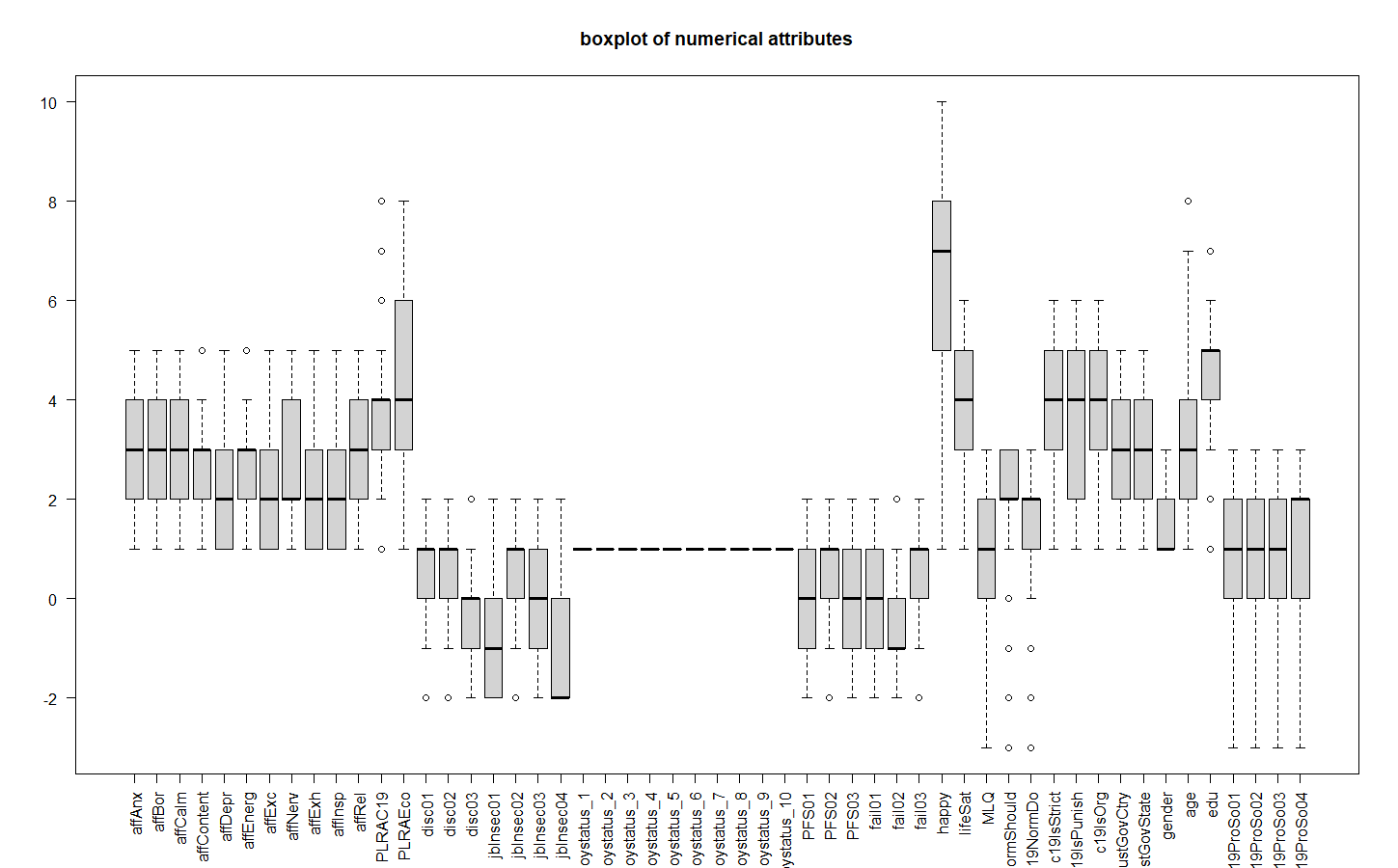
**Name: Kim Ze Lam**

**Student ID: 31860346**

**Question 1a**

The data set contains the information regarding a search on identifying the important predictors of prosocial COVID-19 behaviors among global participants. The original data set contains a total dimension of 64381 rows and 54 columns and only 40000 rows and 54 columns of them are abstracted randomly from the original data set.

The data types in the data set includes integers and characters. Except Country Self Report is character data types, all other attributes are in integer data types.



There are a total of 53 numerical attributes in the data set. They are being classified into 14 groups includes Affect, Likelihood, Societal Discontent, Job Insecurity, Employment Status, Perceived Financial Strain, Disempowerment, Life Satisfaction, Corona Community Injunctive norms, Trust in Government, Gender, Age, Education and Corona ProSocial Behavior.

Boxplot is used to analyze these attributes by showing the distribution. The right-skewed distribution concentrates at the lower end (disagree side) of the distribution and some data widely spread towards the higher end. In contrast, the left-skewed distribution is the majority of data values concentrated towards the higher end and a few data values widely spread towards the lower end. The interquartile range is the range between first and third quartiles of the dataset. The lower end of the box represents the first quartile and the upper end indicates the third quartile. Median is important to measure central tendency since it is less impacted by extreme data values or outliers. It provides the middle location of the data which is a useful summary statistic. If the median is closer to the first quartile, lower half data will cluster denser than the upper half and vice versa.

Overall, the majority of the participants are given quite positive responses in the survey. This indicates that most of them had accepted and adapted to the new lifestyle. The impact of the epidemic has been reduced in terms of emotional willingness to help others who suffer from COVID-19, job opportunities, life satisfaction and so on. However, most of the people still think that they should still abide by the epidemic prevention rules which is a good sign of awareness of epidemic prevention.

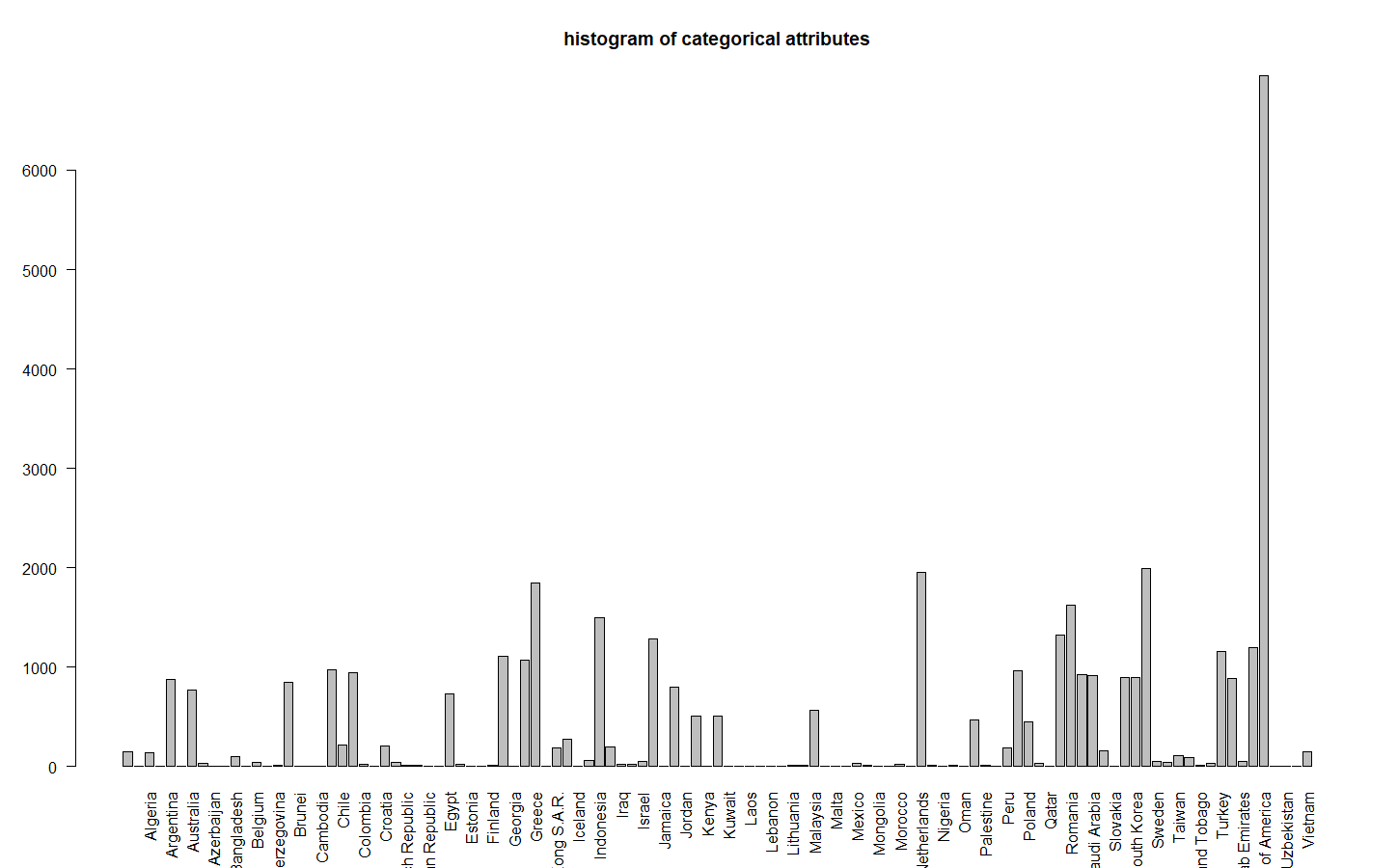
The distribution of people thinking they should do self-isolation and engage in social distancing which is attribute c19NormShould give a left skewed. The median of the attribute is in 2. Almost every participants agree on this statement. The range of the distribution is from 1 to 3 which covers all the positive responses. There are outliers in this attributes who response negatively on importance of self-isolation and engage in social distancing. After affected by the COVID-19 pandemic, most of the people had realized the severity of the outbreak and importance of epidermic prevention, pay attention to the means of epidemic prevention.

The distribution of people rating on they have a clear sense of purpose which is MLQ attribute is skewed to the left. The range cover from -3 to 3 which is quite a large range that says the high variability of dataset and wide spread of range of value. Majority of the people are still having clear sense of purpose which occupied 75% of them. There had already 25% of the people had disagree on the statement. After experienced two years of epidermic, most of the people had recovered or get used to the new environment. They had regained or redefined their sense of purpose from being destroyed by the devastation of the epidemic.

There is no information from the distribution for all employment status attributes because only 1 and NA data throughout the columns of the employment status attributes. The NA in the column need to be preprocess before analyzing the data.

The gender attributes show a right-skewed distribution. Most of them participants are female which is 50% and male and others are both 25%. This unbalance proportion of gender may affect the accuracy of the result collected.

The Corona ProSocial Behavior show a left-skewed distribution for all its attributes. The range of the distribution is from -3 to 3 which is a large range. This indicates the data values are spread out over a wide range of value and the dataset had a high variability. There are 75% of the people are willing to participate in Corona ProSocial and only 25% of them disagree. This suggest that there is a general inclination among the public to help others during the pandemic, which could have positive implications for public health and well-being.



The only non-numerical attributes exist in the data set is country self-report (coded\_country) which means the country the participant come from. There are 110 different countries exist in the dataset and 152 empty string. The country exists the most frequently is “United States of America” which occur 6948 times. There are 18 countries occurs the least amount of time which is 1 includes Armenia, Benin, Cambodia, Dominican Republic, Estonia, Ethiopia, Kenya, Krygyzstan, Laos, Latvia, Malta, Mauritius, Mongolia, Nepal, Oman, Qatar, Uruguay and Uzbekistan. Since the proportion of country is having a large difference, the survey finding will more fitted to the country that the most participants involved.

There is a total of 425902 missing in the data set. The missing values can be separated into two groups which is useful and redundant. The useful missing values carries some information that we can use to analyze the data set but we need to preprocess them before analyzing. For instance, all the attributes of employment status are only filled with 1 and missing value for disagree. We can convert the missing value as 0 to represent disagree. For the redundant missing values, we can check the number of missing values of each columns. If the proportion of missing values are too high, we could remove the entire column. Then, we could remove the rows which contain missing values from the analysis.

**Question 1b**

**Pre-processing**

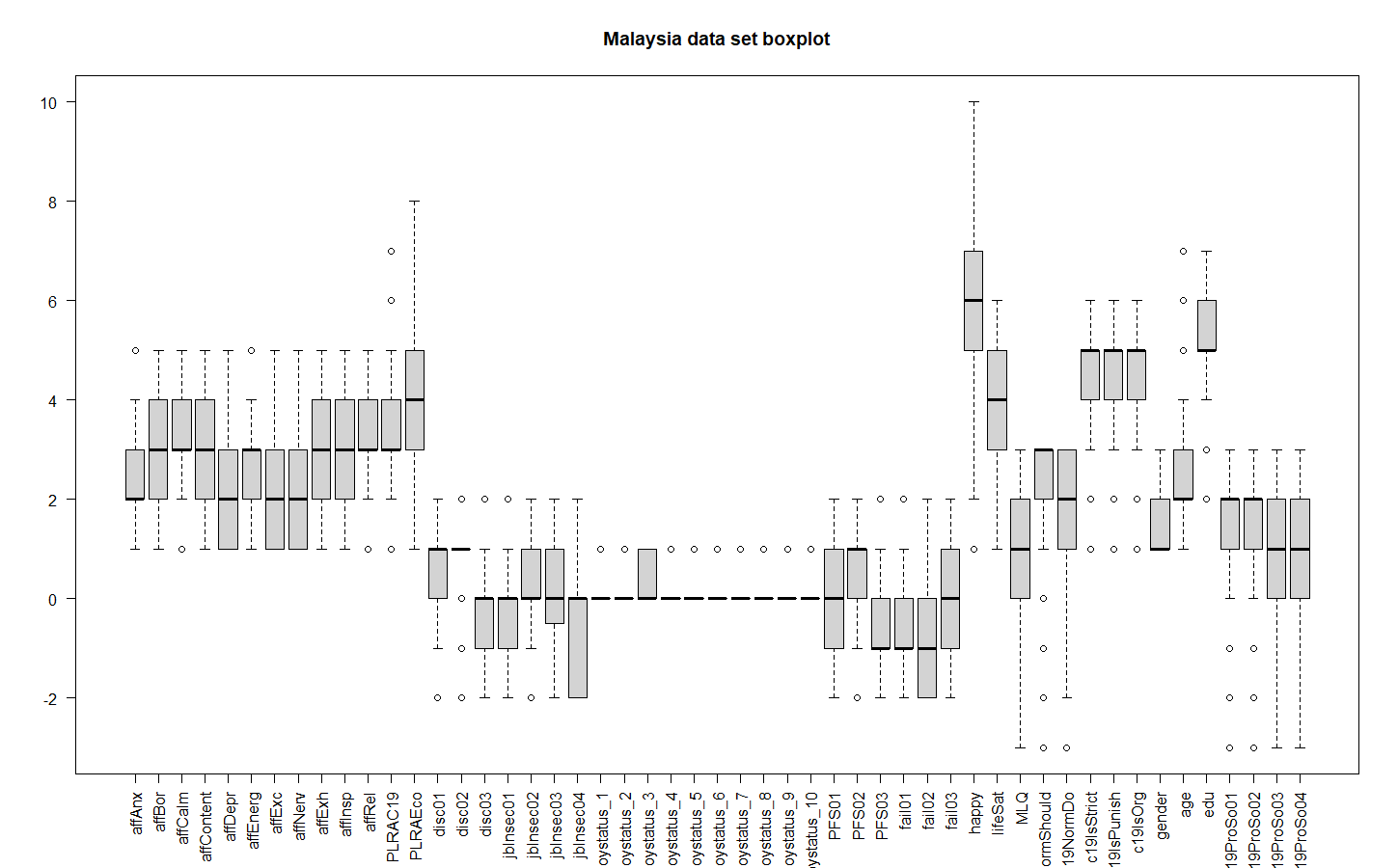
* 1. Changing NA of all attributes of Employment Status (employstatus\_1 to employstatus\_10) to 0. The NAs in Employment Status means NO which is not redundant data.
  2. NA is one of the option in job insecurity (jbInsec01 to jbInsec04). The option means the question is not applicable to the participants but it does not mean redundant. The NA should be change to 0 (Neither agree nor disagree) because it is a useful data.

