University of Canberra

Faculty of Science and

Technology FINAL ASSESMENT

SEMESTER 2, 2021

UNIT NAME: Introduction to Data Science

UNIT NUMBER: 11372

Due date: 7th of November, 23:59

EXAMINER'S NAME: Dr. Ibrahim Radwan

CONTACT details: ibrahim.radwan@canberra.edu.au

INSTRUCTIONS FOR STUDENTS

1. **This is a take-home assignment.**
2. **The assignment is organised into two parts, where the first part is composed of four general questions to assess your understanding of the Data Science principles. This part is expected to be delivered as a PDF file. The other part is code-based tasks, which you will need to submit R-code script and a final report. The final report should contain the output of running your code for those questions that are asking you to generate summaries or graphs from the data.**
3. **The PDF files and R script are expected to be submitted as a one compressed file (e.g. *.ZIP) on the Canvas website of the unit by the due date of the assignment.**
4. **The submitted file should be renamed as [studentID_lastname_final_assessment.zip], where "studentID" is your university ID and "lastname" is your lastname.**
5. **The assignment will be open from Tuesday, the 19th of October (17:30) until Sunday, the 7th of November (23:59).**
6. **This assignment has 100 marks in total and weighs 40% of the final grade of the unit, where 50% of the assignment marks is compulsory to pass the unit.**
7. **The assignment will cover all the learning outcomes and the taught contents of the unit.**

There are no errors deliberately placed in any of the questions in this assignment, unless explicitly stated. If you think you have identified any errors, please contact the supervisor

Task 4: Documentation and Reporting: (10 marks)

You are required to build a report (e.g. using MS Word) for the results of Task 2 and Task 3. The report is basically composed of the answers to those questions that asked you to generate or print summaries or asked you to build graphs for some variables. Then you need to export this MS word file into a PDF file to be submitted.

Answer:

Task 2:

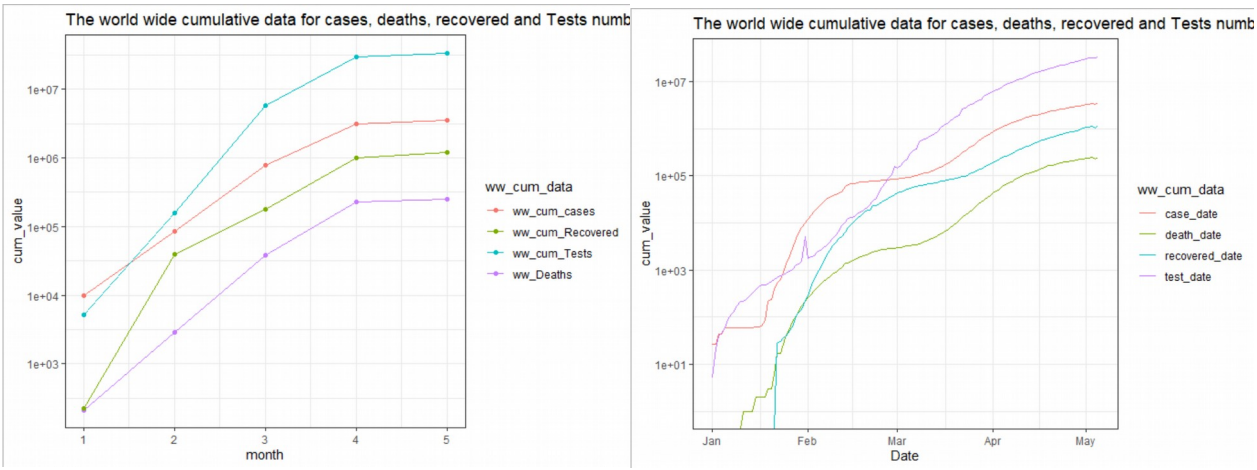


Figure 1: The World wide values for cumulative data of cases, deaths, recovered and tests (For month and Date)

Interpretation: As is illustrated in the above figure, it is obvious that all the variables have a going trend, which can be considered as a tremendous increase. Significantly, there is a skyrocket trend in the number of recovered cases in the first month. This is also recorded as a significant sign in the recovered number as in Date table. Turning now to the tests figure in Date table, it is recorded that there is a stand-out increase within the range from January to February. Overall, all the variables trend remain increasing steadily, and become more stable from April to May. This is regarded as a positive result since the first vaccine trials had been conducted at this stage (Callaway, 2020).

