FIT3152 Data Analytics
Semester 1 2023
Assignment 1

Name: Tatiana Sutulova

Student ID: 30806151

Part 1: Descriptive analysis and pre-processing

1.1 Descriptive Analysis

During the early stages of COVID-19 outbreak, researchers surveyed participants around the globe, aiming to identify the most important predictors of pro-social COVID-19 behaviors. In the following research, my focus is to analyze the country-level differences in predictors of pro-social behaviors, reported by participants and summarized into a dataset named “PsyCoronaBaselineExtract”.

The analysis starts with describing the dimensions of the dataset, which is the number of columns that represent the attributes or features that are collected for each observation. Each column in a dataset contains a different type of information, such as numerical data or text data. In our case, the total number of columns is 54, however the coded_country column is not considered to be a dimension, since it is the main identifier for which all the survey data is collected, thus the total number of dimensions is 53 and all of this data is stored as numerical data type (int), whereas the coded_country is of a text data type (char).

During the descriptive analysis, the distribution of numerical values was examined and is depicted in Figure 1 and Figure 2. It is important to note that the attributes of employstatus_1 - employstatus_10 were not included in the analysis because they only contain values of 1 and NA, as participants were required to choose one of the 10 fields related to their employment status. The data distribution of numerical values ranging from -3 to 10 is shown in the figures. The scaling of several question types mostly ranges from 1 to 5, such as Affect, 1 to 6, such as Corona Community injunctive norms, -2 to 2, such as Societal Disconnect, Job insecurity, Perceived Financial Strain, Disempowerment, and -3 to 1, such as Corona ProSocial Behaviour. Additionally, outliers are represented as circles in Figure 1 and Figure 2, which aids in detecting any unusual or extreme values that may impact further analysis.

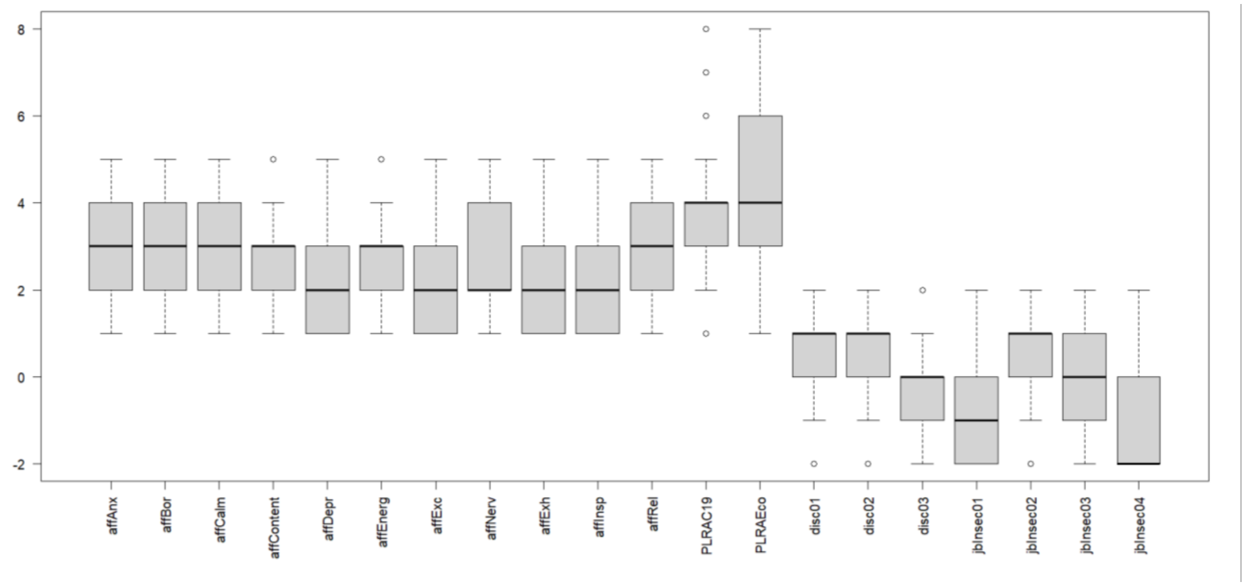


Figure 1. Distribution of numerical values of columns affAnx - jbInsec04

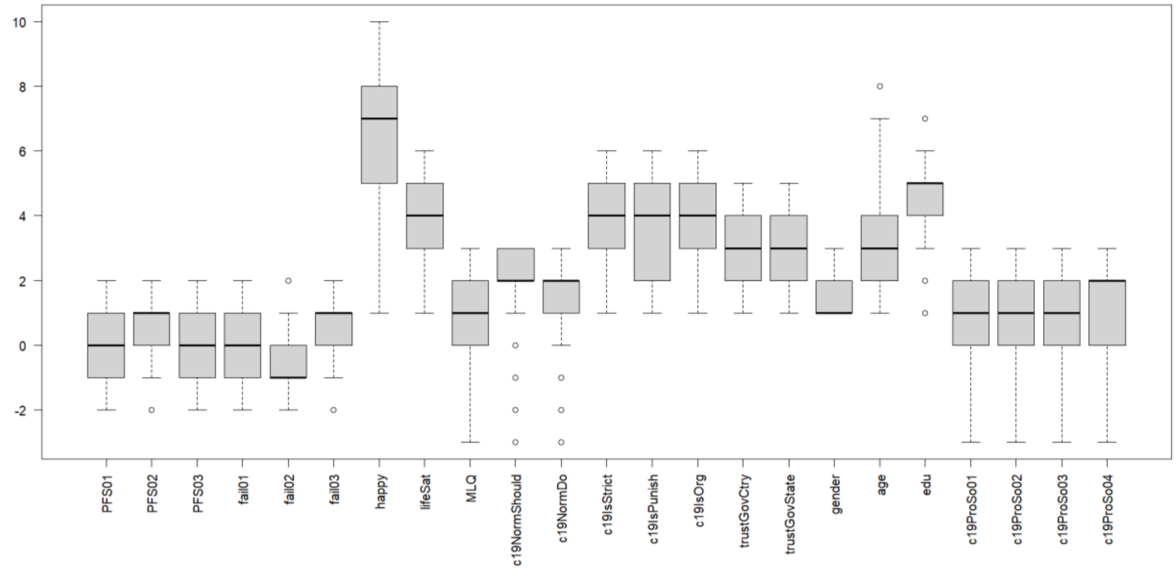
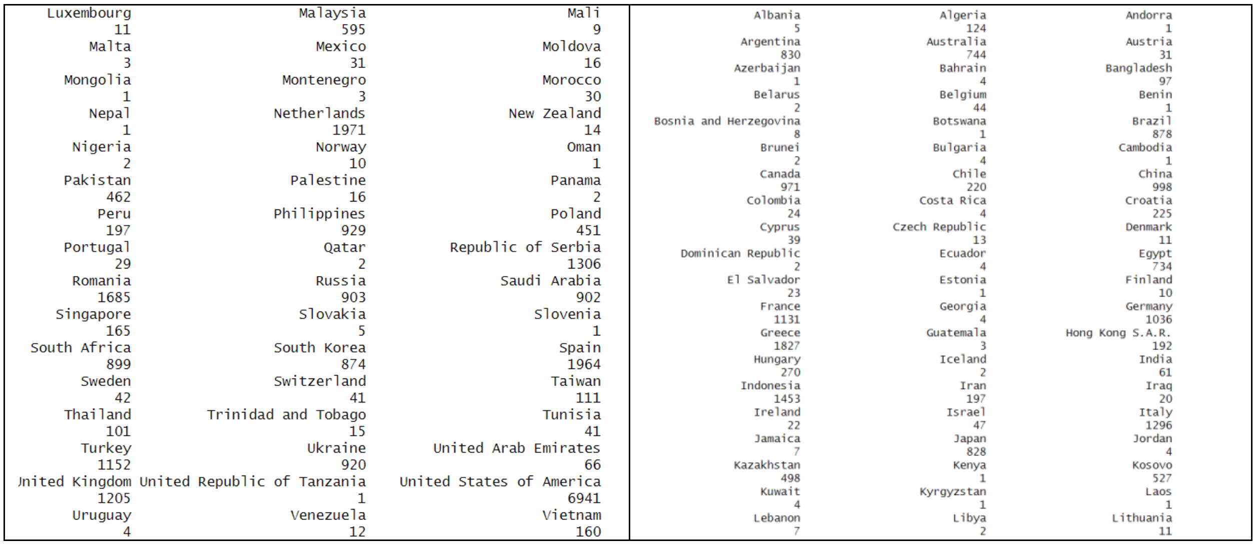


