Monash University

Faculty of Information Technology

FIT3152 - Data Analytics

Assignment 1, Semester 1, 2024

Name: Rhyme Bulbul

Student ID: 31865224

Project: Analysis of country-level predictors of pro-social behaviours to reduce the spread of COVID-19 during the early stages of the pandemic

AI statement: Generative AI was not used in this assignment

Task 1: Descriptive analysis and pre-processing

1(a) A condensed version of the data gathered for the PsyCorona baseline study, a psychological survey investigating pro-social behaviours in several nations during the COVID-19 epidemic, is contained in the file `PsyCoronaBaselineExtract.csv`, by Van Lissa et al. (2002).

We start by taking a unique sample of the dataset based on my student ID, and attaching the data to the R search path for ease of variable use

```r
rm(list = ls())
set.seed(31865224)
cvbase = read.csv("PsyCoronaBaselineExtract.csv")
cvbase <- cvbase[sample(nrow(cvbase), 40000), ]
```

We will require the help of a few libraries, so let's import those

```r
library(ggplot2)
library(dplyr)
library(tidyr)
```

It is helpful to learn about a dataset's features and properties before working on it

```r
dim(cvbase)
as.data.frame(sapply(cvbase, class))
summary(cvbase, na.rm = TRUE)
```

In this case, we will run the dimensions method, to find out the data frame has 40,000 rows, given we have sampled it to be so when reading it in, and 52 columns.

The only text attributes in the dataframe appear to be `coded_country` and the `Rank Order of Life` columns, while the rest are integer data

According to the codebook extract, every column aside from `employstatus`, `gender`, `age`, `edu`, and `coded_country` contains ordinal data in the form of numbers that represent degrees of agreement for things like age group, education level, and level of agreement. Different `gender`, `age`, and `education` categories are coded by the integer values in their respective columns. Each record may only have a maximum of one employstatus column with a value of 1, indicating the employee's employment status.

We may infer that the numerical attributes have different ranges from the `summary()` output. Survey questions evaluating a two-sided degree of agreement vary from a negative number to its modulus, while those measuring a one-sided degree of agreement go from 1 to a larger positive number, such 4 or 5.

We are able to use the following, to understand `coded_country` better

```r
sort(unique(cvbase$coded_country))
table(cvbase$coded_country)
max(table(cvbase$coded_country))
which(table(cvbase$coded_country) == max(table(cvbase$coded_country)))
```

```r
min(table(cvbase$coded_country))
which(table(cvbase$coded_country) == min(table(cvbase$coded_country)))
```

Based on the outputs, there seems to be 110 unique countries inclusive of NA values, with each country having a widely different number of entries, the Croatia being the highest at 6987, and host of other countries having much less.

Missing values are present in all columns, however this is the norm as surveys of this nature do not require participants to answer all questions. The `employstatus` columns appear to be the biggest culprit, given each participant will only pick one out of the 10 categories. In this dataframe, `employstatus_3` seems to have the least missing values, while `employstatus_8` has the highest number. Potentially, this could imply that the majority of participants are working 40 hours or more, while a small number of people who are disabled may be out of work.

Another point to note is the mean age group in this dataframe amounts to 2.905, which indicates that the majority of participants are likely to be aged between 35-44 years. This could indicate working-class adults with concise lifestyle, and are studied accordingly to research the outcomes covid had on people.

1(b)

This dataset is relatively tidy and without too many missing values, hence preprocessing should not be required. However, the missing values in the `employstatus` column should be replaced with 0, as it would be easier to transform and process the data in binary format, which might be required for linear regression involving these attributes down the track.

```r
cvbase[is.na(cvbase)] <- 0
```

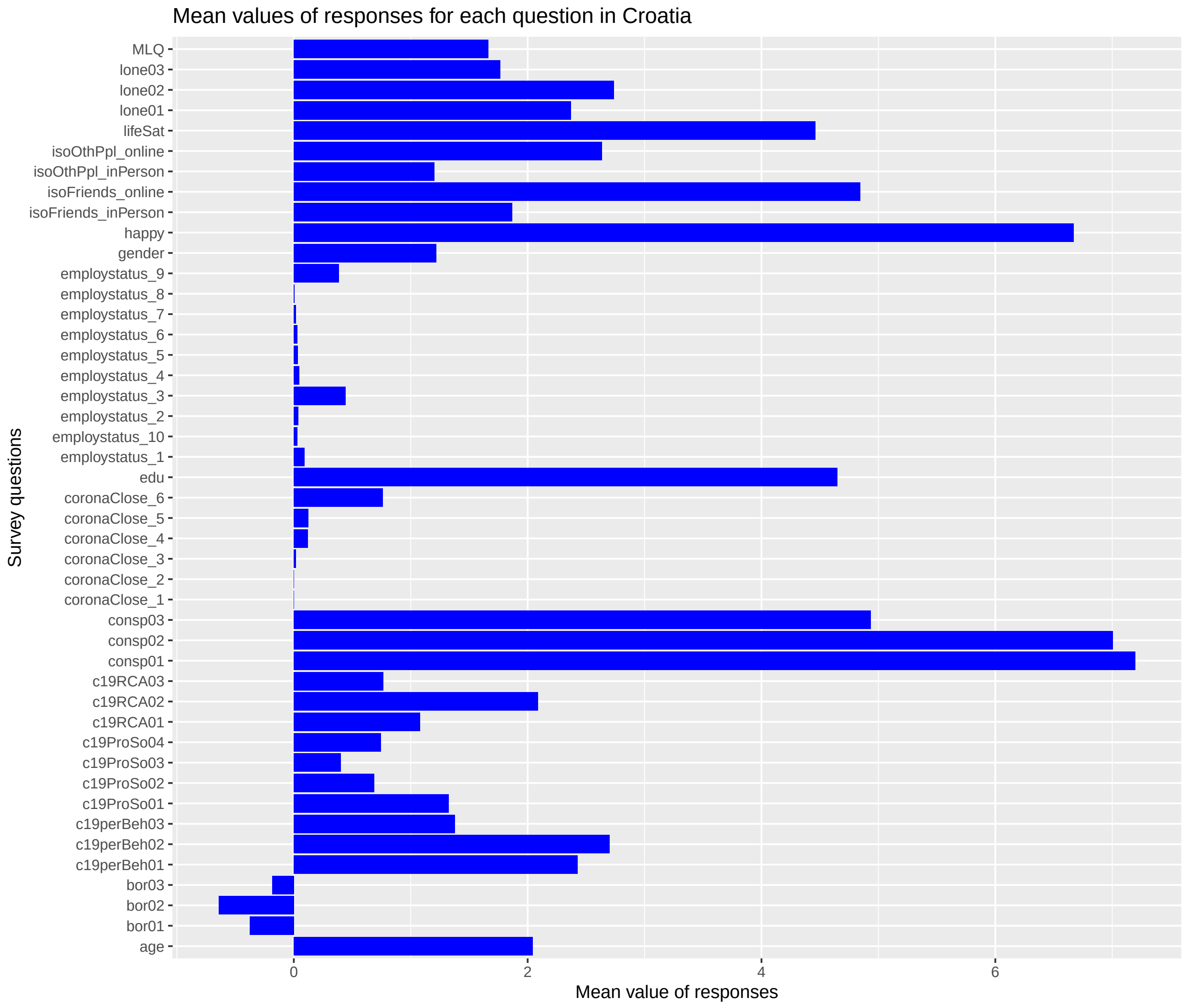
Task 2: Focus country vs all other countries as a group