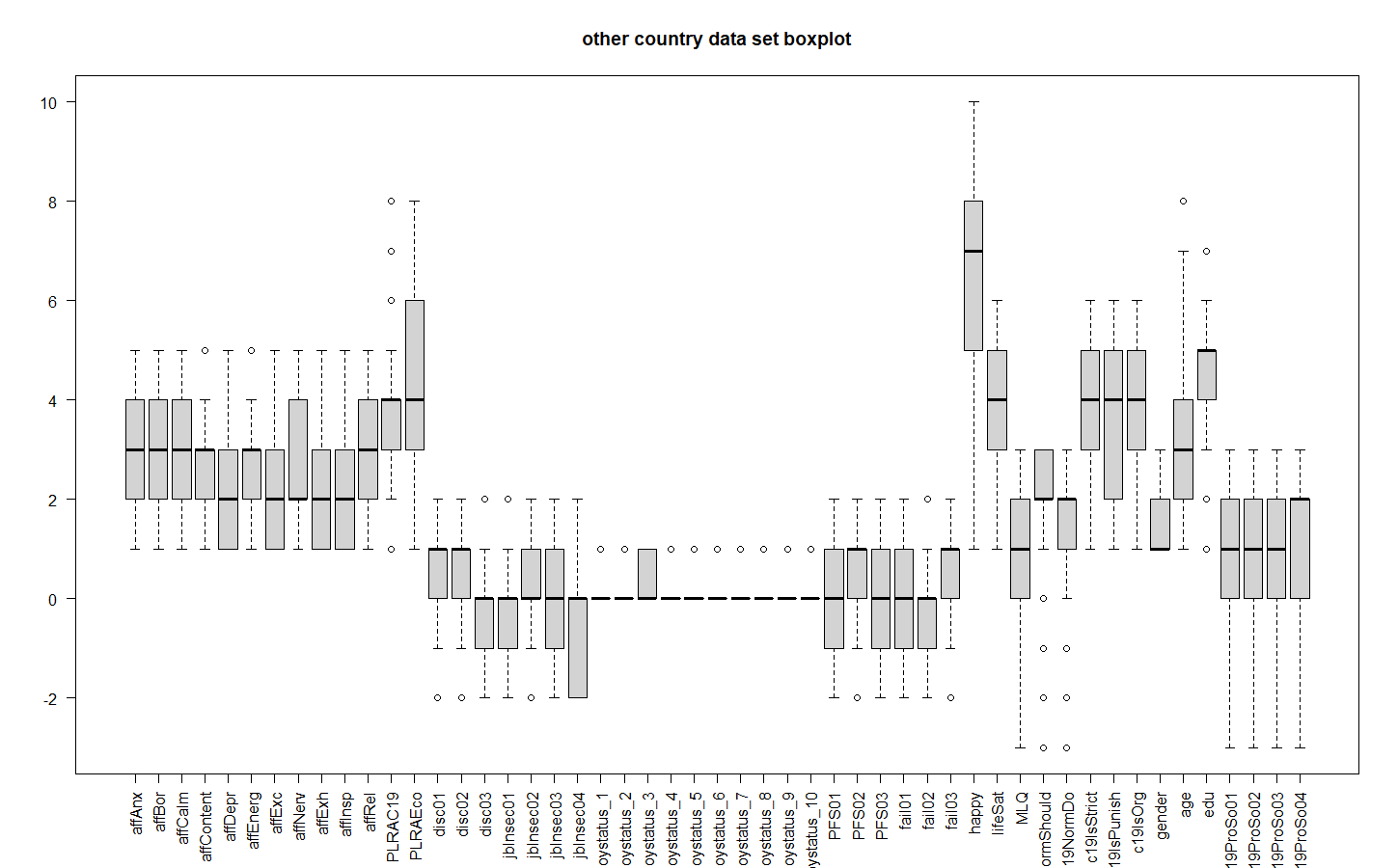
because there is part of population that does not suitable to answer this question or choose not to answer this question. For example, students do not have any job hence they could not answer anything related to job insecurity and employment status. We can only group them to no preference to avoid bias on the distribution, resulting in inaccurate interpretation. Besides, some people refuse to

1. All the columns with missing value more than 20% should be removed. This is because the large portion of missing values will make the predictors highly inaccurate and losing its predictability. As a result, the predictors become redundant. The trustGovCtry and trustGovState attributes are both having over 20% of missing values in the data sets and hence should be dropped from the analysis.
2. All the rows with missing value (after pre-processing (1) ) are removed from the data because other attributes mean not given input and should be removed from the analysis.

After processing, the data sets remaining has a dimension of 37936 rows and 52 columns.

**Question 2a**





The country self-report columns of Malaysia datasets and other country datasets are removed. This is because the value of self-reported column is all Malaysia and hence the column is redundant. The value of country in other country datasets are different but they work as a group and hence the country data is also redundant. The dimension of Malaysia data set is now 552 rows and 51 columns and 37384 rows and 51 columns.

I plotted one boxplot for each datasets for Malaysia and other country to compare the difference in responses of Malaysia participants with other countries in the survey. The most significant difference that we could observed in the attributes of in c19ProSo01 and c19ProSo02 pro-social attitudes of Malaysia participants are overall higher than other countries. The median of c19ProSo01 and c19ProSo02 of Malaysia dataset are in 2 while the other country’s median on those predictors is in 3. The range of both predictors in Malaysia datasets are from 0 to 5 while in other countries have a wide spread of range from -3 to 5. We can conclude from the dataset that 50% of the participants are agree and very agree on participate in the pro-Social activities and another 50% are at least neutral with it. However, there are some outlier on the negative responses in both predictors but they are extremely small number. However, there are 50% of people in the range of very agree to somewhat agree. Another 50% of the people are from neutral to strongly disagree which is quite a large spread of responses all over the choices. After comparing, we could see that almost every participants are having a neutral to positive pro-social attitudes while other countries only having 75% of the people are at least neutral or showing positive pro-social attitudes in c19ProSo01 and c19ProSo02. We could see that Malaysia people are helpful brought by the mutual tolerance and accommodation among multiracial.

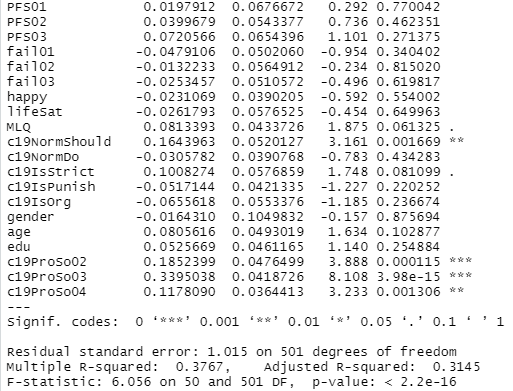
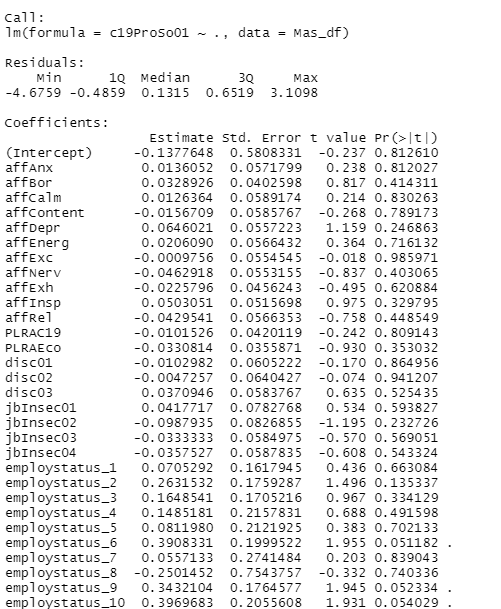
Besides, the overall emotional affect on Malaysia participant is lower compared to other countries participants. From the two boxplot, we could see that the affAnx predictors in Malaysia datasets have a lower rating compared to other countries. The range of Malaysia is from 1 to 4 while other countries cover from 1 to 5. The median of Malaysia datasets is at 2 while other countries is at 3. We can interpret that 50% of Malaysia participant feels lower anxious but others country only 25% of participants in this sector. Another 50% the Malaysia participants spread on 2 to 4 which is which is Moderate to quite a bit while other countries having another 50% on moderate to very anxious. This may occur due to the government's policy to deal with the epidemic is very sound makes the citizen of Malaysia trust the country. This could be support by the predictor trustGovCty. We can observe that there are 50% of the participant of the Malaysia extremely trust the government are taking right measure to deal with COVID-19 pandemic while other country only having 25% of them are extremely trust their government.

In the Corona Community Injunctive norms, Malaysia participant prefer a stricter norm and epidermic prevention measure compared to other countries. From the boxplot, we can see that the median of the predictors of c19IsStrict, c19IsPunish and c19IsOrg are all at 5 in Malaysia but only 4 in other countries. There are half of the people extremely support a stricter norm and epidermic prevention measures which vote 5 to 6. The range of the three predictors are all from 3 to 6 which shows the majority of them are wishing that the Corona Community Injunctive norms could be stricter. However, there are some outliers on the negative responses which not that likely to support the stricter Corona Community Injunctive norms. Comparing to other countries, other countries datasets give a median of 4 and a range from 1 to 6. This means that the 50% will be spread out to 4 compared to Malaysia datasets which reduce the overall willingness. There are 50% giving a low support to the stricter Corona Community Injunctive norms means that there are still a large portion of participants from other countries are not wishing a stricter Corona Community Injunctive norms. I think this phenomena happens because Malaysia participants are having a higher education level compared to other country and also caused by the Malaysia participant is facing the epidemic with an optimistic attitude. From the boxplot, we could see that the median of education level of Malaysia participant is 5 and other countries is 4. The range of Malaysia participant education is from 4 to 7 while other countries are in 1 to 7. Malaysia participants having at least education level of 5 but other countries are at least 4 for the higher 50% of the data. For the other 50%, Malaysia participants have at least an education level of 4 but other countries only have at least 1 for the other 50%. Besides, Malaysia participants have some outliers on lower education level which means only very little participants have a low education level. The education level of other countries participants is widely spread throughout the scale but Malaysia participants are focus on higher education level. The high education level cause by the country pay attention on citizen’s education and resulting in a country having most of the people able to judge things rationally.

**Question 2b**

Throughout the whole predictors’ selection, only the Malaysia datasets are being used to build the linear regression model.

C19ProSo01:

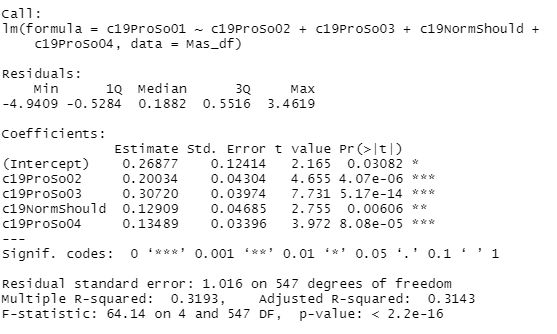


The correlation of the attributes C19ProSo01 with all other predictors is overall performed quite badly. Without including the other 3 pro-social attitudes predictors, the highest positive correlation is only 0.17 with C19NormShould and C19IsStrict attributes. This means when C19NormShould or C19IsStrict attributes increase by 1 unit, the C19ProSo01 will increase by 0.17 and it has the strongest relationship with C19ProSO01. There are even two predictors given 0 relationship with the attributes. The highest negative correlation is fail01 predictors which has a correlation of -0.13. The negative correlation means that when the fail01 predictor increases by 1 unit, C19ProSo01 will decrease by 0.13 units.

After other pro-social attitudes predictors are included, there exist some predictors where the correlation is relatively stronger. All of the pro-social attitude’s predictors give a relatively good correlation with C19ProSo01 attributes compared to the other attributes. c19ProSo03 shows the best correlation among the predictors which has a correlation of 0.5 and followed by c19ProSo02 with 0.43 and c19ProSo04 with 0.35 correlation. This means that they are very strong predictors compared to others because a slight increase in these predictors will affect C19ProSo01 quite a lot.

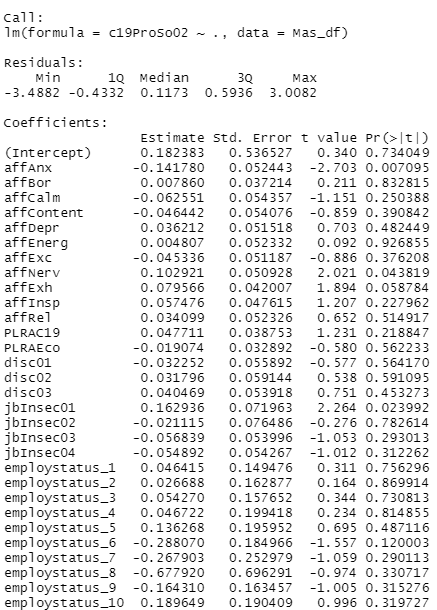
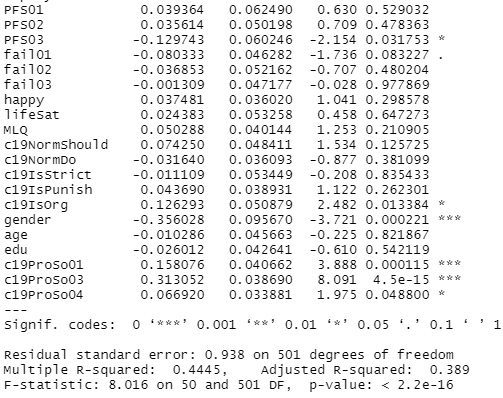
When we plot the linear regression models, the best predictors are c19ProSo02(\*\*\*) and c19ProSo03(\*\*\*) which having the 3 stars and followed by c19NormShould(\*\*) and c19ProSo04(\*\*) which has 2 stars. 3 stars means that the p-value of c19ProSo02 and c19ProSo03 are very small and 2 stars are also small enough. All of the predictors c19ProSo02, c19ProSo03, c19NormShould and c19ProSo04 are selected as best predictors because their p-value are 0.000115, 3.98e-15, 0.001669 and 0.001306 respectively which is all smaller than 0.05. The small p-value indicates the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. The R-squared value is 0.3767 tells us that the independent variables of the linear regression model can explain 37.67% of the variance in the dependent variables.

Another linear regression model is built using the best predictor chosen and using the R-squared value to judge how much variability the linear model could explain.



The R-squared value is now 0.3193 which performs worse than the initial R-squared value of 0.3767. One of the reason that the situation happens is because overfitting may happen. The linear models may include too many predictors and fit the noise rather than the underlying pattern. Hence, it will lead to a poor performance to predict other data. Besides, multicollinearity may happen due to some predictors are highly correlated with each other, predicting the unique effect of each attributes will be difficult and result in inaccurate regression coefficient and causes R-squared value decrease. Another reason could be occurrence of large number of outliers. They will highly affect the linear line and hence resulting in a worser R-squared.

