FIT3152 Data analytics – 2023: Assignment 1

Lee Zhen Xuan 31860532

Q1 Descriptive analysis and pre-processing

Dimension, data types, distribution of numerical attributes, variety of non-numerical (text) attributes

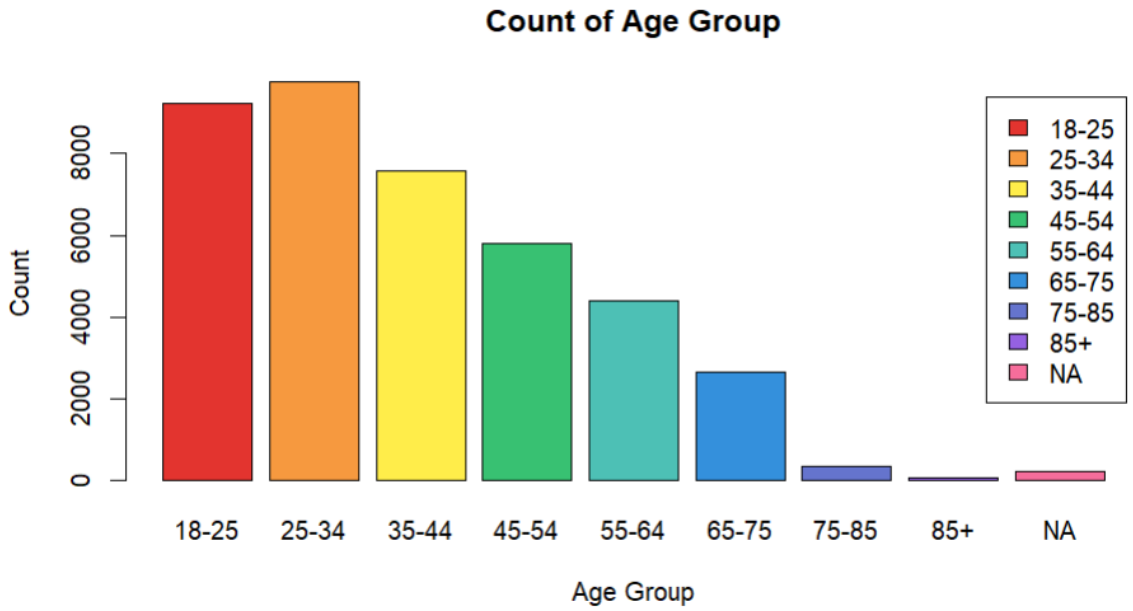
a) The dataset contains 54 columns, and 40000 rows after creating the individual data. Most of the column's datatypes are numerical values as integers, except for the column "coded_country". These numerical values are mostly scale ratings for each of the variables. The only non-numerical variable is "coded_country", which describes the country where the respondents are from.

Missing values

There is some missing values/"NA" present in the dataset, particularly in the "Job insecurity" concept, which consists of the columns "jbInsec01", "jbInsec02", "jbInsec03" and "jbInsec04", and the "Employment status" concept which consists of the column from "employstatus_1" to "employstatus_10".

The missing values particularly in "Job insecurity" and "Employment Status" columns can be explained as follows:

- "Job insecurity": Most of the missing values comes from the Age Group of 18-25, and 25-34 years old, where these respondents may not be working yet. This can be further elaborated by the numbers where these columns from the "Job insecurity" variables from the age group of 18-25 and 25-34 occupies most of the NA values, which are a total of 5561, 4868, 3948 and 6427 respondents for the columns "jbInsec01", "jbInsec02", "jbInsec03" and "jbInsec04" respectively.



- "Employment Status": The employment status from 1 to 10 are mutually exclusive, where if one of these variables have a "1", the other 9 variables would have "NA" values. This could be further elaborated after grouping them based on the "1" present in each of the rows for every employment status of the respondents. From the table below, these values add up exactly to the number of respondents from the dataset after grouping them, which is 40000 respondents. A table can be plotted to show the count for each of the employment status:

employment_status	COUNT
1	6066
2	6554
3	10528
4	3306
5	1870
6	2082
7	3240
8	502
9	5695
10	157

Pre-processing or data manipulation

b) One of the pre-processing/data manipulations done for the dataset is grouping the "Employment status" based on the responses. The employment status of the dataset from "employstatus_1" to "employstatus_10" are mutually exclusive, where for each of the responses, if there exists a "1" in one of the employment statuses, the other nine should have a "NA" value. Therefore, a new column with the name "employment_status" has been created, which are integer/numerical values that ranges from 1 to 10 to specify the employment status for the respondents. For example, if one of the responses at the "employstatus_2" have a value of "1", the value at the newly created column "employment_status" would have a value of "2". After removing all the employstatus from 1 to 10 and adding a new "employment_status" column, there would be a total of 45 columns.

Q2 Focus country vs all other countries as a group

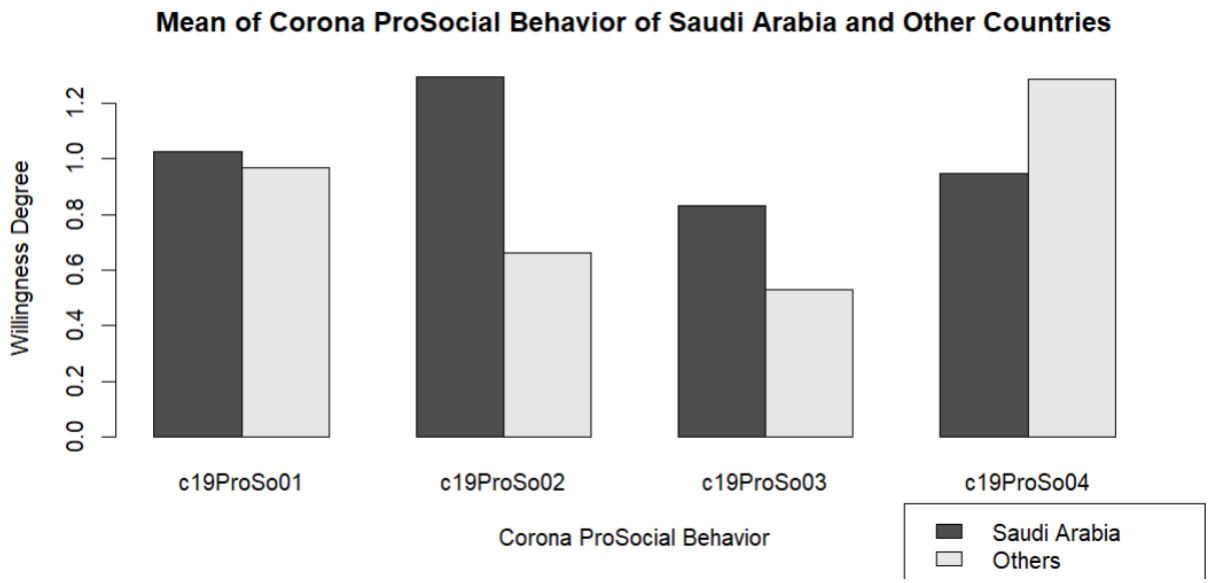
Participant responses for Saudi Arabia differ from the other countries in the survey as a group

a) The focus country for this assignment is "Saudi Arabia". Therefore, a filter is applied on the "coded_country" column, which selects the responses from "Saudi Arabia". After applying the filter, there are 45 columns, and 904 rows from the Saudi Arabia responses after pre-processing.

   Another filter is used to exclude all rows that contain "Saudi Arabia", which results in 39096 responses from the other countries, and 45 columns after applying the filter.