Date	max_deaths_toll_per_day
<date>	*<dbl>*
2020-04-16	10520

Figure 2: Recorded Date for the highest death-toll

As shown in this figure, the above interpretation can be supported since the highest death number was witnessed in the middle of April. Afterward, this number has been kept more stable onwards.

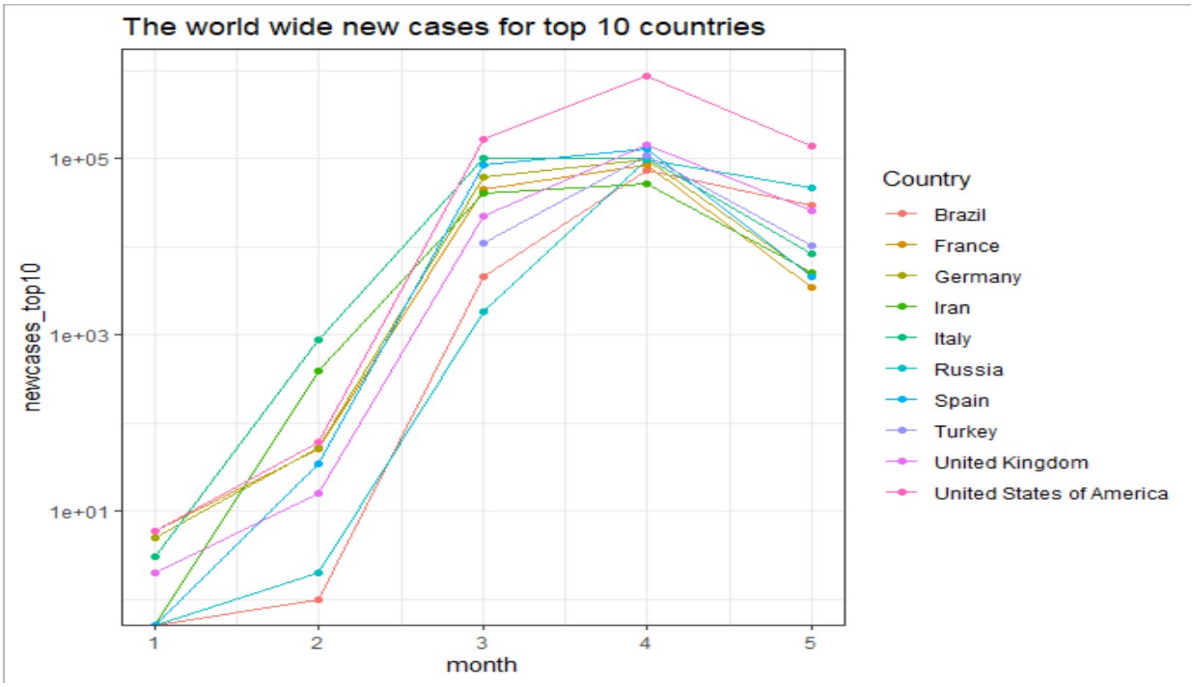
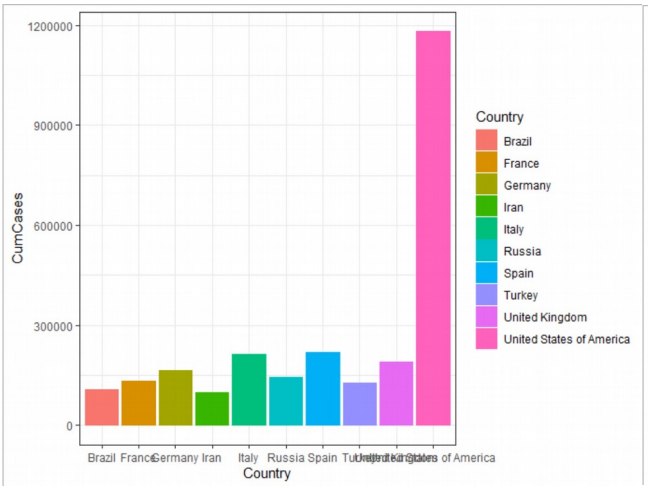
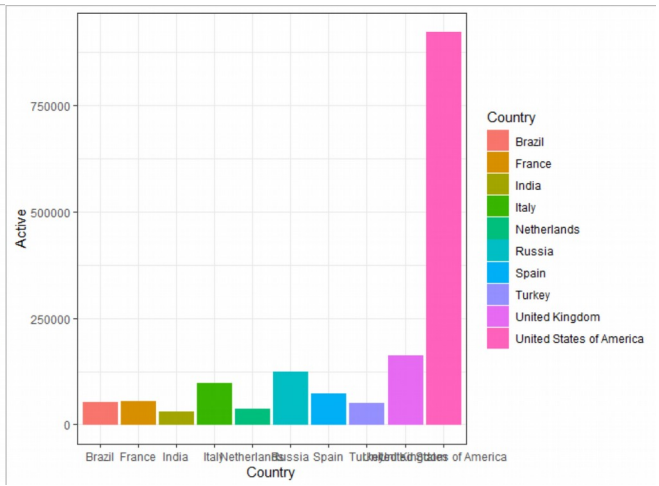


