Tutorial - R - Data Wrangling

Tidying / Cleaning / Scoring

Psychology Tutorial Series - Emily Towner

May 10, 2021

Contents

Setup

Install and load the necessary packages.

This will vary based on the tasks you will do in R for this script.

```r
# This installs each package (only need to do this once).
# install.packages('tidyverse')
# install.packages('psych')
# install.packages('knitr')

# This loads each package (need to do this every time you restart R).
 library(tidyverse)
 library(psych)
 library(knitr)
```

Data Import

Data should usually be in .csv format.

```r
# This reads in from your working directory - if using R-Markdown will be the root folder
data_raw <- read.csv("data_raw.csv")

# Save the working data separately for wrangling
data <- data_raw

# Create a new scored data frame with just the participant ID's for now
data_scored <- as.data.frame(data$participant)

# Rename the participant variable
colnames(data_scored) <- "participant"
```

Define functions

```r
# Create a function for checking if a subset of the data is more than seventy percent complete
more_than_seventy_percent_complete <- function(x) {
  ifelse(rowMeans(is.na(x)) < 0.3, TRUE, FALSE)
}

# Create a function for imputing means for 70% or more complete
mean_imputation <- function(data) {
  data_copy <- data
  data_means <- data_copy %>% rowMeans(na.rm=T)
  data_complete <- more_than_seventy_percent_complete(data_copy)
  for (rowNum in 1:nrow(data_copy)){
  if(data_complete[rowNum]){
  for (columnNum in 1:ncol(data_copy)){
        data_copy[rowNum, columnNum] = ifelse(is.na(data_copy[rowNum, columnNum]), data_means[rowNum], d
  return(data_copy)
}
```

Somatic

Somatic Symptoms Questionnaire (ss)

Scoring

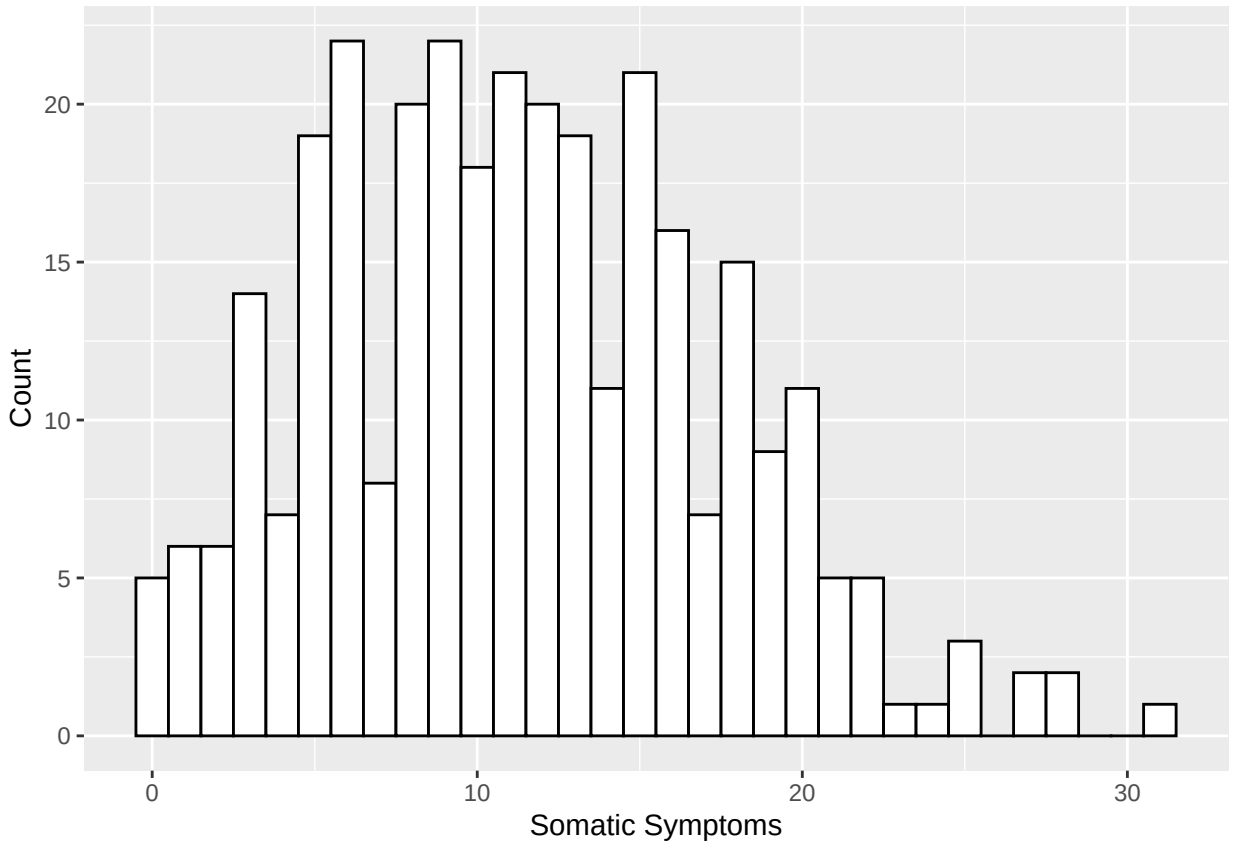
```r
# Subset the ss variables
ss <- data %>%
  dplyr::select(contains('somatic_symptoms'))

# Add the variables to create a sum variable
ss_score <- round(rowSums(ss), digits = 0)

# Bind the sum score with the other scored data
data_scored <- cbind(data_scored, ss = ss_score)
```