   Based on the responses comparing the Corona ProSocial Behaviour for Saudi Arabia with the other countries as a group, Saudi Arabia have a more positive response. It can be seen by the mean for each of the Corona ProSocial Behaviour variables, namely c19ProSo01, c19ProSo02, and c19ProSo03, which show a higher positive response rate for Saudi Arabia compared to the other countries.

   In order to create a linear regression model, the missing values are omitted, and all the values are converted into numerical values to give a better prediction while still retaining the original meaning of the responses.



How well do participant responses (attributes) predict pro-social attitudes (c19ProSo01,2,3 and 4) for Saudi Arabia?

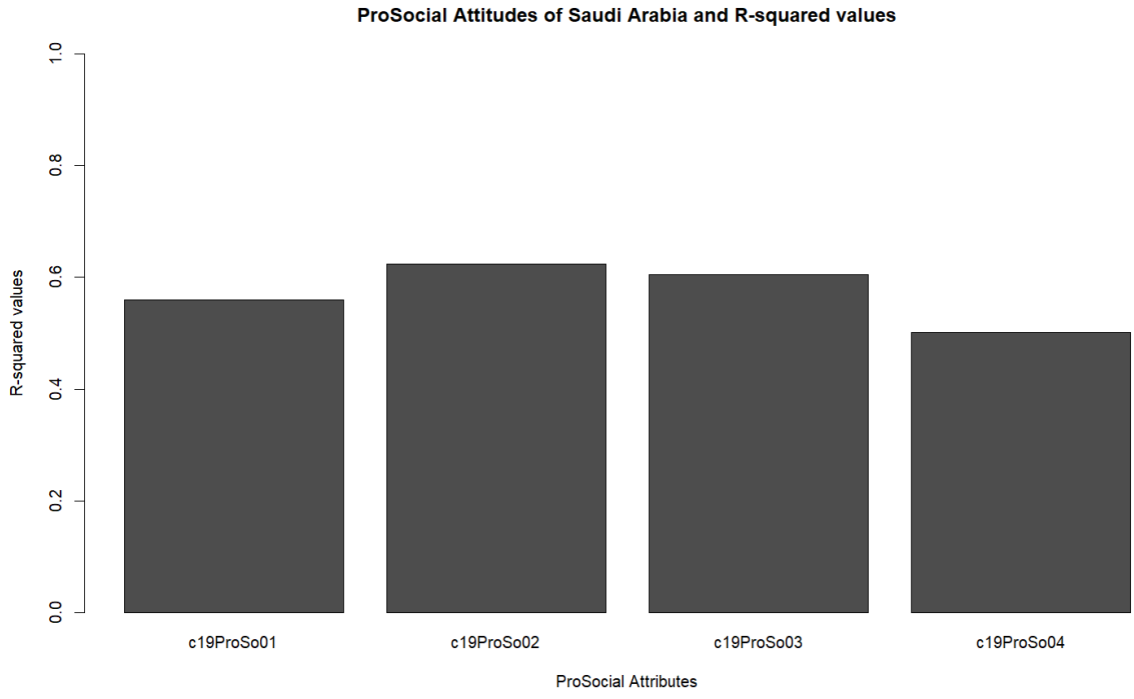
For c19ProSo01 from the responses of Saudi Arabia, the R-squared value is 0.5594, which means that 55.94% of the variability in the Saudi Arabia respondents' responses in the willingness to help others who suffer from coronavirus can be explained by the independent variables. The participants responses moderately predict the pro-social attitudes c19ProSo01. The attributes that are the best predictors for this variable other than the pro-social attitudes are "fail01" and "affEnerg". This is because these values have a p-value lower than 0.05, which are 0.0143 and 0.0470 respectively. This means that these attributes have a stronger relationship with the dependent variable, c19ProSo01.

For c19ProSo02 from the responses of Saudi Arabia, the R-squared value is 0.6235, which has a stronger relationship than c19ProSo01 variable. Approximately 62.35% of the variability in the Saudi Arabia respondents' responses in the degree of making donations to help others that suffer from coronavirus can be explained by their independent variables. The participants responses also moderately predict the pro-social attitudes c19ProSo02. The attributes that are the best predictors for the c19ProSo02 variable other than the pro-social attitudes are "c19NormShould", "affRel", "disc01" and "c19IsPunish". These values also have a p-value lower than 0.05, where "c19NormShould" being the lowest excluding the other pro-social attitudes. This means that these attributes are also statistically significant and are good predictors towards the dependent variable, which is c19ProSo02, in which case the null hypothesis should be rejected.

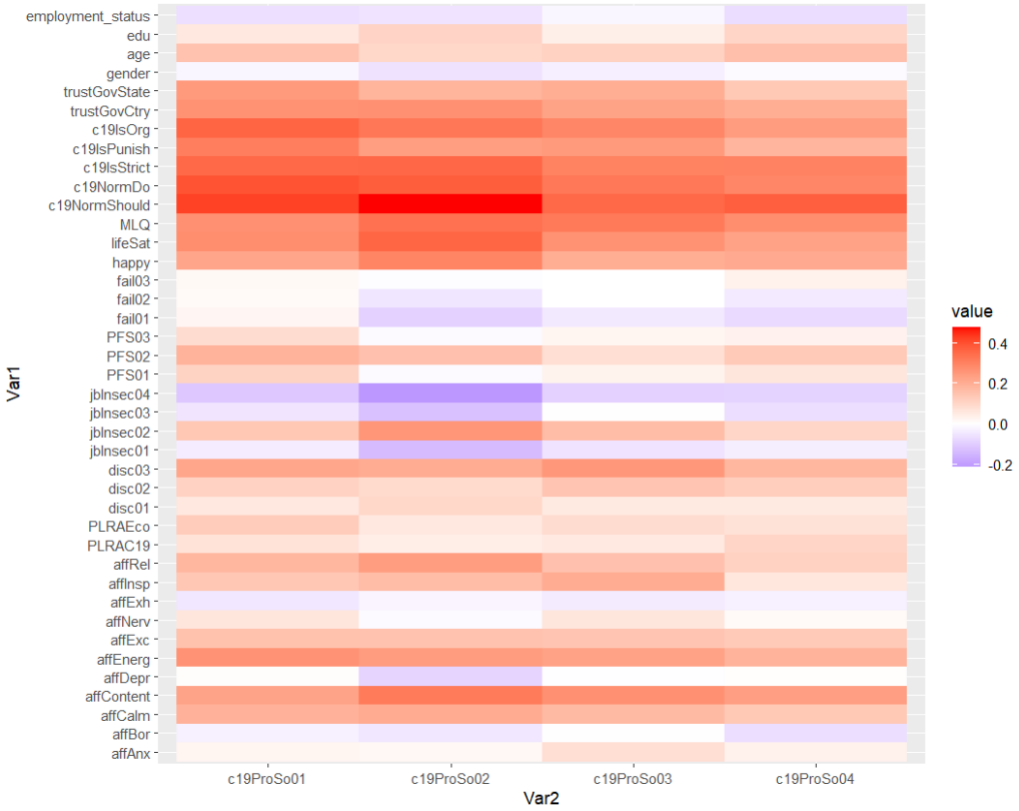
The R-squared value for c19ProSo03 from the responses of Saudi Arabia is 0.6042. This means that 60.42% of the variability observed in the Saudi Arabia respondents' responses by their degree of willingness in protecting vulnerable groups from coronavirus even at their own expense can be explained by the other variables. The participants responses also moderately predict the pro-social attitudes c19ProSo03. The attributes that are the best predictors for the c19ProSo03 variable other than the pro-social attitudes are "jbInsec03", "affInsp" and "MLQ". These variables also have a p-value of less than 0.05, which indicates that these variables also have a stronger relationship in predicting the c19ProSo03 dependent variable,

The R-squared value for c19ProSo04 from the responses of Saudi Arabia, however, is lower with the value of 0.5008. This means that 50.08% of the variability observed in the Saudi Arabia respondents' responses to making personal sacrifices to prevent the spread of coronavirus can be explained by the other independent variables. The participants responses also moderately predict the pro-social attitudes of c19ProSo04. The attributes that are the best predictors for c19ProSo04 variable are "affInsp" and "c19IsStrict", which also have a p-value of less than 0.05, which indicates that these two variables also have a stronger relationship with the dependent variable c19ProSo04.