Figure 3: The total number of new cases within 5 months for top 10 cases countries

Interpretation: Overall, these 10 countries all have an increasing trend and the number of new cases went down within April. Specifically, USA was the dominant country in the outburst of new cases throughout the witnessed time frame. To have a better view about the efficiency of these countries in covid19 pandemic, it is indispensable to analyse their cumulative cases, active cases, fatality rate and tests as shown in below figures.





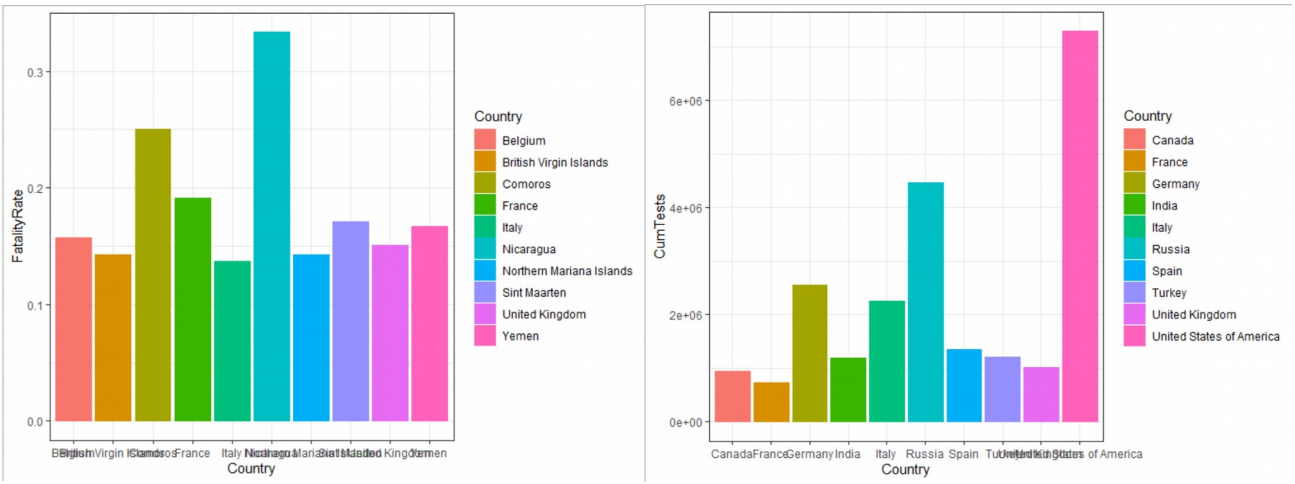


Figure 4: Top 10 countries for different categories (Cumulative cases, Cumulative active cases, Fatality rate, and Cumulative tests)

Interpretation: The above statistics can illustrate how well a country took actions toward the fight with Covid19. USA had the highest cumulative cases, the tests and active cases in May, but the Fatality rate was much lower and did not appear in top 10 for Fatality Rate table. Therefore, it was a sign of improvement in the covid19 management in USA as the Fatality was recorded as low compared to the European countries and United Kingdom.