2(a)

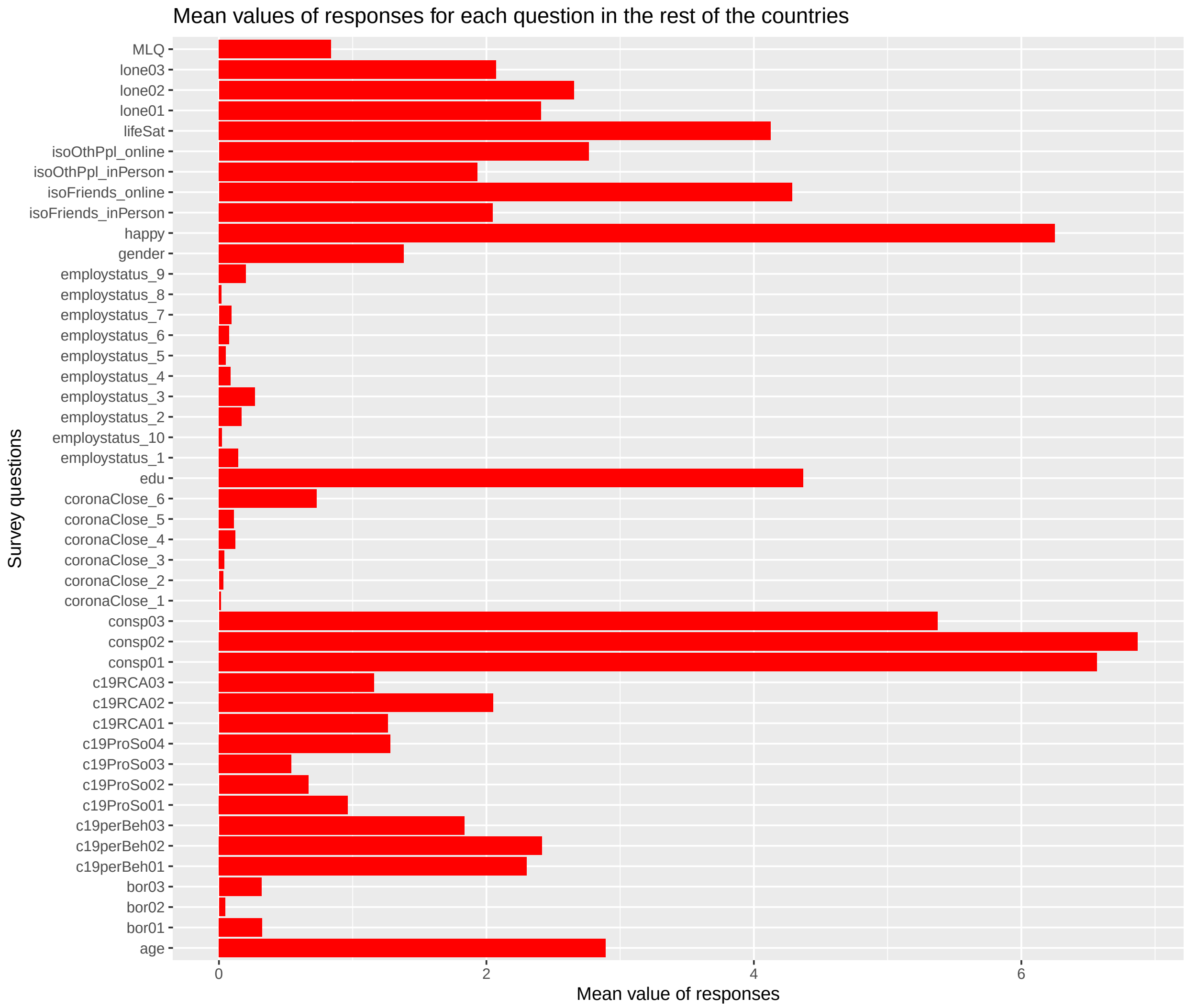
My designated country is Croatia. We will start by creating bar charts for each group of countries, where the y-axis represents the survey questions, and the x-axis represents the mean of each questions' response. Below, we will create data frames for the mean values utilizing `ggplot2`, excluding non-numerical attributes such as `coded_country`.

```r
croatia <- cvbase[cvbase$coded_country == "Croatia", ]
others <- anti_join(cvbase, croatia)
```

```r
numeric_cols <- sapply(croatia, is.numeric)  # Find numeric columns
means <- colMeans(croatia[, numeric_cols], na.rm = TRUE)  # Calculate means

#means <- colMeans(croatia[, !names(croatia) %in% c("coded_country")], na.rm = TRUE)
croatia_means <- data.frame(mean = means)

numeric_cols <- sapply(others, is.numeric)  # Find numeric columns
others_means <- colMeans(others[, numeric_cols], na.rm = TRUE)  # Calculate means
#means <- colMeans(rem[, !names(rem) %in% c("coded_country")], na.rm = TRUE)
others_means <- data.frame(mean = others_means)

croatia_plot <- ggplot(croatia_means) +
  geom_bar(mapping = aes(x = rownames(croatia_means), y = mean), stat = "identity",
    fill = "blue") +
  coord_flip() +
  labs(x = "Survey questions", y = "Mean value of responses",
    title = "Mean values of responses for each question in Croatia")

world_plot <- ggplot(others_means) +
  geom_bar(mapping = aes(x = rownames(others_means), y = mean), stat = "identity",
    fill = "red") +
  coord_flip() +
  labs(x = "Survey questions", y = "Mean value of responses",
    title = "Mean values of responses for each question in the rest of the countries")

croatia_plot
```