The R-squared values for the ProSocial Attributes for Saudi Arabia can be seen from the graph below:



The heatmap below shows the correlation between all the variables against the ProSocial attitudes of Saudi Arabia:



How well do participant responses (attributes) predict pro-social attitudes (c19ProSo01,2,3 and 4) for other countries as a group?

However, the overall R-squared value of the ProSocial Attitudes tends to be lower for the other countries compared to Saudi Arabia.

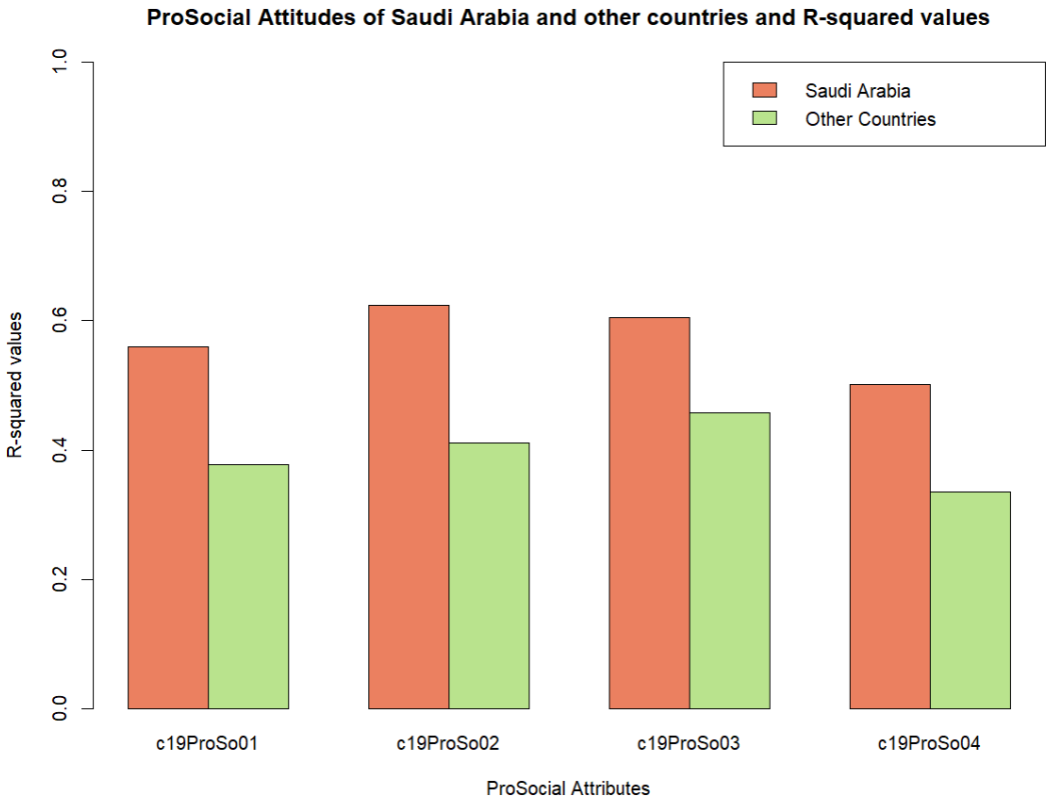
The R-squared value for c19ProSo01 from the responses of other countries is 0.3771. This means that only 37.71% of the variability observed in the other countries respondents' responses in helping others who suffer from coronavirus can be explained by the independent variables. The R-squared value of 0.3771 is considered low, where it suggests that the c19ProSo01 variable may not be a strong predictor of the respondents' pro-social attitudes. However, there tends to be more attributes that seem to be the best predictors for the c19ProSo01 for other countries other than the pro-social attitudes, which are "affAnx", "affContent", "affEnerg", "affExh", "PLRAC19", "PLRAEco", "disc02", "jbInsec04", "PFS01", "fail03", "MLQ", "c19NormDo", "c19IsOrg", "trustGovState", "gender" and "age". All of these attributes have a p-value of less than 0.05, which indicates that they have a stronger relationship with the dependent variable c19ProSo01.

The R-squared value for c19ProSo02 from the responses of other countries is 0.4104. This means that 41.04% of the variability observed in the other countries respondents' responses to making donations to help other that suffer from coronavirus can be explained by the independent variables. This R-squared value is also considered low, where it also suggests that the c19ProSo02 variable may be not a strong predictor of the respondent's pro-social attitudes as well. There are also more attributes that seem to be the best predictors for the c10ProSo02 for other countries compared to Saudi Arabia, which are "affAnx", "affBor", "affCalm", "affExc", "affInsp", "PLRAC19", "PLRAEco", "disc02", "jbInsec02", "jbInsec03", "PFS01", "PFS02", "fail01", "happy", "MLQ", "c19NormShould", "trustGovCtry", "trustGovState", "gender", "age", "edu", "c19ProSo01", "c19ProSo03", "c19ProSo04" and "employment_status". These attributes have a p-value of less than 0.05, which indicates they have a stronger relationship with the dependent variable c19ProSo02.

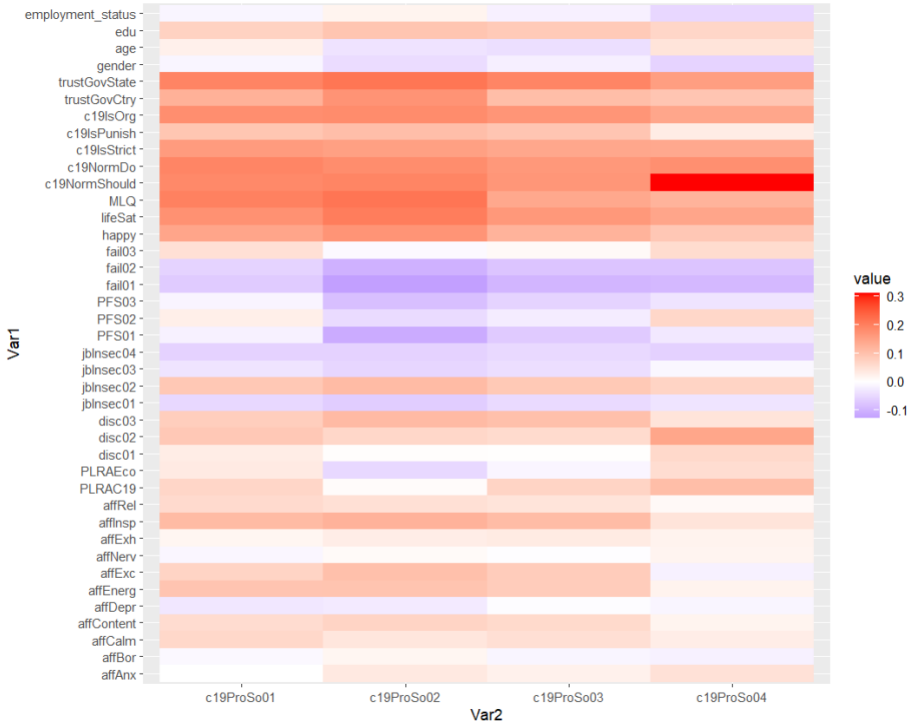
The R-squared value for c19ProSo03 from the responses of other countries is 0.4578. This means that 45.78% of the variability observed in the other countries respondents' responses in their willingness to protect vulnerable groups from coronavirus even at their own expense can be explained by the independent variables. This R-squared value is also considered moderately low, where it also suggests that the c19ProSo03 variable may be not a strong predictor of the respondent's pro-social attitudes as well. The best predictors for the c10ProSo02 other than the pro-social attitudes are "affBor", "affDepr", "affExc", "affNerv", "PLRAC19", "disc03", "PFS02", "lifeSat", "MLQ", "c19NormShould", "c19NormDo", "c19IsPunish", "c19IsOrg", "trustGovCtry", "trustGovState", "gender", "age" and "edu" These attributes have a p-value of less than 0.05, which indicates they have a stronger relationship with the dependent variable c19ProSo03.

