Group exercise 1: Survey research
DATA5207: Data Analysis in the Social Sciences

Emily Wong and Sarah Sweeting

Lab 1: Creating Predictors

Introduction

In this study, data from The World Value Survey will be explored to explain quality of life. Part 1 will explore potential predictors to predict higher/lower quality of life with supported theory. Part 2 will tests these predictors through the creation of a dependent variable and predictive models.

Understanding survey data

```
survey.data <- read.csv("wvs_data.csv")
```

Chosen Predictors

Quality of life is difficult to quantify and can be affected by numerous factors in ones life. Another issue is that 'quality' for an individual could be determined by numerous things such as satisfaction, health and wealth.

To help guide our choices of predictors, we will use the World Health Organisations definition of quality of life (QoL) as an "individuals' perceptions of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns".

[https://www.who.int/tools/whoqol]

We believe the factors that affect an indiviudal's perception of life spans further than physical health and wellbeing (however this is obviously also a factor to consider), it can include psychological, environmental, societal and spiritual factors.

Employment - as means of wealth   Q279: Are you employed now or not? If yes, about how many hours a week do you work? If you have more than one job, please tell us about your main job only.

Education - as means of standard of living   Q275: What is your highest educational level that you have attained?

Security   Q131: how secure do you feel these days?

Human Rights   Q253: How much respect is there for individual human rights nowadays in this country? Do you feel there is…

Health Levels   Q53: "In the last 12 months, how often have you or your family gone without medicine or medical treatment that you needed?

Social & Personal Connectivity Q2. For each of the following, indicate how important it is in your life. How important is... Family, Friends

Theory

Education and employment: When an individual has higher educational attainments and are employed, they are able to have a better quality of life. This is due to higher economic well being and financial security to meet basic needs such as healthcare access and participating in leisure activites. It also can lead to a greater sense of purpose and personal development.

Confounding factor here may be income, since these factors are indicators of what sort of job an individual has an how much they are getting paid as a result (higher education levels = greater income, more hours in employment = higher income)

Security: When an individual is able to live without fear or risk and feels stable, their quality of life will increase. Security can be an impact on the quality of life as it is a measure of both physical safety and economic safety, physical safety affecting individuals health and mental well being, while as economic security to have social safety net to tackle financial challenges and access basic needs.

Human rights: When society upholds and protects human rights, individuals tend to experience better QoL. It is a basic need but also a measure of psychological well-being because of its ability to foster belonging within communities and equality within a just society. It also could promote quality of opportunity, a fair chance for success in society. can you add a blurb for health levels and social and personal connectivity

Health Levels: Health is a foundational element of quality of life. Good health enables individuals to engage actively in various aspects of life including work, social interactions, and leisure activities, thus directly influencing their overall life satisfaction and well-being. Therefore, having a lack of access to necessary medical intervention can lead to decreased physical capabilities, psychological stress, and financial burdens, all of which negatively affect one's quality of life.

Social & Personal Connectivity: Human beings are inherently social creatures, and the quality of our social interactions can significantly impact our mental and emotional well-being. Strong connections with family and friends provide emotional support, reduce stress, and contribute to a sense of belonging and happiness. Furthermore, placing value in personal relationships can influence one's self-esteem and provide opportunities for meaningful engagement in community activities, thereby enhancing an individual's overall quality of life.

Exploring Predictors

```
glimpse(colnames(survey.data))
```

```
##  chr [1:317] "ID" "Mode" "State" "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" ...
```