Mean values of responses for each question in Croatia
Survey questions
MLQ
lone03
lone02
lone01
lifeSat
isoOthPpl_online
isoOthPpl_inPerson
isoFriends_online
isoFriends_inPerson
happy
gender
employstatus_9
employstatus_8
employstatus_7
employstatus_6
employstatus_5
employstatus_4
employstatus_3
employstatus_2
employstatus_10
employstatus_1
edu
coronaClose_6
coronaClose_5
coronaClose_4
coronaClose_3
coronaClose_2
coronaClose_1
consp03
consp02
consp01
c19RCA03
c19RCA02
c19RCA01
c19ProSo04
c19ProSo03
c19ProSo02
c19ProSo01
c19perBeh03
c19perBeh02
c19perBeh01
bor03
bor02
bor01
age
0
2
4
6
Mean value of responses

```
world_plot
```



Looking at both graphs of Croatia, and all other countries in comparison, most responses seem to be quite similar, except for Boredom, `bor01`,`bor02` and `bor03`. While the worldwide mean is between 0 (Neither agree nor disagree) and 1 (Somewhat agree), in Croatia, the mean is negative and closer to 1 (Somewhat disagree). This leads us to believe people in Croatia could be less bored than other countries worldwide, albeit slightly. Additionally, there seems to be no Corona Proximity for participants themselves, as well as members of their family in this dataset.

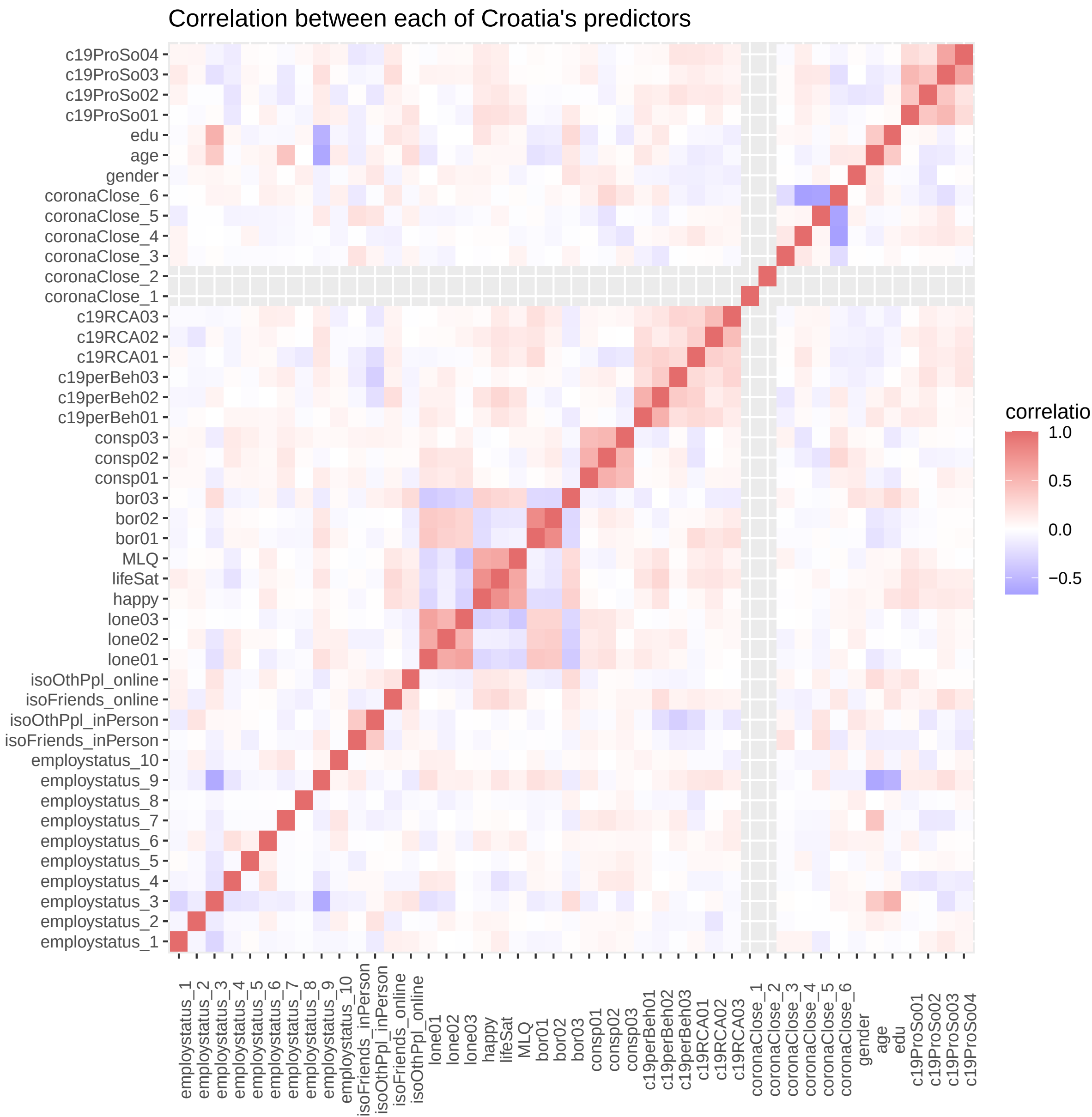
2(b)