Task 3:

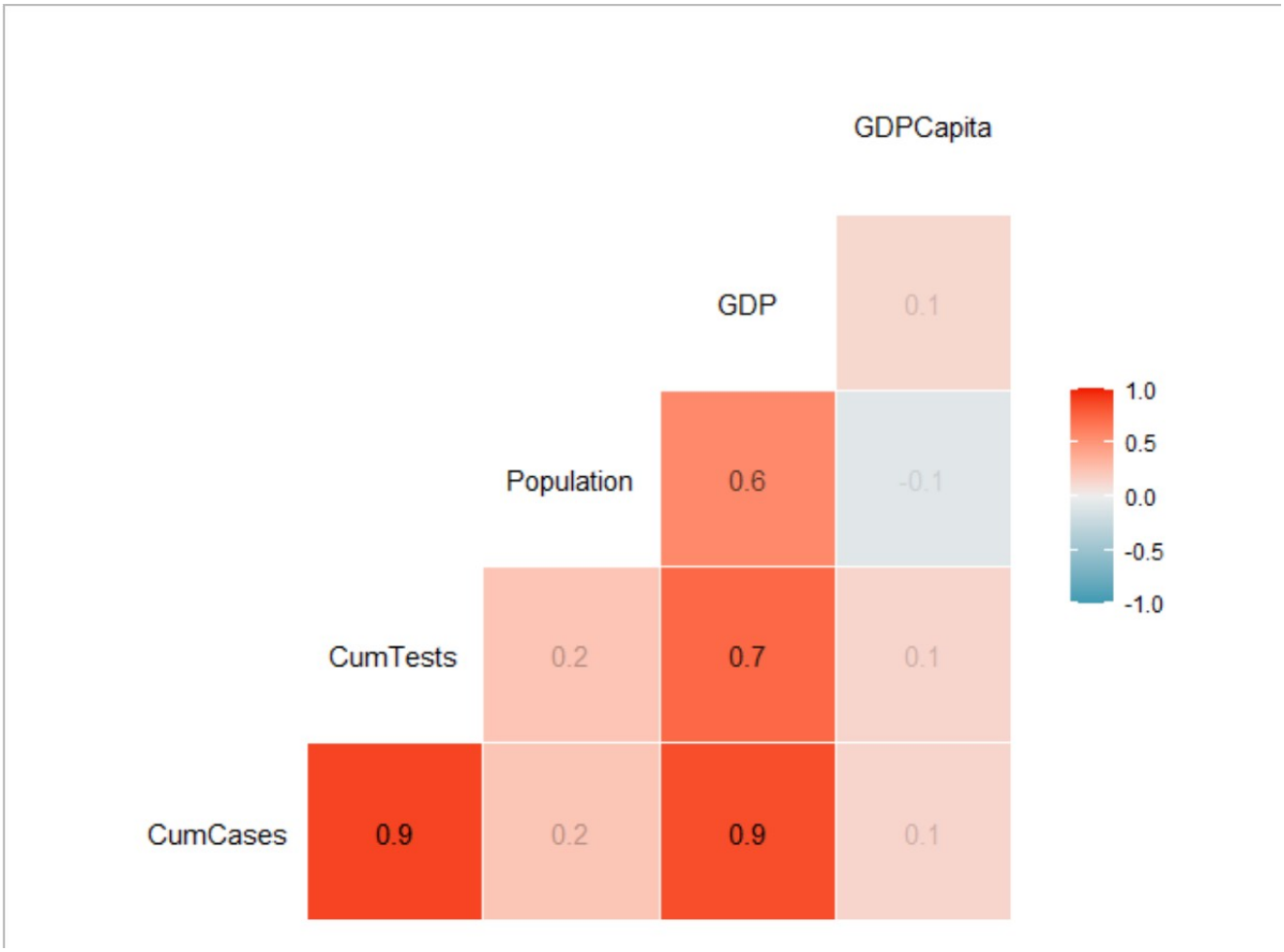


Figure 5: Correlation matrix between variables CumCases, CumTests, Population and GDP.

Interpretation: There is a high correlation between Cumulative cases and GDP at 0.9. Meanwhile, the other correlation statistics show that there are low correlation between Cumcases and GDP, GDP Capita. \

Figure 6: Statistics for single model with GDP

```
Call:
lm(formula = CumCases ~ GDP, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-263165    -315    1294    1590  144053

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.559e+03  2.827e+03  -0.551    0.582
GDP          5.746e-02  1.576e-03  36.447   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32330 on 135 degrees of freedom
Multiple R-squared:  0.9077,    Adjusted R-squared:  0.9071
F-statistic:  1328 on 1 and 135 DF,  p-value: < 2.2e-16
```

Interpretation:
There are multiple significant statistics in the figure which should be discussed as below:

- R-squared: 0.9077; Adjusted R-squared: 0.9071

These numbers illustrate that there is estimated that 90% of the variation of CumCases can be interpreted by the variable GDP.