The column names aren't entirely that helpful. We re-code them for convenience.

```
categories = c("Employment", "Education", "Security", "Human Rights", "Treatment Levels",  "Friendship
questions = c("Q279", "Q275", "Q131", "Q253", "Q53", "Q2")
key = c("V249", "Q275", "V170", "V142", "V190", "V5")
```

```
v249_employment_status <- c(
  "No answer",
  "Full time employee (30 hours a week or more)",
  "Part time employee (less than 30 hours a week)",
  "Self-employed",
  "Retired / On a pension",
  "Home duties, not otherwise employed",
  "Student",
  "Unemployed",
  "Other (please specify)"
```

```
)

values.249 = c(-2,1,2,3,4,5,6,7,8)

q275_education = c("No answer",
"No formal education",
"Primary education only",
"Lower secondary education (i.e. Year 9 or less)",
"Upper secondary education (i.e. between Year 10 and Year 12)",
"Post-secondary non-tertiary education (e.g. apprenticeship or certificate)",
"Bachelor or equivalent",
"Master or equivalent",
"Doctoral or equivalent")

values.275 = c(-2,1,2,3,4,5,7,8,9)

v170_how_secure_are_you = c(
"No answer",
"Very secure",
"Quite secure",
"Not very secure",
"Not at all secure"
)

values.170 = c(-2,1,2,3,4)

v142_respect_human_rights = c(
"No answer",
"A great deal of respect",
"Some respect",
"Not much respect",
"No respect at all"
)

values.142 = c(-2,1,2,3,4)


v190_gone_without_medicine = c(
"No answer",
"Often",
"Sometimes",
"Rarely",
"Never"
)

values.190 = c(-2,1,2,3,4)

v5_friends = c(
"No answer",
"Very important",
"Rather important",
"Not very important",
"Not at all important"
```

```r
)

values.5 = c(-2,1,2,3,4)

categories <- c("Employment", "Education", "Security", "Human Rights", "Treatment Levels", "Friendship
questions <- c("Q279", "Q275", "Q131", "Q253", "Q53", "Q2")
key <- c("V249", "Q275", "V170", "V142", "V190", "V5")

# Define vectors for each category
key <- c("V249", "Q275", "V170", "V142", "V190", "V5")

vectors <- list(
  list(
    values = c(-2, 1, 2, 3, 4, 5, 6, 7, 8),
    chars = c(
      "No answer",
      "Full time employee (30 hours a week or more)",
      "Part time employee (less than 30 hours a week)",
      "Self-employed",
      "Retired / On a pension",
      "Home duties, not otherwise employed",
      "Student",
      "Unemployed",
      "Other (please specify)"
    )
  ),
  list(
    values = c(-2, 1, 2, 3, 4, 5, 7, 8, 9),
    chars = c(
      "No answer",
      "No formal education",
      "Primary education only",
      "Lower secondary education (i.e. Year 9 or less)",
      "Upper secondary education (i.e. between Year 10 and Year 12)",
      "Post-secondary non-tertiary education (e.g. apprenticeship or certificate)",
      "Bachelor or equivalent",
      "Master or equivalent",
      "Doctoral or equivalent"
    )
  ),
  list(
    values = c(-2, 1, 2, 3, 4),
    chars = c(
      "No answer",
      "Very secure",
      "Quite secure",
      "Not very secure",
      "Not at all secure"
    )
  ),
  list(
    values = c(-2, 1, 2, 3, 4),
    chars = c(
      "No answer",
```

```r
      "A great deal of respect",
      "Some respect",
      "Not much respect",
      "No respect at all"
    )
  ),
  list(
    values = c(-2, 1, 2, 3, 4),
    chars = c(
      "No answer",
      "Often",
      "Sometimes",
      "Rarely",
      "Never"
    )
  ),
  list(
    values = c(-2, 1, 2, 3, 4),
    chars = c(
      "No answer",
      "Very important",
      "Rather important",
      "Not very important",
      "Not at all important"
    )
  )
)

# Combine data into a dataframe
Question.key <- bind_rows(lapply(1:length(categories), function(i) {
  data.frame(
    Categories = rep(categories[i], length(vectors[[i]]$values)),
    Question.Number = rep(questions[i], length(vectors[[i]]$values)),
    Codebook = rep(key[i], length(vectors[[i]]$values)),
    Question.values = vectors[[i]]$values,
    Question.chars = vectors[[i]]$chars
  )
}))
```

```r
survey.predictors <- data.frame(
  education = survey.data$Q275,
  employment = survey.data$V249,
  security = survey.data$V170,
  rights = survey.data$V142,
  health = survey.data$V190,
  social = survey.data$V5,
  satisfaction = survey.data$V23
)
```

```r
summary = summary(survey.predictors)
print(summary)
```

```
##    education       employment        security          rights
##  Min.   :-2.00   Min.   :-2.000   Min.   :-2.000   Min.   :-2.000
```

```
##   1st Qu.: 4.00    1st Qu.: 1.000    1st Qu.: 2.000    1st Qu.: 1.000
##   Median : 5.00    Median : 2.000    Median : 2.000    Median : 2.000
##   Mean   : 5.36    Mean   : 2.709    Mean   : 1.977    Mean   : 1.882
##   3rd Qu.: 7.00    3rd Qu.: 4.000    3rd Qu.: 2.000    3rd Qu.: 2.000
##   Max.   : 9.00    Max.   : 8.000    Max.   : 4.000    Max.   : 4.000
##       health           social         satisfaction
##   Min.   :-2.000   Min.   :-2.000   Min.   :-2.000
##   1st Qu.: 4.000   1st Qu.: 1.000   1st Qu.: 7.000
##   Median : 4.000   Median : 1.000   Median : 8.000
##   Mean   : 3.638   Mean   : 1.389   Mean   : 7.443
##   3rd Qu.: 4.000   3rd Qu.: 2.000   3rd Qu.: 9.000
##   Max.   : 4.000   Max.   : 4.000   Max.   :10.000
```

Distribution of predictors

```r
plot_barplots <- function(column, name, values_fill) {
  ggplot(survey.data, aes(x = factor(!!column), fill = factor(!!column))) +
    geom_bar() +
    labs(title = "Distribution", x = name, fill = name, y = "Count") +
    scale_fill_discrete(labels = values_fill) +
    theme_minimal()
}

# Get unique codebooks
unique_codebooks <- unique(Question.key$Codebook)
# Loop through each unique codebook and plot bar plot
for (codebook in unique_codebooks) {
  subset_data <- filter(Question.key, Codebook == codebook)
  print(plot_barplots(column = sym(codebook), name = unique(subset_data$Categories), values_fill = uniqu
}
```