Descriptives

```r
# Summarize the ss score
descriptives <- as.data.frame(describe(data_scored$ss))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
descriptives <- round(descriptives, 2)
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'SS'
kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
SS	317	11.39	6.01	11	0	31	31	0.36	-0.2

```r
# View the histogram for this variable
ggplot(data_scored, aes(x=ss)) +
  geom_histogram(color = "black", fill = "white", binwidth = 1) +
  labs(x = "Somatic Symptoms", y = "Count")
```



Early Life Stress

Childhood Trauma Questionnaire (ctq)

Scoring

```
# Subset the ctq variables
ctq <- data %>%
  dplyr::select(contains('ctq'))

# Recode scale to 1-5
ctq <- ctq %>%
     mutate_all(funs(recode(., `0` = 1, `1` = 2, `2` = 3, `3` = 4, `4` = 5)))

# Reverse score items
ctq$ctq_2 <- recode(ctq$ctq_2, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_5 <- recode(ctq$ctq_5, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_7 <- recode(ctq$ctq_7, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_13 <- recode(ctq$ctq_13, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_19 <- recode(ctq$ctq_19, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_26 <- recode(ctq$ctq_26, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_28 <- recode(ctq$ctq_28, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)

# Reverse score the minimization/denial subset
ctq$ctq_10 <- recode(ctq$ctq_10, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
ctq$ctq_16 <- recode(ctq$ctq_16, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)
```

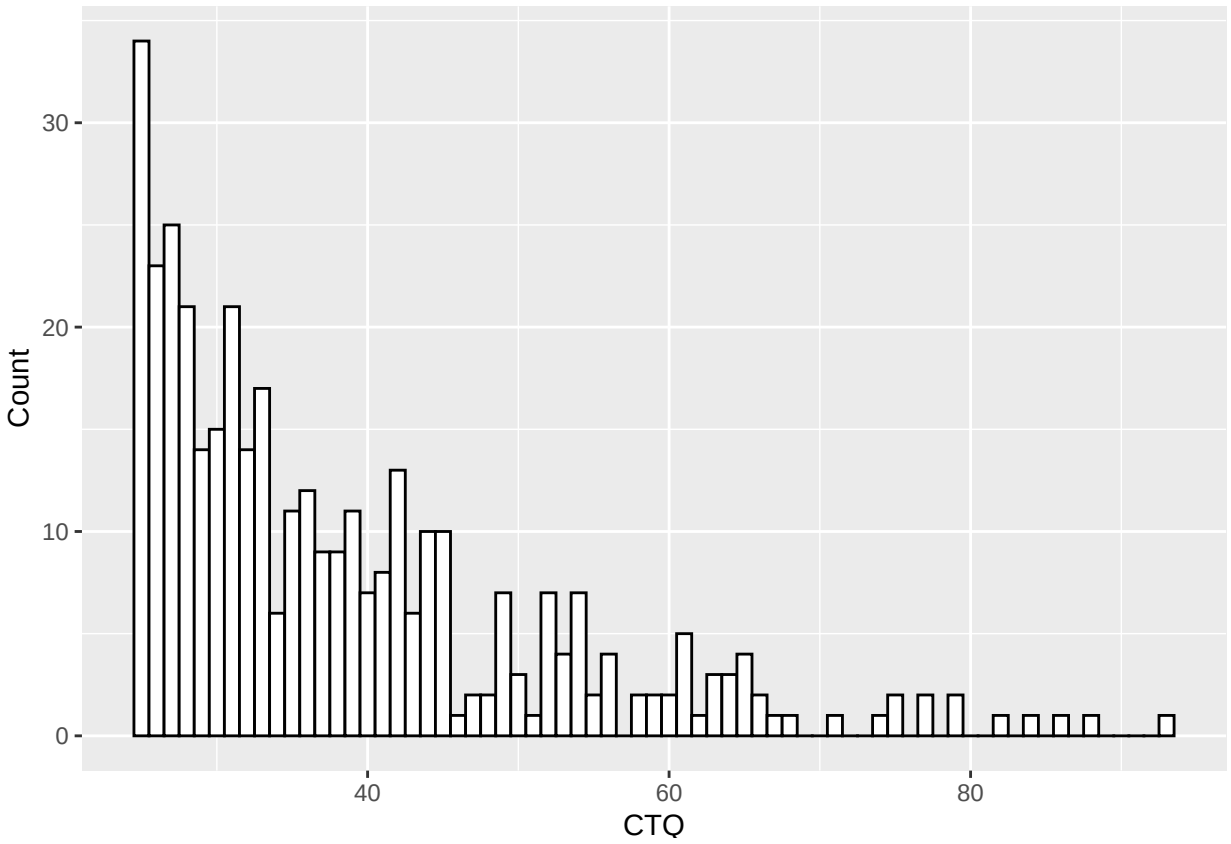
```r
ctq$ctq_22 <- recode(ctq$ctq_22, `1` = 5, `2` = 4, `3` = 3, `4` = 2, `5` = 1)

