FIT3152 Assignment 1

Chenlongjie Weng 29334152

1. Descriptive analysis and pre-processing. (6 Marks)

(a). We can see this data frame we analysis for have 64381 rows and 54 columns. But this time we only use the 40000 rows.

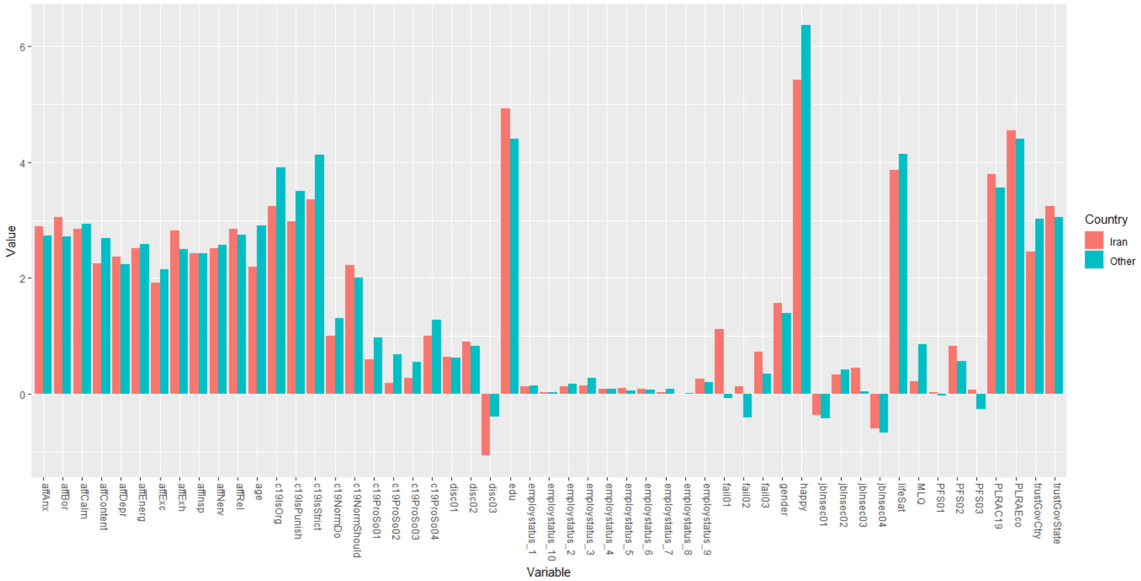
Most of variables in data is int, there only variable coded_country is character type (this one is a lots of the countries name). Also, we can see there are a lot of the NA value in our data (employstatus_8 have most of NA's is about 39320, but it is easy to understand, because employstatus_1 - employstatus_10 is optional, only one or two option will be choose by each person but is also can be some answers missing). We can see these variables have different response level some is from 1 to 5, some is 1 to 8, some is form -2 to 2. They are a lot of different response level. There is some interest thing in trustGovCtry and trustGovState, these two variables also have high amount of Na's value, it is looks like people do not like to answer this question or we can say it hard choice for the peoples (the mean is 3 for both).

(b). This time we need to do some pre-processing and data manipulation. So first we need to remove NA value which will have bad effect for us analysis result. Here I use median number to instead it in most of variable, this preserves some of the characteristics of the original data. for all the affect and likelihood concept variable I use the median value to instead the NA's. then for the Societal Discontent and Job Insecurity concept variable I use the 0(Neither agree nor disagree) to instead the NA's, this will have less effect about the original data, NA just like they choose the noting so the middle option is more similar for that. Employstatus this variable is special, it is like a optional question, so everyone here only choose the one option, some may choose two but that’s ok. So here I change the Na’s to the 0 which means how many people not choose this option. Then we will get the means is the rate of people choose this option. Then we do same things as before until to the happy. In happy we set mean value to instead the NAs, and it will until to the gender. Gender is different here female is 1, male is 2, other is 3. NAs is not option be selected so we set is as other gender which is 3 here. Then age is the period here, like 18-24+25-34+35-44+45-54+55-64+65-75+75-85+85+, there is 8 period, we might set is as median due to it is most common period, education is same as well. Then set rest as 0(Neither agree nor disagree). So, we remove all the NAs, we can do next things.