Distribution

Distribution

Distribution

Distribution

Distribution

Distribution
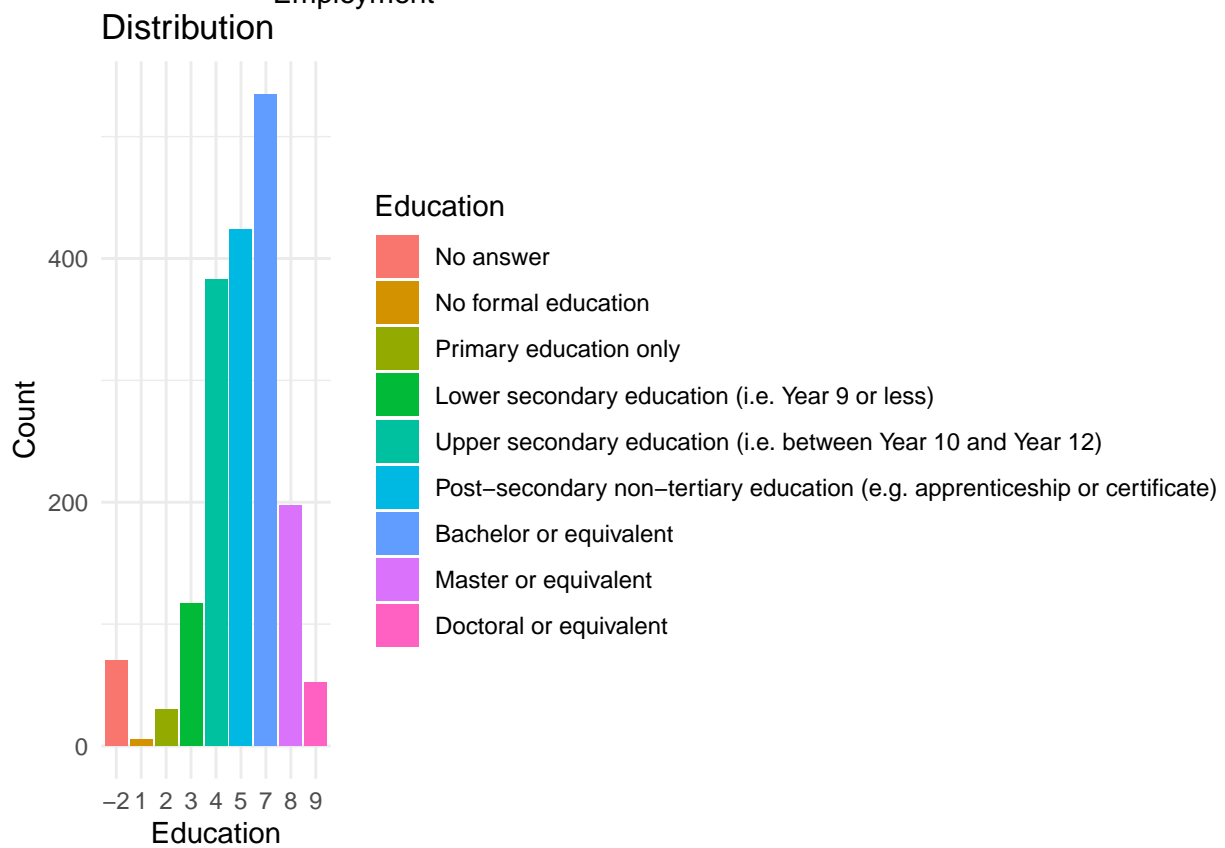