The R-squared value for c19ProSo04 from the responses of other countries is 0.336. This means that only 33.6% of the variability observed in the other countries respondents' responses in their willingness to make personal sacrifices to prevent the spread of coronavirus can be explained by the independent variables. This R-squared value is also considered significantly low, where it also suggests that the c19ProSo04 variable may be not a strong predictor of the respondent's pro-social attitudes as well. The best predictors for the c10ProSo04 for other countries compared to Saudi Arabia, which are "affCalm", "affExc", "PLRAC19", "disc02", "jbInsec02", "PFS02", "PFS03", "fail01", "fail02", "fail03", "lifeSat", "c19NormShould", "c19NormDo", "c19IsStrict", "c19IsPunish", "c19IsOrg", "trustGovState", "age", "edu" and "employment_status". These attributes have a p-value of less than 0.05, which indicates they have a stronger relationship with the dependent variable c19ProSo04.

A comparison plot of the R-squared values for ProSocial Attitudes from Saudi Arabia between Other countries is shown below:



The graph below shows the correlation between all the variables and the ProSocial Attitudes of the other countries:



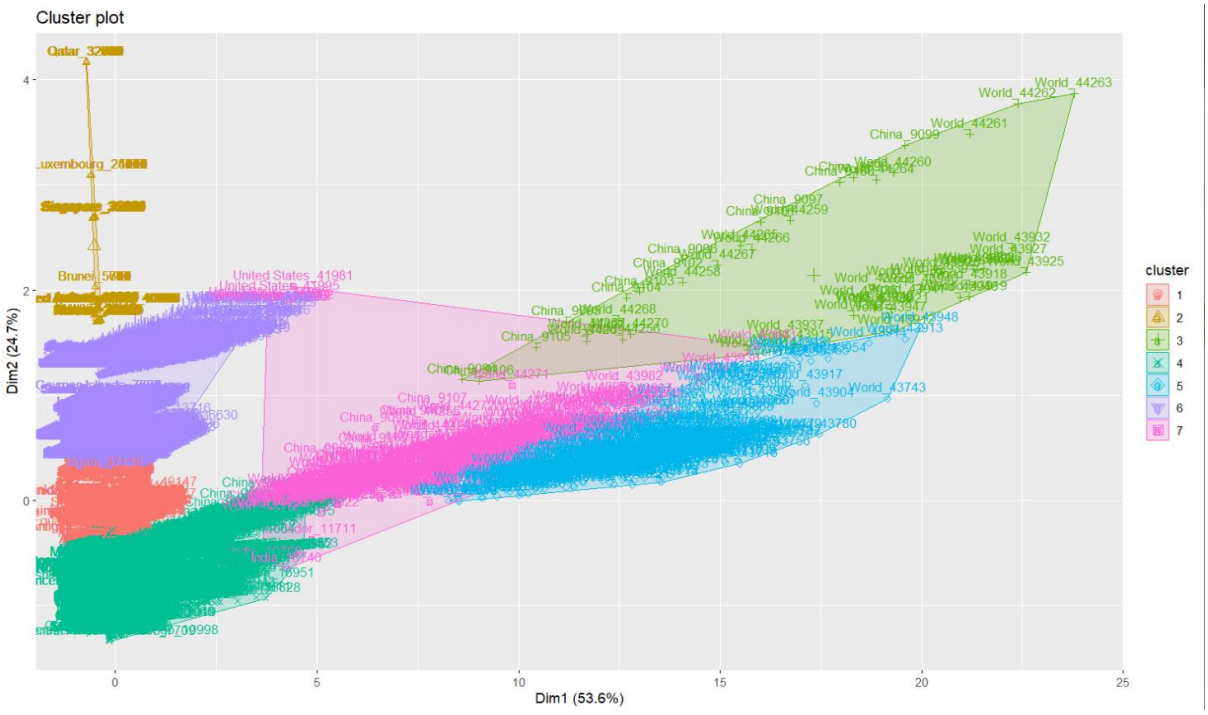
Overall, the ProSocial Attitudes are better explained by responses from Saudi Arabia compared to other countries. This also indicates that the participants responses from Saudi Arabia have better prediction for Pro-Social attitudes compared to other countries.

Q3 Focus country vs cluster of similar countries

<u>Identify countries (in the baseline data) that are similar to your focus country using clustering</u>

a) The dataset chosen to find the similar countries with Saudi Arabia is sourced from https://ourworldindata.org/covid-cases, which contains the data of the confirmed COVID-19 cases for all countries from year 2020 to year 2023. Based on the dataset, some important variables are chosen to find the clusters, which includes the new cases, deaths and vaccinations counts, as well as the GDP per capita. After clustering using k-means with 7 central points using a random seed, there are 22 countries that are found to be in the same cluster as Saudi Arabia. However, only 6 of these countries were chosen. These 6 countries were chosen as after repeating the k-means algorithm without the random seed, these countries seem to appear the most consistently in the same cluster as Saudi Arabia. These 6 countries include: Australia, Canada, Germany, Japan, United Kingdom, United States of America.

The cluster plot below shows the resulting cluster using the k-means cluster method:



How well do participant responses predict pro-social attitudes (c19ProSo01,2,3 and 4) for this cluster of similar countries?