2. Focus country vs all other countries as a group. (12 Marks)

(a). Here we can see in the bar chart, we compare the Iran (focus country) with other countries using the mean of each variable. In some parts we can say what other countries represent is the world average. So, we can see the Iran’s people c19ProSo01-c19ProSo04 values is lower than other countries which means that they are not as willing to help people infected with the coronavirus as people in other countries around the world. There are quite different in variable fail01, fail02, fail03 which are how people feeling about them country. For all these three variables, Iran’s people have higher value than people in other countries around the world. So, the people live in Iran feel life is worse than other countries people. Especially in fail01 and fail02 the other countries even have negative value, but the Iran have quite high value for these, so we can see the Iran people life is not quite good. So, we might can understand why they have low c19ProSo01-c19ProSo04 values. Because people need to live their own lives well to have spare energy and be more willing to help others. Then we can see the value of happy variable the Iran's people not quite happy compared to people from other countries. Now look at the job insecurity, there is a big different in jbInsec03 from Iran and other countries. The Iran's people more worried about the future of them job.

To summarize our findings, we can see that Iranians are more pessimistic about life, which indirectly reflects the poor living standards and social welfare in Iran. Or rather, their lives are not as good as the average of other countries in the world.



(b). Here we use one of c19ProSo01,2,3 and 4 to find best predictor to each of them in focus country data, then we can see some variables have lowest p-value for each of them (3 * with them). Then we find for c19ProSo01, the variables c19NormDo, c19ProSo02 and c19ProSo04 are best predictor with lowest p-values. And we can see the coef of these variables is positive. So, these are most significant predictor to c19ProSo01. Then we look at the mean of these variables, the people who will help others who suffer from coronavirus, they will also make donation for it and make personal sacrifices to prevent the spread of coronavirus and do self-isolate to prevent coronavirus.

For c19ProSo02, the variables c19IsPunish, C19ProSo01 and c19ProSo03 are best predictor with lowest p-values. And we can see the coef of these variables is positive. So, these are most significant predictor to c19ProSo02. Then we look at the mean of these variables, the people who will make donations to help others who suffer from coronavirus, they will also help others and protect vulnerable groups from coronavirus even at their own expense and they more likely live in the place which community punishing people who deviate from the rules that have been put in place in response to the Coronavirus.

For c19ProSo03, the variables C19ProSo02 and c19ProSo04 are best predictor with lowest p-values. And we can see the coef of these variables is positive. So, these are most significant predictor to c19ProSo03. Then we look at the mean of these variables, the people who will help others who suffer from coronavirus by themselves expense, they will also make donation for it and make personal sacrifices to prevent the spread of coronavirus.

For c19ProSo04, the variables C19ProSo01 and c19ProSo03 are best predictor with lowest p-values. And we can see the coef of these variables is positive. So, these are most significant predictor to c19ProSo04. Then we look at the mean of these variables, the people who will make personal sacrifices to prevent the spread of coronavirus, they will also help others and protect vulnerable groups from coronavirus even at their own expense.

(c). Now we are looking for the best predictor of each pro-social attitude's variables in other countries data.

First, we look at the c19ProSo01. Here we can find a lots of variables with 3 *, but we are looking for the strongest predictors. So, we need to find the variables with lowest p-value, then we can see the c19ProSo02, c19ProSo03 and c19ProSo04 are both have lowest p-value $2e^{-16}$. So, these three variables are best predictors for the c19ProSo01.

Secondly, we look at the c19ProSo02. Continue to find the lowest p-value variables which would be affAnx, PFS01, MLQ, C19NormShould, c19ProSo01 and c19ProSo03.

Then, we concentrate to the c19ProSo03. To do the same thing as before, we find the best predictors of c19ProSo03 are age, c19ProSo01, c19ProSo02 and c19ProSo04.

Lastly, for the c19ProSo04. Same as before, we can see the strongest predictors of c19ProSo04 are PLRAC19, disc02, C19NormShould, C19IsPunish, age, c19ProSo01 and c19ProSo03.

To compare these variables with the variables in out focus country. For the c19ProSo01, the similar thing is they all have c19ProSo02 and c19ProSo04 as best predicator. The different is for our focus country we have c19NormDo as best predictor, but in other countries data it uses the c19ProSo03 as best predictor.

For the c19ProSo02, the similar are they both have c19ProSo01 and c19ProSo03, then different is for focus country we have c19IsPunish as predictor. But in other countries data, it uses affAnx, PFS01, MLQ and C19NormShould.