Let's start by taking a peek at the correlation between predictors for pro-social attitude for Croatia. The `cor()` function is used and the correlation matrix is visualised with a heatmap.

```r
# # Select only numeric columns from the dataset
numeric_croatia <- croatia[sapply(croatia, is.numeric)]

# # Calculate correlation matrix for Croatia
croatia_cor <- cor(numeric_croatia, use = "complete.obs")
```

```
## Warning in cor(numeric_croatia, use = "complete.obs"): the standard deviation
## is zero
```

```r
# Remove diagonal elements from correlation matrix
#diag(croatia_cor) <- NA

# Reshape matrix to long format for plotting
croatia_melted <- reshape2::melt(croatia_cor, na.rm = TRUE)

# Create correlation plot for Croatia
croatia_cor_plot <- ggplot(data = croatia_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "#6b74ff", mid = "white", high = "#e46c6c", midpoint = 0) +
  labs(title = "Correlation between each of Croatia's predictors", x = "", y = "",
       fill = "correlation") +
  theme(axis.text.x = element_text(angle = 90))

croatia_cor_plot
```



Tiles that are red or blue denote positive or negative correlation, respectively, and turn white as correlation gets closer to zero. Numerous examples of substantial association between predictors can be seen in this heatmap; nevertheless, the portion displaying the correlation between pro-social views and all other traits is quite weak. This suggests that the characteristics might not be a very good indicator of pro-social sentiments in Croatia.

We can determine how well the replies predict the pro-social attitude question by fitting a linear regression model of each pro-social attitude against the qualities. Additionally, we will be able to identify the most accurate predictions.

A linear model of each pro-social attitude versus the qualities is fitted by the code that follows. Each model's R-squared values, significant predictors with p-values less than 0.001, and corresponding coefficients are summarised using a function and a for loop. The vectors `predictors` and `each_model` will be utilised in a subsequent table to compare the strong predictors for every model.

```r
predictors <- NULL
each_model <- NULL

model_eval <- function(model) {
  rsqr <- summary(model)$r.squared
  a_rsqr <- summary(model)$adj.r.squared
  sig <- which(summary(model)$coefficients[-1, 4] < 0.001) + 1
  preds <- rownames(summary(model)$coefficients[sig, , drop = FALSE])
  coefs <- summary(model)$coefficients[sig, 1]

  return(list(rsqr, a_rsqr, preds, coefs))
}

fitted_croatia1 <- lm(c19ProSo01 ~ .,
  data = subset(croatia, select = -c(coded_country, c19ProSo02, c19ProSo03, c19ProSo04)))
fitted_croatia2 <- lm(c19ProSo02 ~ .,
  data = subset(croatia, select = -c(coded_country, c19ProSo01, c19ProSo03, c19ProSo04)))
fitted_croatia3 <- lm(c19ProSo03 ~ .,
  data = subset(croatia, select = -c(coded_country, c19ProSo01, c19ProSo02, c19ProSo04)))
fitted_croatia4 <- lm(c19ProSo04 ~ .,
  data = subset(croatia, select = -c(coded_country, c19ProSo01, c19ProSo02, c19ProSo03)))

cat("Pro-social attitudes in Croatia predictors model summary\n\n")
```

```
## Pro-social attitudes in Croatia predictors model summary
```

```r
counter <- 1
for (model in list(fitted_croatia1, fitted_croatia2, fitted_croatia3, fitted_croatia4)) {
  cat("C19ProSo0", counter, "\n", sep = "")
  res <- model_eval(model)
  cat("R-squared value:", res[[1]], "\n")
  cat("Adjusted R-squared value:", res[[2]], "\n")
  cat("Significant predictors with p-value < 0.001:\n")
  cat(res[[3]], "\n")
  cat("Coefficients of predictors:\n")
  cat(res[[4]], "\n")
  cat("\n")
  for (pred in res[[3]]) {
    each_model <- c(each_model, paste0("Croatia_C19ProSo0", counter))
  }
  predictors <- c(predictors, res[[3]])
  counter <- counter + 1
}
```

```
## C19ProSo01
## R-squared value: 0.4686053
## Adjusted R-squared value: 0.2519858
## Significant predictors with p-value < 0.001:
##
## Coefficients of predictors:
##
##
## C19ProSo02
## R-squared value: 0.3355854
## Adjusted R-squared value: 0.06474127
## Significant predictors with p-value < 0.001:
##
## Coefficients of predictors:
##
##
## C19ProSo03
## R-squared value: 0.4872678
## Adjusted R-squared value: 0.278256
## Significant predictors with p-value < 0.001:
## isoFriends_online
## Coefficients of predictors:
## 0.1795645
##
## C19ProSo04
## R-squared value: 0.3617789
## Adjusted R-squared value: 0.1016124
## Significant predictors with p-value < 0.001:
## rankOrdLife_4D
## Coefficients of predictors:
## 6.462742
```

`C19ProSo04` has the greatest adjusted R-squared value of 0.2113349 out of all the models, indicating that the responses best predict it. `disc02`, `MLQ`, `c19NormShould`, and `c19IsPunish` are its best predictors. The model for `C19ProSo03` has the lowest adjusted R-squared value, 0.08190663, and `PLRAC19`, `MLQ`, `c19NormShould`, `trustGovState`, and `edu` are its best predictors.

The fact that the majority of the survey items are deemed subjective makes the models' arguably low R-squared values predictable. For instance, different individuals interpret financial hardship differently and perceive various levels of serenity differently. Since Croatia is a large, populated nation with a wide range of living standards, its several regions are like independent nations with their own economies, healthcare systems, and general levels of satisfaction. Because of this, it is challenging to forecast the pro-social attitude reactions with consistency.