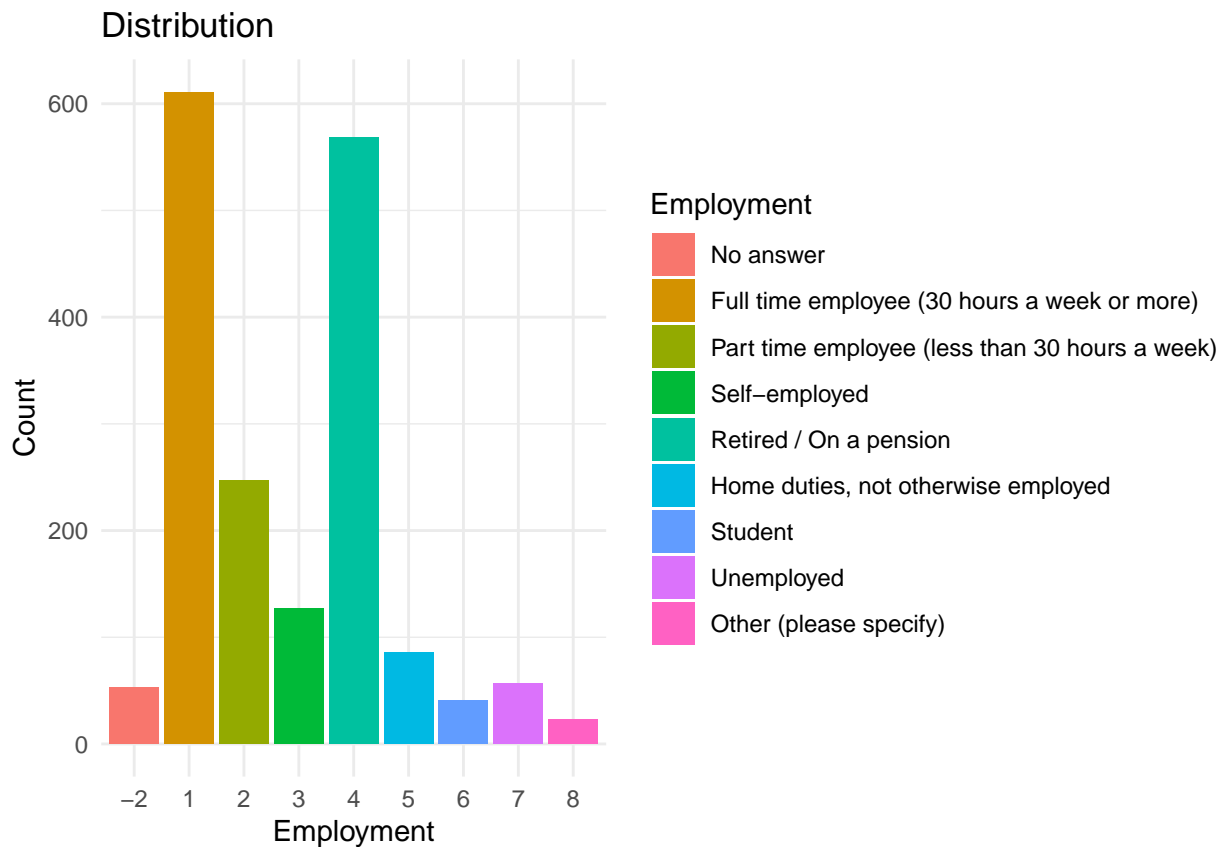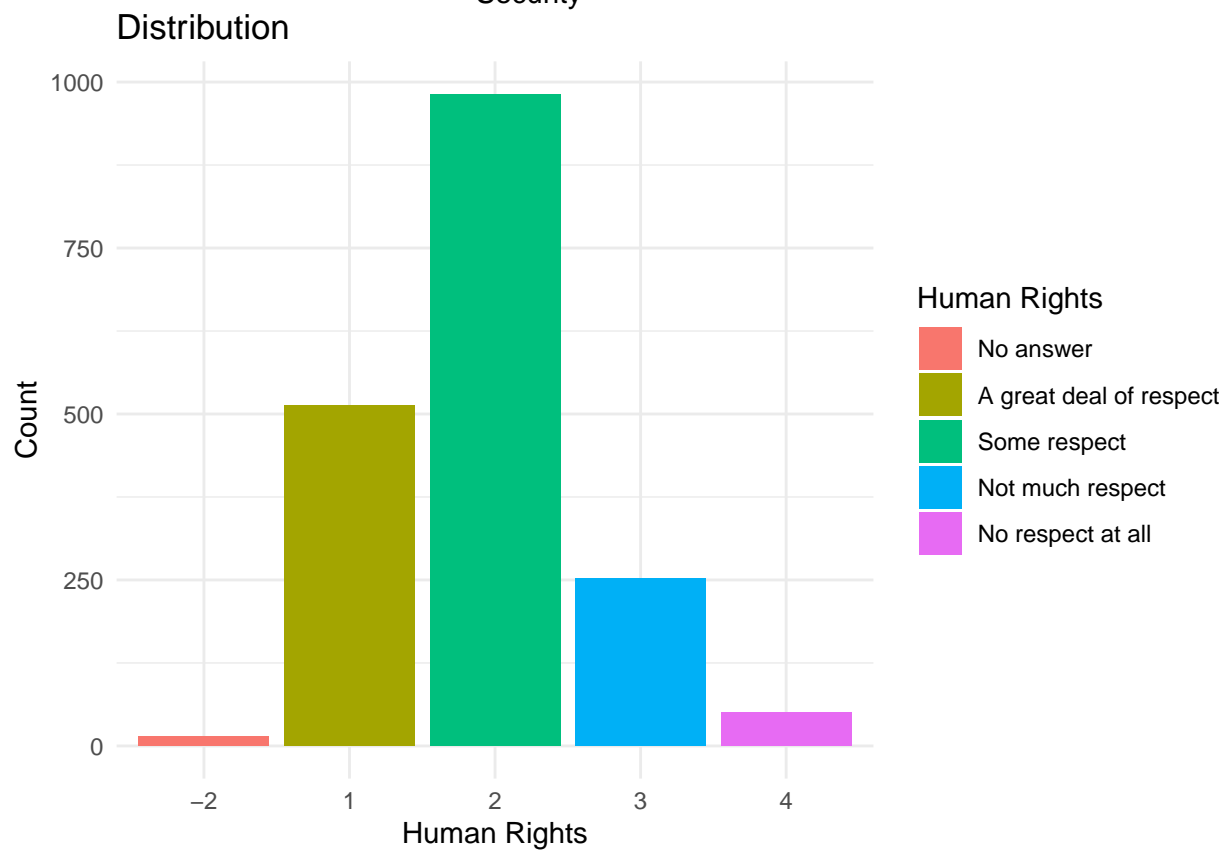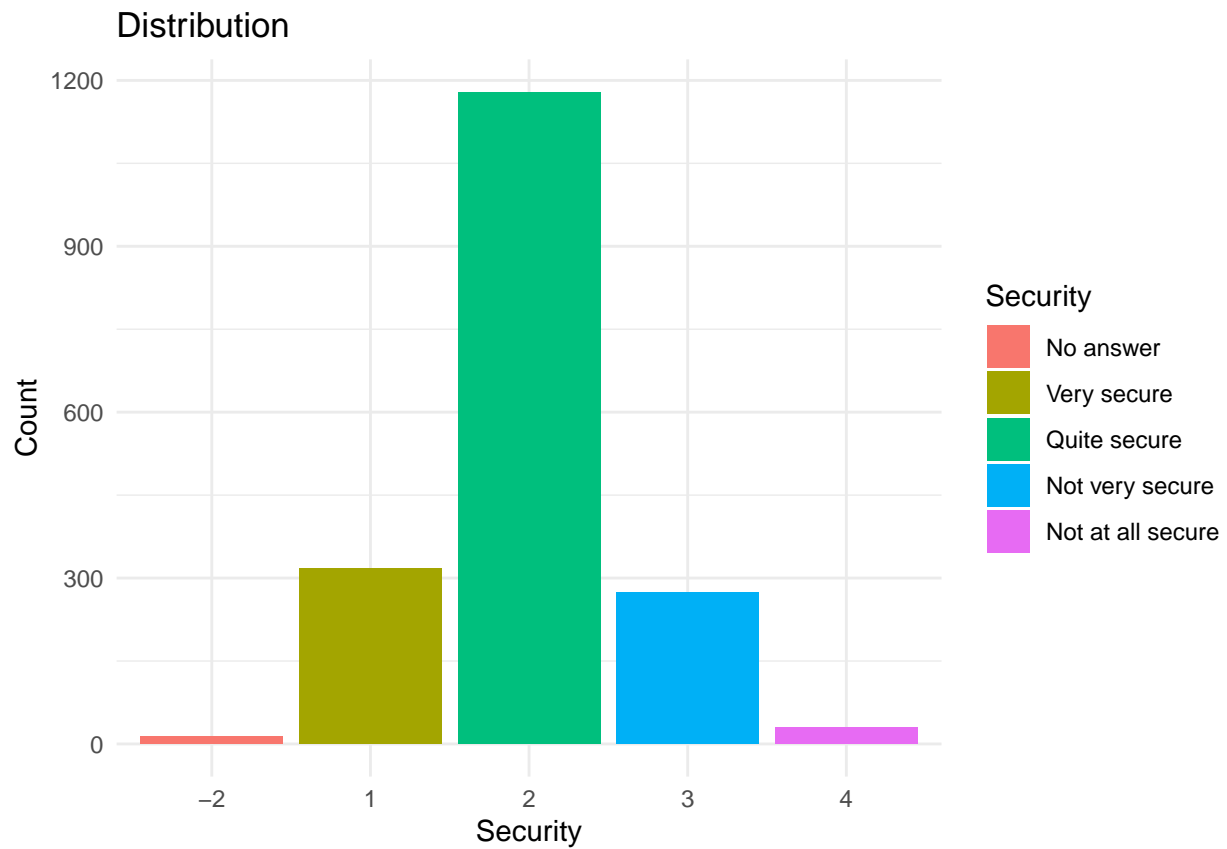