C19ProSo02:

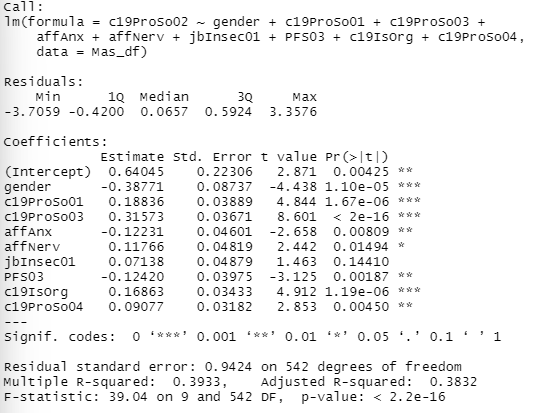
 

The correlation of the attributes C19ProSo02 with all other predictors are overall performed quite badly but better than C19ProSo01. Without including the other 3 pro-social attitudes predictors, the best positive correlation is 0.28 on C19ISOrg predictors. This means that one unit change in C19ISOrg, C19ProSo02 increase by 1 unit. There are quite a few predictors having a correlation around 0.2. Even it is still bad but improve from C19ProSo01. The strongest predictor with negative correlation is fail01 with -0.23 correlation. This explain that 1 unit increase in fail01 will cause C19ProSo01 drop by 0.23.

When other 3 pro-social attitudes predictors are included, some predictors with better correlation exist. All the pro-social attitudes predictors are showing the relatively high positive correlation with C19ProSo02 which is 0.43 for C19ProSo01, 0.52 for C19ProSo03 and 0.33 for C19ProSo04. The high correlation indicates that they are the best predictor since a slight change of these predictors will result in a large impact on C19ProSo02.

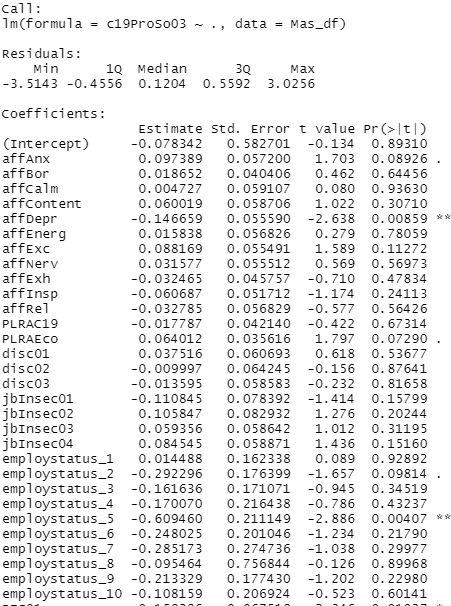
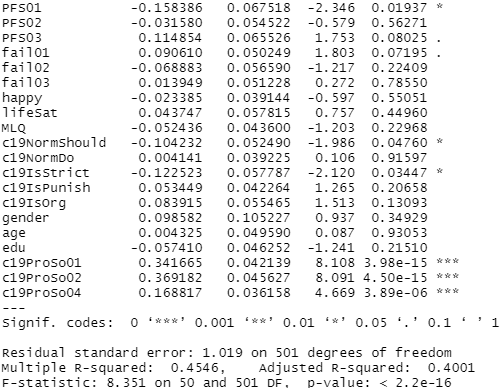
A linear regression model is plotted with Malaysia datasets. The best predictors selected are gender(\*\*\*), c19ProSo01(\*\*\*) and c19ProSo03(\*\*\*) with 3 stars, affAnx(\*\*) with two stars and affNerv(\*), jbInsec01(\*), PFS03(\*), c19IsOrg(\*), c19ProSo04(\*) with one stars. 3 stars means that the attributes are extremely important, 2 stars indicates that the attributes are very important and 1 star represent the predictors are important. They are selected because they are having stars behind the p-value. Besides, their p-value are smaller than 0.05 which had proved that they are important predictors. The small p-value indicates that the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. The R-squared value of the linear regression model is 0.4445 which indicates the independent variables of the linear regression model can explain 44.45% of the variance in the dependent variables.

I then plotted another linear using the best predictor chosen and using the R-squared value to judge how much variability the linear model could explain.



The R-squared value is now 0.3933 which is worser than the initial 0.4445. One of the reason that the situation happens is because overfitting may happen. The linear models may include too many predictors and fit the noise rather than the underlying pattern. Hence, it will lead to a poor performance to predict other data. Besides, multicollinearity may happen due to some predictors are highly correlated with each other, predicting the unique effect of each attributes will be difficult and result in inaccurate regression coefficient and causes R-squared value decrease. Another reason could be occurrence of large number of outliers. They will highly affect the linear line and hence resulting in a worser R-squared.

C19ProSo03:

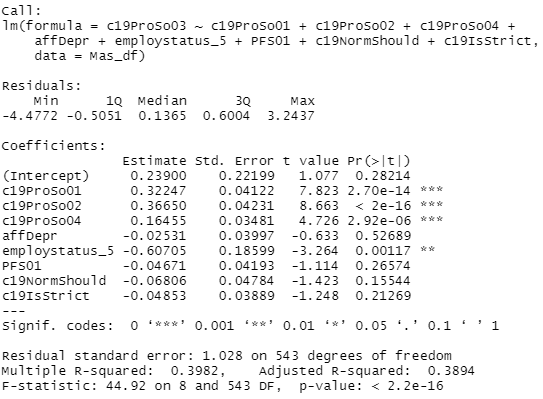
The correlation of the attributes C19ProSo03 with all other predictors are overall performed as bad as C19ProSo01. Excluding the 3 pro-social attitudes predictors, the highest correlation is 0.19 with the predictors C19IsOrg. This means that the highest positive correlation can affect predictors C19ProSo03 by 0.19 with 1 unit increase of it. The highest negative correlation is -0.13 with the predictors employstatus\_5 which indicates that one unit increase in emplystatus\_5 will cause the predictor C19ProSo03 decrease by 0.13 units.

When we include the other pro-social attitudes predictors, the 3 pro-social attitudes show the highest correlation with the dependent C19ProSo03 which are 0.5 for C19ProSo01, 0.52 for C19ProSo02 and 0.37 for C19ProSo04 correlation. These attributes are very stronger predictors compare other attributes since a small increase in these predictors increase the C19ProSo03 dependent by a large portion.

We plot the linear regression models, we could observe that the best predictors are c19ProSo01(\*\*\*), c19ProSo02(\*\*\*) and c19ProSo04(\*\*\*), followed by affDepr(\*\*) and employstatus\_5(\*\*) as well as PFS01(\*), c19NormShould(\*) and c19IsStrict(\*). 3 stars indicates the predictors is very significant, 2 stars are very important and important enough with 1 star.

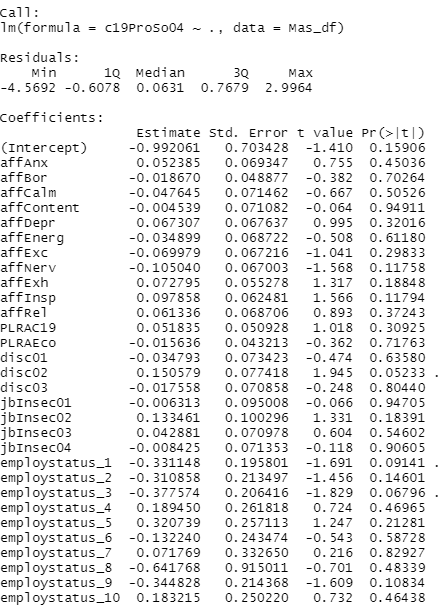
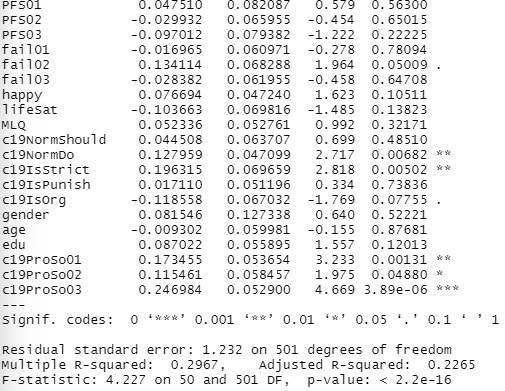
From the Malaysia data set, we can see that the c19ProSo01and c19ProSo02are the most significant predictors to predict the value of C19ProSo03 because they have the lowest p-value which is 0. 1.34e-07and 1.36e-09 respectively. They are important because they are having very small p-value which is smaller than 0.05. The small p-value indicates the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. The R-squared value is 0.4546 meaning that the independent variables of the linear regression model can explain 45.46% of the variance in the dependent variables.

Another linear regression model is built using the best predictor chosen and compare the R-squared value with the initial linear regression model to see whether the new linear model could explain larger variability compared to the initial linear model.



The R-squared value decrease from initial 0.4546 to 0.3982 which perform worser than including all predictors. One of the reason that the situation happens is because overfitting may happen. The linear models may include too many predictors and fit the noise rather than the underlying pattern. Hence, it will lead to a poor performance to predict other data. Besides, multicollinearity may happen due to some predictors are highly correlated with each other, predicting the unique effect of each attributes will be difficult and result in inaccurate regression coefficient and causes R-squared value decrease. Another reason could be occurrence of large number of outliers. They will highly affect the linear line and hence resulting in a worser R-squared.

C19ProSo04:

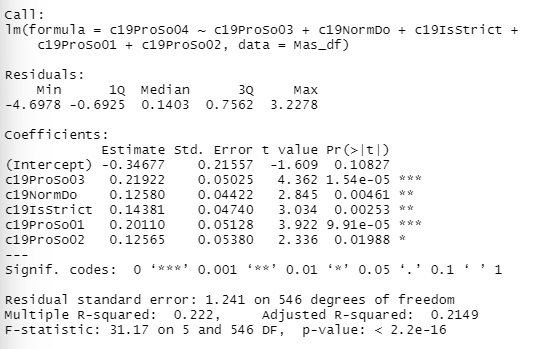
 

The correlation of the attributes C19ProSo04 with all other predictors are overall performed as bad as C19ProSo01. The strongest predictor without considering the 3 other pro-social attitudes is 0.23 with c19IsStrict. This shows that a unit increase in c19IsStrict will cause the dependent attribute C19ProSo04 increase by 0.23 units. The negative correlation is a lot smaller and hence no strong correlation with negative correlation.

When we include the 3 other pro-social attitudes, there exist a slightly better correlation. The 3 other pro-social attitudes give a relatively stronger correlation which are 0.35 for C19ProSo01, 0.33 for C19ProSo02 and 0.37 for C19ProSo03 with C19ProSo04. The impact of these predictors is slightly higher compared to other predictors.

When we plot the linear regression models, the important predictors are c19ProSo03(\*\*\*), followed by c19NormDo(\*\*), c19IsStrict(\*\*) and c19ProSo01(\*\*) as well as c19ProSo02(\*). The stars indicate how importance the predictors are. 3 stars is the most important, 2 stars means very important and 1 start mean important. All the predictors c19ProSo03, c19NormDo, c19IsStrict, c19ProSo01 and c19ProSo02 having p-value of 3.89e-06, 0.00682, 0.00502, 0.00131 and 0.04880 respectively which is all smaller than 0.05 represent the attributes are important. The small p-value indicates the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. The R-square 0.2967 tells us that the independent variables of the linear regression model can explain 29.67% of the variance in the dependent variables.

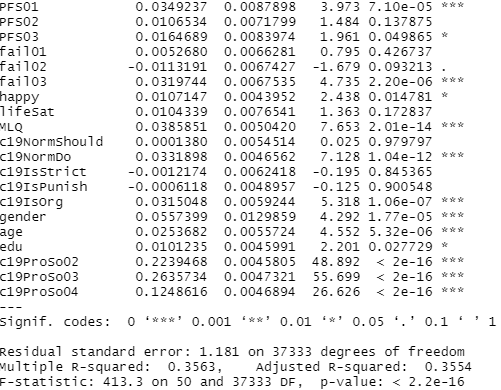
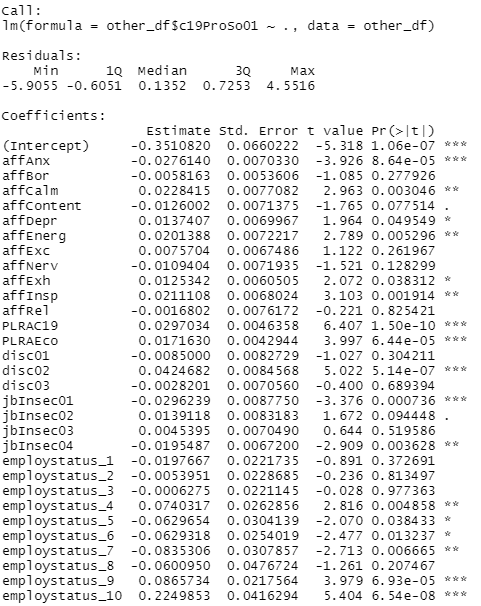
A linear regression model is built with the predictor chosen and using the R-squared value to judge how much variability the linear model could explain.



The R-squared value is now 0.222 which has a worser performance compared to the initial linear model with 0.2967. One of the reason that the situation happens is because overfitting may happen. The linear models may include too many predictors and fit the noise rather than the underlying pattern. Hence, it will lead to a poor performance to predict other data. Besides, multicollinearity may happen due to some predictors are highly correlated with each other, predicting the unique effect of each attributes will be difficult and result in inaccurate regression coefficient and causes R-squared value decrease. Another reason could be occurrence of large number of outliers. They will highly affect the linear line and hence resulting in a worser R-squared.

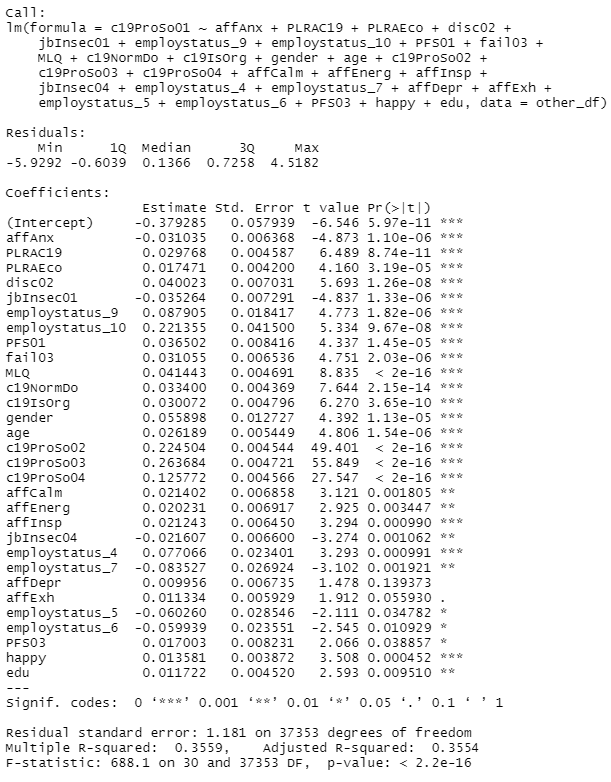
**Question 2c**

C19ProSo01:



We build a linear model linear regression model to help us in choosing the best predictors. affAnx, PLRAC19, PLRAECO, disc02, jbInsec01, employstatus\_9, employstatus\_10, PFS01, fail03, MLQ, c19NormDo, c19IsOrg, gender, age, c19ProSo02, c19ProSo03 and c19ProSo04 are all the strongest predictors. All of them having 3 stars of significant rating. They all have extremely small p-value which means much smaller than 0.05. The small p-value indicates the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. Then, affCalm, affEnerg, affInsp, jbInsec04, employstatus\_4, employstatus\_7, are also strong predictors. They have 2 stars of significant rating. They have very small p-value compared to 0.05. This also indicates that the predictors are important in the linear mode. There also exist predictors with a star of significant rating which are affDepr, affExh, employstatus\_5, employstatus\_6, PFS\_03, happy and edu. Even though they do not have a very small p-value, but their p-value are already small enough to become an important attributes. The R-squared value of the linear model is 0.3563 which means the independent variables of the linear regression model can explain 35.63% of the variance in the dependent variables.