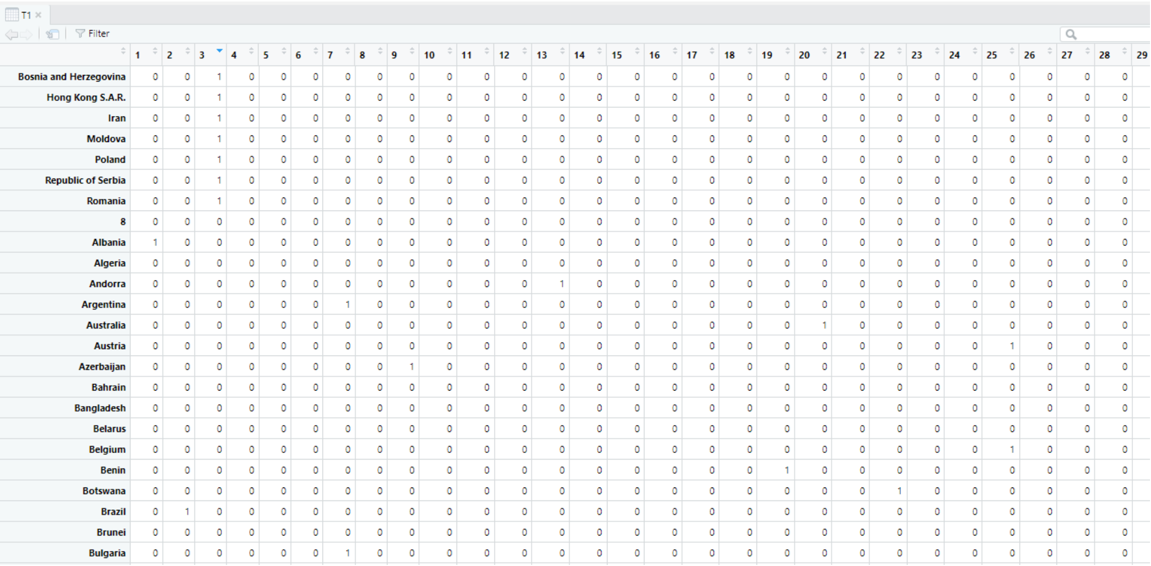
For the c19ProSo03, the similar are they both have c19ProSo02 and c19ProSo04, the different is in other countries data, it uses age and c19ProSo01 also.

Lastly, consider about c19ProSo04, focus country data only have two best predictors which are c19ProSo01 and c19ProSo03. But other countries data have PLRAC19, disc02, C19NormShould, C19IsPunish, age, c19ProSo01 and c19ProSo03. They both have c19ProSo01 and c19ProSo03, but other countries data have more best predictor.

To conclusion, the pro-social attitudes variables are strong connected to each other and they sometimes will have strong connected with another variable also. For the focus country data, we have lesser number of best predictors than other countries data which might because of the other countries data is mixed a lots of different countries data, so they will have more unexpected influencing factors that might can be some countries have quite higher value in some variables, it will affect in the whole data.

3. Focus country vs cluster of similar countries. (10 Marks)

(a). Here I use the variables (c19ProSo01, c19ProSo02, c19ProSo03, c19ProSo04, jbInsec03, fail01, fail02, fail03, disc03, age, c19NormDo) to find the most similar countries for focus country (Iran). We can see in the 2a, focus country have big difference with other countries in these variables so we choose these variables is more useful to find similar countries. Then after the choose the variables, I group it by country names and calculate each variable (we chooses) means and scale the data (not include the char type which is coded_country). The set the seed to make sure every time we run this have same output. Then we make a table to fit the cluster what we use in kmeans. Then look at the table we can see Iran (focus country) is in cluster 3, and then there are 6 similar countries with it which are Bosnia and Herzegovina, Hong Kong S.A.R., Moldova, Poland, Republic of Serbia, Romania.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Bosnia and Herzegovina	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Hong Kong S.A.R.	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Iran	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Moldova	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Poland	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Republic of Serbia	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Romania	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Albania	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Algeria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Andorra	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Argentina	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Australia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
Austria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
Azerbaijan	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Bahrain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Bangladesh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Belarus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Belgium	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
Benin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
Botswana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
Brazil	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Brunei	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Bulgaria	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

.

(b).