# Remove the minimization/denial subset from dataset
ctq <- ctq %>% dplyr::select(-ctq_10, -ctq_16, -ctq_22)

# Add the variables to create a sum variable
ctq_score <- round(rowSums(ctq), digits = 0)

# Bind the sum score with the other scored data
data_scored <- cbind(data_scored, ctq = ctq_score)
```

Descriptives

```r
# Summarize the ctq
descriptives <- as.data.frame(describe(data_scored$ctq))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
descriptives <- round(descriptives, 2)
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'CTQ'
kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
CTQ	375	38.38	13.38	34	25	93	68	1.42	1.84

```r
# View the histogram for this variable
ggplot(data_scored, aes(x=ctq)) +
  geom_histogram(color = "black", fill = "white", binwidth = 1) +
  labs(x = "CTQ", y = "Count")
```



Adverse Childhood Experiences Survey (ace)

Scoring

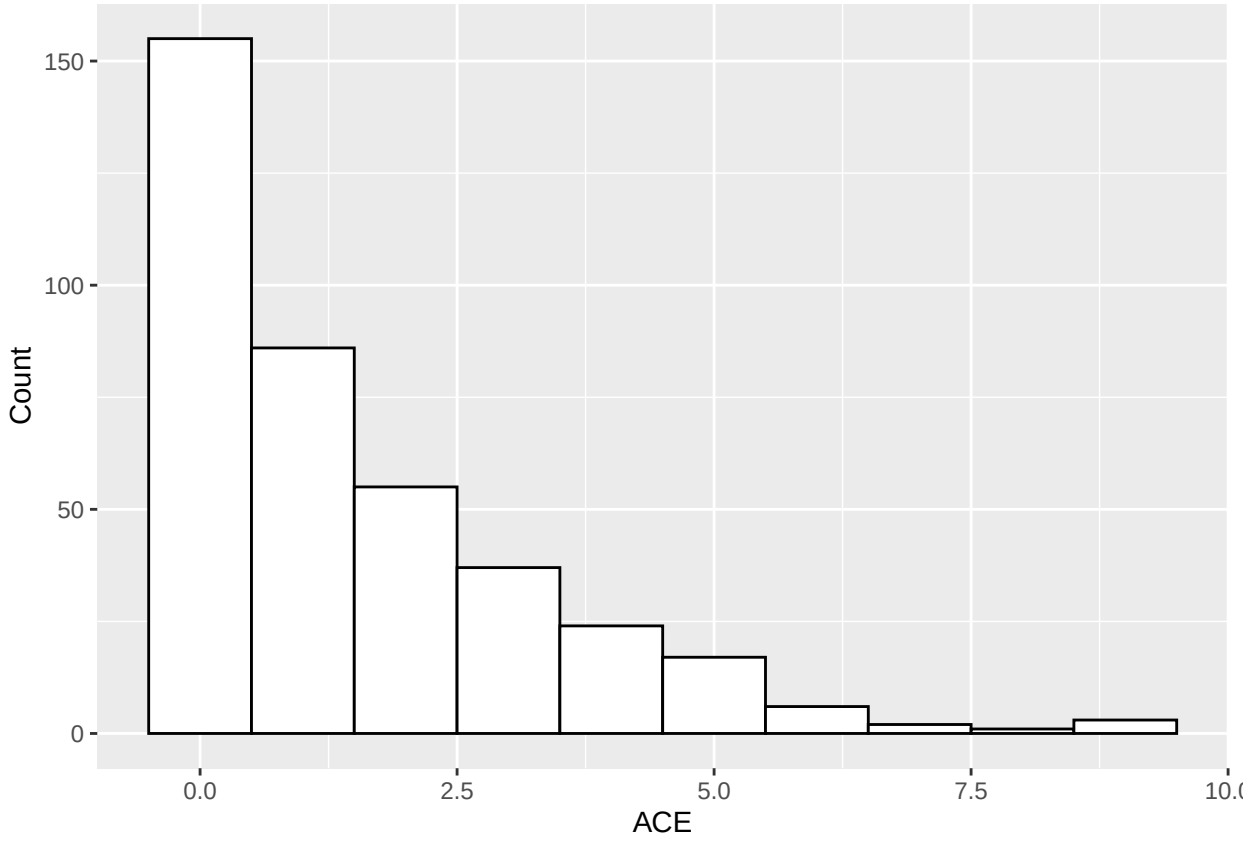
```
# Subset the ace variables
ace <- data %>%
  dplyr::select(contains('ace'))

# Add the variables to create a sum variable
ace_score <- rowSums(ace, na.rm = F)

# Bind the sum score with the other scored data
data_scored <- cbind(data_scored, ace = ace_score)
```

Descriptives

```
# Summarize the ace
descriptives <- as.data.frame(describe(data_scored$ace))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
descriptives <- round(descriptives, 2)
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'CTQ'
kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
CTQ	386	1.48	1.78	1	0	9	9	1.46	2.23

```r
# View the histogram for this variable
ggplot(data_scored, aes(x=ace)) +
  geom_histogram(color = "black", fill = "white", binwidth = 1) +
  labs(x = "ACE", y = "Count")
```



Mental Health

Beck Depression Inventory II (bdi_ii)

Scoring

```r
# Subset the bdi variables
bdi <- data %>%
  dplyr::select(contains('bdi'))

# Fix the incorrectly coded responses for the question 2
bdi$bdi_ii_2 <- recode(bdi$bdi_ii_2, `1` = 1, `2` = 2, `3` = 3, `5` = 4)