Figure 2. Distribution of numerical values of columns PFS01 – c19ProSo04

The next step involved analyzing the different non-numerical (text) attributes. Since the only non-numerical column in the dataset is coded_country, it was discovered that this column has 108 distinct values, indicating that data was collected from 108 different countries. Figure 3 displays the total number of records in the dataset for each of the 108 countries, while Figure 4 illustrates the top 20 countries with the most information provided in the dataset.



```
  Luxembourg                    Malaysia                      Mali
          11                         595                         9
       Malta                      Mexico                   Moldova
           3                          31                        16
    Mongolia                  Montenegro                   Morocco
           1                           3                        30
       Nepal                 Netherlands               New Zealand
           1                        1971                        14
     Nigeria                      Norway                      Oman
           2                          10                         1
    Pakistan                   Palestine                    Panama
         462                          16                         2
        Peru                 Philippines                    Poland
         197                         929                       451
    Portugal                       Qatar       Republic of Serbia
          29                           2                      1306
     Romania                      Russia              Saudi Arabia
        1685                         903                       902
   Singapore                    Slovakia                  Slovenia
         165                           5                         1
South Africa                 South Korea                     Spain
         899                         874                      1964
      Sweden                 Switzerland                    Taiwan
          42                          41                       111
    Thailand         Trinidad and Tobago                   Tunisia
         101                          15                        41
      Turkey                     Ukraine      United Arab Emirates
        1152                         920                        66
Jnited Kingdom United Republic of Tanzania  United States of America
        1205                           1                      6941
     Uruguay                   Venezuela                   Vietnam
           4                          12                       160
```

```
                Albania               Algeria              Andorra
                      5                   124                    1
              Argentina             Australia              Austria
                    830                   744                   31
             Azerbaijan               Bahrain           Bangladesh
                      1                     4                   97
                Belarus               Belgium                Benin
                      2                    44                    1
 Bosnia and Herzegovina              Botswana               Brazil
                      8                     1                  878
                 Brunei              Bulgaria             Cambodia
                      2                     4                    1
                 Canada                 Chile                China
                    971                   220                  998
               Colombia            Costa Rica              Croatia
                     24                     4                  225
                 Cyprus        Czech Republic              Denmark
                     39                    13                   11
     Dominican Republic               Ecuador                Egypt
                      2                     4                  734
            El Salvador               Estonia              Finland
                     23                     1                   10
                 France               Georgia              Germany
                   1131                     4                 1036
                 Greece             Guatemala     Hong Kong S.A.R.
                   1827                     3                  192
                Hungary               Iceland                India
                    270                     2                   61
              Indonesia                  Iran                 Iraq
                   1453                   197                   20
                Ireland                Israel                Italy
                     22                    47                 1296
                Jamaica                 Japan               Jordan
                      7                   828                    4
             Kazakhstan                 Kenya               Kosovo
                    498                     1                  527
                 Kuwait            Kyrgyzstan                 Laos
                      4                     1                    1
                Lebanon                 Libya            Lithuania
                      7                     2                   11
```

Figure 3: Total number of records in the dataset for each of 108 countries

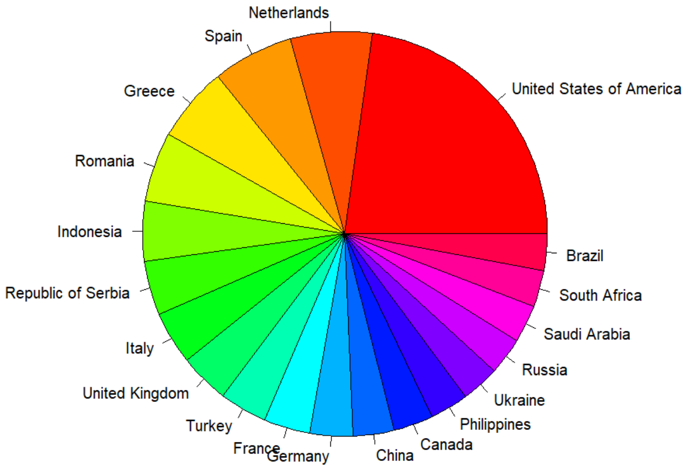


Figure 4: Top 20 countries that have the most records.

Finally, an analysis of missing values was conducted, and the outcomes are presented in Figure 5. The Missing Values Map in Figure 5 demonstrates that there is a high proportion of missing values in the attributes of the Job Insecurity, Employment Status, and Trust in Government subgroups, while the remaining attributes have a very insignificant number of missing values. This implies that further analysis can be conducted productively on the majority of the attributes.