Now we are looking for the best predictor of each pro-social attitude's variables in similar countries data which are combination for the countries like focus country.

First, we look at the c19ProSo01 for similar countries. There are some variables with 3 * and we need to find the best predictors. So, the variables with lowest p-value are what we need, then we can see the c19ProSo02, c19ProSo03 are both have lowest p-value 2e^{-16}. And c19ProSo04's p-value are 3.08e^{-14} which are also quite low. These variables are best predictors for the c19ProSo01.

Secondly, we look at the c19ProSo02. Continue to find the lowest p-value variables which would be c19ProSo01 and c19ProSo03.

Then, we concentrate to the c19ProSo03. To do the same thing as before, we find the best predictors of c19ProSo03 are c19ProSo01, c19ProSo02 and c19ProSo04.

Lastly, for the c19ProSo04. Same as before, we can see the strongest predictors of c19ProSo04 are c19ProSo01 and c19ProSo03.

To compare these variables with the variables in our focus country. For the c19ProSo01, the similar thing is they all have c19ProSo02 and c19ProSo04 as best predicator. The different is for our focus country we have c19NormDo as best predictor, but in similar countries data it uses the c19ProSo03 as best predictor.

For the c19ProSo02, the similar are they both have c19ProSo01 and c19ProSo03, then different is for focus country we have one more c19IsPunish as predictor.

For the c19ProSo03, the similar are they both have c19ProSo02 and c19ProSo04, the different is in similar countries data, it uses c19ProSo01 also.

Lastly, consider about c19ProSo04, they both have c19ProSo01 and c19ProSo03 as best predictor and with no different.

To conclusion, for pro-social attitudes variables the cluster of similar countries give the better match. Because for other countries data have much best predictor for the focus country data. but that situation is not happened in similar countries data. Also, we can see that for c19ProSo01-4, for c19ProSo01 there is only one different, for c19ProSo02 different is focus country data have one extra best predictor than similar countries data. For c19ProSo03, similar countries data have extra one. And is totally same in c19ProSo04. Then we can see that similar countries data are more match to focus country data than the other countries data.

This situation happen I think is due to the similar countries data is similar with focus data in lots of parts these parts also contain the c19ProSo01-4.