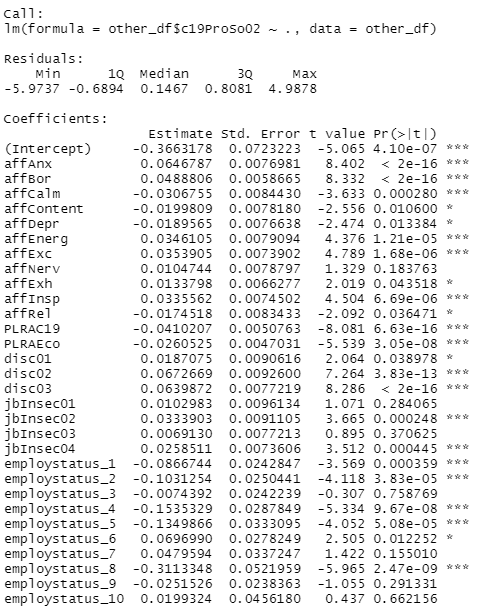
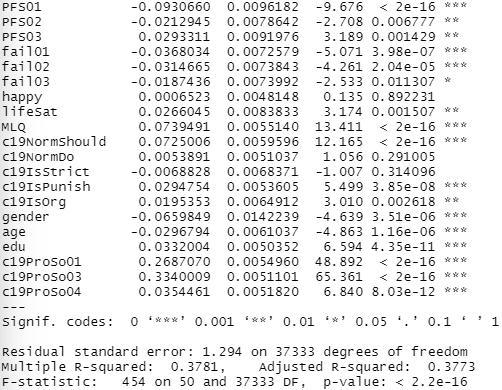
Another linear regression model is built using the best predictor chosen and using the R-squared value to judge how much variability the linear model could explain.



The R-squared value of the linear model built with the best attributes gives a slightly better R-squared value which is 0.3554 compared to initial R-square value 0.3563. This means that the new linear model with all good predictors able to explain a larger variability compared to the initial models.

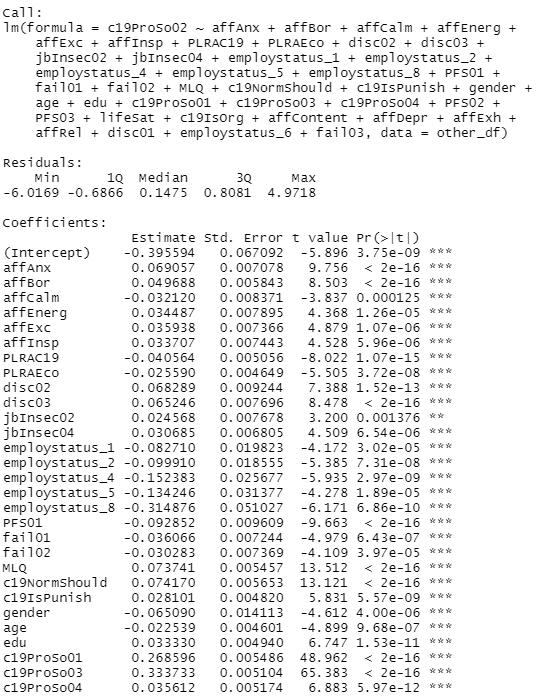
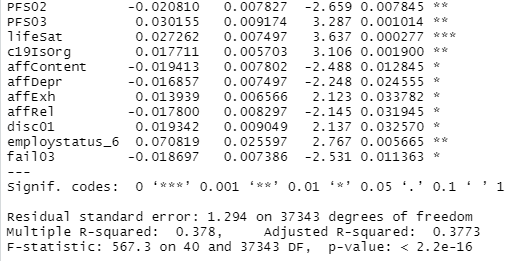
There are a lot more predictors being selected as important predictors when using other countries initial linear regression model in which 30 predictors compare to 5 from Malaysia initial linear regression model. The quality of the predictors other countries is better compared to Malaysia. A large number of predictors being selected will have a risk of over-fitting in which the model will also learn the noise as well. The most important predictors (3 stars or p-value is extremely small) in Malaysia model occupied half of the chosen predictors but other country having almost 66% of most important predictors. However, when we compare the R-squared value of linear model with significant attributes of Malaysia and other countries, other countries’ model gives a better value which is 0.3559 compared to Malaysia with only 0.3193. This means that Malaysia lost a lot of important information with the selected predictors. Selected attributes of other countries can explain 35.59% of variability of C19ProSo01 which is higher than Malaysia. One of the reason that this may occur is other countries are having much more data in their datasets compare to Malaysia’s datasets. The insufficient data makes to train the model will affect the accuracy of predicting the significance of each attributes. Hence, resulting in selecting wrong attributes as important predictors.

C19ProSo02:

A linear model is build with all predictos to find the best predictors to predict the dependent variable C19ProSo02. The most important attributes we choose are affAnx, affBor, affCalm, affEnerg, affExc, affInsp, PLRAC19, PLRAEco, disc02, disc03, jbInsec02, jbInsec04, employstatus\_1, employstatus\_2, employstatus\_4, employstatus\_5, employstatus\_8, PFS01, fail01, fail02, MLQ, c19NormShould, c19IsPunish, gender, age, edu, c19ProSo01, c19ProSo03 and c19ProSo04 with 3 stars rating of importance. Followed by 2 stars rating of importance predictors including PFS02, PFS03, lifeSat and c19IsOrg and 1 stars rating of importance predictor which are affContent, affDepr, affExh, affRel, disc01, employstatus\_6 and fail03. All of them are having small p-value which is smaller than 0.05. The smaller the p-value, the more important the predictors are. The 3 stars rating attributes are having the extremely small p-value and hence are the most important attributes. The 2 stars rating attributes are very small but larger than 3 stars predictors. 1 star rating attributes having a relatively bigger p-value but smaller enough become an important predictors which is still smaller than 0.05. The R-squared value of this linear model is 0.3781 imply the independent variables of the linear regression model can explain 37.81% of the variance in the dependent variables.

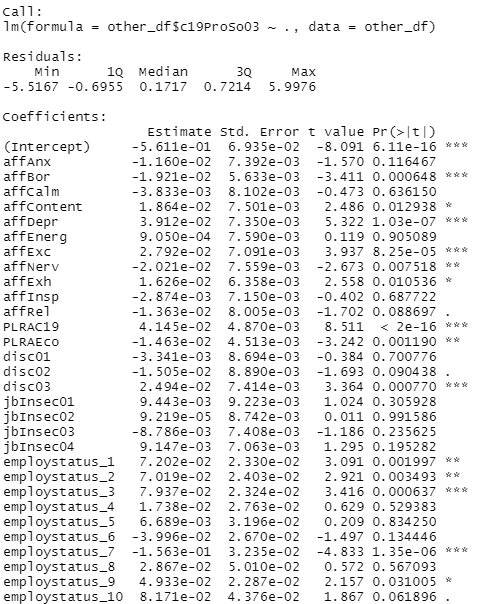
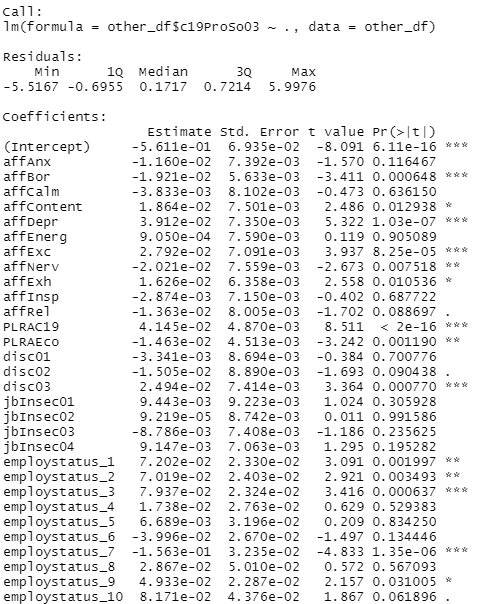
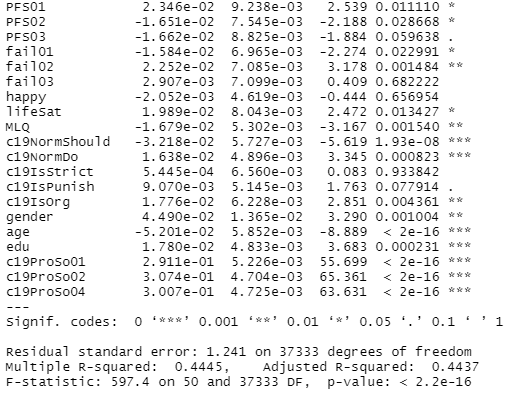
I then plotted another linear using the best predictor chosen and using the R-squared value to judge how much variability the linear model could explain.

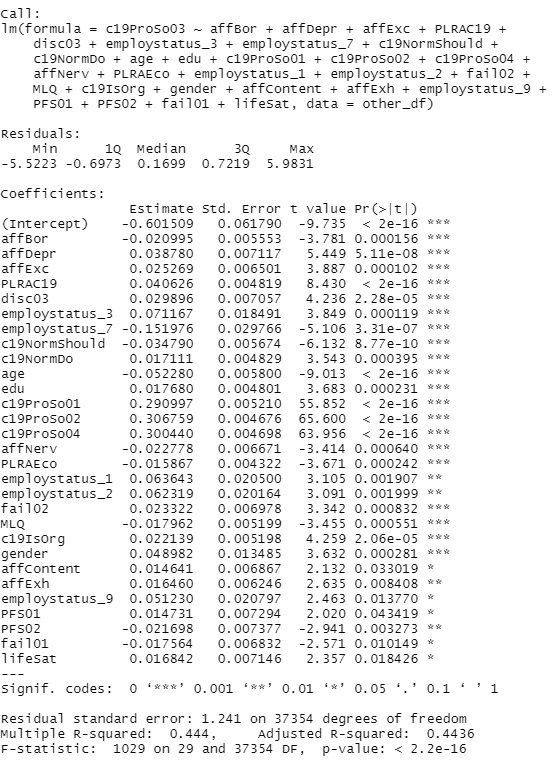
The R-squared value of the linear regression model with all significant predictors 0.378 which is almost similar with the initial linear model with R-squared value of 0.3781. The new combination of predictors does not perform better or worser compare to including every attributes. This may caused by the selected attributes were highly correlated to the removed predictors. The removed attributes may not cause a substantial loss of information and hence does not affect the R-squared value.

A large number of significant predictors are being selected from the other coutries compare to Malaysia datasets in which 40 predictors are considered as significant attributes comparing to 9 predictors for Malaysia. The selection of large predictors will have a high possible of fitting the noise to the model and makes the prediction inaccurate. The quality of the predictors of other countries are much more better because they have 75% of predictors are extremely significant(extremely small p-value) compare to only 33% for Malaysia. However, the R-squared value of Malaysia model with only selected predictors performed better than other countries which is 0.3933 compare to other countries with only 0.378. This means that predictors selected by Malaysia models able to explain 39.33% of the variability of C19ProSo02 variables. The reason that Malaysia models perform better is because other countries may be overfitting the data with selecting too many predictors. This will lead the model to learn noises which is rarely occur again. Hence, the predictability of the model decreases.

C19ProSo03:

We plot the linear regression models, we could observe that the best predictors are affBor, affDepr, affExc, PLRAC19, disc03, employstatus\_3, employstatus\_7, c19NormShould, c19NormDo, age, edu, c19ProSo01, c19ProSo02 and c19ProSo04 with 3 stars significant rating. Then, followed by affNerv, PLRAEco, employstatus\_1, employstatus\_2, fail02, MLQ, c19IsOrg and gender with 2 stars significant rating and affContent, affExh, employstatus\_9, PFS01, PFS02, fail01 as well as lifeSat with a star significant rating. The higher the number stars of a predictor, the lower the p-value, the more significant it is. Although 1 star having larger p-value compared to 2 stars and 3 stars, they are still small enough to be important predictors since they are still smaller than 0.05. Smaller p-value means the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. The R-squared value is 0.4445 which means that the independent variables of the linear regression model can explain 44.45% of the variance in the dependent variables.

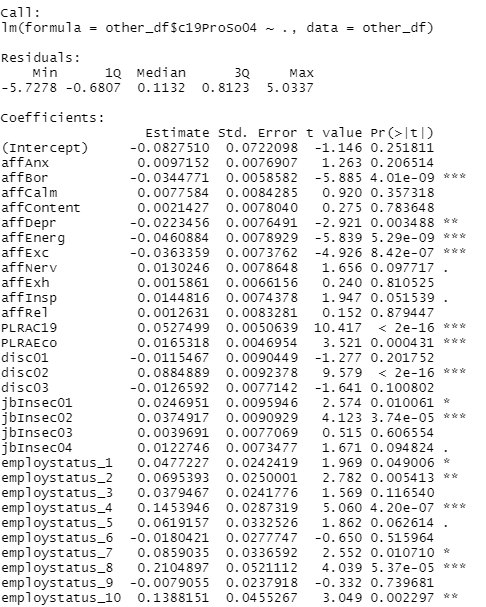
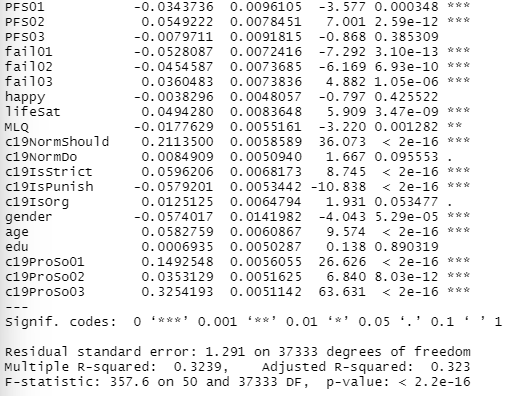


Another linear regression model is built using the best predictor chosen and compare the R-squared value with the initial linear regression model to see whether the new linear model could explain larger variability compared to the initial linear model.

The new linear regression model with only most significant attributes shows 0.444 R-squared value which is almost the same as the initial linear regression of 0.4445. The possible reason of this situation may occur is the selected attributes were highly correlated to the removed predictors. The removed attributes may not cause a substantial loss of information and hence does not affect the R-squared value.

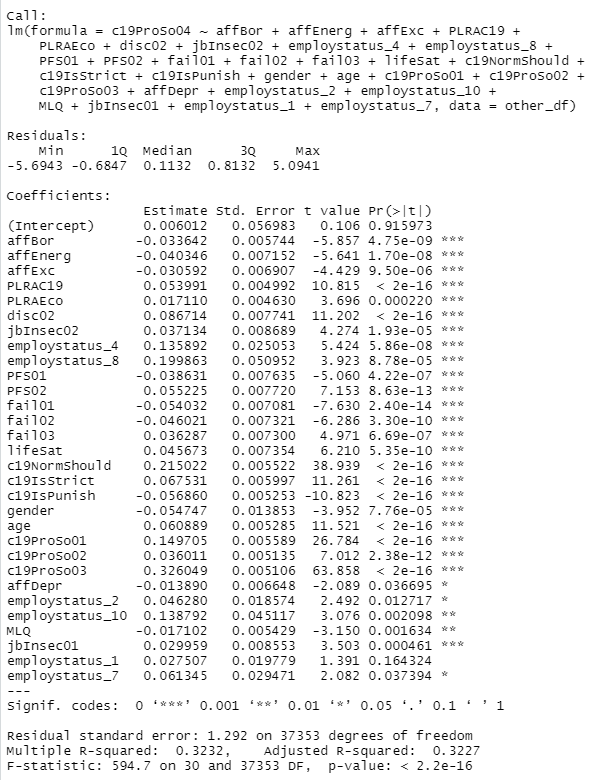
The number of predictors selected by the model other countries is 29 which is a lot more compared to Malaysia model with only 8. A large portion of predictor being chosen will result in inaccuracy prediction. This is because the model will overfitting the data given and learn noises. The noises are rarely repeated and hence will decrease the predictability. The quality of the predictors selected by other countries model are better than Malaysia model in which other country has about 51% of predictors with extremely significant attributes compare to 37.5% in Malaysia model. The higher the quality of predictors, the better the model should be. As what we expected, other countries model has a higher R-squared value which is 0.444 compared to Malaysia model which is 0.3982. This means that predictors selected by Malaysia models able to explain 44.4% of the variability of C19ProSo03 variables which is better than 39.82 in Malaysia model. One of the reason could be other country model is given a larger datasets. A larger datasets will give more information to capture the relationships between independent variables and dependent variables. Hence, models are able to select the important predictors better according to the relationship.