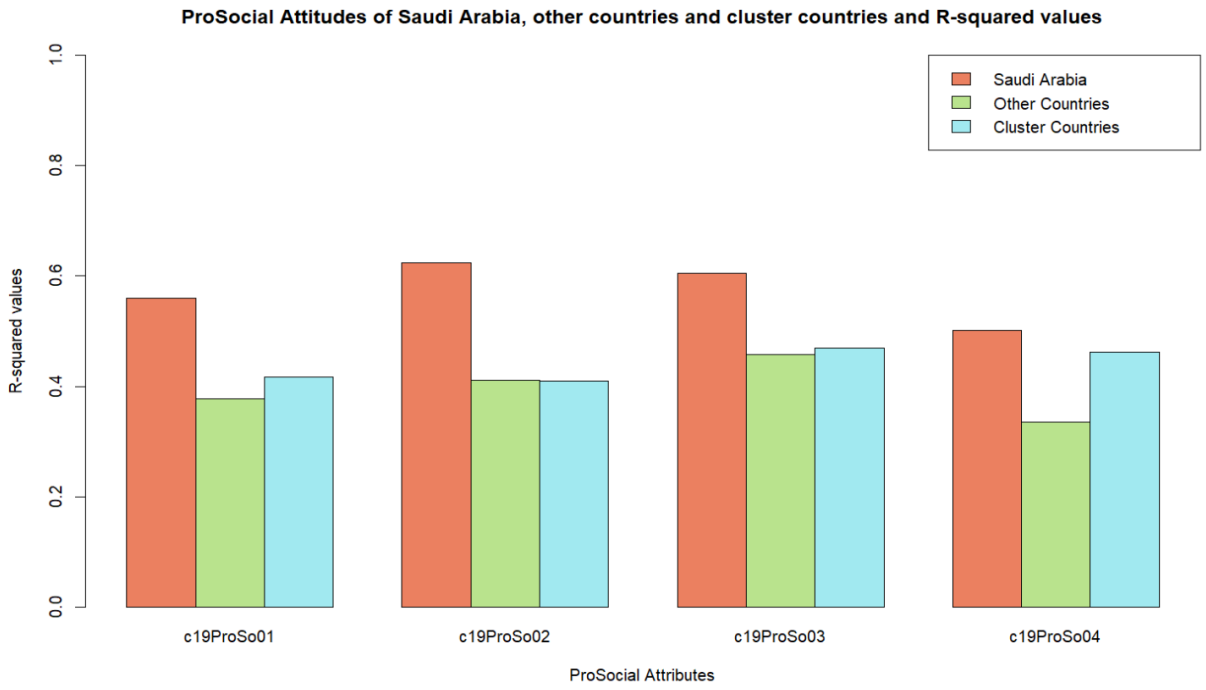
b) The R-squared value for c19ProSo01 from the responses of the cluster countries that are similar to Saudi Arabia is 0.4172. This means that only 41.72% of the variability observed in the cluster countries respondents in the degree of willingness to help others who suffer from coronavirus can be explained by the independent variables used in the dataset. The R-squared value of 0.4172 is also considered low, where it suggests that the c19ProSo01 variable may not be a strong predictor of the respondents' pro-social attitudes. The best predictors for the c19ProSo01 for the cluster countries, which are "coded_countryGermany", "c19NormDo", "affCalm", "affEnerg", "age", "fail02", "PFS01", "gender", "coded_countryJapan", "PFS02", "c19IsPunish", "c19NormShould", "disc02" and "trustGovState". All these attributes have a p-value of less than 0.05, which indicates that they have a stronger relationship with the dependent variable c19ProSo01.

The R-squared value for c19ProSo02 from the responses of the cluster countries that are similar to Saudi Arabia is 0.4095. This means that only 40.95% of the variability observed in the cluster countries respondents' responses to make donations to help others that suffer from coronavirus can be explained by the independent variables. The R-squared value of 0.4095 is also considered low, where it suggests that the c19ProSo02 variable may not be a strong predictor of the respondents' pro-social attitudes. The best predictors for the c19ProSo02 for the cluster countries, which are "trustGovState", "PFS01", "edu", "coded_countryGermany", "coded_countryJapan", "affCalm", "jbInsec03", "c19IsPunish", "affDepr", "happy", "PLRAEco", "MLQ", "coded_countryUnited States of America", "affNerv", "fail01", "disc02", "jbInsec01" and "gender". All these attributes have a p-value of less than 0.05, which indicates that they have a stronger relationship with the dependent variable c19ProSo02.

The R-squared value for c19ProSo03 from the responses of the cluster countries that are similar to Saudi Arabia is 0.4698. This means that only 46.98% of the variability observed in the cluster countries respondents' responses to protect vulnerable groups from the coronavirus even at their own expense can be explained by the independent variables of the dataset, which represents the participants responses. The R-squared value of 0.4698 is also considered low, where it suggests that the c19ProSo03 variable may not be a strong predictor of the respondents' pro-social attitudes. However, there tends to be more attributes that seem to be the best predictors for the c19ProSo02 for the cluster countries, which are "age", "PFS02", "coded_countryJapan", "PLRAC19", "c19IsPunish", "affInsp", "affExh", "c19NormShould", "edu" and "affNerv". All these attributes have a p-value of less than 0.05, which indicates that they have a stronger relationship with the dependent variable c19ProSo03.

The R-squared value for c19ProSo04 from the responses of the cluster countries that are similar to Saudi Arabia is 0.4622. This means that only 46.22% of the variability observed in the other countries respondents' responses in making personal sacrifices to prevent the spread of the coronavirus can be explained by the participants' responses. The R-squared value of 0.4622 is also considered low, where it suggests that the c19ProSo04 variable may not be a strong predictor of the respondents' pro-social attitudes. However, there tends to be more attributes that seem to be the best predictors for the c19ProSo03 for the cluster countries, which are "c19NormShould", "coded_countryJapan", "c19IsPunish", "PFS02", "age", "disc02", "PFS03", "fail03", "employment_status", "c19IsStrict", "gender", "trustGovState", "coded_countryGermany", "edu" and "affDepr". All these attributes have a p-value of less than 0.05, which indicates that they have a stronger relationship with the dependent variable c19ProSo04.

Here is the comparison plot of R-squared values of Saudi Arabia, Other countries and the cluster countries:



The heatmap below shows the correlation between all the variables and the ProSocial Attitudes of the cluster countries:

Overall, the participants responses towards Saudi Arabia are the best for predicting the pro-social attributes compared to the other countries, as well as the cluster countries. This suggests that the independent variables of pro-social attitudes have a stronger association with the other dependent variables, as well as the participants' responses for Saudi Arabia compared to the other countries. This might also be because there might be more variability for a larger dataset. This can be seen as the R-squared values tends to be higher if the countries associated are less. For example, the Saudi Arabia r-squared values for pro-social attitudes are higher compared to the cluster countries, then followed by the other countries as a whole group.

By comparing the heatmap for the Saudi Arabia, other countries as well as cluster countries, we can also see that the correlations between the ProSocial Attitudes and the other variables tends to be higher for Saudi Arabia compared to other countries, and the correlation is the least for the cluster countries. The correlation for these 3 groups of countries tends to be higher for the "Trust in Government" concept, "Corona Community Injunctive norms" concept and , "Live Satisfaction" concept, whereas the least correlations are the "Disempowerment" concept, "Perceived Financial Strain" concept as well as the "Job Insecurity" concept.