Appendix

Rcode

```
#install and import libraries
#install only used when you not have these packages.
#install.packages("ggplot2")
#install.packages('dplyr')
#install.packages("reshape")
#install.packages("tidyverse")

library(ggplot2)
library(dplyr)
library(reshape2)
library(cluster)
library(tidyverse)

rm(list = ls()) #clear data in environment
setwd("C:/Users/WENGCHENLONGJIE/Desktop/FIT3152/Assignment1") #set directory
set.seed(29334152) # seed is student ID
cvbase = read.csv("PsyCoronaBaselineExtract.csv")
cvbase <- cvbase[sample(nrow(cvbase), 40000), ] # 40000 rows

# 1(a)
dim(cvbase)
str(cvbase) #see the variable type and some head data
```

```
summary(cvbase) #more detail about data.


##########################################################################################
#

# 1(b)

#now remove the na value for each variable use the median to instead it.
#affAnx
median_col1 <- median(cvbase$affAnx, na.rm = TRUE)
cvbase$affAnx[is.na(cvbase$affAnx)] <- median_col1
#affCalm
median_col1 <- median(cvbase$affCalm, na.rm = TRUE)
cvbase$affCalm[is.na(cvbase$affCalm)] <- median_col1
#affContent
median_col1 <- median(cvbase$affContent, na.rm = TRUE)
cvbase$affContent[is.na(cvbase$affContent)] <- median_col1
#affBor
median_col1 <- median(cvbase$affBor, na.rm = TRUE)
cvbase$affBor[is.na(cvbase$affBor)] <- median_col1
#affEnerg
median_col1 <- median(cvbase$affEnerg, na.rm = TRUE)
cvbase$affEnerg[is.na(cvbase$affEnerg)] <- median_col1
#affDepr
median_col1 <- median(cvbase$affDepr, na.rm = TRUE)
cvbase$affDepr[is.na(cvbase$affDepr)] <- median_col1
#affExc
median_col1 <- median(cvbase$affExc, na.rm = TRUE)
cvbase$affExc[is.na(cvbase$affExc)] <- median_col1
#affNerv
median_col1 <- median(cvbase$affNerv, na.rm = TRUE)
cvbase$affNerv[is.na(cvbase$affNerv)] <- median_col1
#affExh
median_col1 <- median(cvbase$affExh, na.rm = TRUE)
cvbase$affExh[is.na(cvbase$affExh)] <- median_col1
#affInsp
median_col1 <- median(cvbase$affInsp, na.rm = TRUE)
cvbase$affInsp[is.na(cvbase$affInsp)] <- median_col1
#affRel
median_col1 <- median(cvbase$affRel, na.rm = TRUE)
cvbase$affRel[is.na(cvbase$affRel)] <- median_col1
#PLRAC19
median_col1 <- median(cvbase$PLRAC19, na.rm = TRUE)
cvbase$PLRAC19[is.na(cvbase$PLRAC19)] <- median_col1
#PLRAEco
median_col1 <- median(cvbase$PLRAEco, na.rm = TRUE)
cvbase$PLRAEco[is.na(cvbase$PLRAEco)] <- median_col1
#disc01
cvbase$disc01[is.na(cvbase$disc01)] <- 0
#disc02
cvbase$disc02[is.na(cvbase$disc02)] <- 0
#disc03
```

```
cvbase$disc03[is.na(cvbase$disc03)] <- 0
#jbInsec01
cvbase$jbInsec01[is.na(cvbase$jbInsec01)] <- 0
#jbInsec02
cvbase$jbInsec02[is.na(cvbase$jbInsec02)] <- 0
#jbInsec03
cvbase$jbInsec03[is.na(cvbase$jbInsec03)] <- 0
#jbInsec04
cvbase$jbInsec04[is.na(cvbase$jbInsec04)] <- 0
#employstatus_1
cvbase$employstatus_1[is.na(cvbase$employstatus_1)] <- 0
#employstatus_2
cvbase$employstatus_2[is.na(cvbase$employstatus_2)] <- 0
#employstatus_3
cvbase$employstatus_3[is.na(cvbase$employstatus_3)] <- 0
#employstatus_4
cvbase$employstatus_4[is.na(cvbase$employstatus_4)] <- 0
#employstatus_5
cvbase$employstatus_5[is.na(cvbase$employstatus_5)] <- 0
#employstatus_6
cvbase$employstatus_6[is.na(cvbase$employstatus_6)] <- 0
#employstatus_7
cvbase$employstatus_7[is.na(cvbase$employstatus_7)] <- 0
#employstatus_8
cvbase$employstatus_8[is.na(cvbase$employstatus_8)] <- 0
#employstatus_9
cvbase$employstatus_9[is.na(cvbase$employstatus_9)] <- 0
#employstatus_10
cvbase$employstatus_10[is.na(cvbase$employstatus_10)] <- 0
#PFS01
cvbase$PFS01[is.na(cvbase$PFS01)] <- 0
#PFS02
cvbase$PFS02[is.na(cvbase$PFS02)] <- 0
#PFS03
cvbase$PFS03[is.na(cvbase$PFS03)] <- 0
#fail01
cvbase$fail01[is.na(cvbase$fail01)] <- 0
#fail02
cvbase$fail02[is.na(cvbase$fail02)] <- 0
#fail03
cvbase$fail03[is.na(cvbase$fail03)] <- 0
#happy
median_col1 <- median(cvbase$happy, na.rm = TRUE)
cvbase$happy[is.na(cvbase$happy)] <- median_col1
#lifeSat
median_col1 <- median(cvbase$lifeSat, na.rm = TRUE)
cvbase$lifeSat[is.na(cvbase$lifeSat)] <- median_col1
#MLQ
median_col1 <- median(cvbase$MLQ, na.rm = TRUE)