C19ProSo04:

**** ****

When we plot the linear regression models, the important predictors are affBor, affEnerg, affExc, PLRAC19, PLRAEco, disc02, jbInsec02, employstatus\_4, employstatus\_8, PFS01, PFS02, fail01, fail02, fail03, lifeSat, c19NormShould, c19IsStrict, c19IsPunish, gender, age, c19ProSo01, c19ProSo02 and c19ProSo03 with 3 stars of significant rating which indicates extremely important predictors. Then, affDepr, employstatus\_2, employstatus\_10, MLQ are given 2 stars of significant rating represent very important predictors. The 1 stars of significant rating includes jbInsec01, employstatus\_1 and employstatus\_7 mean that the predictors are important but less important than 2 stars and 3 stars predictors. This is because the p-value of 1-star significant rating predictors is larger than 2 stars and 3 stars. The smaller the p-value, the more important the predictor is. The small p-value indicates the probability of seeing an association between a predictor as strong as the one we have observed, or stronger, just by chance, even if there was no association at the population level. Even though the 1-star predictors have larger p-value, but the p-value are small enough to be selected as a good predictor which is smaller than 0.05. Hence all 1 to 3 stars significant rating are chosen as significant predictors. The R-squared value of this model is 0.3239 tell us that the independent variables of the linear regression model can explain 32.39% of the variance in the dependent variables.

A linear regression model is built with the predictor chosen and using the R-squared value to judge how much variability the linear model could explain.



The R-squared value of the new linear regression model with only all significant attributes is 0.3232 which is almost equal to the initial R-squared value of 0.3239. The percentage explanation of the two linear models is almost the same which is 32.3%. One of the possible reason is the attributes that we did not include does not contribute as much to the variance in the C19ProSo04 variable.

Comparing the predictors selected by Malaysia models, predictors selected by other country is 30 which a lot more comparing to 5 in Malaysia. A large quantity of predictors being selected will cause the model to overfit the datasets. This will make the models learn the noise of the datasets which will low possibility repeated in future. Hence, the predictability of the model will decrease and with the affect of noise. The quality of the predictors in other countries are a lot better compared to Malaysia model because other countries having more than 66% of the predictors are extremely important compared to Malaysia having only 20%. A better quality of predictors in a model will often perform better than worser quality. The R-squared value of other country model built with significant predictors is better than Malaysia supports my point of view. Other country model gives 0.3232 R-squared which is higher than Malaysia model with only 0.222. This shows that the predictors selected by other country model able to explain 32.32% of the variability of C19ProSo04 variables which is better than 22.2% in Malaysia model. This could be caused by a bigger datasets are given in other countries model to predict the most important attributes. A larger datasets will be able to give more information to have a better capture of the relationships between C19ProSo04 and other predictors. Hence, other country model able to select better predictors compare to Malaysia model.

**Comparing predictors of Malaysia (focus country) datasets and other country datasets:**

Similar predictors between focus country and other country datasets:

c19ProSo01: c19ProSo02, c19ProSo03 , c19ProSo04

c19ProSo02: c19ProSo01, c19ProSo03 , c19ProSo04, gender, c19IsOrg, PFS03

c19ProSo03: c19ProSo01, c19ProSo02 , c19ProSo04, c19NormShould, affDepr

c19ProSo04: c19ProSo01, c19ProSo02 , c19ProSo03, c19IsStrict

From the table above, we could observe that the Corona ProSocial Behavior series attributes are being selected as important attributes. This is because the Corona ProSocial Behavior attributes are identifying the similar characteristic on the people. If someone giving high rating on one of these predictors, that means that the person is kind and helpful. The other predictors are often given the similar responses. Hence, we could know that Corona ProSocial Behavior series attributes are highly intra-correlated with each other. They will become each other’s strongest predictors when one of them is selected as dependent variable.

Comparing the overall R-squared value, the R-squared value of focus country data will decrease by a larger portion compared to other country data after comparing the R-squared value of initial linear regression model with the R-squared value of linear regression model build with the selected significant predictors. This means that the predictors of other countries will perform better compared to focus country. This situation occur may be caused by large variation in the size of data size. The dimension of data for focus country is 552 rows and 51 columns while other countries data is 37384 rows and 51 columns after preprocessing. A linear model needs data to learn and capture the underlying pattern in the data. Other countries are having a bigger data size compared to Malaysia. The larger data size will provide other countries linear model to learn and capture the relationship between predictors. Hence, the prediction of other countries linear model will be more accurate. However, the number of data given is very limited. The model is having high opportunity to underfitting. This will result in model having high bias and make the prediction consistently wrong.

In overall, other countries data are selecting a large number of predictors as strongest attributes. The large number of predictors selected will gives the model a lot of information about the datasets and the model able to fit the data well. Even though there are some extreme data given, the large number of datasets are able to neutralize most of the bias. In contrast, the focus country data are limited. If there are some data giving extreme wrong value, the small amount of data is not able to drag the model back on track. As a result, there will be a larger bias on linear model with trained by small amount of data and makes the prediction far from the true value compared to other countries data. of Hence, the other countries linear regression model able to predict the Corona ProSocial Behavior more accurate compared to Malaysia datasets.

**Question 3a**

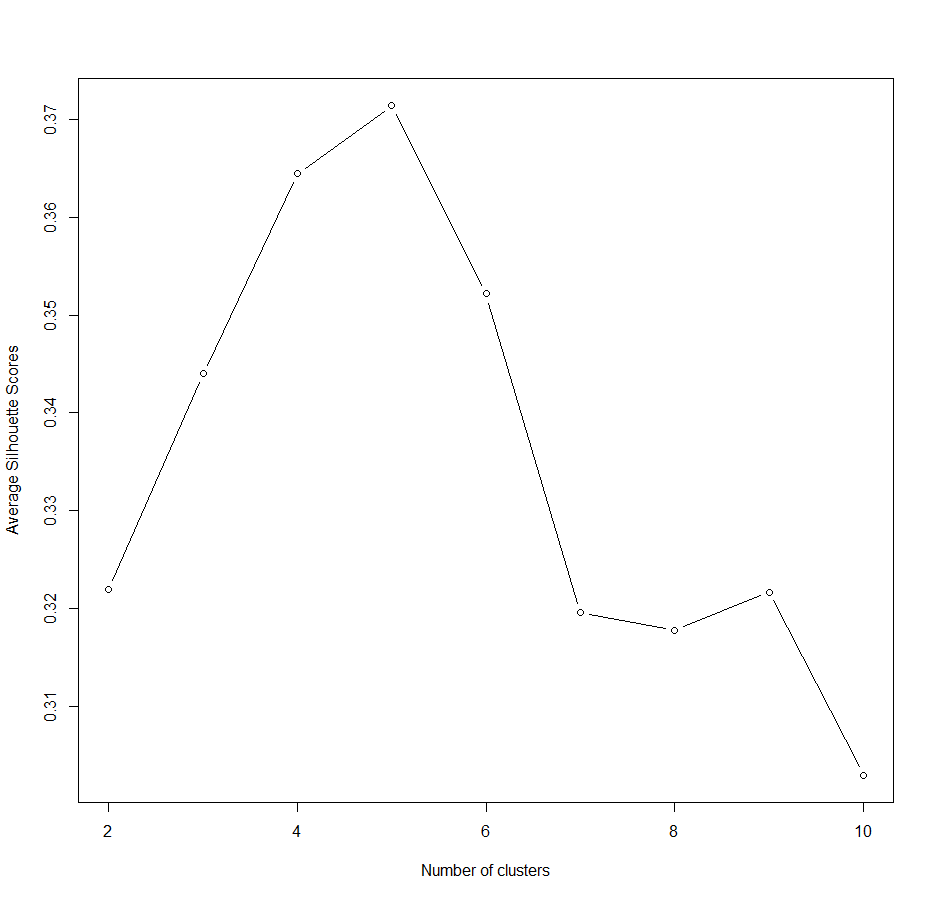
The data that I use to do the clustering is some from the link given in the instruction and some from other sources. The indicators that I choose are GDP from economic indicator, person fully vaccinated per 100 population (fullyVaccinatedper100p) and total cumulative death per 100000 population (totalCumDeathper100000pop) from social and health indicators and government effectiveness (govEffectiveness) in political indicators. In this data, there are no data provided regarding Hong Kong and Taiwan. Hence, I assume that the data of these countries are included under China. Besides, there is a country called Palestine country does not have any data regarding from all the datasets and hence should be removed. There is also a missing value for government effectiveness predictor for country Kosovo. To maintain the fairness, I am not able to find a single value to fit in the missing value and have no choice but remove the whole row. The dimension for the new datasets is 105 rows and 5 columns.

The reason GDP is chosen as a predictor is because GDP is a measure of country’s economic output which it is highly related to the income and career security of its citizen. A bad GDP means most of the people are facing difficulty in economic due to lower income or job losing. This will make people hard to engage in prosocial behavior. In this case, people will have to prioritize their economic needs rather than engage in prosocial behaviors.

The fullyVaccinatedper100p predictors is selected because vaccination will increase the immunity of people and make them brave enough to help other without any worries. People with fully vaccination will receive a higher immunity to COVID-19 virus in which will not heavily sick even if they are infected by COVID-19 virus. With the assurance of vaccination, they are able to participate in prosocial activities with rest assured.

The totalCumDeathper100000pop attribute is chosen because a high cumulative death will probably affect the determination of people engage in prosocial behavior. The high number of cumulative death shows how dangerous and fatal the virus is. After looking at the statistics of people death due to COVID-19, people will probably worry about their own safety even with those initially decide to engage in prosocial behavior will have a high possibility of changing their mind. People engaging in the prosocial behavior will highly affected.

The govEffectiveness predictor is also being selected because government’s effectiveness will build the confident of people to engage in prosocial behavior. A government with bad effectiveness will not take care of their people but only benefit themselves. They will allow the spread of virus faster and in a wider range until the situation is uncontrollable. They know that the situation will be out of control anytime and people will not be willing to put themselves in danger meaninglessly. If they participate in prosocial attributes will only put themselves in the risk and nothing will be better. Hence, it shows the importance of government effectiveness on engaging people in prosocial behavior.



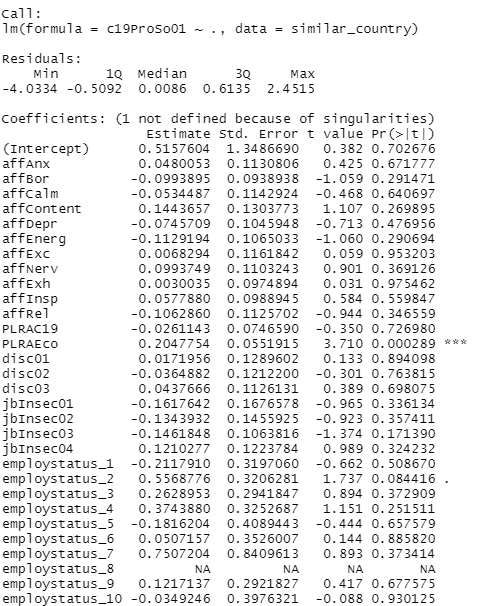
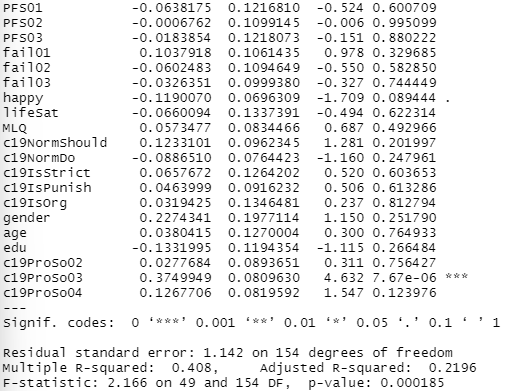
The average silhouette coefficient estimates the average distance between clusters. The silhouette plot shows the silhouette coefficient over values of k in the range of 1 to 10. The plot shows the highest average silhouette coefficient occurring when k = 5. We then choose 5 as the number of clustering. However, there are too many countries that are inside the same cluster as my focus country. To decrease the number of countries in the same cluster to avoid wrongly selected predictors, I increase the number of clusters to 13.

k-mean clustering method is being used to find the country similar to focus country. We then set the parameter with the new datasets we created cluster as 13 and repeated 25 times. The number of cluster is obtained after a few times of trial and error and select a reasonable number of countries in the same cluster. Then we observed the matrix and see which column the focus country is in and find all other countries that are also having 1 in the same column. There are 5 other countries that are in the same clusters with my focus countries which are Vietnam, Brunei, Mauritius, Qatar, United Arab Emirates . The BSS/TSS ratio of the k-mean clustering is as high as 88.3% which is a very good fit to the data. The ratio of BSS and TSS indicates the proportion of total variability that is accounted for by the clustering selected. It is used to measure the effectiveness of the cluster in which higher value means the cluster is more distinct and well-separated.

**Question 3b**

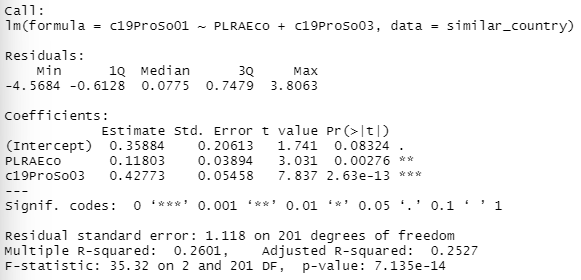
To determine how well the participant responses predict pro-social attitudes from the same cluster, we first create a new datasets which only includes all the predictors from the countries that we selected from the clustering. I then remove the column of coded\_country since the data sets now works as a group.

c19ProSp01:

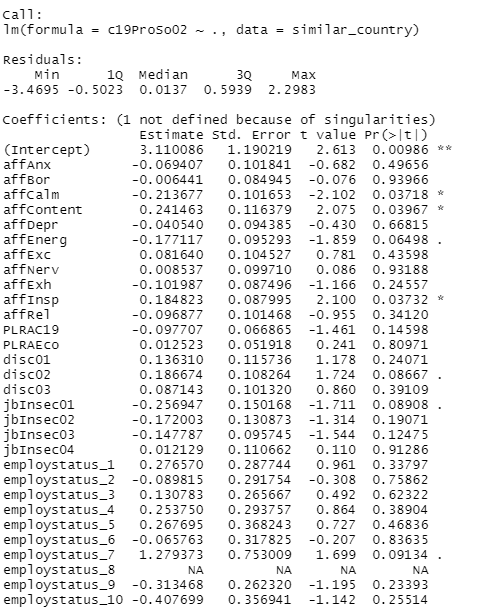
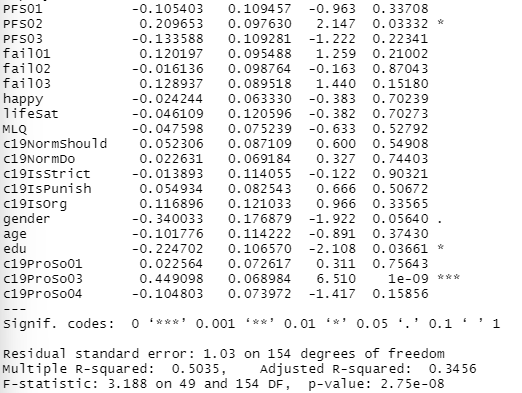
A model is now built with all attributes in the datasets with c19ProSo01 as dependent variables and other attributes as independent variables. From the model, we could select the most important attributes based on the number of stars and also p-value. From the model, we could find that there are only two predictors that are significant which are PLRAECo and c19ProSp03 which are both 3 stars. The p-value of P. The other attributes’ p-value are all larger than 0.05 which does not fulfill the requirement of being significant predictors. The R-squared value of the linear model is 0.408 which mean the attributes of this linear model are able to explain 40.8% of the variability in the dependent variables which is having intermediate explanatory.