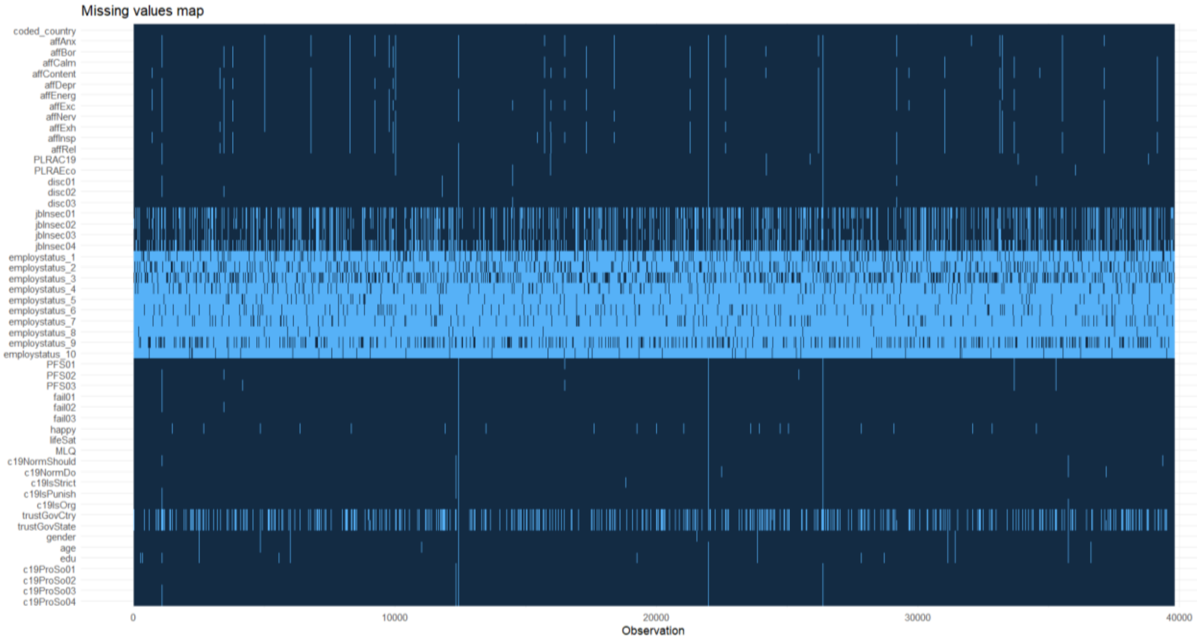


Figure 5: Missing values map

1.2 Pre-processing

In order to perform accurate and reliable descriptive analysis on a dataset, it is crucial to first preprocess the raw data. This involves cleaning and transforming the data to remove any errors, inconsistencies, and redundancies, and to make it more suitable for analysis. The preprocessing step can significantly improve the quality, accuracy and reliability of the results by ensuring that the data is interpretable and ready for use with analytical tools.

When working with a large dataset, such as the one that is used for our analysis, it may be necessary to perform additional pre-processing steps. To begin with, sampling is to be done to reduce the size of the dataset and make it more manageable for analysis. This involves taking a subset of the data that is representative of the entire dataset. Another important preprocessing step is to remove any duplicate rows, which do not provide any additional value to the analysis. Moreover, it is necessary to ensure that all required data is present, such as coded_country column, which acts as the main identifier for the survey. In the provided dataset, it has been noticed that some values in the coded_country column are missing. This is a significant issue as the surveyed data of these rows will not be associated with any country and thus holds no value. Therefore, it is essential to address this issue by removing the rows with missing values to ensure that the remaining data is valid and reliable for analysis. Same was performed with rows where all the data was missing except the country name, since this data is not informative as well. Lastly, it is helpful to structure the data in a way that is easy to work with during analysis. For example, in this case, placing coded_country column at the beginning of the dataset simplifies the code and facilitates the analysis.

Part 2: Focus country vs all other countries as a group

2.1 Extraction of "focus" country and "others"

To begin with, I focus on Indonesia as the country of interest and separate it from the rest of the dataset. This yields two distinct groups for analysis. To effectively compare the survey responses of Indonesia with those of other countries, I create boxplots for all numerical variables (excluding employment status), as illustrated in Figures 6 and 7.

First, an analysis of the boxplots depicting affect values reveals that the emotional state in Indonesia is generally comparable to that of the rest of the world. Most attributes have average values, and depression is reported at low levels. However, Indonesia stands out from the rest of the world in that the respondents report higher levels of excitement, energy, contentment, and inspiration. In terms of job insecurity and financial strain, Indonesia's situation is similar to that of the rest of the world. The only difference is that more respondents in Indonesia feel financially strained. At the same time, they are more optimistic about retaining their jobs. When examining the Corona Community Injective norms, it becomes evident that Indonesians tend to view their rules as strict but organized in responding to the coronavirus. In contrast, the worldwide community believes that more people should be self-isolating and practicing social distancing. With regards to demographics, Indonesian participants have a higher average level of education and are relatively older. Lastly, when analysing Corona ProSocial Behaviour, Indonesia's situation is also quite similar to that of the rest of the world. However, Indonesian respondents express a greater interest in making donations and protecting those who are suffering from coronavirus.

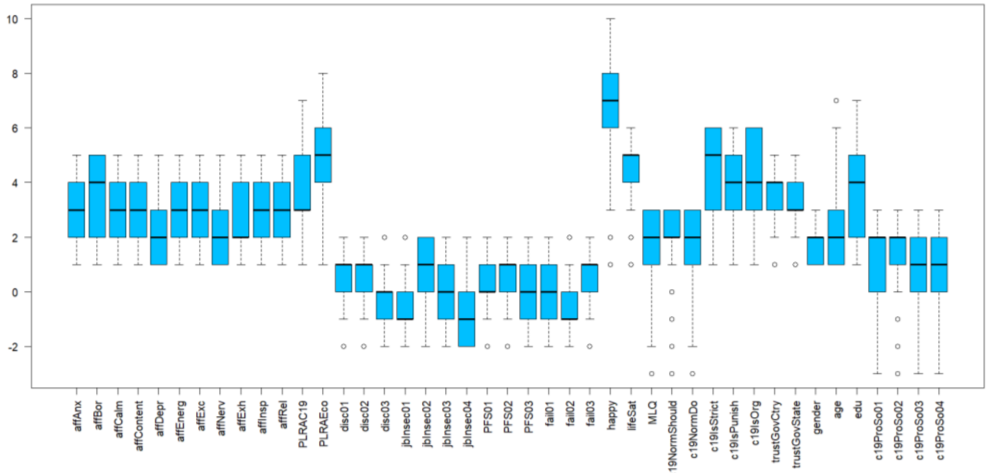


Figure 6. Distribution of numerical values for a focus country (Indonesia)