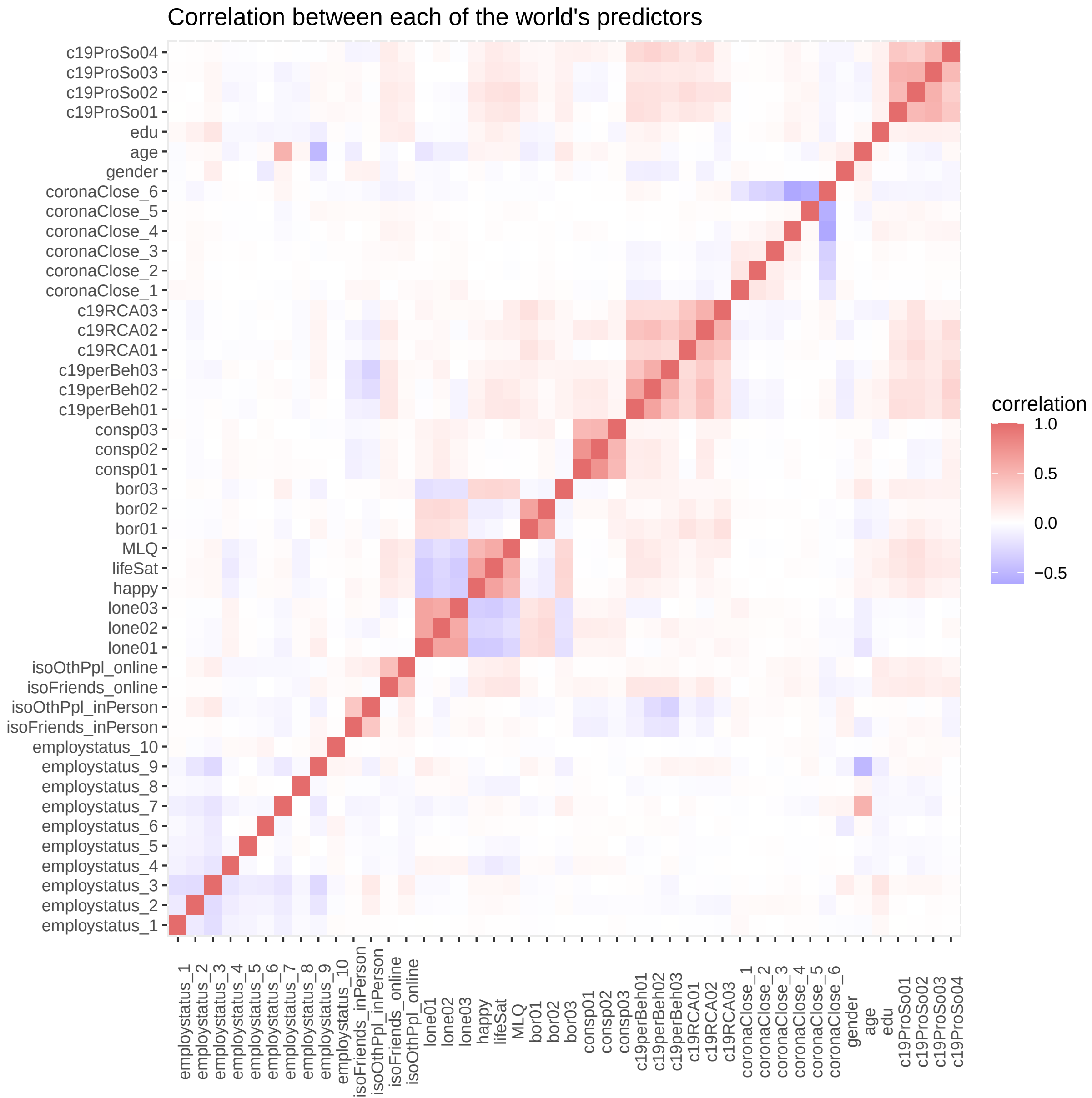
Although each model has a unique set of important predictors, some predictors can be regarded as generally more reliable because they are more frequently found in all of the models. The best illustration would be `c19NormShould`, a highly predictive variable for each of the four models. This makes sense since someone who is eager to help society during a pandemic would want the best for it and would advise people to withdraw from social interactions and seclusion. The Centers for Disease Control and Prevention (CDC) in Croatia recommend these steps to stop the spread of viruses, and since Croatia is a developed country with a highly educated populace, people who aspire to be pro-social generally abide by these recommendations. Conversely, a person devoid of pro-social attitudes would not care about adhering to new rules or showing any interest in societal behaviors. Those who disagree with social distancing policies and believe that doing so benefits society as a whole may also have an impact on the predictive power of `c19NormShould`. During the epidemic, lock down protests were prevalent in Croatia, demonstrating the validity of this viewpoint.

`disc02`, `MLQ`, and `trustGovState` are additional variables that predict three of the models effectively. People are more likely to be pro-social if they care about the future of society, if they have a purpose in life, and if they think society can come to an agreement on how to handle the pandemic.

2(c)

The previous code blocks for 2(b) are modified, but with the `rem` data set instead of `croatia`, to produce a similar correlation matrix for the rest of the world. To keep the report concise, variants of code that have been modified further on, can be found in the **Appendix**.

```
global_cor_plot
```



When we compare the two heatmaps we currently have, we see that `cro_cor_plot` has tiles coloured in a deeper shade of red, which suggests a larger overall correlation between the predictors. Apart from having lighter tiles, `rem_cor_plot` appears "cleaner" due to a reduced dispersion of coloured tiles. However, since the subsections in both plots appear fairly similar, we can make an initial guess that the attributes for both groups of data should predict pro-social attitudes with roughly similar performance by concentrating on the heatmap subsections that show the correlation between pro-social attitudes and all other attributes.

```
## Pro-social attitudes in the world predictors model summary
```

```
## C19ProSo01
## R-squared value: 0.1280411
## Adjusted R-squared value: 0.12646
## Significant predictors with p-value < 0.001:
## employstatus_6 employstatus_7 employstatus_8 employstatus_10 isoFriends_inPerson isoOthPpl_inPerson i
## Coefficients of predictors:
## -0.09356577 -0.2272962 -0.1897203 0.3730428 0.01719855 0.02667376 0.01609172 0.01721033 0.0606518 -0
##
## C19ProSo02
## R-squared value: 0.1766909
## Adjusted R-squared value: 0.1751979
## Significant predictors with p-value < 0.001:
## employstatus_3 employstatus_4 employstatus_5 employstatus_8 employstatus_10 isoFriends_inPerson isoFr
## Coefficients of predictors:
## 0.1072481 -0.1977689 -0.1317748 -0.3161566 0.2091906 0.03416122 0.0152765 0.02820841 0.05811839 -0.04
##
## C19ProSo03
## R-squared value: 0.1199772
## Adjusted R-squared value: 0.1183814
## Significant predictors with p-value < 0.001:
## employstatus_3 employstatus_7 employstatus_10 isoFriends_inPerson isoOthPpl_inPerson isoOthPpl_online
## Coefficients of predictors:
## 0.1555728 -0.2152103 0.3369801 0.02274039 0.02160123 0.02621581 0.066345 0.09423445 0.0588311 0.0282
##
## C19ProSo04
## R-squared value: 0.17178
## Adjusted R-squared value: 0.1702782
## Significant predictors with p-value < 0.001:
## employstatus_2 employstatus_3 employstatus_10 isoFriends_online lone02 lone03 lifeSat MLQ bor02 bor03
## Coefficients of predictors:
## 0.09869336 0.1272541 0.2820308 0.01248981 0.03724033 0.03827886 0.07587075 0.02828067 0.02557799 0.04
```