I then plot another linear regression model which only include selected significant attributes to observe whether the new linear regression perform better.

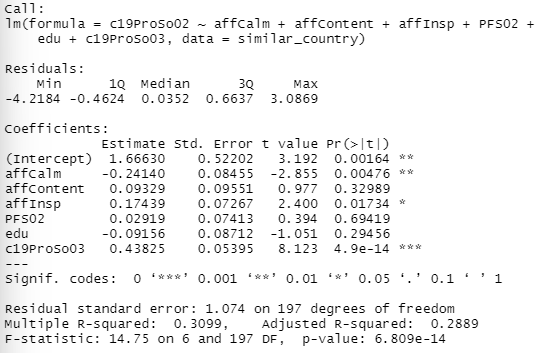


The new linear model gives only 0.2601 R-squared value which perform worse than the initial linear model which have 0.408. The predictors can only explain 26.01% of the variables in dependent variable. The reason could be the linear model had included too few data which cause underfitting. This means that the model does not learn or include enough information to capture underlying patterns in the data. Hence, high bias occurs in the model and will consistently predict wrong values. Besides, there would be another problem

c19ProSp02:

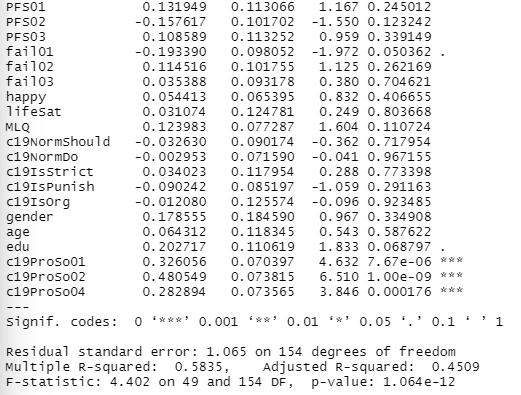
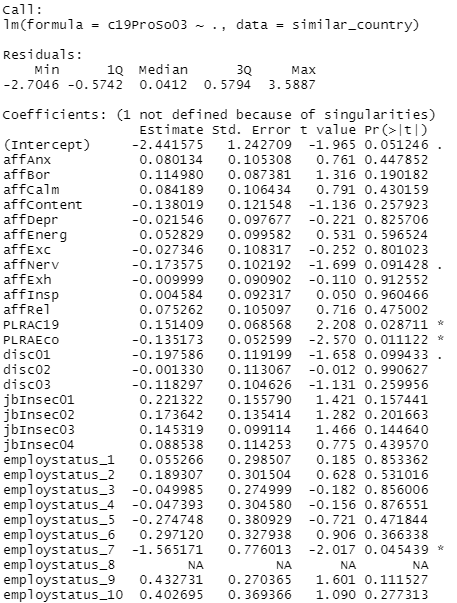
 

A linear model is built with c19ProSo02 as dependent variables while other attributes becoming independent variables to find the strongest predictors. From the table, we could easily select the strongest predictors which are the predictors with stars. The higher the number of stars, the more significant the predictor is. Besides, we can also verify the importance of predictors by their p-values. As long as the p-value smaller than 0.05 they had step over the baseline of important predictors. The smaller the p-value, the more significant the predictor is. From the graph, there is only 1 extremely important predictor which is c19ProSp02 because it has 3 stars on significant rating support by extremely small p-value. Then followed by 1 stars rating attributes which include affCalm, affContent, affInsp, PFS02 and edu supported with slightly smaller p-value from the 0.05 baseline. Although the p-values are not very small but had small enough to become important predictors. There are no 2 stars attributes exist in this linear model. The R-squared value is 0.5035 which indicates the independent variables able to explain about 50% of the variability of c19ProSo02.

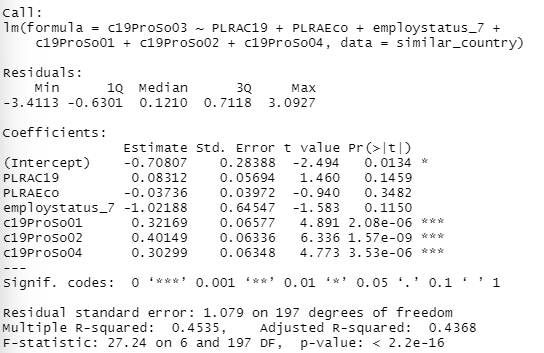
Another linear model is built using selected attributes to test their performance. 

The R-squared value of the new linear regression model is 0.3099 which means the independent variables can only explain 30.99 of the dependent variables. The performance is quite a lot lower than the initial linear model. This situation may cause by multicollinearity may happen due to some predictors are highly correlated with each other, predicting the unique effect of each attributes will be difficult and result in inaccurate regression coefficient and causes R-squared value decrease. Another reason could be occurrence of large number of outliers. They will highly affect the linear line and hence resulting in a worser R-squared.

c19ProSp03:

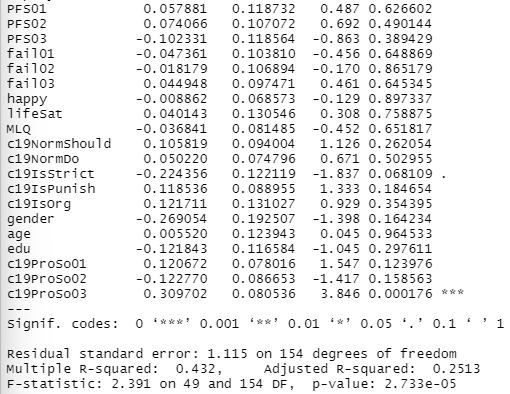
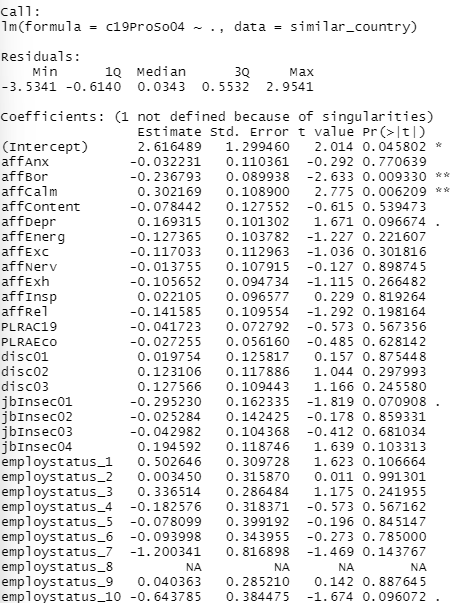


A linear model is built to choose the best predictors. The extremely significant predictors that exist in this model when c19ProSo03 is dependent variable are c19ProSo01, c19ProSo02 and c19ProSo04. They are having the highest rate which is 3 stars which indicate the most important predictors. The extremely small p-value also support the argument of they are strongest predictors. Then, there are PLRAC19, PLRAECo and employstatus\_7 as important predictor with 1-star significant rating and bigger but smaller than 0.05 p-value to indicate the predictors are relatively weaker. The R-squared value is 0.5835 says the independent variables able to explain 58.35% of dependent variables which is c19ProSo03.

A linear model is built using the selected significant predictors to determine their performance. 

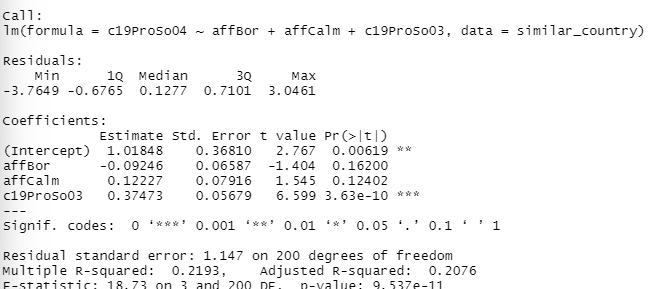
The R-squared value of the new linear regression model is 0.4535 which is a bit worser than the R-squared value of initial linear regression which is 0.5835. The independent variable of new linear regression model able to describe 45.35% of the variability in the c19ProSo03 variables. The reason that this situation occurs because multicollinearity may happen due to some predictors are highly correlated with each other, predicting the unique effect of each attributes will be difficult and result in inaccurate regression coefficient and causes R-squared value decrease. Another reason could be occurrence of large number of outliers. They will highly affect the linear line and hence resulting in a worser R-squared.

c19ProSp04:



A linear regression is created to search for the strongest predictors when c19ProSp04 act as a dependent variable. From the table, we could observe that there is a 3-star significant rating for the predictor c19ProSp03. Its significant could also be supported by its extremely small p-value. The 2-stars significant rating includes affBor and addCalm which have very small p-value to support that they are significant predictors. The R-squared value is 0.432 which means the independent variables of this linear model able to explain 43.2% of the variability in the dependent variable c19ProSp04.

To see the performance of selected predictors, we need to create a new linear model.



The new linear regression model is giving a value of 0.2193 which is almost half worser than the initial R-squared value. The independent variables of the new linear model can only explain 21.93% of the variability of dependent variable. This might be caused by omitting variable bias. When we remove some variables, there is a chance that we will omitting some important information that is related to the outcome and hence, may cause biased and accuracy reduced.

**Comparing clustered countries to focus country:**

c19ProSp01: c19ProSp03

c19ProSp02: c19ProSp03

c19ProSp03: c19ProSp01, c19ProSp02, c19ProSp04

c19ProSp04: c19ProSp03

From the table above we could observe that the datasets selected by clusteters works quite differently with the focus country even though we are using clustering to choose the countries that works similar the the focus countries. However, when we compare the two linear regression, they are quite different in their performance. Malaysia’s linear model with selected strongest predictors, the R-squared value drop a bit compare to its initial linear model but similar countries give a significant decrease in the R-squared value. This had actually against the expectation of countries under the same cluster will have a similar performance on the focus country. One of the possible reason may be insufficient predictors included into the new dataset created. As we know that clustering is classifying the data according to their similar characteristics. Since the performance is so difference, this means that the characteristics of each data are not clearly separated and hence the cluster may include some country that should not be included.

From the similar attributes, we could observe that Corona ProSocial Behavior is still showing the strongest correlation with each other but in this case, not all of them are included. From the obsevation, the predictors c19ProSp03 is always the strongest predictors because It always give a very low p-value when being independent variables. The low p-value indicates the importance of predictors. The smaller the p-value, the stronger the attributes.

The overall quantity of strongest predictors in focus country is more than similar countries. The larger amount of strongest predictor affects the performance of linear model a lot because the linear model will gain more information and data about the data sets and able to capture the underlying patterns in the data accurately. Even though there are some extreme data given, the large number of datasets are able to neutralize most of the bias. In contrast, the focus country data are limited. If there are some data giving extreme wrong value, the small amount of data is not able to drag the model back on track. As a result, there will be a larger bias on linear model with trained by small amount of data and makes the prediction far from the true value compared to other countries data. of Hence, the focus countries linear regression model able to predict the Corona ProSocial Behavior more accurate compared to similar countries datasets.

The quality of the strongest predictors of Malaysia datasets are also better than similar countries. The p-value of Malaysia datasets are smaller compared to similar countries. The reason that this could happen is because the difference in size of datasets given. Malaysia has 552 rows while similar countries have only 204 datasets. The larger amount of data will enable the linear model to learn and capture the underlying pattern in the data. The larger amount of data will enable the linear model to neutralize the error in the data better than smaller data. Other data sets able to pull the model back to the right track if some of the data are in extreme.

**Comparing the difference between clustered countries to focus country and 2(c) :**

From the previous comparison on clustered countries and similar countries from clustering the effectiveness and performance in predicting the behavior of the focus country, Malaysia, the differences is quite large in which the difference between clustered countries with focus country is very large compare to other countries. When the strongest predictors are selected and linear models are built, the differences become more obvious. All country have a minor decrease range which is 0.005 only. However, the drop range of similar countries can be up to 0.3 which is a huge number. The large drop range indicates that the linear model does not perform any good. In overall, others countries also having the higher r-square complare to similar countries which means the better explainary on the variability on the dependent variables.

Next, we would compare about the datasets because datasets are important in training the linear model. A bigger data size will enable a linear model to learn and capture the underlying patterns in the data. As a result it will become more accurate and fit the data sets. The data sets of all others data are 37384 rows while the data sets for similar country is only 204. In this case the all other data is significantly having larger datasets compare to similar country. This means that if there are some extreme or wrong data, the impact on the model does not as high as similar country because the other datasets will neutralize the error bit by bit. If there is an error data in the smaller data sets, there will be a large impact on it.

After doing much comparing between the two groups, I found that the all other groups are the datasets that has a better prediction on the prosocial attitutes in focus country. Regardless the truth performance and accuracy, all other countries have a more similar attributes when comparing the significant attributes with the focus country and hence have a relatively better match with it. Besides, the R-squared value of all other countries also perform closer to focus country compared to the similar countries.

Theorethically, the similar countries that cluster with the focus countries should have a better performance and predict the correct prosocial behaviors and R-squared value compare to all other countries but however, the all others countries have a better match to the important predictors for prosocial behaviors in focus country, Malaysia.

**Appendix**

World Bank Group (2023). World Bank National accounts data and OECD National Accounts data files: GDP.