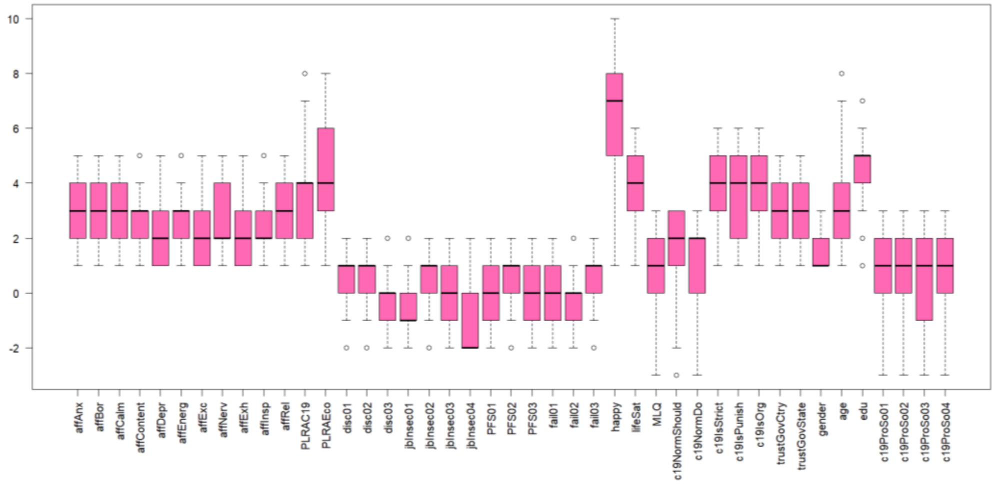


Figure 7. Distribution of numerical values for other countries

2.2 Evaluation of the predictive quality of attributes for “focus” country

To analyze the predictive power of participant responses in relation to pro-social attitudes (c19ProSo01, 2, 3, 4) for Indonesia, which is the assigned country, the dataset must be cleaned of any missing values (NA) and linear regression must be performed.

The Employment Status attributes are pre-processed by replacing any NA fields with 0 since participants were asked to select one of the 10 attributes that best describes their employment status during the last week, allowing these attributes to be treated as Booleans (0 for False, 1 for True). To ensure the accuracy and reliability of the linear regression analysis, all rows containing NA values for any attribute other than Employment Status are removed. It's important to note that other Corona ProSocial Behaviour attributes are not considered when determining the success of the prediction.

The p-value for the t-test is used to determine the most significant predictors. If the p-value is less than a certain significance level, the predictor variable is considered to have a statistically significant relationship with the response variable in the model. Additionally, to assess the overall quality of prediction, the values of Multiple R-squared and Adjusted R-squared are analyzed, with higher values indicating a better fit of the model to the data.

The analysis begins with performing linear regression on c19ProSo01, which reveals that the most significant predictor is jbInsec04 with a Pr(>|t|) value of 0.00252. However, the multiple R-squared value and Adjusted R-squared values are 0.19 and 0.13 respectively, indicating that the participant responses do not predict pro-social attitudes accurately.

Moving on to c19ProSo02, the analysis shows that the most significant predictor is disc02 with a Pr(>|t|) value of 0.000468. Although the multiple R-squared value and Adjusted R-squared values are 0.23 and 0.18 respectively, which is considered relatively good.

The analysis of c19ProSo03 reveals that the most significant predictor is trustGovState with a Pr(>|t|) value of 3.76e-06. However, the multiple R-squared value and Adjusted R-squared values are 0.14 and 0.08 respectively, indicating a poor result.

Lastly, the analysis of c19ProSo04 shows that the most significant predictor is trustGovState with a Pr(>|t|) value of 0.000108. Although the multiple R-squared value and Adjusted R-squared values are 0.16 and 0.11 respectively, they are still considered as a poor result.

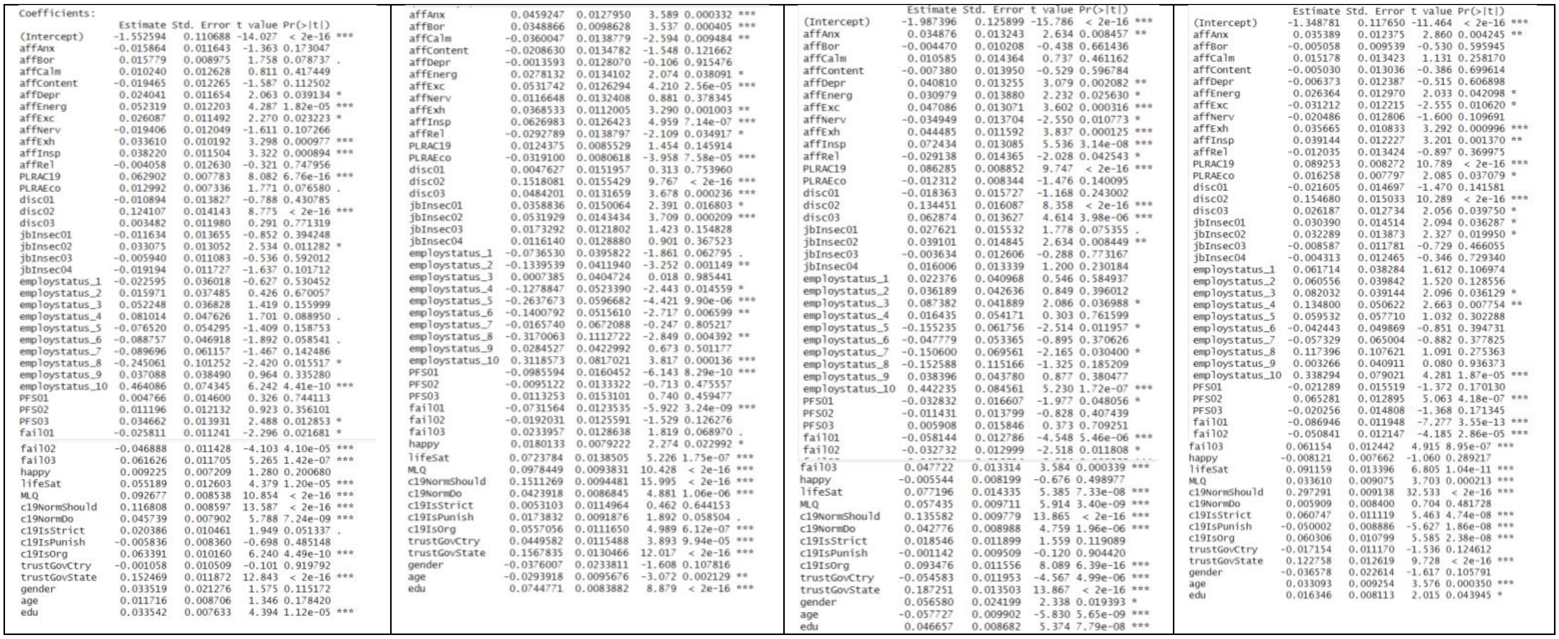
Overall, the analysis suggests that the participant responses (attributes) have low predictive power for pro-social attitudes (c19ProSo01, 2, 3, and 4) in Indonesia.

2.3 Evaluation of the predictive quality of attributes for “other” countries

To analyze the predictive power of participant responses in relation to pro-social attitudes (c19ProSo01, 2, 3, 4) for other countries, the same steps are to be performed, which includes cleaning the dataset of any missing values (NA) and conducting linear regression.