All four models had adjusted R-squared values between 0.12 and 0.17, which is a smaller range than the similar range for the Croatia data set (0.06 - 0.49), according to the summary for the rest of the world. The models' predictors are significantly more important than those of the `croatia` models. The four models are well predicted by `disc02`, `lifeSat`, `c19NormShould`, `c19NormDo`, and `trustGovState`. The majority of the predictors that performed well in all four of the `croatia` models, `c19NormShould`, `disc02`, and `trustGovState` are included in this set. As was already said, Croatia's sheer vastness and diversity make it seem like a collection of independent nations. It follows that strong predictors for Croatia would also apply to other nations collectively.

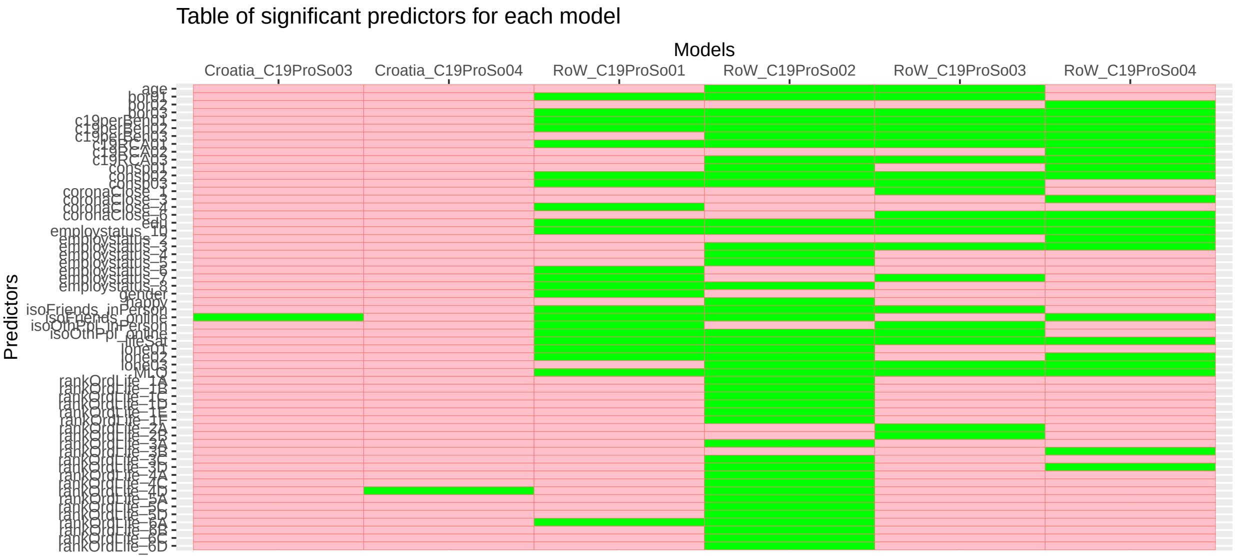
The table below, created with `ggplot2`, displays the results of the best predictors for each pro-social attitude for Croatia and other nations collectively.

TODO LATER!!!!

```r
summ_table <- table(predictors = predictors, models = each_model)

colnames(summ_table)
```

```
## [1] "Croatia_C19ProSo03" "Croatia_C19ProSo04" "RoW_C19ProSo01"
## [4] "RoW_C19ProSo02"     "RoW_C19ProSo03"     "RoW_C19ProSo04"
```

```r
# reorder the columns
#summ_table <- summ_table[, c("Croatia_C19ProSo01", "Croatia_C19ProSo02", "Croatia_C19ProSo03",
summ_table <- summ_table[, c("Croatia_C19ProSo03",
  "Croatia_C19ProSo04", "RoW_C19ProSo01", "RoW_C19ProSo02", "RoW_C19ProSo03",
  "RoW_C19ProSo04")]

summ_table_vis <- ggplot(data = as.data.frame(summ_table)) +
  geom_tile(mapping = aes(x = models, y = predictors, fill = Freq, colour = "black")) +
  scale_fill_gradientn(colours = c("pink", "green")) +
  theme(legend.position = "none") +
  scale_x_discrete(position = "top") +
  scale_y_discrete(limits = rev) +
  labs(x = "Models", y = "Predictors",
    title = "Table of significant predictors for each model")

summ_table_vis
```



Task 3: Focus country vs cluster of similar countries

3(a)

TODO NOW!!!!

Some additional socioeconomic and health data have been collected from other websites in addition to the indicators available in the sources mentioned in the references. Eight indicators make up the final data table I have compiled (in **Appendix**) for use in clustering: `HDI}, {GHS}`,freedom,political_stability,happiness,total_vax_per_hu `and`total_deaths_per_mil‘. The **Appendix** contains information and explanations regarding each indicator as well as its sources.

Using k-means clustering, nations that are comparable to Croatia are found. For the `kmeans()` function to function, countries with NA values must first be eliminated. This has no bearing on our findings because the majority of these nations—such as Afghanistan and Syria—do not initially appear in the baseline data and differ greatly from Croatia in terms of development and data transparency. After scaling the data, K-means clustering is carried out using 15 random beginnings.

```
# collected <- read.csv("task3.csv")
# collected_clean <- na.omit(collected)
# collected_clean[, 2:9] <- scale(collected_clean[, 2:9])
#
# kfit <- kmeans(collected_clean[, 2:9], round(nrow(collected_clean) / 5), nstart = 15)
# clusters <- data.frame(country = collected_clean[[1]], cluster = kfit$cluster)
#
# target <- filter(clusters, country == "Croatia")$cluster
# similar <- filter(clusters, cluster == target)
# similar
```

The clustering indicates that Belgium, the Czech Republic, Lithuania, Slovenia, and the United Kingdom are comparable to Croatia.

