FIT3152 – Data Analytics

MONASH University

COVID-19 Prosocial Behaviour Analysis

Analysing country-level predictors of prosocial behaviour in the pandemic's early stages

Joanna Moy

Introduction

During the COVID-19 pandemic, understanding behaviours that help control the virus spread is crucial. This report analyses pro-social behaviours—voluntary actions aimed at benefiting others—observed in the pandemic's early stages. Such behaviours are vital as they influence public health measures and the effectiveness of virus response strategies. Using initial survey data, this analysis identifies key predictors of pro-social behaviours, including willingness to assist those affected, make donations, protect vulnerable groups, and make personal sacrifices. Although the study takes a global view, it specifically examines these behaviours within **Indonesia** to explore cultural and regional differences in pandemic responses, aiming to guide the design of effective public health campaigns and policies.

Generative AI was not used in the assignment.

1. Descriptive analysis and pre-processing

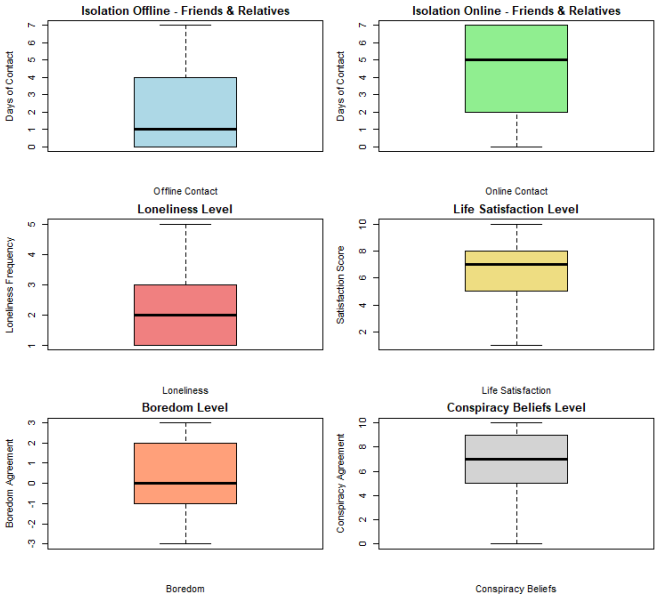
a) To ensure reproducibility, the complete R script used in this analysis is included in Appendix A.

The sampled dataset consists of 40,000 rows and 52 columns, detailing attributes such as gender, age, education, and employment status, among others related to behaviour during the pandemic. The main analysis will focus on the country-level predictors of pro-social behaviours aimed at mitigating the spread of COVID-19 during its early stages. The data types and preliminary insights into missing data are provided by the str() function. A structured data frame in Table A1 of Appendix A elaborates on the data types, column names, and missing values, with most variables being integers, indicative of their origin from multiple-choice responses. From Table A1, it is evident that employment status varies in terms of missing data, which could impact pro-social behaviours as employment influences one's capacity and opportunity for engagement. Isolation variables show low missing data, suggesting a potential influence on pro-social actions, especially for those more isolated. Loneliness variables, with few missing values, indicate that loneliness could drive pro-social activities as a form of seeking connection. The proximity to COVID-19 variables, moderately incomplete, are crucial for understanding pro-social behaviour as personal impact may heighten responsiveness to the pandemic.

Numerical attributes are summarized using the summary() function, offering statistics like minimum, maximum, and quartiles. For better visualization, boxplots and histograms were generated.

Figure 1

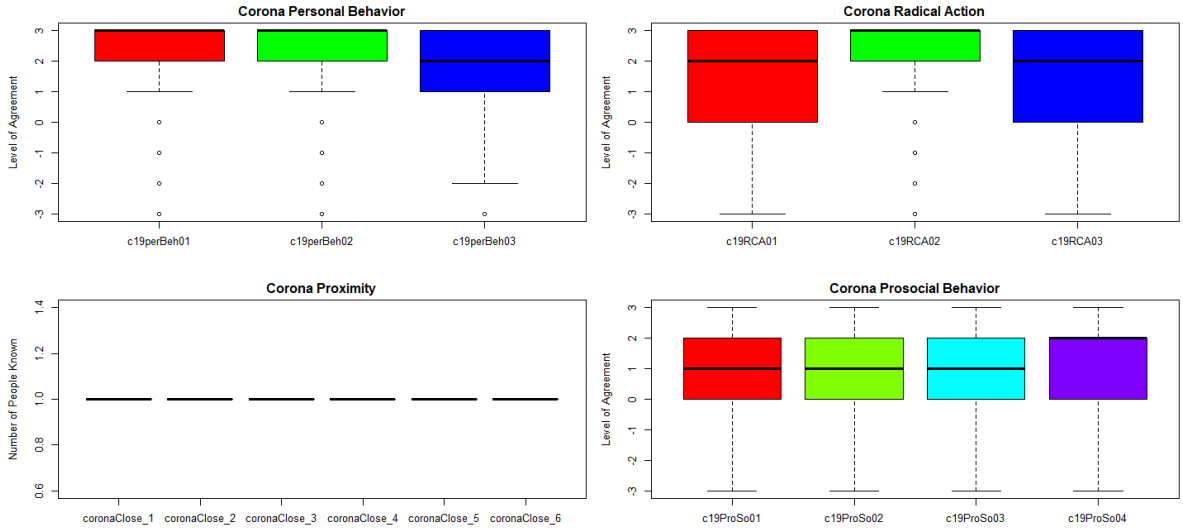
Boxplots of Psychological and Behavioural Responses During the Pandemic



Firstly, six individual boxplots were created to show survey participants' feelings—covering offline and online contact days with friends and relatives during isolation, levels of loneliness, life satisfaction, boredom, and agreement with conspiracy theories. The **offline isolation** boxplot illustrates the distribution of in-person contact days, likely influenced by public health policies that restricted physical gatherings, leading to a lower median and narrow Interquartile Range (IQR). This reflects enforced social distancing to curb virus transmission. The **online isolation** boxplot shows a higher median and slightly wider IQR for online contact, a result of the push towards digital communication due to restrictions on in-person gatherings. Variations may reflect differing access to technology or preferences for digital interactions. The **loneliness** boxplot indicates occasional high levels of loneliness, possibly exacerbated by lockdown-induced social isolation, though some managed well due to strong online social networks or successful adaptation strategies. The **life satisfaction** boxplot reveals a generally positive median satisfaction level, suggesting effective coping mechanisms during the pandemic like new hobbies or more family time. However, experiences varied widely, influenced by economic and health stresses. The **boredom** boxplot suggests low overall boredom levels, with participants engaging in activities like reading, cooking, or exercising, though some faced limitations due to space or resources. The **conspiracy theory** boxplot shows a median skewed towards disagreement with conspiracy theories, yet with a widespread indicating varied beliefs, from strong disagreement to agreement among different individuals.