Appendix: R code

```
install.packages("dplyr")
install.packages("gridExtra")
install.packages("grid")
install.packages("ggplot2")
install.packages("patchwork")
install.packages("reshape2")
library(ggplot2)
library(gridExtra)
library(grid)
library(dplyr)
library(factoextra)
library(tidyr)
library(reshape2)

cvbase = read.csv("PsyCoronaBaselineExtract.csv")
dim(cvbase)
rm(list = ls())
set.seed(31860532) # XXXXXXXX = your student ID
cvbase = read.csv("PsyCoronaBaselineExtract.csv")
cvbase <- cvbase[sample(nrow(cvbase), 40000), ] # 40000 rows

#Q1 Descriptive analysis and pre-processing

#Q1a
#Dimension
dim(cvbase)
ncol(cvbase) #There are 54 columns
nrow(cvbase) #There are 40000 rows
str(cvbase)

age_group = cvbase %>% group_by(age) %>% summarise(COUNT = n()) #count for
the age groups
age_group <- as.data.frame(age_group, row.names = NULL, optional = FALSE)
age_group_bar = age_group$COUNT
names(age_group_bar) <- c("18-25", "25-34", "35-44", "45-54", "55-64", "65-75",
"75-85", "85+", "NA")
age_group_bar

#Bar plot for the count of age groups
barplot(age_group_bar, main = "Count of Age Group", xlab = "Age Group", ylab =
"Count", col = c("#e3342f" ,"#f6993f", "#ffed4a","#38c172","#4dc0b5", "#3490dc",
"#6574cd", "#9561e2", "#f66d9b" ), legend=TRUE)
```

```
#frequency for each age and Pro-Social Attribute Willingness
rename(count(cvbase, age,jbInsec01 ), Freq = n)
rename(count(cvbase, age,jbInsec02), Freq = n)
rename(count(cvbase, age,jbInsec03), Freq = n)
rename(count(cvbase, age,jbInsec04), Freq = n)



employment = c("employstatus_1", "employstatus_2","employstatus_3",
"employstatus_4","employstatus_5","employstatus_6","employstatus_7","employsta
tus_8","employstatus_9","employstatus_10")
#Creating a new column, with the value that has a "1"/not "NA" for the employstatus
cvbase$employment_status <-
names(cvbase[,employment])[max.col(!is.na(cvbase[,employment]) , ties.method =
"first")]
cvbase$employment_status <- as.numeric(apply(cvbase["employment_status"],1,
function(x) gsub("employstatus_", "",x))) #Creating a new column specifying the
employment status

#Grouping the data based on the employment status, with the count
employment_count = cvbase %>% group_by(employment_status) %>%
summarise(COUNT = n())
employment_count


employment_count <- as.data.frame(employment_count, row.names = NULL,
optional = FALSE)
employment_count_bar = employment_count$COUNT
names(employment_count_bar) <- employment_count$employment_status
grid.table(employment_count)
barplot(employment_count_bar, main = "Count of Employment Group", xlab =
"Employment Group", ylab = "Count", col = c("#e3342f" ,"#f6993f",
"#ffed4a","#38c172","#4dc0b5", "#3490dc", "#6574cd", "#9561e2",
"#f66d9b","#bad80a" ))

#Dropping the columns of employstatus, since the employment status has already
been stated in employment_status
cvbase = subset(cvbase, select = -c(employstatus_1, employstatus_2,employstatus_3,
employstatus_4,employstatus_5,employstatus_6,employstatus_7,employstatus_8,e
mploystatus_9,employstatus_10) )


#Q2 Focus country vs all other countries as a group
```

```
#Q2a
#Filtering the coded countries with Saudi Arabia
saudi_arabia = cvbase %>% filter( coded_country == "Saudi Arabia")

#Dimensions of the Saudi Arabia responses
dim(saudi_arabia)
ncol(saudi_arabia) #There are 45 columns
nrow(saudi_arabia) #There are 904 rows


#Filtering the country group that is not Saudi Arabia
not_sa = cvbase %>% filter(coded_country != "Saudi Arabia")

#There are 39096 rows and 45 columns of data with coded_country excluding Saudi Arabia
dim(not_sa)

#Creating a datafrmae to compare the pro-social attitudes of Saudi Arabia and other countries
prosocial <- data.frame(Country = c("Saudi Arabia", "Others"), c19ProSo01 = c(mean(saudi_arabia$c19ProSo01, na.rm = TRUE), mean(not_sa$c19ProSo01, na.rm = TRUE)),
                        c19ProSo02 = c(mean(saudi_arabia$c19ProSo02, na.rm = TRUE), mean(not_sa$c19ProSo02, na.rm = TRUE)),
                        c19ProSo03 = c(mean(saudi_arabia$c19ProSo03, na.rm = TRUE), mean(not_sa$c19ProSo03, na.rm = TRUE)),
                        c19ProSo04 = c(mean(saudi_arabia$c19ProSo04, na.rm = TRUE), mean(not_sa$c19ProSo04, na.rm = TRUE)))
row.names(prosocial) <- prosocial$Country
prosocial <- prosocial[ , 2:ncol(prosocial)]
colnames(prosocial) <- c("c19ProSo01", "c19ProSo02", "c19ProSo03", "c19ProSo04")
prosocial <- as.matrix(prosocial)
prosocial
barplot(main = "Mean of Corona ProSocial Behavior of Saudi Arabia and Other Countries", xlab = "Corona ProSocial Behavior", ylab = "Willingness Degree", height = prosocial, beside = TRUE, legend.text = TRUE, args.legend = list(x = "bottomright", inset = c(-0.05, -0.4)))



#Best predictors
saudi_arabia_num <- unlist(lapply(saudi_arabia, is.numeric))
saudi_arabia_num <- saudi_arabia[ , saudi_arabia_num]
```

```
saudi_arabia_num = na.omit(saudi_arabia_num)
saudi_arabia_num

prosocial_sa <- data.frame(Country = c("Saudi Arabia"), c19ProSo01 =
c(mean(saudi_arabia$c19ProSo01, na.rm = TRUE)),
                  c19ProSo02 = c(mean(saudi_arabia$c19ProSo02, na.rm = TRUE)),
                  c19ProSo03 = c(mean(saudi_arabia$c19ProSo03, na.rm = TRUE)),
                  c19ProSo04 = c(mean(saudi_arabia$c19ProSo04, na.rm = TRUE)))
row.names(prosocial_sa) <- prosocial_sa$Country
prosocial_sa <- prosocial_sa[ , 2:ncol(prosocial_sa)]
colnames(prosocial_sa) <- c("c19ProSo01", "c19ProSo02", "c19ProSo03",
"c19ProSo04")
prosocial_sa <- as.matrix(prosocial_sa)
prosocial_sa
#The mean values of all Corona ProSocial Behaviors for Saudi Arabia
barplot(main = "Mean of Corona ProSocial Behavior of Saudi Arabia", xlab = "Corona
ProSocial Behavior", ylab = "Willingness Degree", height = prosocial_sa,ylim = c(-3,3))


#Correlation between Corona ProSocial Behaviour with all other values
#Q2b

pro_social_col = c("c19ProSo01", "c19ProSo02", "c19ProSo03", "c19ProSo04")
pro_social = saudi_arabia_num[, pro_social_col]
corr_matrix <- cor(saudi_arabia_num[, !colnames(saudi_arabia_num) %in%
pro_social_col], pro_social)

corr_melt = melt(corr_matrix)
head(corr_melt)
heatmap_sa <- ggplot(corr_melt, aes(Var2, Var1)) +                    # Create heatmap
with ggplot2
  geom_tile(aes(fill = value)) +scale_fill_gradient2(low = "blue", high = "red", mid =
"white", midpoint = 0)
heatmap_sa

#Linear regression between c19ProSo01 and all other attributes from Saudi Arabia
lm_c19ProSo01_sa = lm(c19ProSo01 ~ ., data = saudi_arabia_num)
summary(lm_c19ProSo01_sa)


coef01_sa <- summary(lm_c19ProSo01_sa)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef01_sa[order(coef01_sa[ , 4]), ]
coef01_sa_less = coef01_sa[coef01_sa[,4] < 0.05,]
```

```
coef01_sa_less

#Linear regression between c19ProSo02 and all other attributes from Saudi Arabia
lm_c19ProSo02_sa =lm(c19ProSo02~., data = saudi_arabia_num)
summary(lm_c19ProSo02_sa)
coef02_sa <- summary(lm_c19ProSo02_sa)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef02_sa[order(coef02_sa[ , 4]), ]
coef02_sa_less = coef02_sa[coef02_sa[,4] < 0.05,]
coef02_sa_less

#Linear regression between c19ProSo03 and all other attributes from Saudi Arabia
lm_c19ProSo03_sa =lm(c19ProSo03~., data = saudi_arabia_num)
summary(lm_c19ProSo03_sa)
coef03_sa <- summary(lm_c19ProSo03_sa)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef03_sa[order(coef03_sa[ , 4]), ]

#Linear regression between c19ProSo04 and all other attributes from Saudi Arabia
lm_c19ProSo04_sa =lm(c19ProSo04~., data = saudi_arabia_num)
summary(lm_c19ProSo04_sa)
coef04_sa <- summary(lm_c19ProSo04_sa)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef04_sa[order(coef04_sa[ , 4]), ]

#Creating a dataframe to plot these R-squared values
prosocialattrnames <- c("c19ProSo01","c19ProSo02","c19ProSo03","c19ProSo04")
prosocialattrvalue_sa <- c(summary(lm_c19ProSo01_sa)$r.squared,
summary(lm_c19ProSo02_sa)$r.squared,summary(lm_c19ProSo03_sa)$r.squared,su
mmary(lm_c19ProSo04_sa)$r.squared)
prosocialvalue_df <- data.frame(prosocialattrnames,prosocialattrvalue_sa)


prosocialr2_sa <- data.frame(Country = c("Saudi Arabia"), c19ProSo01 =
summary(lm_c19ProSo01_sa)$r.squared,
                  c19ProSo02 = summary(lm_c19ProSo02_sa)$r.squared,
                  c19ProSo03 = summary(lm_c19ProSo03_sa)$r.squared,
                  c19ProSo04 = summary(lm_c19ProSo04_sa)$r.squared)
row.names(prosocialr2_sa) <- c("Saudi Arabia")
prosocialr2_sa <- prosocialr2_sa[ , 2:ncol(prosocialr2_sa)]
colnames(prosocialr2_sa) <- c("c19ProSo01", "c19ProSo02", "c19ProSo03",
"c19ProSo04")
prosocialr2_sa <- as.matrix(prosocialr2_sa)
prosocialr2_sa
```

```
#Bar plot of the R-squared values for each pro-social attitudes for Saudi Arabia
barplot(height = prosocialr2_sa, main = "ProSocial Attitudes of Saudi Arabia and R-squared values", xlab = "ProSocial Attributes", ylab = "R-squared values", ylim = c(0,1))

#Applying a filter to filter out Saudi Arabia
not_sa_num <- unlist(lapply(not_sa, is.numeric))

not_sa_num <- not_sa[ , not_sa_num]
not_sa_num = na.omit(not_sa_num)
not_sa_ctry = na.omit(not_sa)


pro_social_col = c("c19ProSo01", "c19ProSo02", "c19ProSo03", "c19ProSo04")
pro_social = not_sa_num[, pro_social_col]
corr_matrix <- cor(not_sa_num[, !colnames(saudi_arabia_num) %in% pro_social_col], pro_social)