cvbase$MLQ[is.na(cvbase$MLQ)] <- median_col1
#c19NormShould
median_col1 <- median(cvbase$c19NormShould, na.rm = TRUE)
cvbase$c19NormShould[is.na(cvbase$c19NormShould)] <- median_col1
```

```
#c19NormDo
median_col1 <- median(cvbase$c19NormDo, na.rm = TRUE)
cvbase$c19NormDo[is.na(cvbase$c19NormDo)] <- median_col1
#c19IsStrict
median_col1 <- median(cvbase$c19IsStrict, na.rm = TRUE)
cvbase$c19IsStrict[is.na(cvbase$c19IsStrict)] <- median_col1
#c19IsPunish
median_col1 <- median(cvbase$c19IsPunish, na.rm = TRUE)
cvbase$c19IsPunish[is.na(cvbase$c19IsPunish)] <- median_col1
#c19IsOrg
median_col1 <- median(cvbase$c19IsOrg, na.rm = TRUE)
cvbase$c19IsOrg[is.na(cvbase$c19IsOrg)] <- median_col1
#trustGovCtry
median_col1 <- median(cvbase$trustGovCtry, na.rm = TRUE)
cvbase$trustGovCtry[is.na(cvbase$trustGovCtry)] <- median_col1
#trustGovState
median_col1 <- median(cvbase$trustGovState, na.rm = TRUE)
cvbase$trustGovState[is.na(cvbase$trustGovState)] <- median_col1
#gender
cvbase$gender[is.na(cvbase$gender)] <- 3
#age
median_col1 <- median(cvbase$age, na.rm = TRUE)
cvbase$age[is.na(cvbase$age)] <- median_col1
#edu
median_col1 <- median(cvbase$edu, na.rm = TRUE)
cvbase$edu[is.na(cvbase$edu)] <- median_col1
#c19ProSo01
cvbase$c19ProSo01[is.na(cvbase$c19ProSo01)] <- 0
#c19ProSo02
cvbase$c19ProSo02[is.na(cvbase$c19ProSo02)] <- 0
#c19ProSo03
cvbase$c19ProSo03[is.na(cvbase$c19ProSo03)] <- 0
#c19ProSo04
cvbase$c19ProSo04[is.na(cvbase$c19ProSo04)] <- 0


#######################################################################################
##

# 2(a)
attach(cvbase) #attach this data frame, then we can use the variable directly.

#here use the filter method(%>% is kind of pipe) to filter the focus country Iran as one dataframe and
other countries as other data frame
Focus_country = cvbase %>% filter(coded_country == "Iran")
Other_countries = cvbase %>% filter(coded_country != "Iran")

#this one is to see the how other countries and focus country mean value in different variable
Focus_country = Focus_country %>% summarise_all(.funs = mean, na.rm = TRUE)
Focus_country$coded_country[1]="Iran"
Other_countries = Other_countries %>% summarise_all(.funs = mean, na.rm = TRUE)
Other_countries$coded_country[1]="Other"
merged_df <- rbind(Focus_country, Other_countries)
```

```
# use gather() to make the variable be tidy, make df change be to long format
df_tidy <- merged_df %>% gather(key = "variable", value = "value", -coded_country)

#this is use the ggplot to draw a bar chart, data is df_tidy, each variable is the x-axis, and value of that
variable is the y, fill color with country
ggplot(data = df_tidy, aes(x = variable, y = value, fill = coded_country)) + #use dodge mean the bar is next
to each other which make compare easier
geom_bar(stat = "identity", position = "dodge") + #The stat = "identity" option means to use the actual
value in the data instead of the default count statistics.
labs(x = "Variable", y = "Value", fill = "Country") + #The labs() function is used to add chart titles and axis
labels
theme(axis.text.x = element_text(angle = -90, vjust = 0, hjust=0)) #The theme() function is used to change
the Angle and position of the axis label.

detach()


###########################################################################################
#

# 2(b)
#here use the filter method(%>% is kind of pipe) to filter the focus country Iran as one dataframe and
other countries as other data frame
Focus_country = cvbase %>% filter(coded_country == "Iran")
# Remove the coded_country column as all are Iran. there is two different way
#Focus_country = subset(Focus_country, select = -coded_country)
Focus_country$coded_country = NULL

attach(Focus_country)

c19ProSo01.fit <-lm(c19ProSo01 ~ ., data = Focus_country)
summary(c19ProSo01.fit)
max(coef(c19ProSo01.fit))

c19ProSo02.fit <-lm(c19ProSo02 ~ ., data = Focus_country)
summary(c19ProSo02.fit)
max(coef(c19ProSo02.fit))

c19ProSo03.fit <-lm(c19ProSo03 ~ ., data = Focus_country)
summary(c19ProSo03.fit)
max(coef(c19ProSo03.fit))

c19ProSo04.fit <-lm(c19ProSo04 ~ ., data = Focus_country)