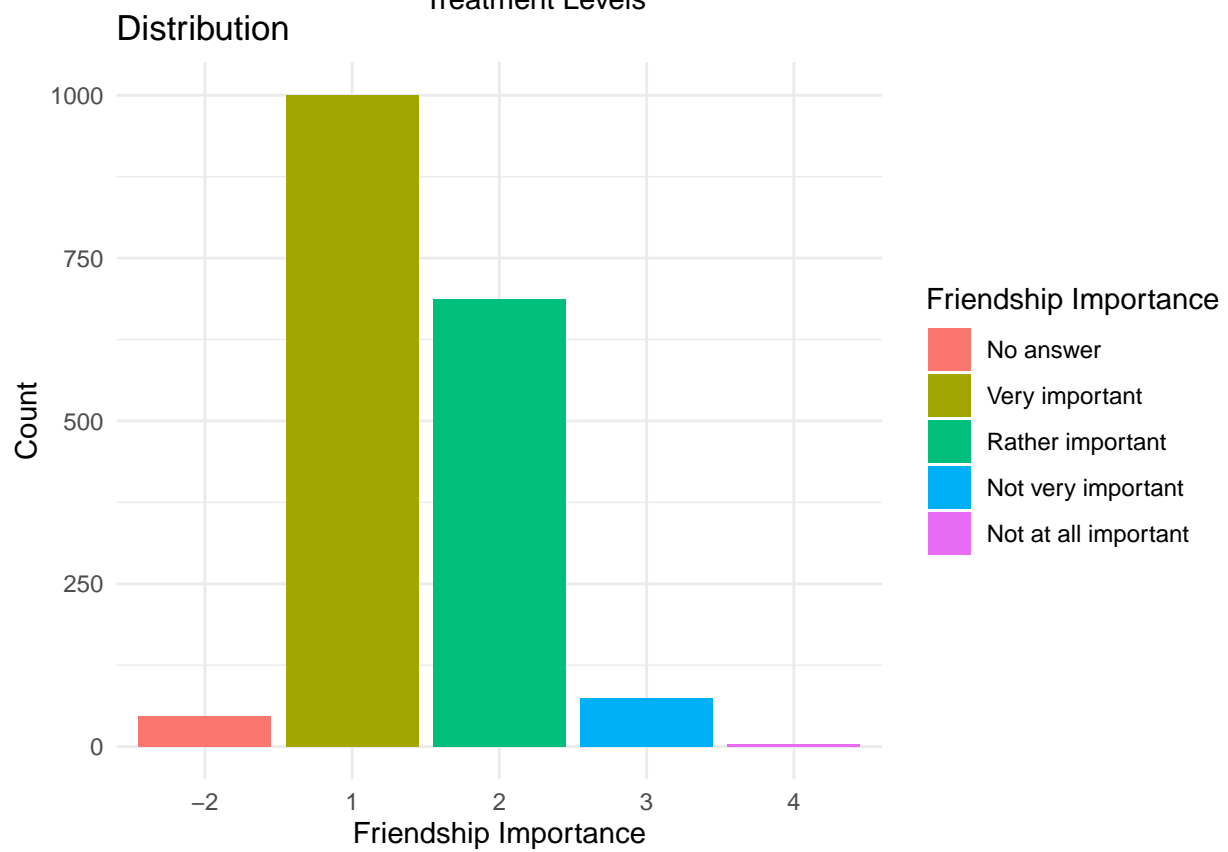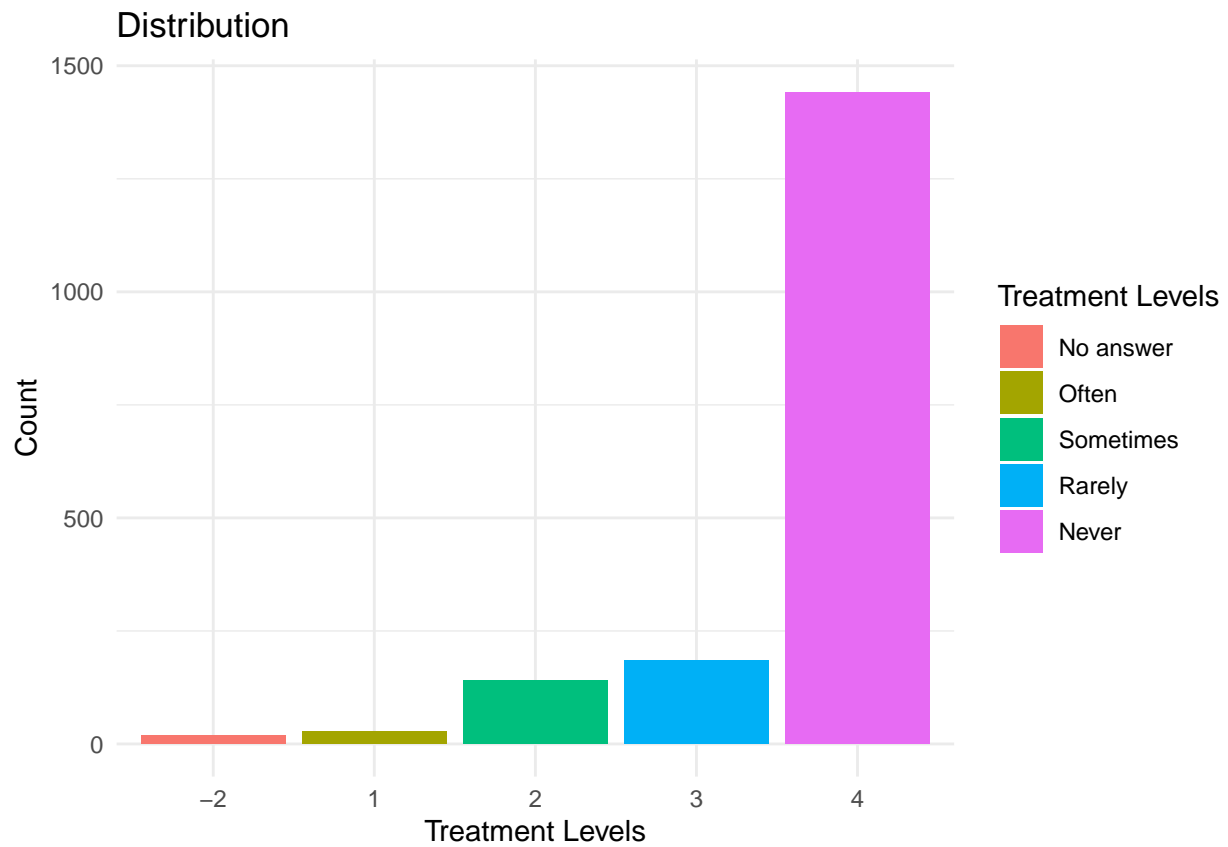
Employment