corr_melt = melt(corr_matrix)
head(corr_melt)
heatmap_oc <- ggplot(corr_melt, aes(Var2, Var1)) +                    # Create heatmap with ggplot2
  geom_tile(aes(fill = value)) +scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0)
heatmap_oc

#Linear regression between c19ProSo01 and all other attributes from other countries
lm_c19ProSo01_oc =lm(c19ProSo01~., data = not_sa_num)
summary(lm_c19ProSo01_oc)
#The R-squared value:
summary(lm_c19ProSo01_oc)$r.squared
coef01_oc <- summary(lm_c19ProSo01_oc)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef01_oc[order(coef01_oc[ , 4]), ]
coef01_oc_less = coef01_oc[coef01_oc[,4] < 0.05,]
coef01_oc[coef01_oc[,4] < 0.05,][,0]

#Linear regression between c19ProSo02 and all other attributes from other countries
lm_c19ProSo02_oc =lm(c19ProSo02~., data = not_sa_num)
summary(lm_c19ProSo02_oc)
coef02_oc <- summary(lm_c19ProSo02_oc)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef02_oc[order(coef02_oc[ , 4]), ]
```

```
coef02_oc_less = coef02_oc[coef02_oc[,4] < 0.05,]
coef02_oc[coef02_oc[,4] < 0.05,][,0]

#Linear regression between c19ProSo03 and all other attributes from other countries
lm_c19ProSo03_oc =lm(c19ProSo03~., data = not_sa_num)
summary(lm_c19ProSo03_oc)
coef03_oc <- summary(lm_c19ProSo03_oc)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef03_oc[order(coef03_oc[ , 4]), ]
coef03_oc_less = coef03_oc[coef03_oc[,4] < 0.05,]
coef03_oc[coef03_oc[,4] < 0.05,][,0]

#Linear regression between c19ProSo04 and all other attributes from other countries
lm_c19ProSo04_oc =lm(c19ProSo04~., data = not_sa_num)
summary(lm_c19ProSo04_oc)
coef04_oc <- summary(lm_c19ProSo04_oc)[["coefficients"]]
#Sorting based on the P-value in ascending order
coef04_oc[order(coef04_oc[ , 4]), ]
coef04_oc_less = coef04_oc[coef04_oc[,4] < 0.05,]
coef04_oc[coef04_oc[,4] < 0.05,][,0]


prosocialr2_oc <- data.frame(Country = c("Saudi Arabia"), c19ProSo01 =
c(summary(lm_c19ProSo01_sa)$r.squared,
summary(lm_c19ProSo01_oc)$r.squared),
                    c19ProSo02 = c(summary(lm_c19ProSo02_sa)$r.squared,
summary(lm_c19ProSo02_oc)$r.squared),
                    c19ProSo03
=c(summary(lm_c19ProSo03_sa)$r.squared,summary(lm_c19ProSo03_oc)$r.squared
),
                    c19ProSo04 =
c(summary(lm_c19ProSo04_sa)$r.squared,summary(lm_c19ProSo04_oc)$r.squared))
row.names(prosocialr2_oc) <- c("Saudi Arabia", "Other Countries")
prosocialr2_oc <- prosocialr2_oc[ , 2:ncol(prosocialr2_oc)]
colnames(prosocialr2_oc) <- c("c19ProSo01", "c19ProSo02", "c19ProSo03",
"c19ProSo04")
prosocialr2_oc <- as.matrix(prosocialr2_oc)
prosocialr2_oc

#A barplot for comparing the pro-social attritudes of Saudi Arabia and other
countries' R-squared values
barplot(height = prosocialr2_oc, main = "ProSocial Attitudes of Saudi Arabia and
other countries and R-squared values", xlab = "ProSocial Attributes", ylab = "R-
squared values", ylim = c(0,1), beside =TRUE, col = c("#eb8060", "#b9e38d"))
```

```
legend("topright",
      legend = c("Saudi Arabia", "Other Countries"), fill =c("#eb8060", "#b9e38d"))



#Q3 clustering
coviddata = read.csv("owid-covid-data.csv")

#There are 255 unique countries
unique(coviddata$location)

#Selecting new_cases, new_deaths, new_vaccinations and gdp
covid_data = coviddata %>% select(new_cases,new_deaths,new_vaccinations,
gdp_per_capita,)

#covid_data but with locations added
covid_data_locations = coviddata %>%
select(location,new_cases,new_deaths,new_vaccinations, gdp_per_capita,)
head(covid_data_locations)
covid_cluster_table = as.data.frame(covid_data_locations)
(head(covid_cluster_table, n =50))

#ignoring all NA values
covid_data = na.omit(covid_data)
covid_data_locations = na.omit(covid_data_locations)

summary(covid_data)

#scaling the data
covid_data_scale = scale(covid_data)

#naming each rows of data with their countries and number separated by "_"
rownames(covid_data_scale) <- paste(covid_data_locations$location,
1:dim(covid_data_locations)[1], sep = "_")

#applying kmeans cluster
set.seed(31860532)
km.out <- kmeans(covid_data_scale, centers = 7, nstart = 50)
km.clusters <- km.out$cluster

#using a dataframe to store the countries and the cluster numbers
cluster_df <- data.frame(Value = as.vector(t(covid_data_locations$location)), Cluster
= km.out$cluster)
cluster_df
```

```
#finding out which cluster the Saudi Arabia country is in
saudi_cluster =unique(subset(cluster_df, grepl("Saudi", Value))$Cluster)

#finding all the countries that are in the same cluster as Saudi Arabia
unique(subset(cluster_df, grepl(saudi_cluster, Cluster))$Value)

#Output the cluster visualisation.
fviz_cluster(list(data =covid_data_scale , cluster = km.clusters),labelsize = 10)

#The cluster countries chosen are: Australia, Canada, Germany, Japan, UK, US
cluster_ctry = c("Australia", "Canada", "Germany", "Japan", "United Kingdom",
"United States of America")
cluster_countries = not_sa_ctry[not_sa_ctry$coded_country %in% cluster_ctry,]
cluster_countries
unique(cluster_countries$coded_country)



pro_social_col = c("c19ProSo01", "c19ProSo02", "c19ProSo03", "c19ProSo04")
pro_social = cluster_countries[, pro_social_col]
corr_matrix <- cor(cluster_countries[, !colnames(saudi_arabia_num) %in%
pro_social_col], pro_social)

corr_melt = melt(corr_matrix)
head(corr_melt)
heatmap_cluster <- ggplot(corr_melt, aes(Var2, Var1)) +                    # Create
heatmap with ggplot2
  geom_tile(aes(fill = value)) +scale_fill_gradient2(low = "blue", high = "red", mid =
"white", midpoint = 0)
heatmap_cluster

#Linear regression of cluster countries with c19ProSo01 dependent variable
lm_c19ProSo01_cluster =lm(c19ProSo01~., data = cluster_countries )
summary(lm_c19ProSo01_cluster)
coef01_oc_cluster <- summary(lm_c19ProSo01_cluster)[["coefficients"]]
coef01_oc_cluster
#Sorting based on the P-value in ascending order, and finding p-values less than 0.05
ordered_cluster01 = coef01_oc_cluster[order(coef01_oc_cluster[ , 4]), ]
ordered_cluster01_less = ordered_cluster01[ordered_cluster01[,4] < 0.05,]
ordered_cluster01_less[,0]

#Linear regression of cluster countries with c19ProSo02 dependent variable
lm_c19ProSo02_cluster =lm(c19ProSo02~., data = cluster_countries )
```

```
summary(lm_c19ProSo02_cluster)
coef02_oc_cluster <- summary(lm_c19ProSo02_cluster)[["coefficients"]]
coef02_oc_cluster
#Sorting based on the P-value in ascending order, and finding p-values less than 0.05
ordered_cluster02 = coef02_oc_cluster[order(coef02_oc_cluster[ , 4]), ]
ordered_cluster02_less = ordered_cluster02[ordered_cluster02[,4] < 0.05,]
ordered_cluster02_less[,0]

#Linear regression of cluster countries with c19ProSo03 dependent variable
lm_c19ProSo03_cluster =lm(c19ProSo03~., data = cluster_countries )
summary(lm_c19ProSo03_cluster)
coef03_oc_cluster <- summary(lm_c19ProSo03_cluster)[["coefficients"]]
coef03_oc_cluster
#Sorting based on the P-value in ascending order, and finding p-values less than 0.05
ordered_cluster03 = coef03_oc_cluster[order(coef03_oc_cluster[ , 4]), ]
ordered_cluster03_less = ordered_cluster03[ordered_cluster03[,4] < 0.05,]
ordered_cluster03_less
ordered_cluster03_less[,0]

#Linear regression of cluster countries with c19ProSo04 dependent variable
lm_c19ProSo04_cluster =lm(c19ProSo04~., data = cluster_countries )
summary(lm_c19ProSo04_cluster)
coef04_oc_cluster <- summary(lm_c19ProSo04_cluster)[["coefficients"]]
coef04_oc_cluster
#Sorting based on the P-value in ascending order, and finding p-values less than 0.05
ordered_cluster04 = coef04_oc_cluster[order(coef04_oc_cluster[ , 4]), ]
ordered_cluster04_less = ordered_cluster04[ordered_cluster04[,4] < 0.05,]
ordered_cluster04_less
ordered_cluster04_less[,0]

#Dataframe to plot the bar plot ffor comparing r-squared values for each pro-social
attitudes attributes
prosocialr2_cluster <- data.frame(Country = c("Saudi Arabia"), c19ProSo01 =
c(summary(lm_c19ProSo01_sa)$r.squared,
summary(lm_c19ProSo01_oc)$r.squared,summary(lm_c19ProSo01_cluster)$r.square
d),
                    c19ProSo02 = c(summary(lm_c19ProSo02_sa)$r.squared,
summary(lm_c19ProSo02_oc)$r.squared,summary(lm_c19ProSo02_cluster)$r.square
d),
                    c19ProSo03
=c(summary(lm_c19ProSo03_sa)$r.squared,summary(lm_c19ProSo03_oc)$r.squared,
summary(lm_c19ProSo03_cluster)$r.squared),
```

```
                    c19ProSo04 =
c(summary(lm_c19ProSo04_sa)$r.squared,summary(lm_c19ProSo04_oc)$r.squared,s
ummary(lm_c19ProSo04_cluster)$r.squared))
row.names(prosocialr2_cluster) <- c("Saudi Arabia", "Other Countries", "Cluster
Countries")
prosocialr2_cluster <- prosocialr2_cluster[ , 2:ncol(prosocialr2_cluster)]
colnames(prosocialr2_cluster) <- c("c19ProSo01", "c19ProSo02", "c19ProSo03",
"c19ProSo04")
prosocialr2_cluster <- as.matrix(prosocialr2_cluster)
prosocialr2_cluster
old.par <- par(mar = c(0, 0, 0, 0))
par(old.par)
barplot(height = prosocialr2_cluster, main = "ProSocial Attitudes of Saudi Arabia,
other countries and cluster countries and R-squared values", xlab = "ProSocial
Attributes", ylab = "R-squared values", ylim = c(0,1), beside =TRUE,col = c("#eb8060",
"#b9e38d", "#a1e9f0"))
legend("topright",
      legend = c("Saudi Arabia", "Other Countries", "Cluster Countries"), fill
=c("#eb8060", "#b9e38d", "#a1e9f0"))
```