Figure 2

Boxplots Showing the Variations in COVID-19 Related Behaviours

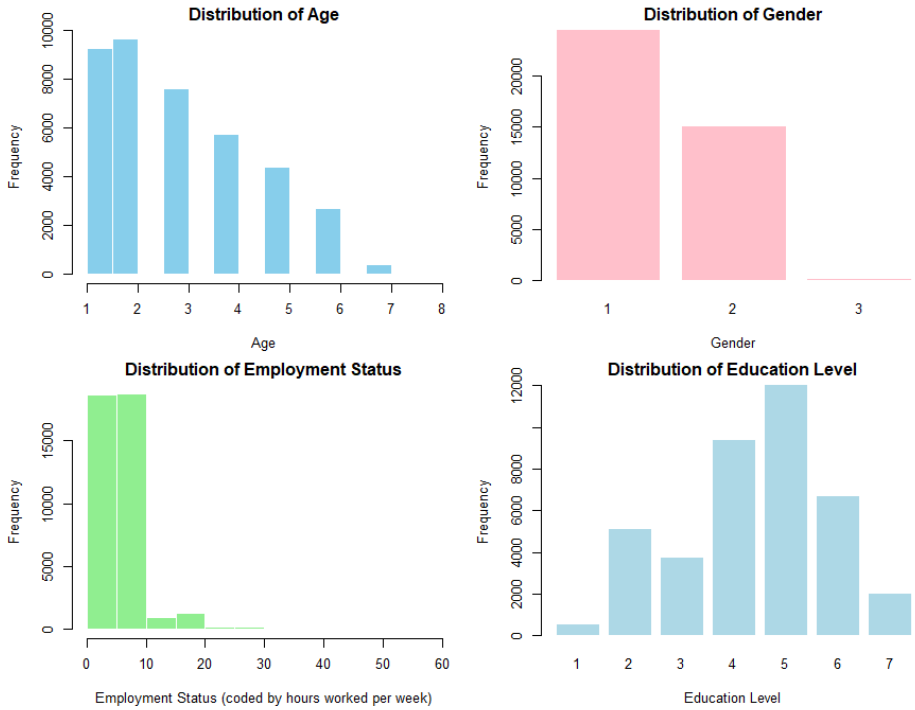


The second group of boxplots illustrates participant responses to various COVID-19 related behaviours. The first boxplot displays **agreement with personal protective behaviours** (like washing hands and avoiding crowds). The red and green boxplots ('c10perBeh01' and 'c19perBeh02') show high median levels, indicating widespread acceptance, influenced by effective public health messaging. Variability in responses could stem from personal circumstances, such as job type or personal beliefs. The lower median and broader spread for self-quarantining reflect its practical challenges, with factors like employment and caregiving responsibilities affecting the ability to quarantine. The second set of boxplots focuses on **stringent measures** like mandatory vaccinations and quarantines. The first boxplot (c19RCA01) reveals high agreement with mandatory vaccination, shown by its compact distribution. The second boxplot (c19RCA02) displays a moderate agreement with mandatory quarantine, with less spread and outliers present. The third (c19RCA03) shows varied opinions on reporting suspected COVID-19 cases, as indicated by its wider spread. The **corona proximity** plot suggests that most participants did not personally know anyone with COVID-19, indicative of the pandemic's early stages, with outliers representing those more directly affected. The final plot examines **willingness to engage in pro-social behaviours** like helping others and making donations. Similar median values in 'c19ProSo01', 'c19ProSo02', and 'c19ProSo03' suggest strong

willingness for supportive actions, while ‘c19ProSo04’ (personal sacrifices) show variability, especially in willingness to make personal sacrifices.

Figure 3

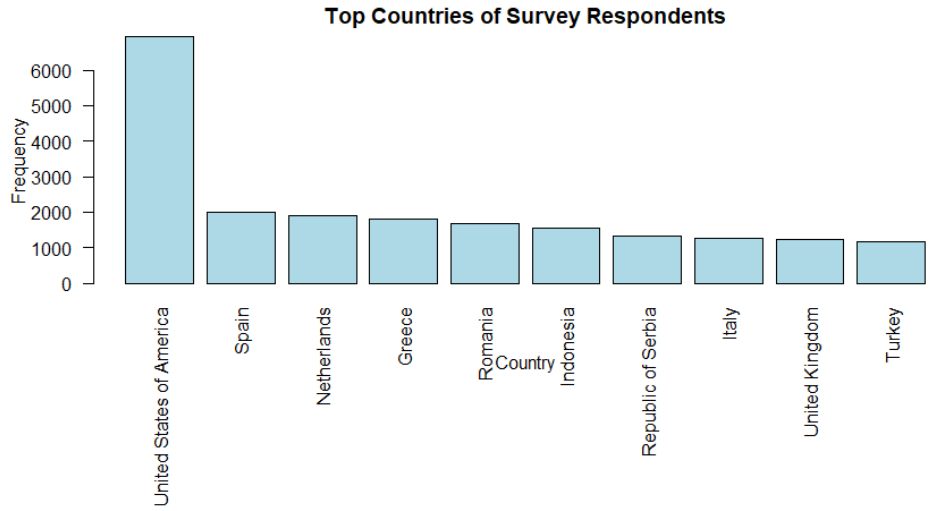
Demographic Profiles of Survey Respondents During the COVID-19 Pandemic



The next group of graphs displays the demographic profiles of respondents as histograms. The first plot charts **age distribution**, with younger respondents predominating as indicated by a left-skewed distribution on the age-axis. This skew could reflect the survey's greater accessibility to or popularity among younger individuals, who may have been more affected by COVID-19 related changes in education and employment, heightening their interest in participating. The second histogram shows **gender distribution** with categories likely representing female, male, and possibly other genders. The highest count is in the first category, indicating a higher representation of female respondents. The representation of other genders is notably lower, which may reflect societal or survey structural biases. The third histogram details **employment status**, focusing on hours worked per week. A significant skew towards fewer working hours suggests that the survey predominantly reached part-time workers, students, or those not in full-time employment, likely influenced by the pandemic’s effects on jobs. The final histogram outlines **education levels**, showing a concentration in mid-range educational attainments (potentially high school and undergraduate levels), with fewer respondents at the very low or high ends of the spectrum. This pattern suggests that the survey appealed more to those with some level of formal education.

Figure 4

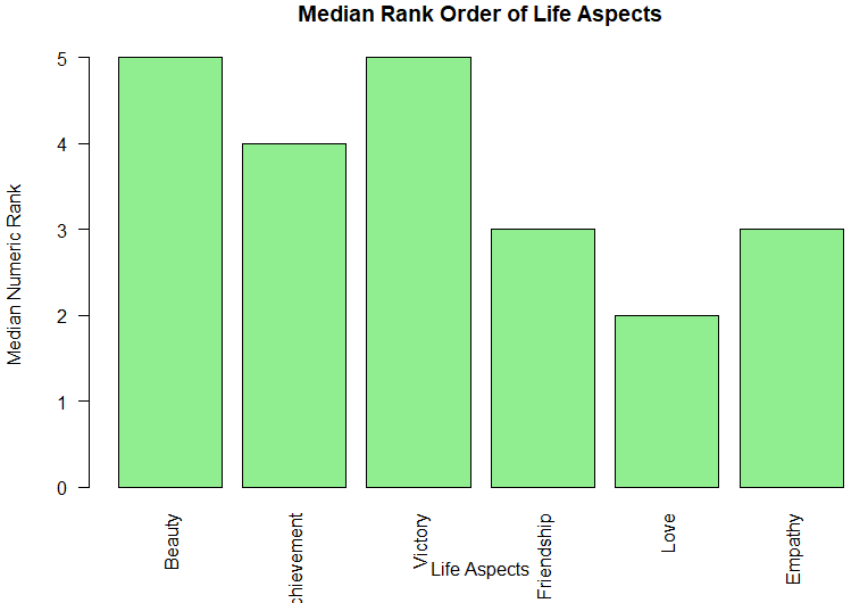
Top Countries of Participant Engagement in COVID-19 Survey



The following histogram represents the distribution of participants by country. As shown in the chart, the United States of America has the highest frequency of respondents, significantly more than any other country listed. This suggests that a large portion of the survey's participants are from the USA, which might reflect the survey's distribution channels, language, or the interest and engagement level. The other countries shown have significantly fewer respondents compared to the USA, but they seem to have a relatively similar representation within this subset.

Figure 5

Median Rank Order of Life Aspects of COVID-19 Survey Respondents



The following chart shows median rankings for various life aspects among survey participants. **Beauty** and **victory** have higher median ranks, indicating they are considered less important compared to other aspects. In contrast, **achievement** ranks slightly lower, suggesting it is somewhat more important but still not a top priority. **Friendship** and **empathy** with a notably lower median rank,

is valued highly, reflecting its importance to respondents. **Love**, with the lowest median ranks, are seen as highly important aspects, demonstrating their significant value to participants.

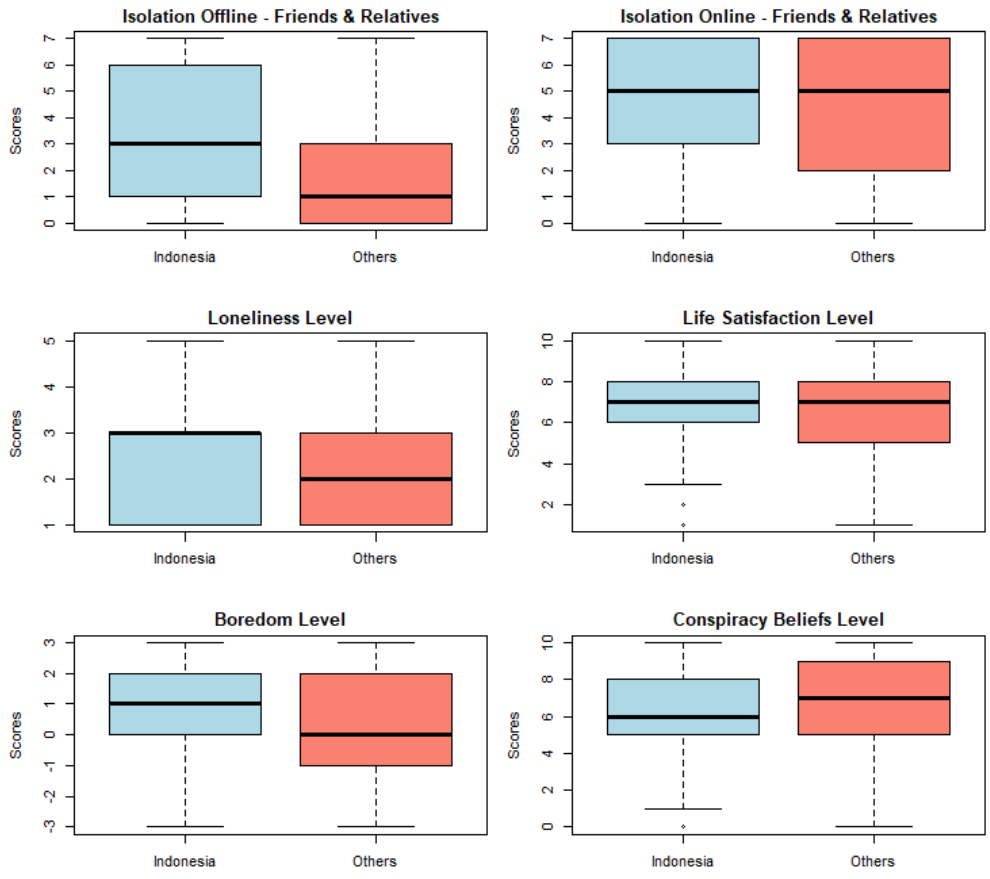
b) In preparing for the analysis of the dataset, there are several areas that require attention to ensure the integrity of the subsequent analysis. Firstly, while the R script accounts for missing values in descriptive statistics, it does not specify how they are handled in the analysis. Additionally, for variables such as 'rankOrdLife', the data transformation from non-numeric categorical data into numeric ranks for analysis is essential for median calculation. Similarly, for the employment status, it assumes a numeric representation based on the hours worked per week, however, this might need to be changed to match the survey's encoding. Other points to consider are managing outliers and ensuring that the graphical parameters after each set of plots are reset.