After performing the steps mentioned above, it was determined that for other countries as a group, there are much more best predictors than just for Indonesia. From the Table 1, it may be observed predictor values determined for “other” countries are much more significant than the ones determined for Indonesia, since some values reach <2e-16, which is significantly smaller than the Pr(>|t|) values determined for any of the four (c19ProSo01, 2, 3, 4) values for Indonesia.

c19ProSo01	c19ProSo02	c19ProSo03	c19ProSo04



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.552594   0.110688 -14.027  < 2e-16 ***
affAnx             -0.015864   0.011643  -1.363 0.173047
affBor              0.015779   0.008975   1.758 0.078737 .
affCalm             0.010240   0.012628   0.811 0.417449
affContent         -0.019465   0.012265  -1.587 0.112502
affDepr             0.024041   0.011654   2.063 0.039134 *
affEnerg            0.052319   0.012203   4.287 1.82e-05 ***
affExc              0.026087   0.011492   2.270 0.023223 *
affNerv            -0.019406   0.012049  -1.611 0.107266
affExh              0.033610   0.010192   3.298 0.000977 ***
affInsp             0.038220   0.011504   3.322 0.000894 ***
affRel             -0.004058   0.012630  -0.321 0.747956
PLRAC19             0.062902   0.007783   8.082 6.76e-16 ***
PLRAEco             0.012992   0.007336   1.771 0.076580 .
disc01             -0.010894   0.013827  -0.788 0.430785
disc02              0.124107   0.014143   8.775  < 2e-16 ***
disc03              0.003482   0.011980   0.291 0.771319
jbInsec01          -0.011634   0.013655  -0.852 0.394248
jbInsec02           0.033075   0.013052   2.534 0.011282 *
jbInsec03          -0.005940   0.011083  -0.536 0.592012
jbInsec04          -0.019194   0.011727  -1.637 0.101712
employstatus_1     -0.022595   0.036018  -0.627 0.530452
employstatus_2      0.015971   0.037485   0.426 0.670057
employstatus_3      0.052248   0.036828   1.419 0.155999
employstatus_4      0.081014   0.047626   1.701 0.088950 .
employstatus_5     -0.076520   0.054295  -1.409 0.158753
employstatus_6     -0.088757   0.046918  -1.892 0.058541 .
employstatus_7     -0.089696   0.061157  -1.467 0.142486
employstatus_8     -0.245061   0.101252  -2.420 0.015517 *
employstatus_9      0.037088   0.038490   0.964 0.335280
employstatus_10     0.464086   0.074345   6.242 4.41e-10 ***
PFS01               0.004766   0.014600   0.326 0.744113
PFS02               0.011196   0.012132   0.923 0.356101
PFS03               0.034662   0.013931   2.488 0.012853 *
fail01             -0.025811   0.011241  -2.296 0.021681 *
fail02             -0.046888   0.011428  -4.103 4.10e-05 ***
fail03              0.061626   0.011705   5.265 1.42e-07 ***
happy               0.009225   0.007209   1.280 0.200680
lifeSat             0.055189   0.012603   4.379 1.20e-05 ***
MLQ                 0.092677   0.008538  10.854  < 2e-16 ***
c19NormShould       0.116808   0.008597  13.587  < 2e-16 ***
c19NormDo           0.045739   0.007902   5.788 7.24e-09 ***
c19IsStrict         0.020386   0.010461   1.949 0.051337 .
c19IsPunish        -0.005836   0.008360  -0.698 0.485148
c19IsOrg            0.063391   0.010160   6.240 4.49e-10 ***
trustGovCtry       -0.001058   0.010509  -0.101 0.919792
trustGovState       0.152469   0.011872  12.843  < 2e-16 ***
gender              0.033519   0.021276   1.575 0.115172
age                 0.011716   0.008706   1.346 0.178420
edu                 0.033542   0.007633   4.394 1.12e-05 ***
```