First few rows of the dataset used for clustering:

	location	new_cases	new_deaths	new_vaccinations	gdp_per_capita
511	Afghanistan	623	14	2859	1803.987
518	Afghanistan	1093	27	4015	1803.987
756	Afghanistan	456	4	6868	1803.987
846	Afghanistan	24	0	383	1803.987
984	Afghanistan	365	0	9447	1803.987
1035	Afghanistan	104	1	36587	1803.987
1049	Afghanistan	109	1	14800	1803.987
2757	Albania	376	6	60	11803.431
2758	Albania	656	5	78	11803.431
2759	Albania	707	4	42	11803.431
2760	Albania	660	5	61	11803.431
2761	Albania	641	4	36	11803.431
2762	Albania	581	5	42	11803.431
2763	Albania	474	7	36	11803.431
2764	Albania	292	4	36	11803.431
2765	Albania	586	6	30	11803.431
2793	Albania	801	15	1348	11803.431
2794	Albania	1075	18	1128	11803.431
2826	Albania	659	4	3461	11803.431
2827	Albania	344	8	2302	11803.431
2828	Albania	303	11	5356	11803.431
2829	Albania	448	15	2900	11803.431
2830	Albania	472	13	1827	11803.431
2831	Albania	449	8	13925	11803.431
2834	Albania	285	6	19525	11803.431
2835	Albania	304	11	16617	11803.431
2836	Albania	434	8	17023	11803.431
2837	Albania	349	6	13010	11803.431
2838	Albania	336	6	7386	11803.431
2839	Albania	341	9	13826	11803.431
2840	Albania	348	9	14880	11803.431
2841	Albania	264	9	10791	11803.431
2842	Albania	141	9	10163	11803.431