- P-value: <2.2e-16: as the typical threshold is 0.05, this value is much lower than the standard p-value. Therefore, it can indicate a statistical significance in the single model where the variable GDP shows a significant difference in the response variable.

Figure 7: Statistics for multi model with GDP, Population, GDPCapita

```
lm(formula = CumCases ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-183526     102    1400    2832  104289

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.707e+02  2.706e+03  -0.248    0.805
CumTests     4.822e-02  5.258e-03   9.169 8.22e-16 ***
Population  -9.689e-05  1.842e-05  -5.259 5.68e-07 ***
GDP          4.199e-02  2.328e-03  18.042  < 2e-16 ***
GDPCapita   -5.008e-02  9.029e-02  -0.555    0.580
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24410 on 132 degrees of freedom
Multiple R-squared:  0.9486,    Adjusted R-squared:  0.947
F-statistic: 608.9 on 4 and 132 DF,  p-value: < 2.2e-16
```

Interpretation:

There are multiple significant statistics in the figure which should be discussed as below:

- R-squared: 0.9486; Adjusted R-squared: 0.947

These numbers illustrate that there is estimated that 94% of the variation of CumCases can be interpreted by multiple variables.

- P-value: <2.2e-16: as the typical threshold is 0.05, this value is much lower than the standard p-value. Therefore, it can indicate a statistical significance in the single model where the variable GDP shows a significant difference in the response variable.

RMSE

- Single-plot: **75431.51**
- Multi-plot: **39875.78**

The smaller the RMSE, the smaller the residual values (the difference between the predicted and actual results). As for this case, the single-plot is estimated to have a larger value compared to the multi-plot.

Residual vs Fitted

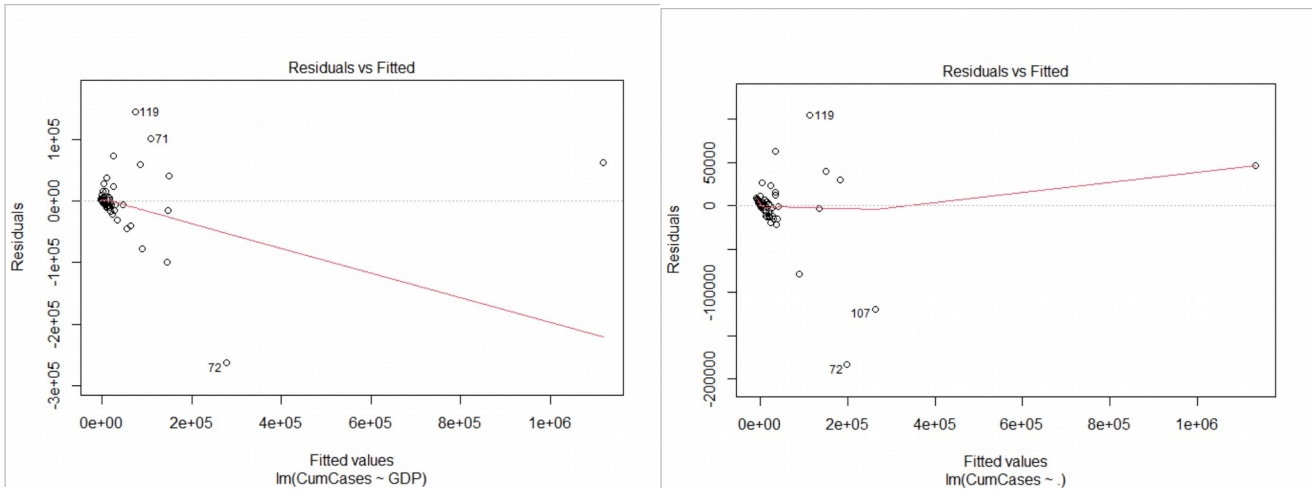


Figure 8: The linearity of the models and the relationship residuals – fitted values of two models

Interpretation: Both of the models show no linearity relationship since both of the red lines are far from the dashed line. All the values seem to gather at the first quarter and spread out lately.