# Recode sleeping changes question
bdi$bdi_ii_16 <- recode(bdi$bdi_ii_16, `1` = 1, `2` = 2, `3` = 2, `4` = 3, `5` = 3, `6` = 4, `7` = 4)

# Recode appetite changes question
bdi$bdi_ii_18 <- recode(bdi$bdi_ii_18, `1` = 1, `2` = 2, `3` = 2, `4` = 3, `5` = 3, `6` = 4, `7` = 4)

# Recode variables with correct numeration
bdi <- bdi %>%
```

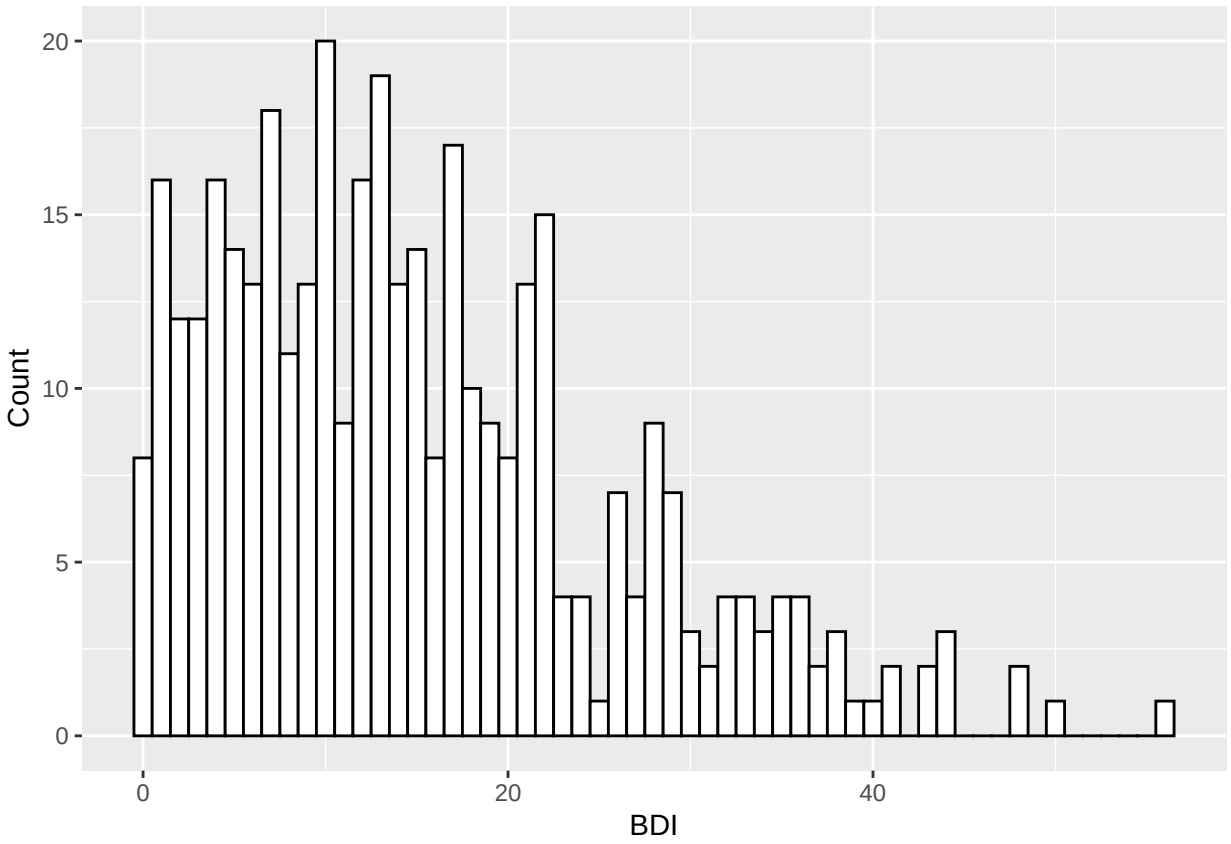
```
    mutate_all(funs(recode(., `1` = 0, `2` = 1, `3` = 2, `4` = 3)))

# Mean impute
bdi<- mean_imputation(bdi)

# Create a vector of means
bdi_means <- bdi %>% rowMeans(na.rm=F)

# Add a variable for bdi_ii_1 (suicide question which was omitted)
bdi$bdi_ii_1 <- bdi_means

# Add the variables to create a sum variable
bdi_score <- round(rowSums(bdi, na.rm = F), digits = 0)

# Bind the sum score with the other scored data
data_scored <- cbind(data_scored, bdi = bdi_score)
```

Descriptives

```
# Summarize
descriptives <- as.data.frame(describe(data_scored$bdi))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
descriptives <- round(descriptives, 2)
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'BDI'
kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
BDI	382	15.2	10.89	13	0	56	56	0.88	0.45

```
# View the histogram for this variable
ggplot(data_scored, aes(x=bdi)) +
  geom_histogram(color = "black", fill = "white", binwidth = 1) +
  labs(x = "BDI", y = "Count")
```



State-Trait Anxiety Inventory (stai)