Descriptive Analysis:

- Full-time employment is most common, suggesting stable work schedules for many.
- Significant numbers of students and part-time employees, indicating a mix of work and study.
- Fewer respondents in self-employed, retired, and home duties categories, showing less representation of these demographics.
- Positively skewed distribution with a majority as full-time employees and diminishing counts towards 'Unemployed' and 'Other'.

Statistical Analysis:

- Mean employment score around 2.709, indicating a skew towards full-time employment.
- Median of 2, reinforcing the prevalence of full-time employment.

Education

Descriptive Analysis:

- Upper secondary education is most prevalent, indicating a common level of educational attainment.
- A considerable proportion of respondents with a bachelor's degree, possibly reflecting the target demographic or societal education trends.
- Lesser counts of post-secondary non-tertiary education, master's, or doctoral degrees.
- A distribution with a primary mode at upper secondary education and a secondary mode at the bachelor level, with tapering counts at the lowest and highest education levels.

Statistical Analysis:

- Mean education level approximately 5.36, showing a skew towards upper secondary education.
- Median value of 5, indicating over half of the respondents completed at least upper secondary education.

Security Perception

Descriptive Analysis:

- Majority of respondents feel quite secure, which might reflect societal stability or personal circumstances.
- Smaller proportions feel very secure or not secure, suggesting fewer extremes in security perception.
- Negatively skewed distribution where a large majority feels 'Quite secure', and fewer responses are on the 'Not very secure' or 'Not at all secure' end.

Statistical Analysis:

- Mean close to 2, leaning towards 'Quite secure'.
- Median of 2, confirming 'Quite secure' as a common sentiment.

Human Rights Perception

Descriptive Analysis:

- Most believe there is 'some respect' for human rights, indicating moderate views.
- Significant perception of 'a great deal of respect', suggesting a positive outlook among many.
- A distribution with a slight negative skew, indicating that most respondents feel there is 'Some respect' for human rights, with a substantial number also feeling there is 'A great deal of respect'.

Statistical Analysis:

- Mean of approximately 1.882, hinting the average perception is close to 'some respect'.
- Median of 2, aligning with the average perception towards human rights.

Health Treatment Levels

Descriptive Analysis:

- Predominant 'never' category might suggest good health or barriers to healthcare.
- 'Sometimes' as the next most common response, indicating occasional health concerns.
- A heavily positively skewed distribution, where most respondents 'Never' seek treatment, with progressively fewer responses for more frequent healthcare utilisation.

Statistical Analysis:

- Mean around 3.638, trending towards infrequently seeking treatment.
- Median of 4, suggesting the middle ground of responses leans towards 'rarely'.

Friendship Importance

Descriptive Analysis:

- Friendship considered 'very important' by many, emphasizing the high value on social relationships.
- 'A strongly negatively skewed distribution, showing that 'Very important' is the predominant response, with 'Not at all important' being the least common.

Statistical Analysis:

- Mean skewed towards 'very important', with a mean value of around 1.839.
- Median value of 1, indicating a majority view friendship as 'very important'.