2. Focus country vs. all other countries as a group

a) To analyse the participant responses for Indonesia compared with other countries, a series of box plots are created as shown below, in which their psychological and behavioural responses are compared with other countries.

Figure 6

Comparative Analysis of Psychological and Behavioural Responses to COVID-19 Between Indonesia and Other Countries



As observed in the first plot (**isolation offline**), the higher median for Indonesia suggests that Indonesian participants had more offline interactions with friends and relatives compared to participants from other countries. This could be due to cultural norms that emphasize community and family ties, or possibly less stringent or differently enforced social distancing measures. The broader IQR for Indonesia indicates more variability in behaviour, which might reflect differences in the

severity of lockdowns across regions within the country or varying personal circumstances, like living arrangements and job requirements that necessitate in-person contact.

For the second plot (**isolation online**), with medians being equal and high IQRs, it shows a significant reliance on online methods for social interaction across all respondents, likely as a substitute for reduced offline contact. The online interaction remained crucial for maintaining social connections during the pandemic. The wider IQR for other countries may suggest access to digital tools or preferences for online communication is more variable outside of Indonesia.

In the third plot (**loneliness level**), the median of 3 for Indonesia indicates a relatively higher reported level of loneliness, which may correlate with the impact of social distancing on well-being. Cultural aspects in Indonesia might make the lack of physical social interaction more pronounced. Other countries reporting a lower median might have varying social norms or greater acceptance or use of digital communication to mitigate loneliness.

In regards to **life satisfaction level**, the similar medians suggest that participants from both groups had comparable levels of life satisfaction, despite the pandemic's challenges. The broader IQR and outliers for Indonesia might indicate a wider range of experiences impacting life satisfaction, such as economic factors or public health impacts.

As shown in the fourth plot for **boredom level**, the lower median for other countries could reflect a greater engagement in activities or adaptation to the lockdown lifestyle, whereas the slightly higher median for Indonesia might suggest a struggle to find engaging activities, possibly due to restrictions or available resources. The broader negative range suggests that some participants experienced less boredom, potentially due to engaging work-from-home setups or other absorbing activities.

In terms of **conspiracy beliefs level**, the slightly lower median for Indonesia could suggest less endorsement of conspiracy beliefs, possibly due to trust in local information sources or public health messaging. The higher median and IQR for other countries indicate a broader range of beliefs, which could be influenced by the diverse nature of information sources and cultural attitudes towards authority and public health directives. Outliers in Indonesia may represent individuals who are sceptical of mainstream narratives or have had different experiences influencing their perceptions.

b) To analyse how well various attributes predict pro-social attitudes in Indonesia, a linear regression model for each pro-social behaviour is used while also selecting appropriate predictors from the dataset. Tables consisting of summary statistics were also created for each of the four pro-social behaviours (see Appendix B).

For **'c19ProSo01'** (willingness to help others suffering from coronavirus), significant predictors include boredom level (**bor03**), support for reporting suspected cases (**c19RCA03**), avoiding crowded spaces (**c19perBeh02**), in-person contact (**isoOthPpl_inPerson**), and support for mandatory vaccination (**c19RCA01**). Higher boredom levels correspond to increased willingness to help, suggesting engagement leads to pro-social behaviour. Strong support for proactive measures like reporting cases and mandatory vaccination also indicates a higher propensity to assist those affected by the virus.

In **'c19ProSo02'** (willingness to make donations), key predictors are avoiding crowded spaces (**c19perBeh02**), happiness (**happy**), support for reporting suspected cases (**c19RCA03**), boredom level (**bor03**), and support for mandatory vaccination (**c19RCA01**). Those avoiding crowds and supporting proactive health measures show a higher likelihood of donating, likely reflecting a community-oriented mindset. Happiness and lower boredom are also linked to a greater inclination to contribute financially during the pandemic.

For **'c19ProSo03'** (willingness to protect vulnerable groups), significant predictors include support for reporting cases (**c19RCA03**), feeling left out (**lone03**), in-person contact (**isoOthPpl_inPerson**), boredom (**bor03**), gender (**males**), and conspiracy beliefs (**consp03**). High support for reporting cases and greater social contact correlate with a willingness to protect vulnerable groups, possibly driven by

a sense of community responsibility. Feelings of exclusion and certain conspiracy beliefs may also motivate individuals to take protective actions for others.

Lastly, **'c19ProSo04'** (willingness to make personal sacrifices to prevent the spread of coronavirus) highlights feeling left out (**lone03**), in-person contact (**isoOthPpl_inPerson**), support for reporting suspected cases (**c19RCA03**), self-quarantine (**c19perBeh03**), and frequent hand washing (**c19perBeh01**) as predictors. Those who feel socially connected or isolated are more willing to make sacrifices, indicating personal experiences significantly influence protective behaviours. Support for reporting and personal hygiene practices also suggest a readiness to take preventive actions for public health.

c) Similarly to 2b), to analyse how well various attributes predict pro-social attitudes in other countries, a linear regression model for each pro-social behaviour is used while also selecting appropriate predictors from the dataset. Tables consisting of summary statistics were also created for each of the four pro-social behaviours (see Appendix B).

It is observed that compared to Indonesia, other countries as a group tend to have more predictors that are deemed statistically significant. In regards to **'c19ProSo01'**, the variable **isoOthPpl_inPerson** (In-Person Contact), was identified as most significant. Similarly to Indonesia, other countries have a strong willingness to help other suffering from coronavirus when engaging with others in-person. However, other factors such as age, conspiracy (consp01 and consp02), education (edu5 and edu6), and gender (gender3), were identified as strongly significant in other countries, but not in Indonesia.

For **'c19ProSo02'**, the strongest predictor that was identified in other countries was the variable **isoFriends_online** (Online Contact with Friends). This suggests that people in other countries who are more connected online may feel more empowered or obliged to contribute financially to causes they learn about through these channels. In contrast, respondents from Indonesia may prioritise personal and direct actions over financial contributions due to economic conditions or personal responsibilities in pandemic response.

In regard to **'c19ProSo03'**, **c19perBeh03** (Self-Quarantine) was identified as the strongest predictor in other countries. This behaviour may correlate strongly with a willingness to protect others in other countries, compared to Indonesia, in which **bor03** (boredom level) was identified as their strongest predictor. The effectiveness and acceptance of such measures can also be influenced by how health regulations are framed and enforced in the country. If public health campaigns in Indonesia emphasise vigilance and reporting as key strategies against COVID-19, individuals may align their behaviours accordingly.

When it comes to **'c19ProSo04'**, the attribute **isoFriends_online** (Online Contact with Friends) was identified to be the most significant variable, while Indonesia's strongest predictor was feeling left out (lone03). The role of social media and online platforms were crucial in spreading information about how individual actions and sacrifices could contribute to communal safety. However, in many cultures, including possibly Indonesia, strong communal bonds and a collective sense of identity are prevalent. In such contexts, feelings of exclusion can prompt individuals to take actions that would reaffirm their commitment to community values, such as making personal sacrifices during a crisis. This offers an interesting glimpse into how different social and psychological factors influence pro-social behaviours across diverse cultural settings.

3. Focus country vs. cluster of similar countries

a) Firstly, datasets from World Health Organisation and United Nations Development Programme were gathered to compare five countries like Indonesia. These countries are **Malaysia, Singapore, Thailand, Vietnam,** and **Philippines**. Among these countries, four components were compared: population density, GDP per capita, political stability index, and human development index (see Appendix C).