Scoring

```
# Subset the variables
stai <- data %>%
  dplyr::select(contains('stai'))

# Recode scale to 1-4
stai <- stai %>%
     mutate_all(funs(recode(., `0` = 1, `1` = 2, `2` = 3, `3` = 4)))

# Reverse score items
stai$stai_1 <- recode(stai$stai_1, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_2 <- recode(stai$stai_2, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_5 <- recode(stai$stai_5, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_8 <- recode(stai$stai_8, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_10 <- recode(stai$stai_10, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_11 <- recode(stai$stai_11, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_15 <- recode(stai$stai_15, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_16 <- recode(stai$stai_16, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_19 <- recode(stai$stai_19, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_20 <- recode(stai$stai_20, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_21 <- recode(stai$stai_21, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_23 <- recode(stai$stai_23, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_26 <- recode(stai$stai_26, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_27 <- recode(stai$stai_27, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
```

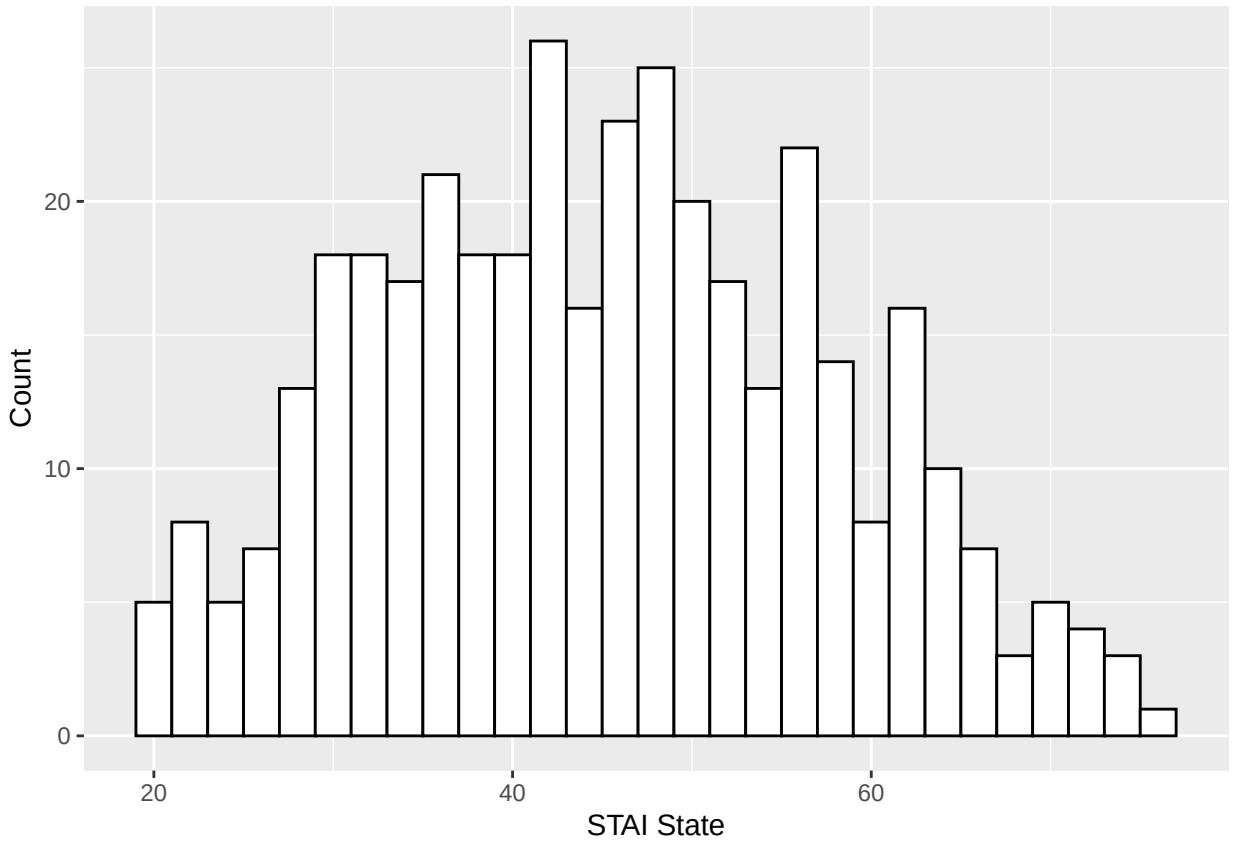
```r
stai$stai_30 <- recode(stai$stai_30, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_33 <- recode(stai$stai_33, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_34 <- recode(stai$stai_34, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_36 <- recode(stai$stai_36, `1` = 4, `2` = 3, `3` = 2, `4` = 1)
stai$stai_39 <- recode(stai$stai_39, `1` = 4, `2` = 3, `3` = 2, `4` = 1)

# Subset the first 20 columns as state
stai_state <- stai[,1:20]

# Subset the second 20 columns as trait
stai_trait <- stai[,21:40]

# Mean impute
stai_state <- mean_imputation(stai_state)
stai_trait <- mean_imputation(stai_trait)

# Add the variables to create a sum variable
stai_state_score <- round(rowSums(stai_state), digits = 0)
stai_trait_score <- round(rowSums(stai_trait), digits = 0)

# Bind the sum score with the other scored data
data_scored <- cbind(data_scored, stai_state = stai_state_score, stai_trait = stai_trait_score)
```