3(b)

After removing the data from Croatia, the baseline data of the cluster's member nations are initially extracted using an inner join of `cvbase` and `similar`. The correlation matrix for this subset of data is then shown, same to what was done for `croatia` and `rem`.

```
# colnames(similar)[colnames(similar) == "country"] <- "coded_country"
# intersect <- merge(cvbase, similar, by = "coded_country", all = FALSE)
# intersect <- intersect[, -ncol(intersect)]
# clus <- filter(intersect, coded_country != "Croatia")
#
# clus_cor <- cor(subset(clus, select = -coded_country), use = "complete.obs")
# clus_melted <- reshape2::melt(clus_cor)
# ```

# ```{r include = FALSE}
# clus_cor_plot <- ggplot(data = clus_melted) +
#   geom_tile(mapping = aes(x = Var1, y = Var2, fill = value)) +
#   scale_fill_gradient2(low = "#6b74ff", mid = "white", high = "#e46c6c", midpoint = 0) +
#   labs(title = "Correlation between predictors for countries similar to Croatia",
#     x = "", y = "", fill = "correlation") +
#   theme(axis.text.x = element_text(angle = 90))
```

```
# clus_cor_plot
```

This heat map's scatter of colored tiles is identical to the Croatia heat map's, showing how similar these two nations are. In comparison to the previous plots, the portion of tiles displaying the correlation between predictors and pro-social views is generally darker, suggesting that the predictors for this cluster of countries may perform better in terms of prediction than the data from the previous two groups.

The same code used in 2(b) and 2(c) is repeated to print a structured summary of the four models in order to determine how participant answers predict pro-social views for this cluster of similar countries.

Based on the results, the models for these comparable nations typically have adjusted R-squared values that are comparable to those of Croatia and all other nations combined. Similar to the Croatia models, the model for `C19ProSo04` has the greatest adjusted R-squared value (0.2478493). With the exception of the model for `C19ProSo04`, whose significant predictors include `disc02` and `PFS02`, none of these models, in contrast to the preceding eight models, had significant predictors with p-values less than 0.001. In the Croatia model, disc02 is also a very good predictor for C19ProSo04, but not for PFS02. However, disc02 and PFS02 are both significant predictors in the rest-of-the-world model for C19ProSo04.

Therefore, with comparable R-squared values and predictors with generally larger p-values, the predictive performance of attributes for this cluster of countries is not substantially better than that of Croatia or the rest of the world. Rather than reflecting true statistically significant links between attribute and pro-social attitude, the previously found substantial association may have been the result of chance or a limited sample size.

We can define a strong predictor in relation to a model's total p-values for comparative purposes. For these novel cluster models, we characterize a strong predictor as one with a p-value of less than 0.05, which is a widely accepted threshold. A new visualization table is constructed and the model_eval function is changed to reflect this (see **Appendix**).

```
# summ_table_vis_2
```

We note that, with a few shared significant predictors, the distribution of strong predictors between the models of similar countries is more akin to that of the Croatian models (i.e., they appear as "sparse" as the US models). With more comparable p-values, the models from the group of all other nations have a great deal more significant predictors in common with the Croatian models. These models do, however, also contain a large number of powerful predictors that are absent from the Croatian models. Consequently, the group of comparable nations may provide a better fit to the critical characteristics needed to predict pro-social sentiments. When more research is conducted or a larger sample size is used, the higher p-values and fewer shared common strong predictors included in their models might no longer be noticeable.

One explanation could be that each country in the cluster differs slightly from the others in terms of socioeconomic aspects that are not included in the clustering indicators, even if they are all similar to Croatia. Their collective performance in forecasting pro-social attitudes differs substantially from Croatia alone when these small variations are taken into account. However, because Croatia's politics, culture, and other aspects of society are complicated, much like those of a group of many countries, its models have many strong predictors in common with the models of the group of all countries. Because the group of all other countries may be excessively big and complex, several significant predictors that are not important in reality may be reported by the models.

Appendix

cvbase head, 1(b)

```
# head(cvbase)
```

rem Correlation matrix, 2(c)

```
# rem_cor <- cor(subset(rem, select = -coded_country), use = "complete.obs")
# rem_melted <- reshape2::melt(rem_cor)
#
# rem_cor_plot <- ggplot(data = rem_melted) +
#   geom_tile(mapping = aes(x = Var1, y = Var2, fill = value)) +
#   scale_fill_gradient2(low = "#6b74ff", mid = "white", high = "#e46c6c", midpoint = 0) +
#   labs(title = "Correlation between predictors for the rest of the world", x = "", y = "",
#     fill = "correlation") +
#   theme(axis.text.x = element_text(angle = 90))
```

rem models, 2(c)

```
# fitted_rem1 <- lm(c19ProSo01 ~ .,
#   data = subset(rem, select = -c(coded_country, c19ProSo02, c19ProSo03, c19ProSo04)))
# fitted_rem2 <- lm(c19ProSo02 ~ .,
#   data = subset(rem, select = -c(coded_country, c19ProSo01, c19ProSo03, c19ProSo04)))
# fitted_rem3 <- lm(c19ProSo03 ~ .,
#   data = subset(rem, select = -c(coded_country, c19ProSo01, c19ProSo02, c19ProSo04)))
# fitted_rem4 <- lm(c19ProSo04 ~ .,
#   data = subset(rem, select = -c(coded_country, c19ProSo01, c19ProSo02, c19ProSo03)))
#
# cat("Summary of models for predicting pro-social attitudes in the rest of the world\n\n")
# counter <- 1
# for (model in list(fitted_rem1, fitted_rem2, fitted_rem3, fitted_rem4)) {
#   cat("C19ProSo0", counter, "\n", sep = "")
#   res <- model_eval(model)
#   cat("R-squared value:", res[[1]], "\n")
#   cat("Adjusted R-squared value:", res[[2]], "\n")
#   cat("Significant predictors with p-value < 0.001:\n")
#   cat(res[[3]], "\n")
#   cat("Coefficients of predictors:\n")
#   cat(res[[4]], "\n")
#   cat("\n")
#   for (pred in res[[3]]) {
#     each_model <- c(each_model, paste0("RoW C19ProSo0", counter))
#   }
#   predictors <- c(predictors, res[[3]])
#   counter <- counter + 1
# }
```