Scatter plots containing the clustering of countries based on various socioeconomic indicators were also developed. The scatter plots depict clustering of the countries based on different indicators: GDP per capita, Population Density, Human Development Index (HDI), and Political Stability. Each point represents a country, and the colour represents the cluster to which the country has been assigned by the k-means algorithm.

Countries in the red cluster tend to have lower GDP per capita and lower HDI relative to the other clusters. They also tend to have a moderate population density and political stability. Indonesia, Vietnam, and the Philippines are in this cluster, suggesting that they share similar socio-economic profiles.

Countries in the blue cluster appears to be a middle ground between the red and green clusters. Countries like Malaysia and Thailand have higher GDP per capita than cluster 1 but lower than Singapore in cluster 3. They have relatively high HDI scores and moderate political stability, indicating a balance between economic capacity and social development.

For countries in the green cluster, Singapore stands out, characterized by significantly higher GDP per capita and HDI. It also has the highest political stability score and an extremely high population density. This cluster represents a higher level of economic and human development as well as political stability.

Appendix A: Descriptive Analysis and Pre-Processing

A.1 R Script for Descriptive Analysis and Pre-Processing

The following code contains the R script used to prepare the dataset for the analysis presented in the report.

```
# Task 1a)

rm(list = ls())

set.seed(32694547) # Student ID number

cvbase = read.csv("C:\\Users\\Joanna
Moy\\Desktop\\Y3S1\\FIT3152\\Assessments\\Assignment
1\\PsyCoronaBaselineExtract.csv")

cvbase <- cvbase[sample(nrow(cvbase), 40000), ]


# Display the dimensions of the dataset

cat("Dimensions of the dataset:", dim(cvbase), "\n\n")


# Shows the structure of the dataset to retrieve the data types and
missing values

cat("Structure of the dataset:\n")

str(cvbase)


# Extracting column names

column_names <- names(cvbase)


# Determining the data types of each column

data_types <- sapply(cvbase, class)


# Counting missing values in each column

missing_counts <- sapply(cvbase, function(x) sum(is.na(x)))


# Creating a data frame to display this information

structure_table <- data.frame(

  Column = column_names,

  DataType = data_types,

  MissingValues = missing_counts,

  stringsAsFactors = FALSE
```

```
)

# View the structure in a table format
print(structure_table)

# Summary statistics for numerical attributes
cat("Summary statistics for numerical attributes: \n")
summary(cvbase)

# Boxplots for psychological and behavioural responses
# Set graphical parameters
par(mfrow = c(3, 2), mar = c(4, 4, 2, 1))

# Boxplot for isolation offline
boxplot(cvbase$isoFriends_inPerson,
        main = "Isolation Offline – Friends & Relatives",
        ylab = "Days of Contact",
        xlab = "Offline Contact",
        col = "lightblue")

# Boxplot for isolation online
boxplot(cvbase$isoFriends_online,
        main = "Isolation Online – Friends & Relatives",
        ylab = "Days of Contact",
        xlab = "Online Contact",
        col = "lightgreen")

# Boxplot for loneliness
boxplot(cvbase$lone01,
        main = "Loneliness Level",
        ylab = "Loneliness Frequency",
        xlab = "Loneliness",
        col = "lightcoral")
```

```
# Boxplot for life satisfaction
boxplot(cvbase$happy,
        main = "Life Satisfaction Level",
        ylab = "Satisfaction Score",
        xlab = "Life Satisfaction",
        col = "lightgoldenrod")

# Boxplot for boredom level
boxplot(cvbase$bor01,
        main = "Boredom Level",
        ylab = "Boredom Agreement",
        xlab = "Boredom",
        col = "lightsalmon")

# Boxplot for conspiracy level
boxplot(cvbase$consp01,
        main = "Conspiracy Beliefs Level",
        ylab = "Conspiracy Agreement",
        xlab = "Conspiracy Beliefs",
        col = "lightgrey")

# Reset graphical parameters back to default
par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)


# Boxplot for corona variables
# Set up the graphical parameters
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))

# Boxplot for Corona Personal Behavior
boxplot(cvbase[, grep("c19perBeh", names(cvbase))],
        main = "Corona Personal Behavior",
        ylab = "Level of Agreement",
        col = rainbow(3))
```

```
# Boxplot for Corona Radical Action
boxplot(cvbase[, grep("c19RCA", names(cvbase))],
        main = "Corona Radical Action",
        ylab = "Level of Agreement",
        col = rainbow(3))

# Boxplot for Corona Proximity
boxplot(cvbase[, grep("coronaClose", names(cvbase))],
        main = "Corona Proximity",
        ylab = "Number of People Known",
        col = rainbow(6))

# Boxplot for Corona Prosocial Behavior
boxplot(cvbase[, grep("c19ProSo", names(cvbase))],
        main = "Corona Prosocial Behavior",
        ylab = "Level of Agreement",
        col = rainbow(4))

# Reset the graphical parameters back to default
par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)


# Histograms for demographic variables
# Set up the graphical parameters
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))

# Histogram for Age
hist(cvbase$age,
     main = "Distribution of Age",
     xlab = "Age",
     col = "skyblue",
     border = "white")
```

```
# Bar Chart for Gender
gender_table <- table(cvbase$gender)
barplot(gender_table,
        main = "Distribution of Gender",
        xlab = "Gender",
        ylab = "Frequency",
        col = "pink",
        border = "white")

# Employment Status as a continuous variable
# (Assuming each employstatus represents a range of hours worked per week)
employment_status <- rowSums(cvbase[,grep("employstatus", names(cvbase))] * 1:10, na.rm = TRUE)
hist(employment_status,
     main = "Distribution of Employment Status",
     xlab = "Employment Status (coded by hours worked per week)",
     col = "lightgreen",
     border = "white")

# Bar Chart for Education
education_table <- table(cvbase$edu)
barplot(education_table,
        main = "Distribution of Education Level",
        xlab = "Education Level",
        ylab = "Frequency",
        col = "lightblue",
        border = "white")

# Reset the graphical parameters back to default
par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)

# Set up the graphical parameters
par(mfrow = c(2, 1), mar = c(4, 4, 2, 1))
```

```
# Bar Chart for Country Self Report (Top 10 countries)
country_counts <- sort(table(cvbase$coded_country), decreasing =
TRUE)
top_countries <- head(country_counts, 10)
barplot(top_countries,
        main = "Top Countries of Survey Respondents",
        xlab = "Country",
        ylab = "Frequency",
        col = "lightblue",
        las = 2) # Orient the axis labels vertically

# Reset the graphical parameters back to default
par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)



# Bar Chart for Rank Order Life Variables
# Set up the graphical parameters
par(mfrow = c(2, 1), mar = c(4, 4, 2, 1))

# Assuming 'rankOrdLife' variables are 15abelled with letters A-F
and include N/A values
rank_vars <- grep("rankOrdLife", names(cvbase), value = TRUE)

# Create a function to convert letters to numeric ranks
convert_to_numeric <- function(x) {
  levels <- c("A", "B", "C", "D", "E", "F") # Specify the levels in
the order of the ranks
  as.numeric(factor(x, levels = levels))
}

# Apply the function to convert the rank data
rank_data_numeric <- sapply(cvbase[, rank_vars], convert_to_numeric)

# Calculate the median for each item, excluding N/A values
```

```
rank_medians <- apply(rank_data_numeric, 2, function(x) median(x,
na.rm = TRUE))


# Plot the barplot with the numeric medians

barplot(rank_medians,

        main = "Median Rank Order of Life Aspects",

        names.arg = c("Beauty", "Achievement", "Victory",
"Friendship", "Love", "Empathy"),

        xlab = "Life Aspects",

        ylab = "Median Numeric Rank",

        col = "lightgreen",

        las = 2) # Orient the axis labels vertically


# Reset the graphical parameters back to default

par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)
```

A.2 An Overview of Each Variable's Data Type and Preliminary Insights into Missing Data

Variables	DataType	MissingValues
employstatus_1	integer	34303
employstatus_2	integer	33229
employstatus_3	integer	29101
employstatus_4	integer	36594
employstatus_5	integer	37933
employstatus_6	integer	36884
employstatus_7	integer	36340
employstatus_8	integer	39262
employstatus_9	integer	31925
employstatus_10	integer	39069
isoFriends_inPerson	integer	324
isoOthPpl_inPerson	integer	516
isoFriends_online	integer	946
isoOthPpl_online	integer	1164
lone01	integer	88
lone02	integer	123
lone03	integer	146
happy	integer	506
lifeSat	integer	117
MLQ	integer	118
bor01	integer	162
bor02	integer	174
bor03	integer	179

consp01	integer	1504
consp02	integer	1532
consp03	integer	1555
rankOrdLife_1	character	1778
rankOrdLife_2	character	1778
rankOrdLife_3	character	1778
rankOrdLife_4	character	1778
rankOrdLife_5	character	1778
rankOrdLife_6	character	1779
c19perBeh01	integer	124
c19perBeh02	integer	132
c19perBeh03	integer	131
c19RCA01	integer	130
c19RCA02	integer	140
c19RCA03	integer	149
coronaClose_1	integer	39478
coronaClose_2	integer	38719
coronaClose_3	integer	38434
coronaClose_4	integer	35109
coronaClose_5	integer	35510
coronaClose_6	integer	10662
gender	integer	212
age	integer	226
edu	integer	269
coded_country	character	0
c19ProSo01	integer	129
c19ProSo02	integer	138
c19ProSo03	integer	144
c19ProSo04	integer	146