Descriptives

```r
# Summarize state
descriptives <- as.data.frame(describe(data_scored$stai_state))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
descriptives <- round(descriptives, 2)
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'STAI State'
kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
STAI State	381	45.62	12.56	46	20	77	57	0.14	-0.66

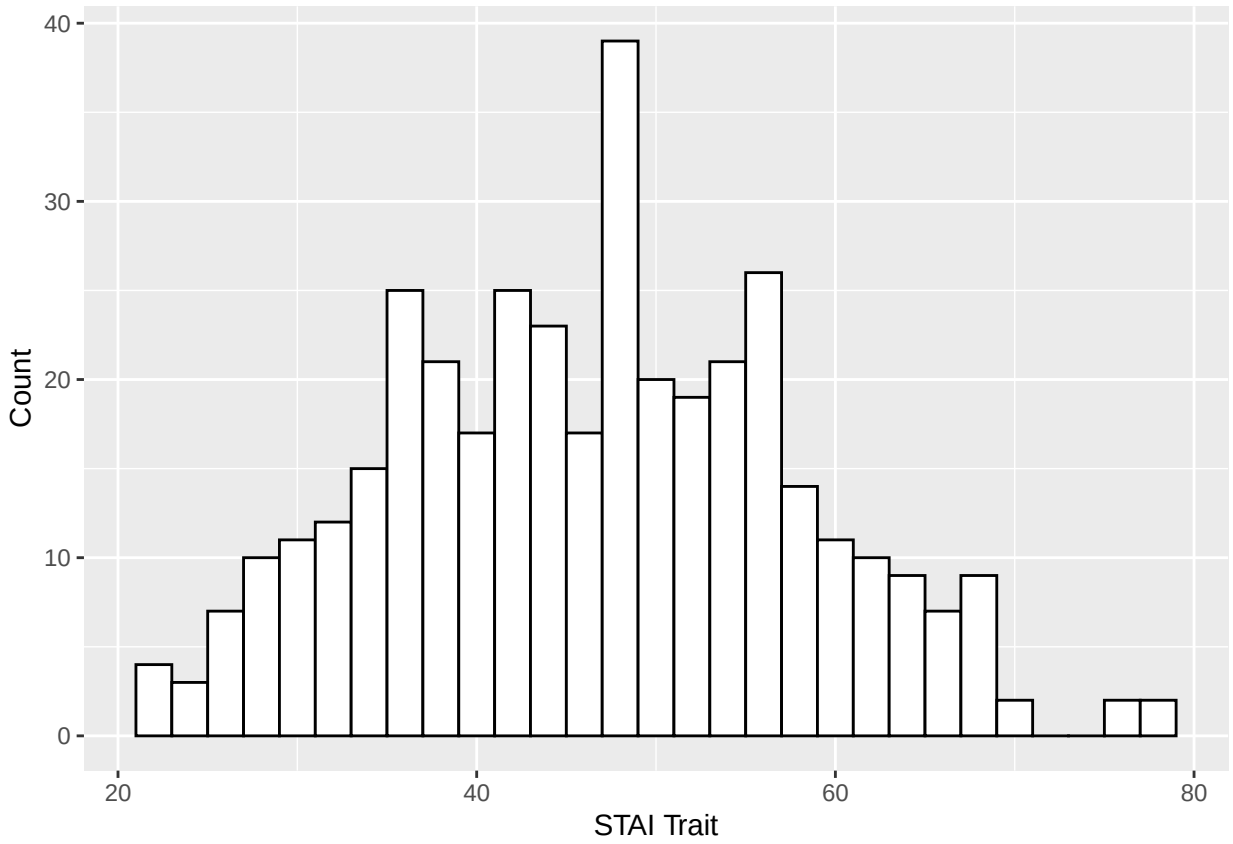
```r
# View the histogram for this variable
ggplot(data_scored, aes(x=stai_state)) +
  geom_histogram(color = "black", fill = "white", binwidth = 2) +
  labs(x = "STAI State", y = "Count")
```




```
# Summarize trait
descriptives <- as.data.frame(describe(data_scored$stai_trait))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
descriptives <- round(descriptives, 2)
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'STAI Trait'
kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
STAI Trait	381	46.98	11.32	48	22	78	56	0.12	-0.48

```
# View the histogram for this variable
ggplot(data_scored, aes(x=stai_trait)) +
  geom_histogram(color = "black", fill = "white", binwidth = 2) +
  labs(x = "STAI Trait", y = "Count")
```



Demographics

Scoring

```
# Subset participant variables
participant <- data %>%
  dplyr::select(contains('participant')) %>%
  dplyr::select(-(participant))

# Factor and label variables
participant$participant_age <- as.numeric(as.character(participant$participant_age, digits = 2))

participant$participant_sex <- factor(participant$participant_sex,
                                levels = c(1, 2, 3),
                                labels = c("Male", "Female", "Other"))

participant$participant_sex_other <- as.character(participant$participant_sex_other)

participant$participant_gender <- factor(participant$participant_gender,
                                levels = c(1, 2, 3, 4),
                                labels = c("Man", "Woman", "Nonbinary", "Other"))

participant$participant_gender_other <- as.character(participant$participant_gender_other)

# Rename and label ethnicity
participant$participant_ethnicity_caucasian <- participant$participant_ethnicity___1
```

```
participant$participant_ethnicity[participant$participant_ethnicity_caucasian == 1 ] <- "Caucasian"

participant$participant_ethnicity_latinx <- participant$participant_ethnicity___2
participant$participant_ethnicity[participant$participant_ethnicity_latinx == 1 ] <- "Latinx/Hispanic"

participant$participant_ethnicity_middle_eastern <- participant$participant_ethnicity___3
participant$participant_ethnicity[participant$participant_ethnicity_middle_eastern == 1 ] <- "Middle Eas

participant$participant_ethnicity_african <- participant$participant_ethnicity___4
participant$participant_ethnicity[participant$participant_ethnicity_african == 1 ] <- "African"

participant$participant_ethnicity_carribean <- participant$participant_ethnicity___5
participant$participant_ethnicity[participant$participant_ethnicity_caribbean == 1 ] <- "Caribbean"

participant$participant_ethnicity_south_asian <- participant$participant_ethnicity___6
participant$participant_ethnicity[participant$participant_ethnicity_south_asian == 1 ] <- "South Asian"

participant$participant_ethnicity_east_asian <- participant$participant_ethnicity___7
participant$participant_ethnicity[participant$participant_ethnicity_east_asian == 1 ] <- "East Asian"

participant$participant_ethnicity_other <- participant$participant_ethnicity___8
participant$participant_ethnicity[participant$participant_ethnicity_other == 1 ] <- "Other"

# Factor multi ethnicities
participant$participant_ethnicity_multi <- {
  ifelse((participant$participant_ethnicity_caucasian + participant$participant_ethnicity_latinx + parti

# Select only the scored variables
participant <- participant %>% select(participant_age, participant_sex, participant_sex_other, participa

# Create a scored dataset with the participant variables
data_scored <- cbind(data_scored, participant)
```