```
affAnx              0.0459247  0.0127950   3.589 0.000332 ***
affBor              0.0348866  0.0098628   3.537 0.000405 ***
affCalm            -0.0360047  0.0138779  -2.594 0.009484 **
affContent         -0.0208630  0.0134782  -1.548 0.121662
affDepr            -0.0013593  0.0128070  -0.106 0.915476
affEnerg            0.0278132  0.0134102   2.074 0.038091 *
affExc              0.0531742  0.0126294   4.210 2.56e-05 ***
affNerv             0.0116648  0.0132408   0.881 0.378345
affExh              0.0368533  0.0112005   3.290 0.001003 **
affInsp             0.0626983  0.0126423   4.959 7.14e-07 ***
affRel             -0.0292789  0.0138797  -2.109 0.034917 *
PLRAC19             0.0124375  0.0085529   1.454 0.145914
PLRAEco            -0.0319100  0.0080618  -3.958 7.58e-05 ***
disc01              0.0047627  0.0151957   0.313 0.753960
disc02              0.1518081  0.0155429   9.767  < 2e-16 ***
disc03              0.0484201  0.0131659   3.678 0.000236 ***
jbInsec01           0.0358836  0.0150064   2.391 0.016803 *
jbInsec02           0.0531929  0.0143434   3.709 0.000209 ***
jbInsec03           0.0173292  0.0121802   1.423 0.154828
jbInsec04           0.0116140  0.0128880   0.901 0.367523
employstatus_1     -0.0736530  0.0395822  -1.861 0.062795 .
employstatus_2     -0.1339539  0.0411940  -3.252 0.001149 **
employstatus_3      0.0007385  0.0404724   0.018 0.985441
employstatus_4     -0.1278847  0.0523390  -2.443 0.014559 *
employstatus_5     -0.2637673  0.0596682  -4.421 9.90e-06 ***
employstatus_6     -0.1400792  0.0515610  -2.717 0.006599 **
employstatus_7     -0.0165740  0.0672088  -0.247 0.805217
employstatus_8     -0.3170063  0.1112722  -2.849 0.004392 **
employstatus_9      0.0284527  0.0422992   0.673 0.501177
employstatus_10     0.3118573  0.0817021   3.817 0.000136 ***
PFS01              -0.0985594  0.0160452  -6.143 8.29e-10 ***
PFS02              -0.0095122  0.0133322  -0.713 0.475557
PFS03               0.0113253  0.0153101   0.740 0.459477
fail01             -0.0731564  0.0123535  -5.922 3.24e-09 ***
fail02             -0.0192031  0.0125591  -1.529 0.126276
fail03              0.0233957  0.0128638   1.819 0.068970 .
happy               0.0180133  0.0079222   2.274 0.022992 *
lifeSat             0.0723784  0.0138505   5.226 1.75e-07 ***
MLQ                 0.0978449  0.0093831  10.428  < 2e-16 ***
c19NormShould       0.1511269  0.0094481  15.995  < 2e-16 ***
c19NormDo           0.0423918  0.0086845   4.881 1.06e-06 ***
c19IsStrict         0.0053103  0.0114964   0.462 0.644153
c19IsPunish         0.0173832  0.0091876   1.892 0.058504 .
c19IsOrg            0.0557056  0.0111650   4.989 6.12e-07 ***
trustGovCtry        0.0449582  0.0115488   3.893 9.94e-05 ***
trustGovState       0.1567835  0.0130466  12.017  < 2e-16 ***
gender             -0.0376007  0.0233811  -1.608 0.107816
age                -0.0293918  0.0095676  -3.072 0.002129 **
edu                 0.0744771  0.0083882   8.879  < 2e-16 ***
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.987396   0.125899 -15.786  < 2e-16 ***
affAnx              0.034876   0.013243   2.634 0.008457 **
affBor             -0.004470   0.010208  -0.438 0.661436
affCalm             0.010585   0.014364   0.737 0.461162
affContent         -0.007380   0.013950  -0.529 0.596784
affDepr             0.040810   0.013255   3.079 0.002082 **
affEnerg            0.030979   0.013880   2.232 0.025630 *
affExc              0.047086   0.013071   3.602 0.000316 ***
affNerv            -0.034949   0.013704  -2.550 0.010773 *
affExh              0.044485   0.011592   3.837 0.000125 ***
affInsp             0.072434   0.013085   5.536 3.14e-08 ***
affRel             -0.029138   0.014365  -2.028 0.042543 *
PLRAC19             0.086285   0.008852   9.747  < 2e-16 ***
PLRAEco            -0.012312   0.008344  -1.476 0.140095
disc01             -0.018363   0.015727  -1.168 0.243002
disc02              0.134451   0.016087   8.358  < 2e-16 ***
disc03              0.062874   0.013627   4.614 3.98e-06 ***
jbInsec01           0.027621   0.015532   1.778 0.075355 .
jbInsec02           0.039101   0.014845   2.634 0.008449 **
jbInsec03          -0.003634   0.012606  -0.288 0.773167
jbInsec04           0.016006   0.013339   1.200 0.230184
employstatus_1      0.022376   0.040968   0.546 0.584937
employstatus_2      0.036189   0.042636   0.849 0.396012
employstatus_3      0.087382   0.041889   2.086 0.036988 *
employstatus_4      0.016435   0.054171   0.303 0.761599
employstatus_5     -0.155235   0.061756  -2.514 0.011957 *
employstatus_6     -0.047779   0.053365  -0.895 0.370626
employstatus_7     -0.150600   0.069561  -2.165 0.030400 *
employstatus_8     -0.152588   0.115166  -1.325 0.185209
employstatus_9      0.038396   0.043780   0.877 0.380477
employstatus_10     0.442235   0.084561   5.230 1.72e-07 ***
PFS01              -0.032832   0.016607  -1.977 0.048056 *
PFS02              -0.011431   0.013799  -0.828 0.407439
PFS03               0.005908   0.015846   0.373 0.709251
fail01             -0.058144   0.012786  -4.548 5.46e-06 ***
fail02             -0.032732   0.012999  -2.518 0.011808 *
fail03              0.047722   0.013314   3.584 0.000339 ***
happy              -0.005544   0.008199  -0.676 0.498977
lifeSat             0.077196   0.014335   5.385 7.33e-08 ***
MLQ                 0.057435   0.009711   5.914 3.40e-09 ***
c19NormShould       0.135582   0.009779  13.865  < 2e-16 ***
c19NormDo           0.042776   0.008988   4.759 1.96e-06 ***
c19IsStrict         0.018546   0.011899   1.559 0.119089
c19IsPunish        -0.001142   0.009509  -0.120 0.904420
c19IsOrg            0.093476   0.011556   8.089 6.39e-16 ***
trustGovCtry       -0.054583   0.011953  -4.567 4.99e-06 ***
trustGovState       0.187251   0.013503  13.867  < 2e-16 ***
gender              0.056580   0.024199   2.338 0.019393 *
age                -0.057727   0.009902  -5.830 5.65e-09 ***
edu                 0.046657   0.008682   5.374 7.79e-08 ***
```

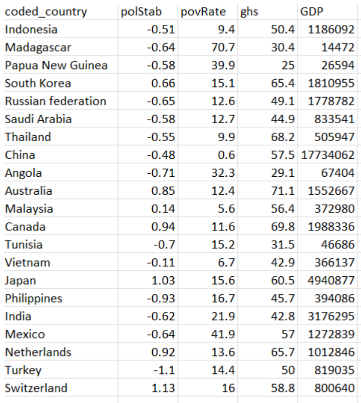
```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.348781   0.117650 -11.464  < 2e-16 ***
affAnx              0.035389   0.012375   2.860 0.004245 **
affBor             -0.005058   0.009539  -0.530 0.595945
affCalm             0.015178   0.013423   1.131 0.258170
affContent         -0.005030   0.013036  -0.386 0.699614
affDepr            -0.006373   0.012387  -0.515 0.606898
affEnerg            0.026364   0.012970   2.033 0.042098 *