Table A2.1. An Overview of Each Variable's Data Type and Preliminary Insights into Missing Data

Appendix B: Focus country vs. all other countries as a group

B.1 R Script for the Comparison of Focus Country vs. Other Countries

```
#######----------TASK 2----------#########

# Task 2a)

# Create a database for Indonesia

IndoDataBase <- cvbase[cvbase$coded_country == "Indonesia",]

# Remove rows that have all "NA" values

IndoDataBase <- IndoDataBase[rowSums(is.na(IndoDataBase)) !=
ncol(IndoDataBase), ]


# Create a database for other countries

NotIndoDataBase <- cvbase[cvbase$coded_country != "Indonesia",]

# Remove rows that have all "NA" values for other countries data as
well

NotIndoDataBase <-
NotIndoDataBase[rowSums(is.na(NotIndoDataBase)) !=
ncol(NotIndoDataBase), ]


# Set graphical parameters

par(mfrow = c(3, 2), mar = c(4, 4, 2, 1))


# Create boxplot functions for comparison

plot_box_comparison <- function(indo_data, other_data, col_names,
main_title) {

  data_combined <- list(Indonesia = indo_data[[col_names]],
Other_Countries = other_data[[col_names]])

  boxplot(data_combined, main = main_title, ylab = "Scores", col =
c("lightblue", "salmon"), names = c("Indonesia", "Others"))

}


# Apply the function to the attributes

attributes <- c("isoFriends_inPerson", "isoFriends_online", "lone01",
"happy", "bor01", "consp01")

titles <- c("Isolation Offline - Friends & Relatives", "Isolation
Online - Friends & Relatives",

            "Loneliness Level", "Life Satisfaction Level", "Boredom
Level", "Conspiracy Beliefs Level")
```

```
for (I in 1:length(attributes)) {
  plot_box_comparison(IndoDataBase, NotIndoDataBase, attributes[i],
titles[i])
}

# Reset graphical parameters back to default
par(mfrow = c(1, 1), mar = c(5, 4, 4, 2) + 0.1)



# Task 2b)
# Converting factors
IndoDataBase$gender <- factor(IndoDataBase$gender)
IndoDataBase$edu <- factor(IndoDataBase$edu)

# Defining predictors: Psychological, Social, Behavioral,
Demographic
psychological_vars <- c("lone01", "lone02", "lone03", "happy",
"lifeSat",
                        "consp01", "consp02", "consp03", "bor01",
"bor02", "bor03")

social_behavior_vars <- c("isoFriends_inPerson",
"isoOthPpl_inPerson",
                          "isoFriends_online", "isoOthPpl_online",
                          "c19perBeh01", "c19perBeh02",
"c19perBeh03",
                          "c19RCA01", "c19RCA02", "c19RCA03")

demographic_vars <- c("age", "gender", "edu")

# Combining all predictors
all_predictors <- c(psychological_vars, social_behavior_vars,
demographic_vars)

# Prepare the complete formula string for regression
```

```
formula_string <- paste("c19ProSo01 ~", paste(all_predictors,
collapse = "+"))


# Running the linear regression model

model_proso01 <- lm(formula_string, data = IndoDataBase)

summary(model_proso01)


# Repeat the regression for each pro-social behavior variable

pro_social_vars <- c("c19ProSo01", "c19ProSo02", "c19ProSo03",
"c19ProSo04")

models <- lapply(pro_social_vars, function(ps_var) {

  formula_string <- paste(ps_var, "~", paste(all_predictors,
collapse = "+"))

  lm(as.formula(formula_string), data = IndoDataBase)

})


# Displaying the summaries of each model

model_summaries <- lapply(models, summary)


print(model_summaries[[1]])  # Summary for c19ProSo01

print(model_summaries[[2]])  # Summary for c19ProSo02

print(model_summaries[[3]])  # Summary for c19ProSo03

print(model_summaries[[4]])  # Summary for c19ProSo04



# Task 2c)

# Converting factors (Ensure categorical variables are treated
appropriately)

NotIndoDataBase$gender <- factor(NotIndoDataBase$gender)

NotIndoDataBase$edu <- factor(NotIndoDataBase$edu)


# Defining predictors: Psychological, Social, Behavioral,
Demographic

psychological_vars <- c("lone01", "lone02", "lone03", "happy",
"lifeSat",
```

```
                        "consp01", "consp02", "consp03", "bor01",
"bor02", "bor03")


social_behavior_vars <- c("isoFriends_inPerson",
"isoOthPpl_inPerson",
                          "isoFriends_online", "isoOthPpl_online",
                          "c19perBeh01", "c19perBeh02",
"c19perBeh03",
                          "c19RCA01", "c19RCA02", "c19RCA03")


demographic_vars <- c("age", "gender", "edu")


# Combining all predictors
all_predictors <- c(psychological_vars, social_behavior_vars,
demographic_vars)


# Running regression models for all four pro-social behavior
variables
pro_social_vars <- c("c19ProSo01", "c19ProSo02", "c19ProSo03",
"c19ProSo04")
models_not_indo <- lapply(pro_social_vars, function(ps_var) {
  formula_string <- paste(ps_var, "~", paste(all_predictors,
collapse = "+"))
  lm(as.formula(formula_string), data = NotIndoDataBase)
})


# Displaying the summaries of each model
model_summaries_not_indo <- lapply(models_not_indo, summary)


# Print the summary of the models
print(model_summaries_not_indo[[1]])  # Summary for c19ProSo01
print(model_summaries_not_indo[[2]])  # Summary for c19ProSo02
print(model_summaries_not_indo[[3]])  # Summary for c19ProSo03
print(model_summaries_not_indo[[4]])  # Summary for c19ProSo04
```

B.2 Summary Statistics on Attributes that Predict Pro-Social Behaviours for Indonesia

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.98366	0.43204	-2.277	0.022949	*
lone01	0.072125	0.040426	1.784	0.074619	.
lone02	-0.03691	0.035323	-1.045	0.296201	
lone03	0.026767	0.038936	0.687	0.491907	
happy	-0.00774	0.025487	-0.304	0.761508	
lifeSat	0.082316	0.043755	1.881	0.060139	.
consp01	0.00178	0.017669	0.101	0.919771	
consp02	0.001325	0.017402	0.076	0.939306	
consp03	0.021189	0.013991	1.514	0.130134	
bor01	-0.01094	0.021335	-0.513	0.608166	
bor02	-0.03089	0.019257	-1.604	0.108907	
bor03	0.096862	0.024315	3.984	7.13E-05	***
isoFriends_inPerson	0.010324	0.013878	0.744	0.457066	
isoOthPpl_inPerson	0.04141	0.016969	2.44	0.014794	*
isoFriends_online	0.030021	0.014962	2.006	0.044997	*
isoOthPpl_online	-0.0066	0.013732	-0.481	0.630618	
c19perBeh01	0.089875	0.051354	1.75	0.080316	.
c19perBeh02	0.143234	0.05431	2.637	0.008447	**
c19perBeh03	0.023746	0.03556	0.668	0.504388	
c19RCA01	0.086163	0.035607	2.42	0.015654	*
c19RCA02	-0.00839	0.04705	-0.178	0.858518	
c19RCA03	0.151416	0.042839	3.535	0.000422	***
age	0.032329	0.028116	1.15	0.250393	
gender2	0.019526	0.067793	0.288	0.773374	
gender3	-1.32894	0.5132	-2.59	0.00971	**
edu2	0.131256	0.337604	0.389	0.697492	
edu3	0.231971	0.360399	0.644	0.519905	
edu4	0.157497	0.368847	0.427	0.669445	
edu5	0.21743	0.339414	0.641	0.521883	
edu6	0.344826	0.34938	0.987	0.323829	
edu7	0.610479	0.385231	1.585	0.113255	