summary(c19ProSo04.fit)
max(coef(c19ProSo04.fit))

detach()

###########################################################################################
#

# 2(c)
# repeat what we do in 2b but now we use the other countries as the data
#here use the filter method(%>% is kind of pipe) to filter the focus country Iran as one dataframe and other countries as other data frame
Other_countries = cvbase %>% filter(coded_country != "Iran")
# Remove the coded_country column as all are not Iran. there is two different way
Other_countries = subset(Other_countries, select = -coded_country)
#Focus_country$coded_country = NULL

attach(Other_countries)

c19ProSo01.fit <-lm(c19ProSo01 ~ ., data = Other_countries)
summary(c19ProSo01.fit)
max(coef(c19ProSo01.fit))

c19ProSo02.fit <-lm(c19ProSo02 ~ ., data = Other_countries)
summary(c19ProSo02.fit)
max(coef(c19ProSo02.fit))

c19ProSo03.fit <-lm(c19ProSo03 ~ ., data = Other_countries)
summary(c19ProSo03.fit)
max(coef(c19ProSo03.fit))

c19ProSo04.fit <-lm(c19ProSo04 ~ ., data = Other_countries)
summary(c19ProSo04.fit)
max(coef(c19ProSo04.fit))

#par(mfrow = c(2,2)) can be help for see the linear relation
#plot(c19ProSo01.fit)
#plot(c19ProSo02.fit)
#plot(c19ProSo03.fit)
#plot(c19ProSo04.fit)

detach()


########################################################################################

# 3(a)
#here is to make coded_country column to be last column
C = cvbase %>% select(-coded_country, everything(), coded_country)
attach(C)
#then here we group it by country names and calculate the each variables(we chooses) means
C = C %>% select(-everything(), c(c19ProSo01, c19ProSo02, c19ProSo03, c19ProSo04, jbInsec03, fail01, fail02, fail03, disc03, age, c19NormDo), coded_country)
C = C %>% group_by(`coded_country`) %>% summarise_all(mean, na.rm = TRUE)
detach()
#then we scaling the C data
C_scaled = scale(C[,-1])
#set the seed for nstart, if not set the seed, every time we do this will different
set.seed(29334152)
#here we need to find 3-7 similar countries from 112 countries, so we set as 35 clusters.
ckfit = kmeans(C_scaled, centers = 35, nstart = 20) #35
#the code below is used to find the cluster which contain the country Iran.
```

```
similar_countries <- C[ckfit$cluster == ckfit$cluster[which(C$coded_country == "Iran")], "coded_country"]
#then we print it out we can see which cluster contain Iran(focus country) and the similar coutries for Iran.
similar_countries

#this code below is print 3 which is the which cluster Iran in.
ckfit$cluster[which(C$coded_country == "Iran")]
T1=table(actual=C$coded_country, fitted=ckfit$cluster)
T1=as.data.frame.matrix(T1) #converts the table T1 into a data frame.
#then we look at the table in column 16 we can also see the similar country.
T1



################################################################################################
#

# 3(b)
# repeat what we do in 2b but now we use the similar countries as the data
#here use the filter method(%>% is kind of pipe) to filter the focus country Iran as one dataframe and
similar countries as other data frame
Similar_countries = cvbase %>% filter(coded_country %in% c("Bosnia and Herzegovina", "Hong Kong
S.A.R.", "Moldova", "Poland", "Republic of Serbia", "Romania"))
# Remove the coded_country column as all are similar countries. there is two different way
Similar_countries = subset(Similar_countries, select = -coded_country)
#Focus_country$coded_country = NULL

attach(Similar_countries)

c19ProSo01_similar.fit <-lm(c19ProSo01 ~ ., data = Similar_countries)
summary(c19ProSo01_similar.fit)
max(coef(c19ProSo01_similar.fit))

c19ProSo02_similar.fit <-lm(c19ProSo02 ~ ., data = Similar_countries)
summary(c19ProSo02_similar.fit)
max(coef(c19ProSo02_similar.fit))

c19ProSo03_similar.fit <-lm(c19ProSo03 ~ ., data = Similar_countries)
summary(c19ProSo03_similar.fit)
max(coef(c19ProSo03_similar.fit))

c19ProSo04_similar.fit <-lm(c19ProSo04 ~ ., data = Similar_countries)
summary(c19ProSo04_similar.fit)
max(coef(c19ProSo04_similar.fit))
detach()
```