Lab 2: Relationship of independent and dependent variables

Recode our variables. As we do this we provide the variables with more intuitive names, to make our work easier. This can be done with the code:

```r
survey.data <- survey.data %>%
  mutate(happiness = dplyr::recode(V10,
                                   '1' = 4,
                                   '2' = 3,
                                   '3' = 2,
                                   '4' = 1,
                                   '-2' = NULL),

         health = dplyr::recode(V11,
                                '1' = 5,
                                '2' = 4,
                                '3' = 3,
                                '4' = 2,
                                '5' = 1,
                                '-2' = NULL),

         finances = dplyr::recode(V59,
                                  '-2' = NULL,
                                  .default = V59),
```

```
            satisfaction = dplyr::recode(V23,
                                        '-2' = NULL,
                                        .default = V23),

            freedom = dplyr::recode(V55,
                                    '-2' = NULL,
                                    .default = V55))
```

Factor Analysis

```
fa.fit <- fa(survey.data[,c("happiness",
                            "health",
                            "finances",
                            "satisfaction",
                            "freedom")],
            nfactors=1)
```

```
## Factor Analysis using method =  minres
## Call: fa(r = survey.data[, c("happiness", "health", "finances", "satisfaction",
##      "freedom")], nfactors = 1)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                MR1   h2   u2 com
## happiness     0.64 0.41 0.59   1
## health        0.52 0.28 0.72   1
## finances      0.63 0.39 0.61   1
## satisfaction 0.93 0.86 0.14   1
## freedom       0.71 0.50 0.50   1
##
##                    MR1
## SS loadings       2.43
## Proportion Var 0.49
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## df null model =  10  with the objective function =  1.73 with Chi Square =  3136.16
## df of  the model are 5  and the objective function was  0.02
##
## The root mean square of the residuals (RMSR) is  0.02
## The df corrected root mean square of the residuals is  0.04
##
## The harmonic n.obs is  1788 with the empirical chi square  22.35  with prob <  0.00045
## The total n.obs was  1813  with Likelihood Chi Square =  33.28  with prob <  3.3e-06
##
## Tucker Lewis Index of factoring reliability =  0.982
## RMSEA index =  0.056  and the 90 % confidence intervals are  0.039 0.075
## BIC =  -4.23
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                     MR1
## Correlation of (regression) scores with factors   0.95
## Multiple R square of scores with factors          0.90
## Minimum correlation of possible factor scores     0.79
```

```
survey.data$life.quality <- as.numeric(fa.fit$scores)
ggplot(survey.data, aes(life.quality)) +
  geom_histogram(fill = 'black') +
  theme_minimal()
```
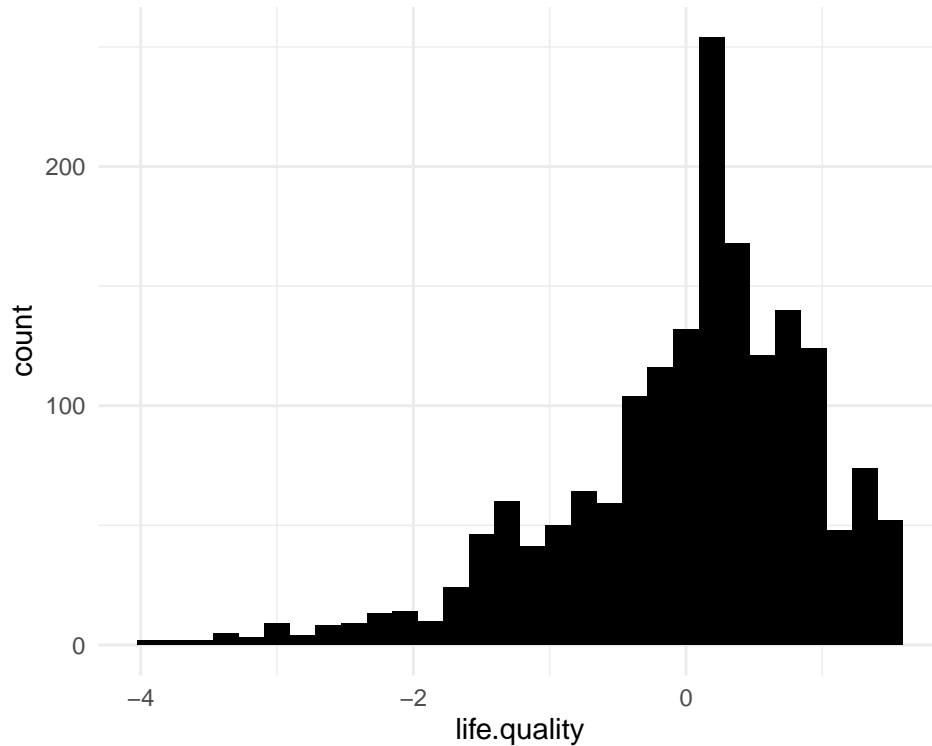