Table B2.1. Summary Statistic for c19ProSo01

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.14346	0.360761	-3.17	0.001559	**
lone01	-0.00022	0.03376	-0.007	0.994747	
lone02	-0.00799	0.029505	-0.271	0.786486	
lone03	0.043692	0.032501	1.344	0.179066	
happy	0.088862	0.021281	4.176	3.15E-05	***
lifeSat	-0.01149	0.036621	-0.314	0.753668	
consp01	0.004919	0.014761	0.333	0.739004	
consp02	0.001967	0.014537	0.135	0.892366	
consp03	-0.00204	0.011694	-0.174	0.861568	
bor01	-0.0107	0.017868	-0.599	0.549323	
bor02	-0.0016	0.016103	-0.1	0.920688	
bor03	0.066293	0.020387	3.252	0.001174	**
isoFriends_inPerson	0.00781	0.011594	0.674	0.500616	
isoOthPpl_inPerson	-0.00441	0.014172	-0.311	0.755629	
isoFriends_online	0.023059	0.012498	1.845	0.065243	.
isoOthPpl_online	-0.00034	0.011462	-0.03	0.97637	
c19perBeh01	0.105512	0.043211	2.442	0.014735	*
c19perBeh02	0.286687	0.045485	6.303	3.89E-10	***
c19perBeh03	-0.01078	0.029751	-0.362	0.717098	
c19RCA01	0.096326	0.029742	3.239	0.001228	**
c19RCA02	0.030916	0.039296	0.787	0.431566	
c19RCA03	0.122973	0.035825	3.433	0.000615	***
age	-0.06062	0.02352	-2.577	0.010056	*
gender2	-0.05345	0.056626	-0.944	0.345355	
gender3	-0.4078	0.470727	-0.866	0.386458	
edu2	0.653063	0.282324	2.313	0.020856	*
edu3	0.555283	0.301433	1.842	0.065663	.
edu4	0.359466	0.308508	1.165	0.244144	
edu5	0.684339	0.283887	2.411	0.016053	*
edu6	0.857468	0.29263	2.93	0.003442	**
edu7	0.906225	0.322238	2.812	0.004987	**

Table B2.2. Summary Statistic for c19ProSo02

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.57773	0.504542	-3.127	0.0018	**
lone01	-0.00249	0.047222	-0.053	0.95804	
lone02	-0.00268	0.041264	-0.065	0.94825	
lone03	0.115472	0.04544	2.541	0.01115	*
happy	-0.01693	0.029762	-0.569	0.56959	
lifeSat	0.076178	0.051096	1.491	1.36E-01	
consp01	-0.0249	0.02064	-1.207	0.22781	
consp02	0.033167	0.020329	1.631	0.10301	
consp03	0.033884	0.016345	2.073	0.03835	*
bor01	-0.00506	0.0249	-0.203	0.83909	
bor02	-0.00764	0.022496	-0.34	0.73416	
bor03	0.063783	0.028403	2.246	0.02488	*
isoFriends_inPerson	-0.01946	0.016212	-1.2	0.23018	
isoOthPpl_inPerson	0.049413	0.019814	2.494	0.01275	*
isoFriends_online	0.009622	0.017475	0.551	0.58198	
isoOthPpl_online	0.030396	0.016033	1.896	0.05818	.
c19perBeh01	0.124043	0.05995	2.069	0.03872	*
c19perBeh02	0.055621	0.063413	0.877	0.38057	
c19perBeh03	0.022439	0.04154	0.54	5.89E-01	
c19RCA01	0.081191	0.041592	1.952	0.05113	.
c19RCA02	-0.05702	0.054955	-1.038	0.29966	
c19RCA03	0.151108	0.050043	3.02	0.00258	**
age	0.017879	0.032843	0.544	0.58627	
gender2	0.16843	0.079187	2.127	0.03359	*
gender3	-0.0056	0.599516	-0.009	0.99254	
edu2	0.479449	0.394379	1.216	0.2243	
edu3	0.520319	0.421016	1.236	0.21671	
edu4	0.513474	0.430886	1.192	0.23359	
edu5	0.511638	0.396503	1.29	0.19713	
edu6	0.592311	0.408145	1.451	0.14694	
edu7	0.628358	0.450026	1.396	0.16285	

Table B2.3. Summary Statistic for c19ProSo03

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.15904	0.50077	-2.315	0.02078	*
lone01	-0.04385	0.046869	-0.936	0.34967	
lone02	0.083047	0.040955	2.028	0.04277	*
lone03	0.13598	0.045101	3.015	0.00262	**
happy	0.059682	0.02954	2.02	0.04353	*
lifeSat	-0.08874	0.050714	-1.75	0.08036	.
consp01	-0.03126	0.020486	-1.526	0.12722	
consp02	0.040099	0.020177	1.987	0.04707	*
consp03	0.011834	0.016223	0.729	0.46584	
bor01	0.009588	0.024714	0.388	0.6981	
bor02	-0.02342	0.022328	-1.049	0.29441	
bor03	0.048611	0.028191	1.724	0.08486	.
isoFriends_inPerson	-0.0164	0.016091	-1.019	0.30829	
isoOthPpl_inPerson	0.051456	0.019666	2.616	0.00898	**
isoFriends_online	0.008637	0.017344	0.498	0.61857	
isoOthPpl_online	-0.00236	0.015913	-0.148	0.88235	
c19perBeh01	0.155254	0.059502	2.609	0.00917	**
c19perBeh02	0.022295	0.062939	0.354	0.72322	
c19perBeh03	0.116737	0.04123	2.831	0.0047	**
c19RCA01	0.060799	0.041281	1.473	0.14102	
c19RCA02	0.051921	0.054545	0.952	0.34131	
c19RCA03	0.125976	0.049669	2.536	0.01131	*
age	0.013395	0.032598	0.411	0.6812	
gender2	0.019538	0.078595	0.249	0.80371	
gender3	-0.0843	0.595035	-0.142	0.88736	
edu2	0.328262	0.391431	0.839	0.40182	
edu3	0.343001	0.417869	0.821	0.41188	
edu4	0.289757	0.427665	0.678	0.49818	
edu5	0.340485	0.393539	0.865	0.38708	
edu6	0.235836	0.405094	0.582	0.56054	
edu7	0.362252	0.446662	0.811	0.41749	

Table B2.4. Summary Statistic for c19ProSo04

B.3 Summary Statistics on Attributes that Predict Pro-Social Behaviours for Other Countries

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.86733	0.085154	-10.185	< 2e-16	***
lone01	0.06002	0.009609	6.246	4.25E-10	***
lone02	-0.02248	0.008428	-2.668	0.007643	**
lone03	0.019644	0.009091	2.161	0.030721	*
happy	0.023466	0.005324	4.407	1.05E-05	***
lifeSat	0.100902	0.008638	11.681	< 2e-16	***
consp01	0.020079	0.003941	5.095	3.51E-07	***
consp02	-0.02241	0.004127	-5.43	5.66E-08	***
consp03	0.012294	0.003133	3.923	8.74E-05	***
bor01	0.026587	0.005486	4.846	1.26E-06	***
bor02	0.010543	0.005503	1.916	0.055409	.
bor03	0.044741	0.004953	9.034	< 2e-16	***
isoFriends_inPerson	0.017004	0.003539	4.805	1.55E-06	***
isoOthPpl_inPerson	0.031829	0.003955	8.049	8.64E-16	***
isoFriends_online	0.032204	0.003524	9.139	< 2e-16	***
isoOthPpl_online	0.018382	0.003229	5.693	1.26E-08	***
c19perBeh01	0.139709	0.008916	15.67	< 2e-16	***
c19perBeh02	0.116417	0.010803	10.776	< 2e-16	***
c19perBeh03	0.003753	0.006305	0.595	0.551692	
c19RCA01	0.063006	0.004745	13.28	< 2e-16	***
c19RCA02	0.03985	0.007688	5.183	2.19E-07	***
c19RCA03	-0.03326	0.005071	-6.558	5.54E-11	***
age	-0.02758	0.00499	-5.528	3.27E-08	***
gender2	0.053693	0.016132	3.328	0.000875	***
gender3	0.222342	0.109796	2.025	0.04287	*
edu2	-0.04688	0.068069	-0.689	0.490969	
edu3	0.058631	0.069006	0.85	0.395526	
edu4	0.021371	0.066423	0.322	0.747646	
edu5	0.159344	0.066085	2.411	0.015906	*
edu6	0.15278	0.067331	2.269	0.023269	*
edu7	0.296767	0.072853	4.074	4.64E-05	***