Descriptives

```
# Summarize the numeric variables
descriptives <- participant %>% dplyr::select("Participant Age" = participant_age)
descriptives <- as.data.frame(describe(descriptives))
descriptives <- descriptives %>% dplyr::select("N" = n, "Mean" = mean, "SD" = sd, "Median" = median, "M
rownames(descriptives)[rownames(descriptives) == 'X1'] <- 'Participant Age'
descriptives <- round(descriptives, 2)

# Count the nominal variables
descriptives_sex <- participant %>%
  group_by(Sex = participant_sex) %>%
  count() %>%
  rename("N" = n)

descriptives_gender <- participant %>%
  group_by(Gender = participant_gender) %>%
  count() %>%
  rename("N" = n)
```

```
descriptives_ethnicity <- participant %>%
  group_by(Ethnicity = participant_ethnicity) %>%
  count() %>%
  rename("N" = n)

kable(descriptives)
```

	N	Mean	SD	Median	Min	Max	Range
Participant Age	397	20.34	2.27	20	17	34	17

```
kable(descriptives_sex)
```

Sex	N
Male	72
Female	325
Other	1
NA	78

```
kable(descriptives_gender)
```

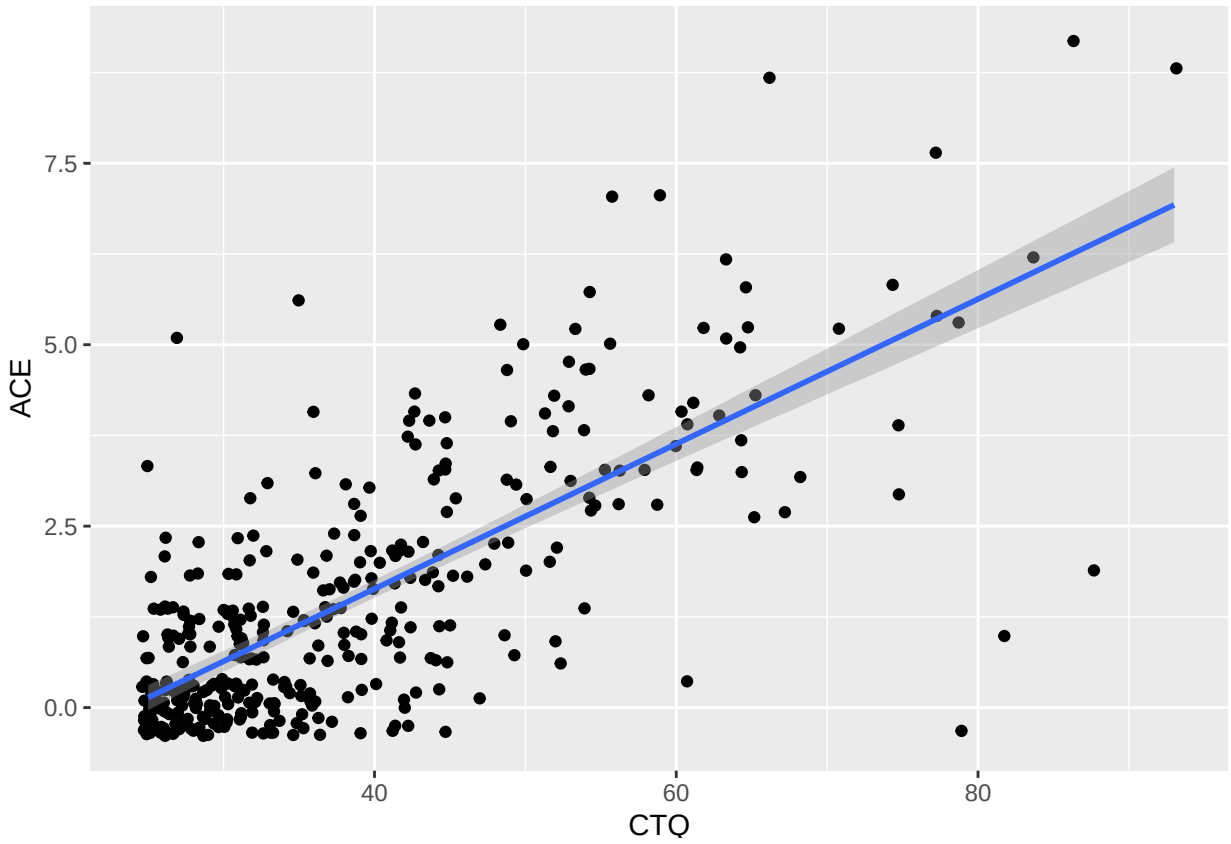
Gender	N
Man	69
Woman	323
Nonbinary	6
NA	78

```
kable(descriptives_ethnicity)
```