affExc             -0.031212   0.012215  -2.555 0.010620 *
affNerv            -0.020486   0.012806  -1.600 0.109691
affExh              0.035665   0.010833   3.292 0.000996 ***
affInsp             0.039144   0.012227   3.201 0.001370 **
affRel             -0.012035   0.013424  -0.897 0.369975
PLRAC19             0.089253   0.008272  10.789  < 2e-16 ***
PLRAEco             0.016258   0.007797   2.085 0.037079 *
disc01             -0.021605   0.014697  -1.470 0.141581
disc02              0.154680   0.015033  10.289  < 2e-16 ***
disc03              0.026187   0.012734   2.056 0.039750 *
jbInsec01           0.030390   0.014514   2.094 0.036287 *
jbInsec02           0.032289   0.013873   2.327 0.019950 *
jbInsec03          -0.008587   0.011781  -0.729 0.466055
jbInsec04          -0.004313   0.012465  -0.346 0.729340
employstatus_1      0.061714   0.038284   1.612 0.106974
employstatus_2      0.060556   0.039842   1.520 0.128556
employstatus_3      0.082032   0.039144   2.096 0.036129 *
employstatus_4      0.134800   0.050622   2.663 0.007754 **
employstatus_5      0.059532   0.057710   1.032 0.302288
employstatus_6     -0.042443   0.049869  -0.851 0.394731
employstatus_7     -0.057329   0.065004  -0.882 0.377825
employstatus_8      0.117396   0.107621   1.091 0.275363
employstatus_9      0.003266   0.040911   0.080 0.936373
employstatus_10     0.338294   0.079021   4.281 1.87e-05 ***
PFS01              -0.021289   0.015519  -1.372 0.170130
PFS02               0.065281   0.012895   5.063 4.18e-07 ***
PFS03              -0.020256   0.014808  -1.368 0.171345
fail01             -0.086946   0.011948  -7.277 3.55e-13 ***
fail02             -0.050841   0.012147  -4.185 2.86e-05 ***
fail03              0.061154   0.012442   4.915 8.95e-07 ***
happy              -0.008121   0.007662  -1.060 0.289217
lifeSat             0.091159   0.013396   6.805 1.04e-11 ***
MLQ                 0.033610   0.009075   3.703 0.000213 ***
c19NormShould       0.297291   0.009138  32.533  < 2e-16 ***
c19NormDo           0.005909   0.008400   0.704 0.481728
c19IsStrict         0.060747   0.011119   5.463 4.74e-08 ***
c19IsPunish        -0.050002   0.008886  -5.627 1.86e-08 ***
c19IsOrg            0.060306   0.010799   5.585 2.38e-08 ***
trustGovCtry       -0.017154   0.011170  -1.536 0.124612
trustGovState       0.122758   0.012619   9.728  < 2e-16 ***
gender             -0.036578   0.022614  -1.617 0.105791
age                 0.033093   0.009254   3.576 0.000350 ***
edu                 0.016346   0.008113   2.015 0.043945 *
```

Table 1: Significant predictors for “other” countries

Part 3: Focus country vs cluster of similar countries

3.1 Identification of indicators and data collection

Various social, economic, health, and political indicators were considered when searching for countries similar to Indonesia. Table 2 displays these indicators, including political stability and freedom from violence/terrorism[1], poverty rate[2], Global Health security index[3], and GPD[4], as they are linked to the data shown in the PsyCoronaBaselineExtract. To perform clustering and determine the seven countries closest to Indonesia, a subset of 20 countries with similarities in one or more of the four indicators were selected.



coded_country	polStab	povRate	ghs	GDP
Indonesia	-0.51	9.4	50.4	1186092
Madagascar	-0.64	70.7	30.4	14472
Papua New Guinea	-0.58	39.9	25	26594
South Korea	0.66	15.1	65.4	1810955
Russian federation	-0.65	12.6	49.1	1778782
Saudi Arabia	-0.58	12.7	44.9	833541
Thailand	-0.55	9.9	68.2	505947
China	-0.48	0.6	57.5	17734062
Angola	-0.71	32.3	29.1	67404
Australia	0.85	12.4	71.1	1552667
Malaysia	0.14	5.6	56.4	372980
Canada	0.94	11.6	69.8	1988336
Tunisia	-0.7	15.2	31.5	46686
Vietnam	-0.11	6.7	42.9	366137
Japan	1.03	15.6	60.5	4940877
Philippines	-0.93	16.7	45.7	394086
India	-0.62	21.9	42.8	3176295
Mexico	-0.64	41.9	57	1272839
Netherlands	0.92	13.6	65.7	1012846
Turkey	-1.1	14.4	50	819035
Switzerland	1.13	16	58.8	800640

Table 2: 20 countries relatively similar to Indonesia in terms of social, health, political and economic situation

We used the K-means clustering method to identify 7 countries similar to Indonesia in terms of social, health, political, and economic factors. K-means clustering is an unsupervised algorithm that groups similar data points into K clusters based on their similarity. To apply K-means clustering, we first scaled the data and determined the appropriate number of clusters K through trial and error. After clustering the countries based on the mentioned factors, we identified countries that share similarities with Indonesia. These countries, which can be seen in Figure 8, are India, Turkey, Russian Federation, Vietnam, Saudi Arabia, Philippines, and Tunisia.

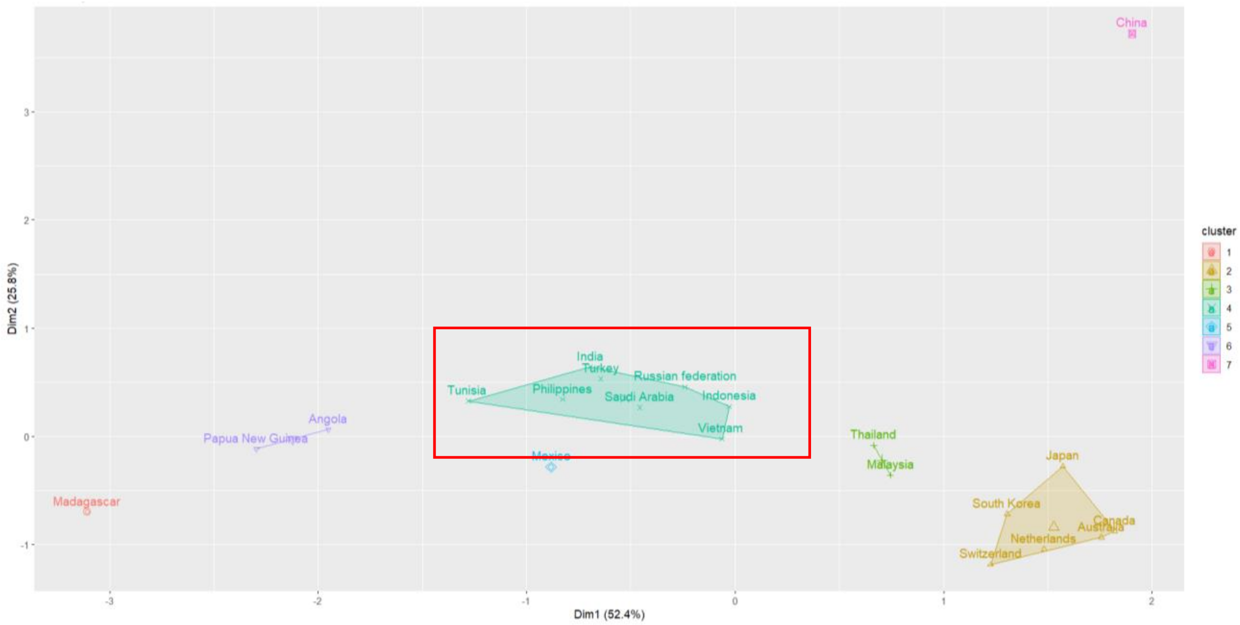


Figure 8: Clusters created by apply k-means clustering method on data in Table 2

3.2 Evaluation of the predictive quality of attributes in cluster

In order to analyse how well participants' answers predict pro-social attitudes (c19ProSo01,2,3 and 4) for the cluster of similar countries, we perform the same steps as in 2.2, which includes data pre-processing and linear regression. As a result, we have determined the values of Multiple R-Squared and Adjusted R-Squared for each of the c19ProSo01, 2, 3 and 4: c19ProSo01 with 0.36 and 0.18, c19ProSo02 with 0.41 and 0.24, c19ProSo03 with 0.38 and 0.21, c19ProSo04 with 0.39 and 0.22 respectively, which may be considered as a good prediction performance, comparing to the values received in 2.2.

As for the strongest predictors, it was determined that for c19ProSo01 they are affCalm(0.00602), affExc(0.00944); for c19ProSo02 - affCalm(0.00181), affExc(0.00861), affRel (0.00244); for c19ProSo03 - affCalm(0.00396); and c19ProSo01 - affAnx(0.00425), affCalm (0.00180), disc02 (0.00672). The strongest predictors determined for the group of clustered countries are distinct from those for Indonesia, whereas the best predictors for all other countries (Part 2.3) include the best predictors for Indonesia. This suggests that the other countries have a better match to the important attributes for prediction pro-social attitudes in Indonesia.

References

[1] https://info.worldbank.org/governance/wgi/

[2] https://worldpopulationreview.com/country-rankings/poverty-rate-by-country

[3] https://www.ghsindex.org/

[4] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?most_recent_value_desc=true

Appendix