<https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2021&start=2021&view=bar>

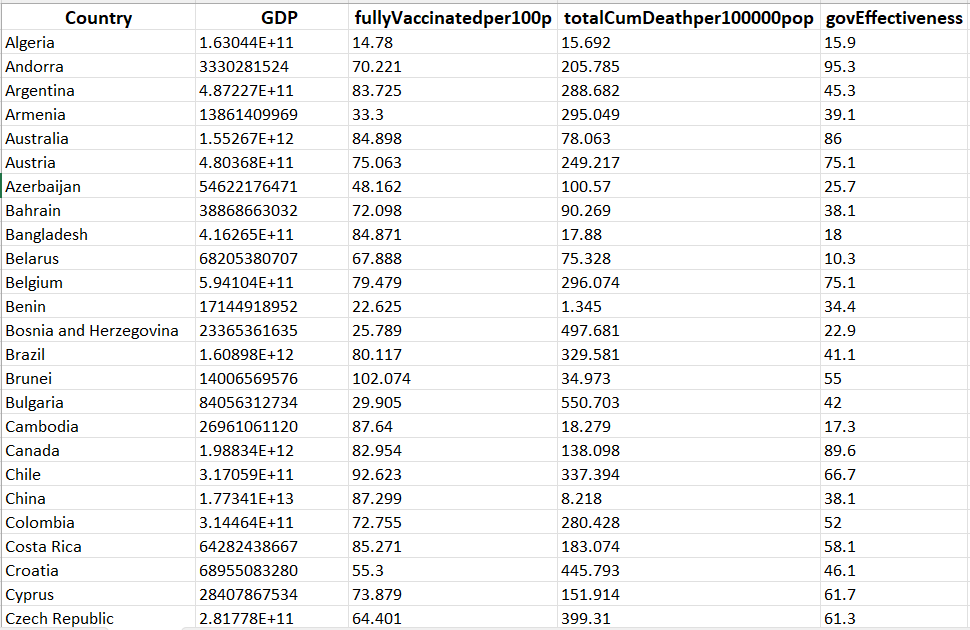
Global Health Security Index: Reports and Data

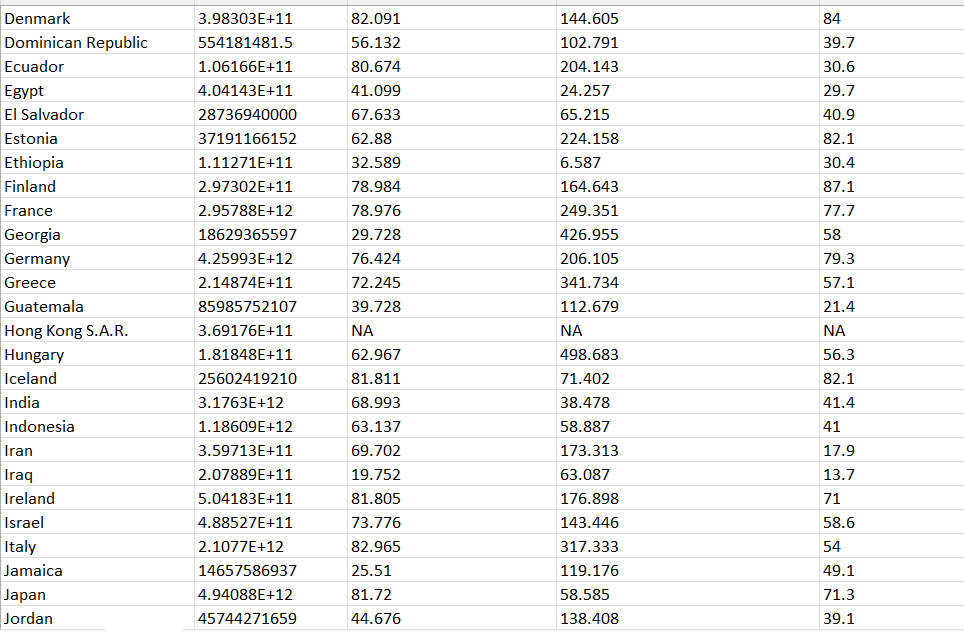
<https://www.ghsindex.org/report-model/>

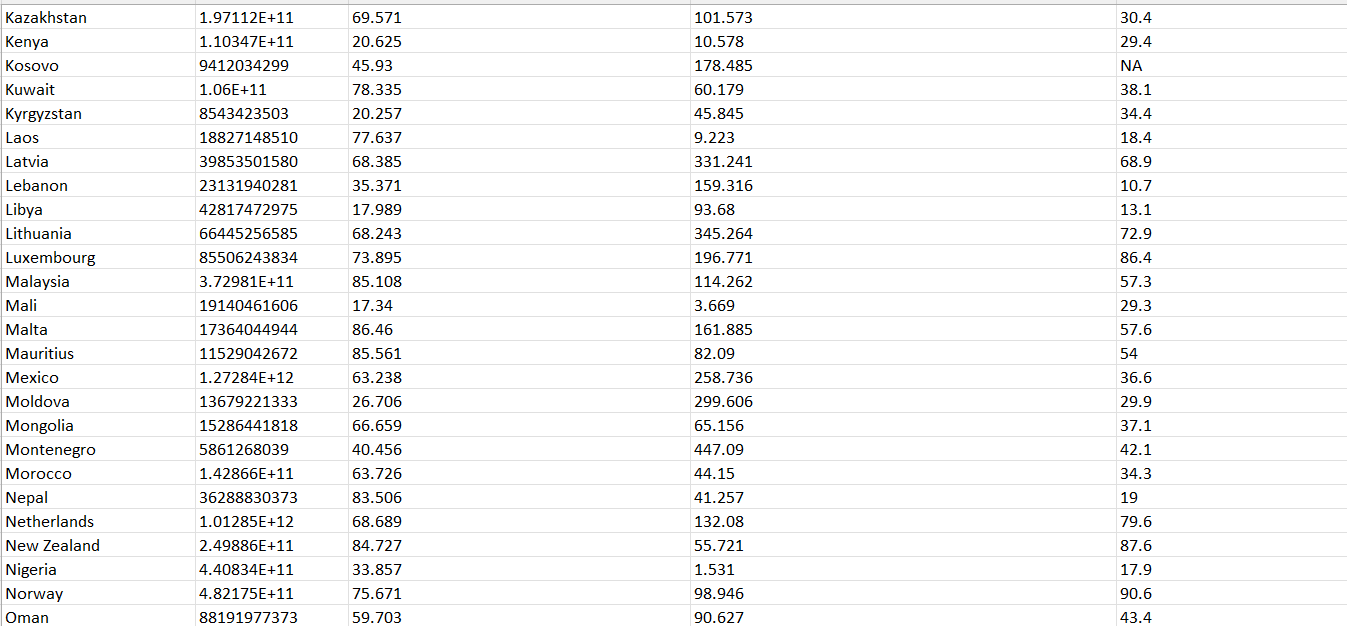
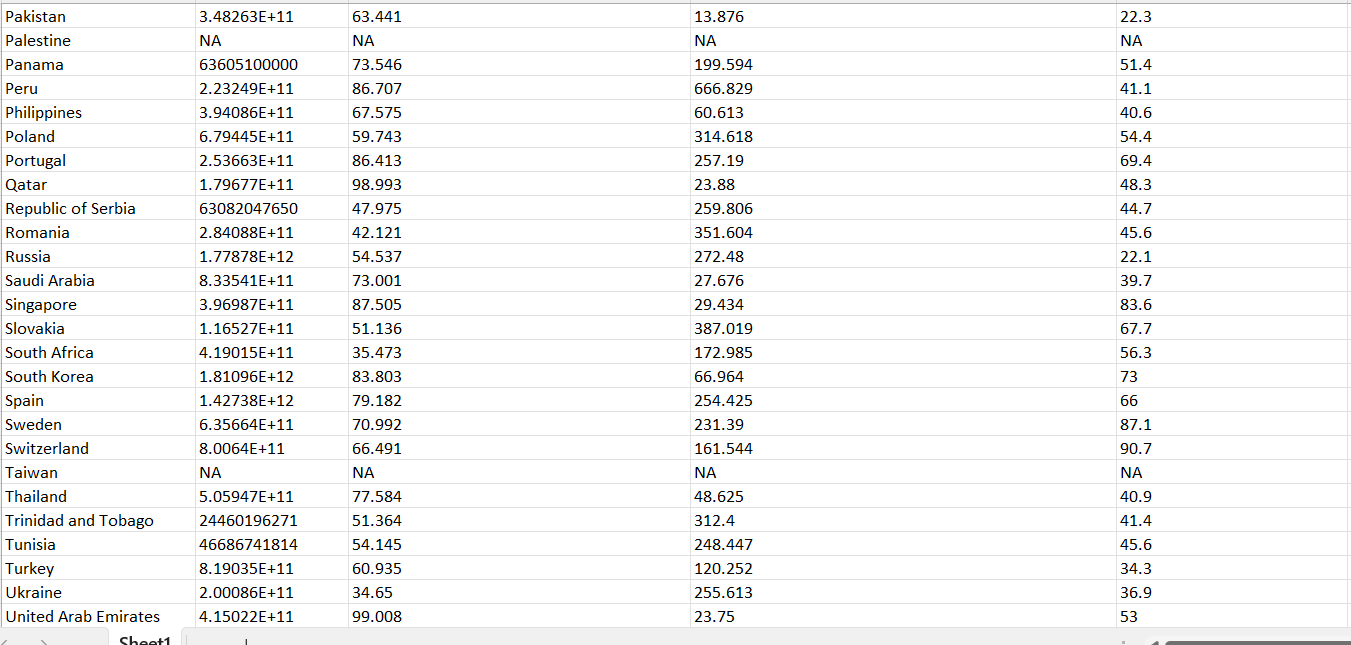
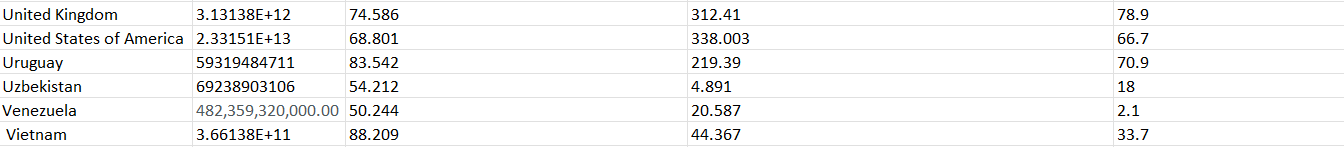
World Health Organization

<https://www.who.int/>

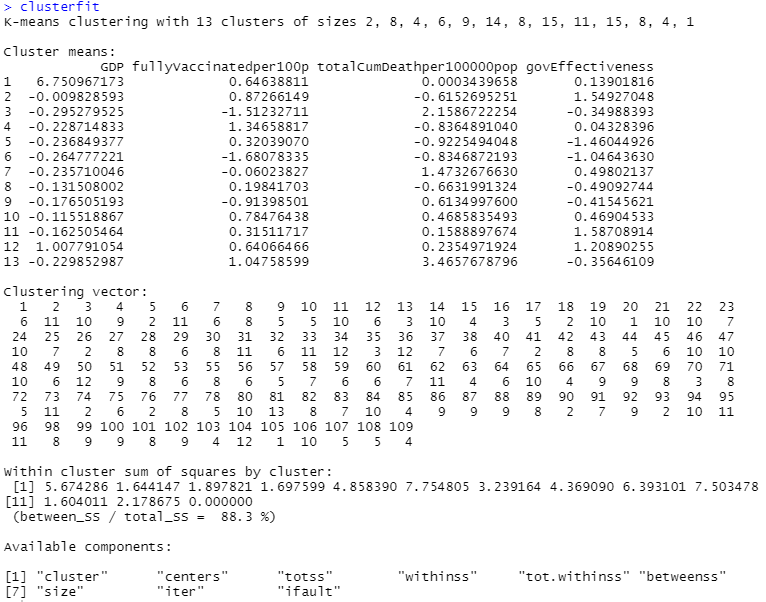
**table for question 3a**



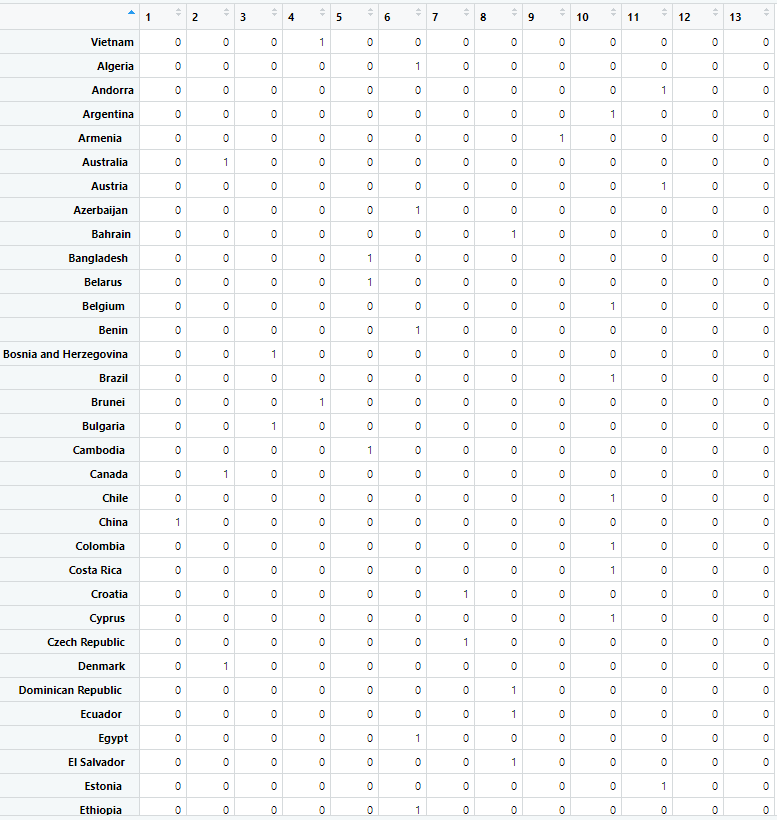
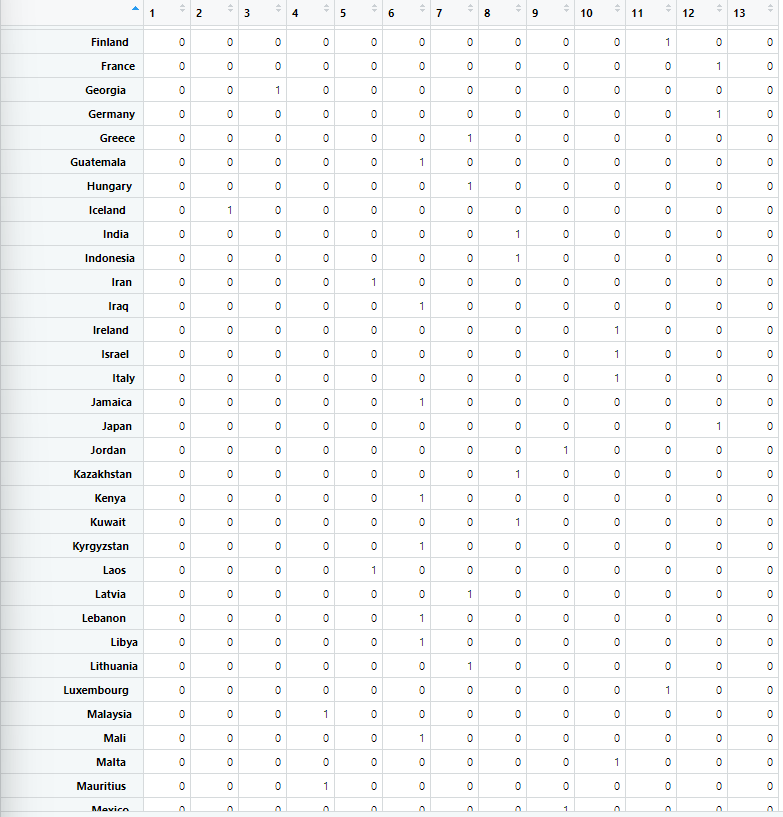