Ethnicity	N
African	10
Caucasian	86
East Asian	122
Latinx/Hispanic	67
Middle Eastern	23
Other	38
South Asian	33
NA	97

Sanity Checks

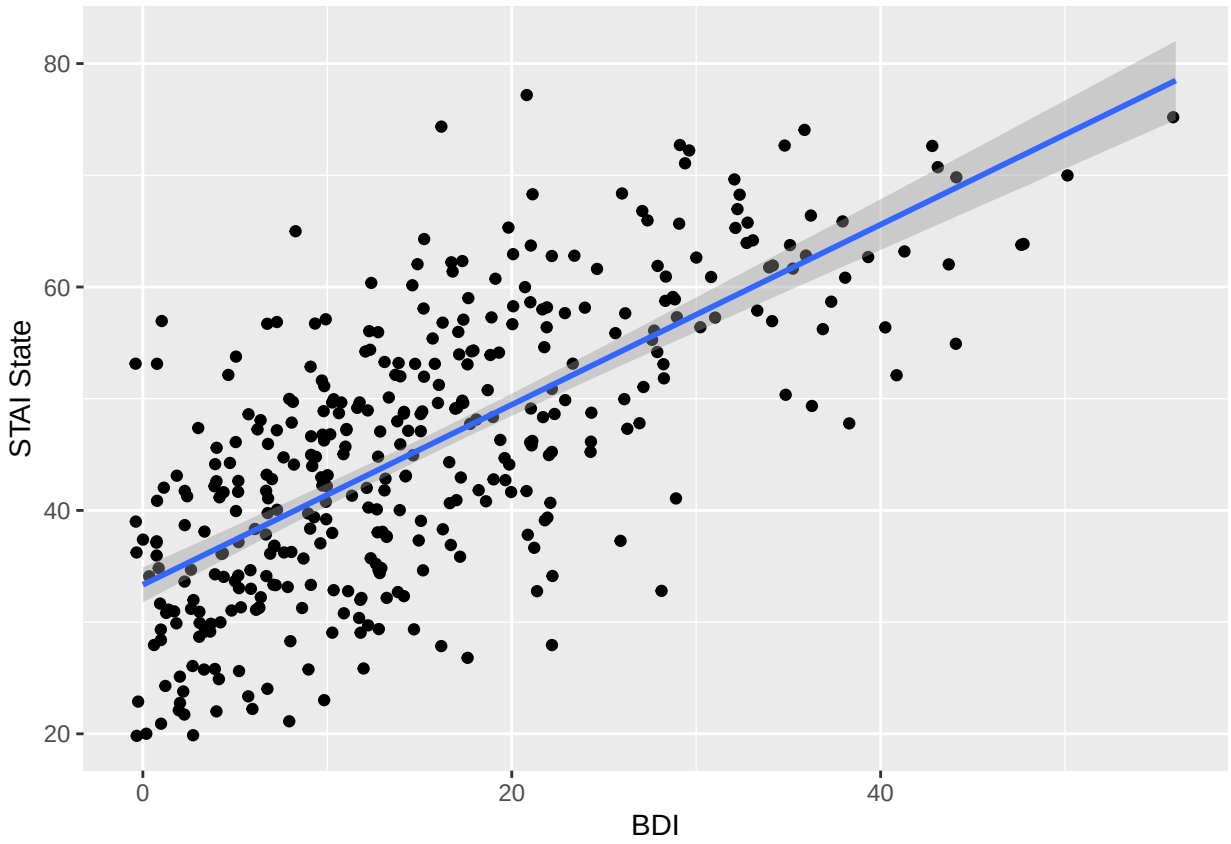
```
# View a scatterplot of the ctq and the ace - two different measures of early life stress
ggplot(data_scored, aes(x=ctq, y=ace)) +
  geom_point(position = "jitter") +
  labs(x = "CTQ", y = "ACE") +
  geom_smooth(method = lm)
```




```
# Check the correlation between the ctq and the ace
cor.test(data_scored$ctq, data_scored$ace)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_scored$ctq and data_scored$ace
## t = 21.437, df = 370, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6952354 0.7864594
## sample estimates:
##       cor
## 0.7442993
```

```
# View a scatterplot of the stai_state and bdi scores - a measure of depression and anxiety
ggplot(data_scored, aes(x=bdi, y=stai_state)) +
  geom_point(position = "jitter") +
  labs(x = "BDI", y = "STAI State") +
  geom_smooth(method = lm)
```




```
# Check the correlation between the ctq and the ace
cor.test(data_scored$bdi, data_scored$stai_state)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_scored$bdi and data_scored$stai_state
## t = 18.944, df = 376, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6433057 0.7470373
## sample estimates:
##       cor
## 0.6988275
```

Data Export

```
# Export the scored data to the root folder
write_csv(data_scored, "data_scored.csv")
```