Clustering data, 3(a)

```
# collected
```

Clustering indicators, 3(a) ###### TODO: WARNING!!!!!!!!########## - `HDI`: Human Development Index (2021); a value between 0 and 1 that measures average achievement in human development based on three dimensions - life expectancy, education and standard of living. (Source: Human Development Reports) - `GHS`: Global Health Security Index (2021); a value between 0 and 100 that benchmarks a country's health security and preparedness in preventing, detecting and responding to health emergencies. (Source: Global Health Security Index: Reports and Data) - `freedom`: Human Freedom Index (2021); a value between 0 and 10 that assesses the level of human freedom in a country. Human freedom is a combination of two distinct dimensions - personal freedom (freedom of religion, speech, sexual orientation, etc.) and economic freedom (size of government, judicial impartiality, freedom to trade, etc.) (Source: World Population Review) - `political_stability`: a value **approximately** between -2.5 and 2.5 that evaluates political stability and absence of violence/terrorism of each country in 2021. (Source: The World Bank Data Collections (and Governance Indicators)) - `happiness`: World Happiness Report score (2021); a value between 0 and 10 that represents happiness of a country's citizens based on several socioeconomic factors. (Source: World Happiness Report) - `total_vax_per_hundred`: latest updated total number of COVID-19 vaccinations administered per 100 people before 2022. - `total_cases_per_mil`: latest updated total number of COVID-19 cases per 1,000,000 people before 2022. - `total_deaths_per_mil`: latest updated total number of COVID-19 cases per 1,000,000 people before 2022.

The last three indicators were sourced from Our World in Data's COVID-19 Github repository.

K-means clustering, 3(a)

```
# library(cluster)
# clusplot(collected_clean, kfit$cluster, color = TRUE, shade = TRUE, labels = 0, lines = 0)
```

`rem` correlation matrix, 3(b)

```
# clus_cor_plot <- ggplot(data = clus_melted) +
#   geom_tile(mapping = aes(x = Var1, y = Var2, fill = value)) +
#   scale_fill_gradient2(low = "#6b74ff", mid = "white", high = "#e46c6c", midpoint = 0) +
#   labs(title = "Correlation between predictors for countries similar to the Croatia",
#     x = "", y = "", fill = "correlation") +
#   theme(axis.text.x = element_text(angle = 90))
```

`clus` model, 3(b)

```
# fitted_clus1 <- lm(c19ProSo01 ~ .,
#   data = subset(clus, select = -c(coded_country, c19ProSo02, c19ProSo03, c19ProSo04)))
# fitted_clus2 <- lm(c19ProSo02 ~ .,
#   data = subset(clus, select = -c(coded_country, c19ProSo01, c19ProSo03, c19ProSo04)))
# fitted_clus3 <- lm(c19ProSo03 ~ .,
#   data = subset(clus, select = -c(coded_country, c19ProSo01, c19ProSo02, c19ProSo04)))
# fitted_clus4 <- lm(c19ProSo04 ~ .,
#   data = subset(clus, select = -c(coded_country, c19ProSo01, c19ProSo02, c19ProSo03)))
#
# cat("Summary of models for predicting pro-social attitudes in countries similar to the US\n\n")
# counter <- 1
# for (model in list(fitted_clus1, fitted_clus2, fitted_clus3, fitted_clus4)) {
#   cat("C19ProSo0", counter, "\n", sep = "")
#   res <- model_eval(model)
#   cat("R-squared value:", res[[1]], "\n")
#   cat("Adjusted R-squared value:", res[[2]], "\n")
#   cat("Significant predictors with p-value < 0.001:\n")
#   cat(res[[3]], "\n")
#   cat("Coefficients of predictors:\n")
#   cat(res[[4]], "\n")
#   cat("\n")
#   counter <- counter + 1
# }
```

`clus` models, updated `model_eval` function with p-value less than 0.05, 3(b)

```
# model_eval_2 <- function(model) {
#   rsqr <- summary(model)$r.squared
#   a_rsqr <- summary(model)$adj.r.squared
#   sig <- which(summary(model)$coefficients[-1, 4] < 0.05) + 1
#   preds <- rownames(summary(model)$coefficients[sig, , drop = FALSE])
#   coefs <- summary(model)$coefficients[sig, 1]
#
#   return(list(rsqr, a_rsqr, preds, coefs))
# }
#
# cat("Summary of models for predicting pro-social attitudes in countries similar to the US\n\n")
# counter <- 1
# for (model in list(fitted_clus1, fitted_clus2, fitted_clus3, fitted_clus4)) {
#   cat("C19ProSo0", counter, "\n", sep = "")
#   res <- model_eval_2(model)
#   cat("R-squared value:", res[[1]], "\n")
#   cat("Adjusted R-squared value:", res[[2]], "\n")
#   cat("Significant predictors with p-value < 0.05:\n")
#   cat(res[[3]], "\n")
#   cat("Coefficients of predictors:\n")
#   cat(res[[4]], "\n")
#   cat("\n")
#   for (pred in res[[3]]) {
#     each_model <- c(each_model, paste0("Similar C19ProSo0", counter))
#   }
#   predictors <- c(predictors, res[[3]])
#   counter <- counter + 1
# }
```

`croatia`, `rem` and `clus` models strongest predictors, 3(b)

```
# summ_table_2 <- table(predictors = predictors, models = each_model)
# summ_table_2 <- summ_table_2[, c("Croatia_C19ProSo01", "Croatia_C19ProSo02", "Croatia_C19ProSo03",
#   "Croatia_C19ProSo04", "RoW_C19ProSo01", "RoW_C19ProSo02", "RoW_C19ProSo03", "RoW_C19ProSo04",
#   "Similar C19ProSo01", "Similar C19ProSo02", "Similar C19ProSo03", "Similar C19ProSo04")]
#
# summ_table_vis_2 <- ggplot(data = as.data.frame(summ_table_2)) +
#   geom_tile(mapping = aes(x = models, y = predictors, fill = Freq, colour = "black")) +
#   scale_fill_gradientn(colours = c("pink", "green")) +
#   theme(legend.position = "none") +
#   scale_x_discrete(position = "top") +
#   scale_y_discrete(limits = rev) +
#   labs(x = "Pro-social attitudes", y = "Predictors",
#     title = "Table of significant predictors for each pro-social attitude") +
#   theme(axis.text.x = element_text(angle = 90))
```