Assumption: As is illustrated in figure 9, all the data variables have tremendous skewness which ends up in many outliers and affecting the prediction process. Therefore, there will be an indispensable requirement that the data has to be cleaned up carefully before applying any models.

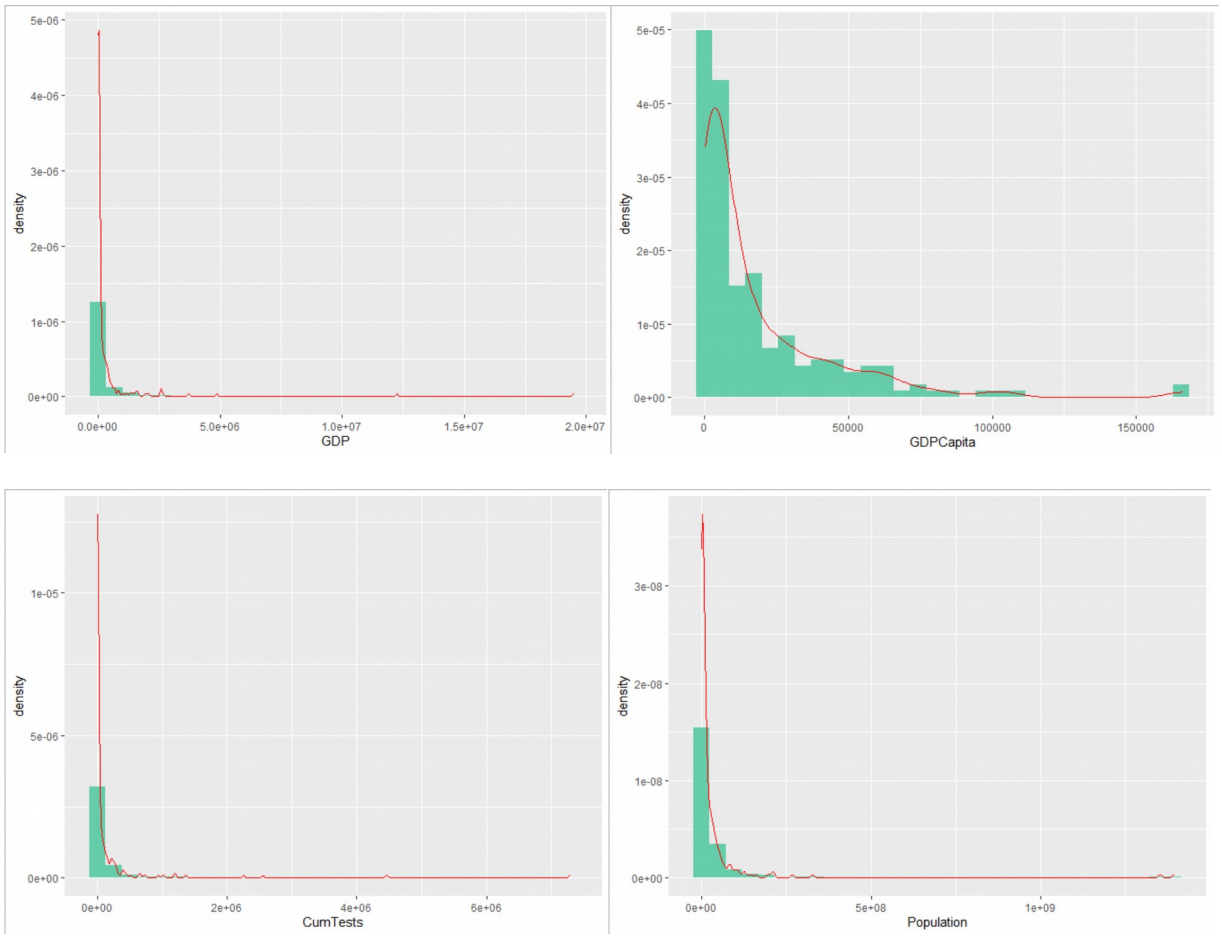


Figure 9: Distribution of multiple variables in cor_data

References:

Callaway, E. (2020). Coronavirus vaccine trials have delivered their first results — but their promise is still unclear. Nature, [online] 581, pp.363–364. Available at: https://www.nature.com/articles/d41586-020-01092-3.