Table B3.1. Summary Statistic for c19ProSo01

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.3245	0.091881	-14.415	< 2e-16	***
lone01	0.055036	0.010374	5.305	1.13E-07	***
lone02	-0.0275	0.009099	-3.023	0.00251	**
lone03	0.03129	0.009815	3.188	0.00143	**
happy	0.031516	0.005748	5.483	4.22E-08	***
lifeSat	0.174407	0.009327	18.7	< 2e-16	***
consp01	-0.01035	0.004255	-2.433	0.01498	*
consp02	-0.03261	0.004456	-7.317	2.58E-13	***
consp03	0.017613	0.003383	5.207	1.93E-07	***
bor01	0.053292	0.005922	8.999	< 2e-16	***
bor02	0.005962	0.00594	1.004	0.31559	
bor03	0.032492	0.005347	6.077	1.24E-09	***
isoFriends_inPerson	0.031429	0.003821	8.226	< 2e-16	***
isoOthPpl_inPerson	0.005393	0.00427	1.263	0.20656	
isoFriends_online	0.029856	0.003805	7.847	4.39E-15	***
isoOthPpl_online	0.029156	0.003486	8.363	< 2e-16	***
c19perBeh01	0.108266	0.009626	11.248	< 2e-16	***
c19perBeh02	0.088311	0.011663	7.572	3.78E-14	***
c19perBeh03	0.04912	0.006807	7.217	5.44E-13	***
c19RCA01	0.111025	0.005123	21.673	< 2e-16	***
c19RCA02	0.011409	0.0083	1.375	0.16927	
c19RCA03	0.04922	0.005475	8.99	< 2e-16	***
age	-0.04527	0.005387	-8.404	< 2e-16	***
gender2	-0.08198	0.017416	-4.707	2.52E-06	***
gender3	-0.084	0.118533	-0.709	0.47856	
edu2	-0.01666	0.073415	-0.227	0.82043	
edu3	-0.03407	0.074427	-0.458	0.64712	
edu4	0.087029	0.071638	1.215	0.22443	
edu5	0.292518	0.071273	4.104	4.07E-05	***
edu6	0.3316	0.072619	4.566	4.98E-06	***
edu7	0.539429	0.07859	6.864	6.82E-12	***

Table B3.2. Summary Statistic for c19ProSo02

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.06959	0.09673	-11.057	< 2e-16	***
lone01	0.02272	0.010922	2.08	0.037507	*
lone02	-0.01079	0.009579	-1.127	0.259925	
lone03	0.068957	0.010331	6.675	2.51E-11	***
happy	0.017476	0.006051	2.888	0.003878	**
lifeSat	0.14019	0.009819	14.278	< 2e-16	***
consp01	0.01036	0.004479	2.313	0.020737	*
consp02	-0.04537	0.004691	-9.672	< 2e-16	***
consp03	0.012639	0.003561	3.549	0.000387	***
bor01	0.021333	0.006235	3.422	0.000623	***
bor02	0.021539	0.006254	3.444	0.000574	***
bor03	0.04835	0.005629	8.59	< 2e-16	***
isoFriends_inPerson	0.019845	0.004022	4.934	8.09E-07	***
isoOthPpl_inPerson	0.029545	0.004495	6.573	5.01E-11	***
isoFriends_online	0.017074	0.004005	4.263	2.02E-05	***
isoOthPpl_online	0.03158	0.00367	8.605	< 2e-16	***
c19perBeh01	0.09751	0.010132	9.624	< 2e-16	***
c19perBeh02	0.11011	0.012278	8.968	< 2e-16	***
c19perBeh03	0.054325	0.007165	7.582	3.50E-14	***
c19RCA01	0.081356	0.005393	15.086	< 2e-16	***
c19RCA02	0.032466	0.008738	3.715	0.000203	***
c19RCA03	-0.04502	0.005764	-7.812	5.81E-15	***
age	-0.09175	0.005671	-16.179	< 2e-16	***
gender2	0.002833	0.018335	0.154	0.877217	
gender3	0.255934	0.124783	2.051	0.040272	*
edu2	0.038361	0.077285	0.496	0.619644	
edu3	0.05061	0.078353	0.646	0.518335	
edu4	0.086633	0.075415	1.149	0.250667	
edu5	0.260238	0.075031	3.468	0.000524	***
edu6	0.291071	0.07645	3.807	0.000141	***
edu7	0.550986	0.08273	6.66	2.78E-11	***

Table B3.3. Summary Statistic for c19ProSo03

Variables	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.73948	0.087804	-8.422	< 2e-16	***
lone01	-0.01743	0.009906	-1.759	0.078534	.
lone02	0.036348	0.008688	4.183	2.88E-05	***
lone03	0.050345	0.009372	5.372	7.84E-08	***
happy	-0.00449	0.005489	-0.818	0.413344	
lifeSat	0.104157	0.008907	11.694	< 2e-16	***
consp01	0.030237	0.004063	7.442	1.01E-13	***
consp02	-0.02338	0.004255	-5.494	3.95E-08	***
consp03	-0.0034	0.00323	-1.052	0.292836	
bor01	-0.0115	0.005656	-2.034	0.041956	*
bor02	0.03373	0.005673	5.945	2.79E-09	***
bor03	0.033065	0.005106	6.476	9.53E-11	***
isoFriends_inPerson	-0.0032	0.003648	-0.876	0.38102	
isoOthPpl_inPerson	0.015624	0.004077	3.832	0.000127	***
isoFriends_online	0.028193	0.003633	7.76	8.72E-15	***
isoOthPpl_online	0.002054	0.003329	0.617	0.537138	
c19perBeh01	0.093416	0.009191	10.164	< 2e-16	***
c19perBeh02	0.204655	0.011138	18.375	< 2e-16	***
c19perBeh03	0.099356	0.0065	15.285	< 2e-16	***
c19RCA01	0.051715	0.004892	10.572	< 2e-16	***
c19RCA02	0.14165	0.007926	17.871	< 2e-16	***
c19RCA03	-0.07212	0.005228	-13.795	< 2e-16	***
age	0.025759	0.005143	5.008	5.52E-07	***
gender2	-0.04352	0.016631	-2.617	0.008885	**
gender3	0.192382	0.113524	1.695	0.090153	.
edu2	-0.04852	0.070173	-0.691	0.489282	
edu3	-0.03108	0.071139	-0.437	0.662213	
edu4	0.035251	0.068478	0.515	0.606715	
edu5	0.132161	0.068129	1.94	0.052405	.
edu6	0.168113	0.069412	2.422	0.015442	*
edu7	0.254508	0.075106	3.389	0.000703	***