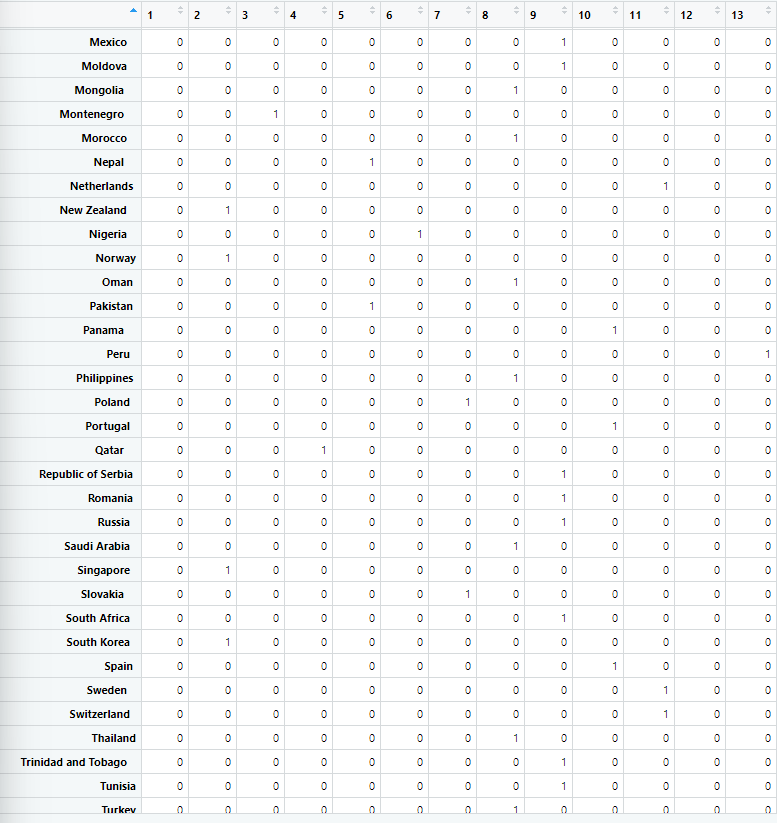
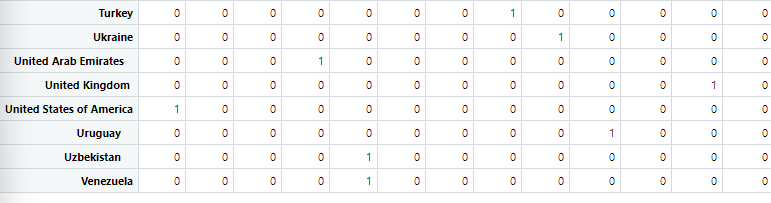
  

K-mean clustering:



k-mean clustering table:

Code:

rm(list = ls())

set.seed(31860346)

setwd("C:/Users/Predator/Desktop/monash/y3s1/fit3152/assignment 1")

cvbase = read.csv("PsyCoronaBaselineExtract.csv")

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

library(ggplot2)

library(dplyr)

#Question 1

#===============================================================================

# can use boxplot to show how distribution of numerical attributes but not only way

#1a)

# count the dimension of the the datasets

dim(cvbase)

# find the data type of each attributes

str(cvbase)

# numerical data set

num\_attr = cvbase[, -c(50)]

# show the distribution of numerical attributes

boxplot(num\_attr, las = 2, main = "boxplot of numerical attributes")

# summary of number of numerical attributes

summary(num\_attr)

# categorical data set

cat\_attr = cvbase[, c(50)]

unique(cat\_attr)

# show the distribution of the coded country

plot(as.factor(cat\_attr), las = 2, main = "histogram of categorical attributes")

# summary of number of occurrence of the coded country

occurrence <- as.factor(cat\_attr)

summary(occurrence)

count\_table = table(cvbase$coded\_country)

sort(count\_table, decreasing = TRUE)

count\_table = as.data.frame(count\_table)

print(count\_table)

# count the total number of NA in the data set

sum(is.na(cvbase))

# 1b)

# replace the NA in employment status to 0

employ = cvbase[,c(21:30)]

employ[is.na(employ)] <- 0

cvbase[,c(21:30)] <- employ

# replace the NA in job insecurity to 0

job\_insec = cvbase[,c(17:20)]

job\_insec[is.na(job\_insec)] <- 0

cvbase[,c(17:20)] <- job\_insec

# checking the proportion of NA in each column

prop\_na <- sort(colMeans(is.na(cvbase)))

print(prop\_na)

# removing the columns of NA which is more than 20%

cvbase$trustGovCtry <- NULL

cvbase$trustGovState <- NULL

# remove all the rows with NA

cvbase <- cvbase[complete.cases(cvbase),]

# dimension after pre-processing

dim(cvbase)

#Question 2

#===============================================================================

# 2a)

# data of Malaysia

Mas\_df <- cvbase[(cvbase$coded\_country == "Malaysia"),]

# data other than Malaysia

other\_df <- cvbase[(cvbase$coded\_country != "Malaysia"),]

# remove coded\_country from Mas\_df as all are Malaysia

Mas\_df$coded\_country <- NULL

#removing coded-country from other\_df as country is not important

other\_df$coded\_country <- NULL

# dimension of malaysia datasets and other country datasets

dim(Mas\_df)

dim(other\_df)

# showing all distribution of Malaysia data set in box-plot

boxplot(Mas\_df, las = 2, main = "Malaysia data set boxplot")

# showing all distribution of other country data set in box-plot

boxplot(other\_df, las = 2, main = "other country data set boxplot")

summary(Mas\_df)

summary(other\_df)

# 2b)

# show the dimension of Mas\_df

dim(Mas\_df)

attach(Mas\_df)

# exclude all the C19ProSo

MY\_df.nonProSocial <- Mas\_df[, -c(48:51)]

# show the dimension of MY\_df.nonProSocial

dim(MY\_df.nonProSocial)

# predict c19ProSo01

# correlation of c19ProSo01 and all other predictors in Malaysia datasets

round(cor(Mas\_df, c19ProSo01), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo01 in Malaysia data set in the model summary

mas.c19ProSo01.fit <- lm(c19ProSo01 ~ ., data = Mas\_df)

summary(mas.c19ProSo01.fit)

# plot the linear regression model with the best predictor chosen

mas.c19ProSo01.betterfit <- lm(c19ProSo01 ~ c19ProSo02 + c19ProSo03 +

c19NormShould + c19ProSo04, data = Mas\_df)

#check the performance of new build linear model

summary(mas.c19ProSo01.betterfit)

# predict c19ProSo02

# correlation of c19ProSo02 and all other predictors in Malaysia datasets

round(cor(Mas\_df, c19ProSo02), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo02 in Malaysia data set in the model summary

mas.c19ProSo02.fit <- lm(c19ProSo02 ~ ., data = Mas\_df)

summary(mas.c19ProSo02.fit)

# plot the linear regression model with the best predictor chosen

mas.c19ProSo02.betterfit <- lm(c19ProSo02 ~ gender + c19ProSo01 + c19ProSo03 +

affAnx + affNerv + jbInsec01 + PFS03 + c19IsOrg

+ c19ProSo04, data = Mas\_df)

#check the performance of new build linear model

summary(mas.c19ProSo02.betterfit)

# predict c19ProSo03

# correlation of c19ProSo03 and all other predictors in Malaysia datasets

round(cor(Mas\_df, c19ProSo03), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo03 in Malaysia data set in the model summary

mas.c19ProSo03.fit <- lm(c19ProSo03 ~ ., data = Mas\_df)

summary(mas.c19ProSo03.fit)

# plot the linear regression model with the best predictor chosen

mas.c19ProSo03.betterfit <- lm(c19ProSo03 ~ c19ProSo01 + c19ProSo02 + c19ProSo04

+ affDepr + employstatus\_5 + PFS01 + c19NormShould

+ c19IsStrict, data = Mas\_df)

#check the performance of new build linear model

summary(mas.c19ProSo03.betterfit)

# predict c19ProSo04

# correlation of c19ProSo04 and all other predictors in Malaysia datasets

round(cor(Mas\_df, c19ProSo04), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo04 in Malaysia data set in the model summary

mas.c19ProSo04.fit <- lm(c19ProSo04 ~ ., data = Mas\_df)

summary(mas.c19ProSo04.fit)

# plot the linear regression model with the best predictor chosen

mas.c19ProSo04.betterfit <-lm(c19ProSo04 ~ c19ProSo03 + c19NormDo + c19IsStrict + c19ProSo01 + c19ProSo02, data = Mas\_df)

#check the performance of new build linear model

summary(mas.c19ProSo04.betterfit)

# 2c)

attach(other\_df)

# predict c19ProSo01

# correlation of c19ProSo01 and all other predictors in other country datasets

round(cor(other\_df, c19ProSo01), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo01 in other countries dataset in the model summary

other.c19ProSo01.fit <- lm(c19ProSo01 ~ ., data = other\_df)

summary(other.c19ProSo01.fit)

# plot the linear regression model with the best predictor chosen

other.c19ProSo01.betterfit <- lm(c19ProSo01 ~

affAnx + PLRAC19 + PLRAEco + disc02 +

jbInsec01 + employstatus\_9 + employstatus\_10 +

PFS01 + fail03 + MLQ + c19NormDo + c19IsOrg +

gender + age + c19ProSo02 + c19ProSo03 + c19ProSo04

+ affCalm + affEnerg + affInsp + jbInsec04 +

employstatus\_4 + employstatus\_7 + affDepr + affExh

+ employstatus\_5 + employstatus\_6 + PFS03 + happy +

edu, data = other\_df)

#check the performance of new build linear model

summary(other.c19ProSo01.betterfit)

# predict c19ProSo02

# correlation of c19ProSo02 and all other predictors in other country datasets

round(cor(other\_df, other\_df$c19ProSo02), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo01 in other countries dataset in the model summary

other.c19ProSo02.fit <- lm(other\_df$c19ProSo02 ~ ., data = other\_df)

summary(other.c19ProSo02.fit)

# plot the linear regression model with the best predictor chosen

other.c19ProSo02.betterfit <- lm(c19ProSo02 ~ affAnx + affBor + affCalm + affEnerg +

affExc + affInsp + PLRAC19 + PLRAEco + disc02 +

disc03 + jbInsec02 + jbInsec04 + employstatus\_1 +

employstatus\_2 + employstatus\_4 + employstatus\_5 +

employstatus\_8 + PFS01 + fail01 + fail02 + MLQ +

c19NormShould + c19IsPunish + gender + age + edu +

c19ProSo01 + c19ProSo03 + c19ProSo04 + PFS02 +

PFS03 + lifeSat + c19IsOrg + affContent + affDepr +

affExh + affRel + disc01 + employstatus\_6 + fail03

, data = other\_df)

#check the performance of new build linear model

summary(other.c19ProSo02.betterfit)

# predict c19ProSo03

# correlation of c19ProSo03 and all other predictors in other country datasets

round(cor(other\_df, other\_df$c19ProSo03), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo01 in other countries dataset in the model summary

other.c19ProSo03.fit <- lm(other\_df$c19ProSo03 ~ ., data = other\_df)

summary(other.c19ProSo03.fit)

# plot the linear regression model with the best predictor chosen

other.c19ProSo03.betterfit <- lm(c19ProSo03 ~ affBor + affDepr + affExc + PLRAC19 +

disc03 + employstatus\_3 + employstatus\_7 +

c19NormShould + c19NormDo + age + edu + c19ProSo01 +

c19ProSo02 + c19ProSo04 + affNerv + PLRAEco +

employstatus\_1 + employstatus\_2 + fail02 + MLQ +

c19IsOrg + gender + affContent + affExh +

employstatus\_9 + PFS01 + PFS02 + fail01 + lifeSat

, data = other\_df)

#check the performance of new build linear model

summary(other.c19ProSo03.betterfit)

# predict c19ProSo04

# correlation of c19ProSo04 and all other predictors in other country datasets

round(cor(other\_df, other\_df$c19ProSo04), digits = 2)

# create a linear regression model to check the significant predictors to

# predict c19ProSo01 in other countries dataset in the model summary

other.c19ProSo04.fit <- lm(other\_df$c19ProSo04 ~ ., data = other\_df)

summary(other.c19ProSo04.fit)

# plot the linear regression model with the best predictor chosen

other.c19ProSo04.betterfit <- lm(c19ProSo04 ~ affBor + affEnerg + affExc + PLRAC19

+ PLRAEco + disc02 + jbInsec02 + employstatus\_4 +

employstatus\_8 + PFS01 + PFS02 + fail01 + fail02

+ fail03 + lifeSat + c19NormShould + c19IsStrict +

c19IsPunish + gender + age + c19ProSo01 +

c19ProSo02 + c19ProSo03 + affDepr + employstatus\_2

+ employstatus\_10 + MLQ + jbInsec01 + employstatus\_1

+ employstatus\_7

, data = other\_df)

#check the performance of new build linear model

summary(other.c19ProSo04.betterfit)

#Question 3

#===============================================================================

# 3a)

set.seed(31860346)

# prepare data for clustering

# find all country in the given datasets

unique(cat\_attr)

cluster\_df = read.csv("3a\_clustering.data.csv", stringsAsFactors = FALSE)

# remove rows with missing values

cluster\_df <- cluster\_df[complete.cases(cluster\_df),]

# convert everthing to whole number

cluster\_df[,2:5] = scale(cluster\_df[,2:5])

# silhouette from lecture 5

library(cluster)

sil\_score <- function(k){

km <- kmeans(cluster\_df[,2:5], centers = k, nstart = 25)

ss <- silhouette(km$cluster, dist(cluster\_df[,2:5]))

mean(ss[,3])

}

k <- 2:10

avg\_sil <- sapply(k, sil\_score)

plot(k, type='b', avg\_sil, xlab = 'Number of clusters',

ylab = 'Average Silhouette Scores')

#using k mean clustering

clusterfit <- kmeans(cluster\_df[,2:5], 13, nstart = 25)

clustertable <- table(actual = cluster\_df$Country, fitted = clusterfit$cluster)

clustertable = as.data.frame.matrix(clustertable)

clusterfit

clustertable

#checking the number of countries that are in the same cluster with the focus country

Mas\_table <- table(clustertable[,4])

Mas\_table

# 3b)

# datasets with only countries in the same cluster as focus country

similar\_country <- cvbase [is.element (cvbase$coded\_country,c('Vietnam', 'Brunei', 'Mauritius', 'Qatar', 'United Arab Emirates')),]

similar\_country <- subset(similar\_country,select = -c(coded\_country))

attach(similar\_country)

# c19ProSo01

# build a linear model to select important prictor for c19ProSo01

cluster.similar\_country.fit1 <- lm(c19ProSo01~., data = similar\_country)

summary(cluster.similar\_country.fit1)

# observe the performance of selected predictors with building a linear models

cluster.similar\_country.betterfit1 <- lm(c19ProSo01~ PLRAEco + c19ProSo03 ,

data = similar\_country)

summary(cluster.similar\_country.betterfit1)

# c19ProSo02

# build a linear model to select important prictor for c19ProSo02

cluster.similar\_country.fit2 <- lm(c19ProSo02~., data = similar\_country)

summary(cluster.similar\_country.fit2)

# observe the performance of selected predictors with building a linear models

cluster.similar\_country.betterfit2 <- lm(c19ProSo02~ affCalm + affContent + affInsp

+ PFS02 + edu + c19ProSo03,

data = similar\_country)

summary(cluster.similar\_country.betterfit2)

#c19ProSo03

# build a linear model to select important prictor for c19ProSo03

cluster.similar\_country.fit3 <- lm(c19ProSo03~ ., data = similar\_country)

summary(cluster.similar\_country.fit3)

# observe the performance of selected predictors with building a linear models

cluster.similar\_country.betterfit3 <- lm(c19ProSo03~ PLRAC19 + PLRAEco + employstatus\_7

+ c19ProSo01 + c19ProSo02 + c19ProSo04

, data = similar\_country)

summary(cluster.similar\_country.betterfit3)

# c19ProSo04

# build a linear model to select important prictor for c19ProSo04

cluster.similar\_country.fit4 <- lm(c19ProSo04~., data = similar\_country)

summary(cluster.similar\_country.fit4)

# observe the performance of selected predictors with building a linear models

cluster.similar\_country.betterfit4 <- lm(c19ProSo04~ affBor + affCalm + c19ProSo03,

data = similar\_country)

summary(cluster.similar\_country.betterfit4)