```
# Author: Sutulova Tatiana, 30806151
# Assignment 1
# Describe the data overall, including things such as dimension, data types,
# distribution of numerical attributes, variety of non-numerical (text)
# attributes, missing values, and anything else of interest or relevance.

# Libraries and packages to be used
library(dplyr)
library(finalfit)
library(tidyverse) # data manipulation
library(cluster) # clustering algorithms
library(factoextra) # clustering algorithms & visualization

# Read the data from the file into a data frame
rm (list = ls())
set.seed(30806151)
cvbase = read.csv("PsyCoronaBaselineExtract.csv", header = TRUE, na.strings = c("", " ",
"NA")) # replacing empty spaces with NA
cvbase <- cvbase [sample(nrow(cvbase), 40000),] # take a sample of 40000

# Pre-processing:
# Clear out duplicates
cvbase <- unique(cvbase)

# Remove rows where there is no country name
cvbase = cvbase[!is.na(cvbase$coded_country),]

# Placing the column coded_country to the front for easier processing
cvbase <- cvbase %>% relocate(coded_country)

# Remove rows where everything except country name is NA
cvbase <- cvbase %>% filter(!if_all(affAnx:c19ProSo04, is.na))


# Task 1:
# Dimensions of data
print(ncol(cvbase))

# Data types
str(cvbase)

# Distribution of numerical attributes
par(mar=c(10,5,2,2))
boxplot(cvbase[2:21], las = 2)
boxplot(cvbase[32:54], las = 2)

# Variety of non-numerical (text) attributes
# Find how many countries are recorded in the data set
```

```r
unique_val <- unique(cvbase$coded_country)
length(unique_val)

# List down how many times each country is met
print(table(cvbase$coded_country))

# Create a pie chart to show top 20 countries
countries <- as.data.frame(table(cvbase$coded_country))

top20_countries <- countries %>%
  arrange(desc(Freq)) %>%
  slice(1:20)

pie(top20_countries$Freq, labels = top20_countries$Var1, col = rainbow(20)) # show a pie chart

# Missing values
cvbase %>% missing_plot() # plot missing values


# Task 2:
# Preparing data
# Extracting responses of Indonesia into a new df
cvbase_focus = cvbase %>% filter(coded_country == c("Indonesia"))
# Extracting responses of all other countries except the focus one into a new df
cvbase_others = filter(cvbase, coded_country != "Indonesia")

# Replace all NA values for employment status with 0
cvbase_focus[c(22:31)][is.na(cvbase_focus[c(22:31)])] <- 0
# Remove all the rows with NA values
cvbase_focus <- na.omit(cvbase_focus)
cvbase_focus$coded_country <- NULL

# Replace all NA values for employment status with 0
cvbase_others[c(22:31)][is.na(cvbase_others[c(22:31)])] <- 0
# Remove all the rows with NA values
cvbase_others <- na.omit(cvbase_others)
cvbase_others$coded_country <- NULL

# Part A
# drawing a box plot to see the distribution for focus country
par(mar=c(10,5,2,2))
boxplot(cvbase_focus[c(1:20, 31:53)], las = 2, col = "deepskyblue")

par(mar=c(10,5,2,2))
boxplot(cvbase_others[c(1:20, 31:53)], las = 2, col = "hotpink")

# Part B
# Calculating the correlation with all the attributes for c19ProSo01,2,3 and 4

# Creating a linear regression model to check significant predictors in the model summary
fit01 <- lm(c19ProSo01~ ., data = cvbase_focus[1:50]) #c19ProSo01
summary(fit01)

fit02 <- lm(c19ProSo02~ ., data = cvbase_focus[c(1:49, 51)]) #c19ProSo02
```

```r
summary(fit02)

fit03 <- lm(c19ProSo03~ ., data = cvbase_focus[c(1:49, 52)]) #c19ProSo03
summary(fit03)

fit04 <- lm(c19ProSo04~ ., data = cvbase_focus[c(1:49, 53)]) #c19ProSo04
summary(fit04)

# Part C
# Calculating the correlation with all the attributes for c19ProSo01,2,3 and 4
fit1.1 <- lm(c19ProSo01~ ., data = cvbase_others[1:50]) #c19ProSo01
summary(fit1.1)

fit1.2 <- lm(c19ProSo02~ ., data = cvbase_others[c(1:49, 51)]) #c19ProSo02
summary(fit1.2 )

fit1.3 <- lm(c19ProSo03~ ., data = cvbase_others[c(1:49, 52)]) #c19ProSo03
summary(fit1.3)

fit1.4 <- lm(c19ProSo04~ ., data = cvbase_others[c(1:49, 53)]) #c19ProSo04
summary(fit1.4)

# Task 3
# Part A
# Reading the database that was created based on several social, economic, health and policial indicators
countriesdf <- read.csv("Countries.csv")

# Make the countries as row names
rownames(countriesdf) <- countriesdf$coded_country
countriesdf$coded_country <- NULL

# Scale numerical data
countriesdf = scale(countriesdf)

# Performing clustering and plotting
countriesdf_kfit = kmeans(countriesdf, 7, nstart = 20)
fviz_cluster(countriesdf_kfit, data = countriesdf)

#Part B
# Extracting responses of similar countries into a new df
cvbase_cluster = cvbase %>% filter(coded_country == c("India","Tukey", "Russia", "Vietnam", "Saudi Arabia", "Philippines", "Tunisia"))

# Prepare "clusters" for linear regression

# Replace all NA values for employment status with 0
cvbase_cluster[c(22:31)][is.na(cvbase_cluster[c(22:31)])] <- 0

# Remove all the rows with NA values
cvbase_cluster <- na.omit(cvbase_cluster)

# Calculating the correlation with all the attributes for c19ProSo01,2,3 and 4
# Remove the coded_country column as all are considered to be in the same group
```

```r
cvbase_cluster$coded_country <- NULL

fit2.1 <- lm(c19ProSo01~ ., data = cvbase_cluster[1:50]) #c19ProSo01
summary(fit2.1)

fit2.2 <- lm(c19ProSo02~ ., data = cvbase_cluster[c(1:49, 51)]) #c19ProSo02
summary(fit2.2)

fit2.3 <- lm(c19ProSo03~ ., data = cvbase_cluster[c(1:49, 52)]) #c19ProSo03
summary(fit2.3)

fit2.4 <- lm(c19ProSo04~ ., data = cvbase_cluster[c(1:49, 53)]) #c19ProSo04
summary(fit2.4)
```