Table B3.4. Summary Statistic for c19ProSo04

Appendix C: Focus country vs. cluster of similar countries

C.1 R Script for the Analysis of 6 Southeast Asian Countries' 4 attributes

```
#######----------TASK 3----------#########

# Task 3a)

# Comparing Population Density

pop_density_data <- read.csv("C:\\Users\\Joanna
Moy\\Desktop\\Y3S1\\FIT3152\\Assessments\\Assignment
1\\Datasets\\API_EN.POP.DNST_DS2_en_csv_v2_321.csv", skip = 4)


# Select the relevant columns

relevant_data <- pop_density_data[, c("Country.Name", "X2020")]


# Clean data by removing NAs

clean_data <- na.omit(relevant_data)


# Find Indonesia's population density

indonesia_density <- clean_data[clean_data$Country.Name ==
"Indonesia", "X2020"]

indonesia_density


# Example countries to compare

countries_to_compare <- c("Indonesia", "Malaysia", "Singapore",
"Thailand", "Viet Nam", "Philippines")


# Extract population densities for the selected countries

selected_densities <- clean_data[clean_data$Country.Name %in%
countries_to_compare, ]


# Print the population densities

print(selected_densities)



# Comparing GDP
```

```
# Load the GDP data

gdp_data <- read.csv("C:\\Users\\Joanna
Moy\\Desktop\\Y3S1\\FIT3152\\Assessments\\Assignment
1\\Datasets\\API_NY.GDP.PCAP.CD_DS2_en_csv_v2_133.csv", skip = 4)


# Select the relevant columns

relevant_gdp_data <- gdp_data[, c("Country.Name", "X2020")]


# Clean data by removing NAs

clean_gdp_data <- na.omit(relevant_gdp_data)


# Example countries to include with Indonesia

countries_to_compare <- c("Indonesia", "Malaysia", "Singapore",
"Thailand", "Viet Nam", "Philippines")


# Extract GDP per capita for the selected countries

selected_gdp <- clean_gdp_data[clean_gdp_data$Country.Name %in%
countries_to_compare, ]


# Print the GDP per capita

print(selected_gdp)




# Comparing Political Stability Index

# Load the Political Stability Index data

political_stability_data <- read.csv("C:\\Users\\Joanna
Moy\\Desktop\\Y3S1\\FIT3152\\Assessments\\Assignment
1\\Datasets\\API_PV.PER.RNK_DS2_en_csv_v2_52652.csv", skip = 4)


# Select the relevant columns

relevant_political_data <- political_stability_data[,
c("Country.Name", "X2020")]


# Clean data by removing NAs

clean_political_data <- na.omit(relevant_political_data)
```

```
# Example countries to include with Indonesia

countries_to_compare <- c("Indonesia", "Malaysia", "Singapore",
"Thailand", "Viet Nam", "Philippines")


# Extract Political Stability Index for the selected countries

selected_political_stability <-
clean_political_data[clean_political_data$Country.Name %in%
countries_to_compare, ]


# Print the Political Stability Index values

print(selected_political_stability)


# Comparing Human Development Index (HDI)

# Load necessary library

library(readxl)


# Load the HDI data from Excel file, skipping the first 4 rows

hdi_data <- read_excel("C:\\Users\\Joanna
Moy\\Desktop\\Y3S1\\FIT3152\\Assessments\\Assignment
1\\Datasets\\HDR23-24_Statistical_Annex_HDI_Trends_Table.xlsx", skip
= 4)


relevant_hdi_data <- hdi_data[, c("Country", "2020")]


# Clean data by removing NAs

clean_hdi_data <- na.omit(relevant_hdi_data)


# Example countries to include with Indonesia

countries_to_compare <- c("Indonesia", "Malaysia", "Singapore",
"Thailand", "Viet Nam", "Philippines")


# Extract HDI for the selected countries

selected_hdi <- clean_hdi_data[clean_hdi_data$Country %in%
countries_to_compare, ]
```

```
# Print the HDI values
print(selected_hdi)



# Perform clustering
countries <- c("Indonesia", "Malaysia", "Philippines", "Singapore",
"Thailand", "Vietnam")
gdp_per_capita <- c(3895.618, 10164.34, 3224.423, 61273.99, 7001.785,
3586.347)
population_density <- c(144.7964, 101.05, 376.2651, 7918.951,
139.9042, 308.3591)
hdi <- c(0.712, 0.802, 0.705, 0.942, 0.8, 0.725)
political_stability <- c(28.30189, 51.88679, 19.81132, 97.16982,
25.9434, 47.16981)


# Combine into a dataframe
country_data <- data.frame(
  Country = countries,
  GDP = gdp_per_capita,
  PopDensity = population_density,
  HDI = hdi,
  PolStability = political_stability
)


# Standardizing the data
country_data_scaled <- as.data.frame(scale(country_data[, -1]))


# Clustering with k-means
set.seed(123)  # Setting seed for reproducibility
kmeans_result <- kmeans(country_data_scaled, centers=3, nstart=25)


# Adding cluster labels back to the original data
country_data$Cluster <- kmeans_result$cluster
```

```
# Find Indonesia's cluster
indonesia_cluster <- country_data$Cluster[country_data$Country ==
"Indonesia"]


# Get similar countries
similar_countries <- country_data$Country[country_data$Cluster ==
indonesia_cluster]
print(similar_countries)


# Display cluster characteristics
aggregate(country_data_scaled, by=list(country_data$Cluster),
FUN=mean)


# Create scatter plots
# GDP vs. Population Density
p1 <- ggplot(country_data, aes(x = GDP, y = PopDensity, color =
Cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = Country), vjust = 1.5, color = "black") +
  labs(title = "GDP vs. Population Density", x = "GDP per Capita", y
= "Population Density") +
  theme_minimal() +
  scale_color_manual(values = c("red", "blue", "green"))


# GDP vs. HDI
p2 <- ggplot(country_data, aes(x = GDP, y = HDI, color = Cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = Country), vjust = 1.5, color = "black") +
  labs(title = "GDP vs. HDI", x = "GDP per Capita", y = "Human
Development Index") +
  theme_minimal() +
  scale_color_manual(values = c("red", "blue", "green"))


# GDP vs. Political Stability
```

```
p3 <- ggplot(country_data, aes(x = GDP, y = PolStability, color =
Cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = Country), vjust = 1.5, color = "black") +
  labs(title = "GDP vs. Political Stability", x = "GDP per Capita",
y = "Political Stability") +
  theme_minimal() +
  scale_color_manual(values = c("red", "blue", "green"))

# HDI vs. Political Stability
p4 <- ggplot(country_data, aes(x = HDI, y = PolStability, color =
Cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = Country), vjust = 1.5, color = "black") +
  labs(title = "HDI vs. Political Stability", x = "Human Development
Index", y = "Political Stability") +
  theme_minimal() +http://127.0.0.1:24705/graphics/256d9ac2-92e1-
414f-8541-439191c90391.png
  scale_color_manual(values = c("red", "blue", "green"))

# Population Density vs. HDI
p5 <- ggplot(country_data, aes(x = PopDensity, y = HDI, color =
Cluster)) +
  geom_point(size = 4) +
  geom_text(aes(label = Country), vjust = 1.5, color = "black") +
  labs(title = "Population Density vs. HDI", x = "Population
Density", y = "Human Development Index") +
  theme_minimal() +
  scale_color_manual(values = c("red", "blue", "green"))

# Display all plots
library(gridExtra)
grid.arrange(p1, p2, p3, p4, p5, ncol = 2)
```

C.2 Tables Consisting of 6 Southeast Asian Countries' 4 attributes

Population density data retrieved from
https://data.worldbank.org/indicator/EN.POP.DNST?end=2021&start=2016

Country Name	2020
Indonesia	144.7964
Malaysia	101.05
Philippines	376.2651
Singapore	7918.951
Thailand	139.9042
Viet Nam	308.3591

Table C.2.1 Population Density of 5 Southeast Asian Countries Compared with Indonesia

Gross domestic per capita data retrieved from
https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Country Name	2020
Indonesia	3895.618
Malaysia	10164.34
Philippines	3224.423
Singapore	61273.99
Thailand	7001.785
Viet Nam	3586.347

Table C.2.2 GDP per Capita of 5 Southeast Asian Countries Compared with Indonesia

Political stability index data retrieved from
https://data.worldbank.org/indicator/PV.PER.RNK?end=2022&start=2019

Country Name	2020
Indonesia	28.30189
Malaysia	51.88679
Philippines	19.81132
Singapore	97.16982
Thailand	25.9434

Viet Nam	47.16981

Table C.2.3 Political Stability Index of 5 Southeast Asian Countries Compared with Indonesia

Human development index data retrieved from

https://hdr.undp.org/data-center/documentation-and-downloads

Country Name	2020
Singapore	0.942
Malaysia	0.802
Thailand	0.8
Indonesia	0.712
Philippines	0.705
Viet Nam	0.725

Table C.2.4 Human Development Index of 5 Southeast Asian Countries Compared with Indonesia

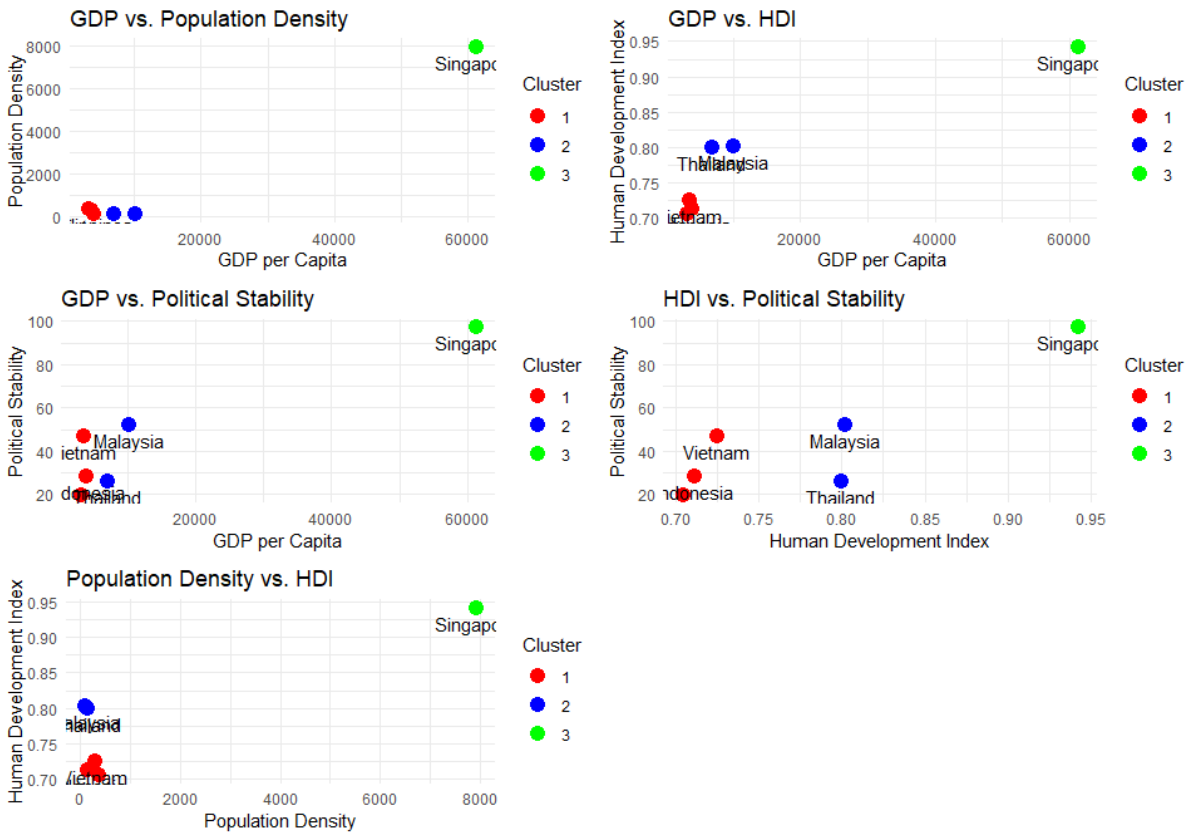


Figure C.2 Scatter Plot of Countries Based on Various Socioeconomic Indicators.