Figure 1: Distribution of quality of life measure produced by factor analysis.

```
survey.data$life.quality <- as.numeric(fa.fit$scores)

survey.predictors2 <- data.frame(
  education = survey.data$Q275,
  employment = survey.data$V249,
  security = survey.data$V170,
  rights = survey.data$V142,
  health = survey.data$V190,
  social = survey.data$V5,
  satisfaction = survey.data$V23,
  life.qual = survey.data$life.quality
)


survey.predictors2 = drop_na(survey.predictors2, life.qual)


regmodel = lm(life.qual ~ as.factor(education) + as.factor(employment) + as.factor(security) + as.facto

summary(regmodel)

##
## Call:
```

```
## lm(formula = life.qual ~ as.factor(education) + as.factor(employment) +
##     as.factor(security) + as.factor(rights) + as.factor(health) +
##     as.factor(social) + as.factor(satisfaction), data = survey.predictors2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8013 -0.1143  0.0041  0.1265  0.6927
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -3.595281   0.113828 -31.585  < 2e-16 ***
## as.factor(education)1      0.171477   0.088289   1.942   0.0523 .
## as.factor(education)2      0.006848   0.043854   0.156   0.8759
## as.factor(education)3      0.001018   0.031283   0.033   0.9740
## as.factor(education)4      0.045801   0.027591   1.660   0.0971 .
## as.factor(education)5      0.051866   0.027658   1.875   0.0609 .
## as.factor(education)7      0.070142   0.027698   2.532   0.0114 *
## as.factor(education)8      0.060915   0.029949   2.034   0.0421 *
## as.factor(education)9      0.060633   0.037634   1.611   0.1073
## as.factor(employment)1     0.025142   0.032101   0.783   0.4336
## as.factor(employment)2     0.016824   0.033676   0.500   0.6174
## as.factor(employment)3     0.054789   0.035662   1.536   0.1246
## as.factor(employment)4     0.014601   0.031504   0.463   0.6431
## as.factor(employment)5     0.044088   0.037022   1.191   0.2339
## as.factor(employment)6     0.049825   0.043066   1.157   0.2475
## as.factor(employment)7    -0.046496   0.040070  -1.160   0.2461
## as.factor(employment)8    -0.110666   0.049977  -2.214   0.0269 *
## as.factor(security)1       0.026989   0.062372   0.433   0.6653
## as.factor(security)2      -0.030717   0.061628  -0.498   0.6182
## as.factor(security)3      -0.065762   0.062446  -1.053   0.2924
## as.factor(security)4       0.014274   0.071921   0.198   0.8427
## as.factor(rights)1        -0.014886   0.059130  -0.252   0.8013
## as.factor(rights)2        -0.018424   0.058826  -0.313   0.7542
## as.factor(rights)3        -0.048204   0.059800  -0.806   0.4203
## as.factor(rights)4        -0.114156   0.064779  -1.762   0.0782 .
## as.factor(health)1        -0.004119   0.067029  -0.061   0.9510
## as.factor(health)2         0.042127   0.057445   0.733   0.4634
## as.factor(health)3         0.029992   0.056863   0.527   0.5980
## as.factor(health)4         0.121479   0.055290   2.197   0.0281 *
## as.factor(social)1         0.059242   0.031498   1.881   0.0602 .
## as.factor(social)2         0.038131   0.031890   1.196   0.2320
## as.factor(social)3         0.007649   0.038190   0.200   0.8413
## as.factor(social)4         0.086517   0.100827   0.858   0.3910
## as.factor(satisfaction)2   0.498978   0.077435   6.444 1.51e-10 ***
## as.factor(satisfaction)3   1.161063   0.068547  16.938  < 2e-16 ***
## as.factor(satisfaction)4   1.646046   0.066596  24.717  < 2e-16 ***
## as.factor(satisfaction)5   2.134779   0.062733  34.029  < 2e-16 ***
## as.factor(satisfaction)6   2.631124   0.062978  41.778  < 2e-16 ***
## as.factor(satisfaction)7   3.143959   0.062235  50.517  < 2e-16 ***
## as.factor(satisfaction)8   3.655695   0.061867  59.090  < 2e-16 ***
## as.factor(satisfaction)9   4.186623   0.062378  67.117  < 2e-16 ***
## as.factor(satisfaction)10  4.705666   0.062940  74.764  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Residual standard error: 0.1879 on 1716 degrees of freedom
## Multiple R-squared:  0.9612, Adjusted R-squared:  0.9602
## F-statistic:  1036 on 41 and 1716 DF,  p-value: < 2.2e-16

survey.predictors$V249 <- recode(survey.data$V249,
                          `-2` = "No answer",
                          `1` = "Full time employee (30 hours a week or more)",
                          `2` = "Part time employee (less than 30 hours a week)",
                          `3` = "Self-employed",
                          `4` = "Retired / On a pension",
                          `5` = "Home duties, not otherwise employed",
                          `6` = "Student",
                          `7` = "Unemployed",
                          `8` = "Other (please specify)")

survey.predictors$Q275 <- recode(survey.data$Q275,
                          `-2` = "No answer",
                          `1` = "No formal education",
                          `2` = "Primary education only",
                          `3` = "Lower secondary education (i.e. Year 9 or less)",
                          `4` = "Upper secondary education (i.e. between Year 10 and Year 12)",
                          `5` = "Post-secondary non-tertiary education (e.g. apprenticeship or certifi
                          `7` = "Bachelor or equivalent",
                          `8` = "Master or equivalent",
                          `9` = "Doctoral or equivalent")

survey.predictors$V170 <- recode(survey.data$V170,
                          `-2` = "No answer",
                          `1` = "Very secure",
                          `2` = "Quite secure",
                          `3` = "Not very secure",
                          `4` = "Not at all secure")

survey.predictors$V142 <- recode(survey.data$V142,
                          `-2` = "No answer",
                          `1` = "A great deal of respect",
                          `2` = "Some respect",
                          `3` = "Not much respect",
                          `4` = "No respect at all")

survey.predictors$V190 <- recode(survey.data$V190,
                          `-2` = "No answer",
                          `1` = "Often",
                          `2` = "Sometimes",
                          `3` = "Rarely",
                          `4` = "Never")

survey.predictors$V5 <- recode(survey.data$V5,
                          `-2` = "No answer",
                          `1` = "Very important",
                          `2` = "Rather important",
                          `3` = "Not very important",
                          `4` = "Not at all important")
```

```
# The survey_data data frame should now have the updated values as per the provided mappings.
```