Working With Data

Table of contents

##Library

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```r
library(descr)
library(knitr)
library(dplyr)
library(haven)
library(ggplot2)
library(Hmisc)
```

```
Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

```r
library(readr)
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

##Data Set Load

```r
file_path <- file.path("..", "data", "36168-0001-Data.sav")
df <- read_sav(file_path)
```

##DATA TABLE

```r
head(df)
```

```
# A tibble: 6 x 59
     ID   AGE SEX        MAR       RACE      SES DX        DX2       DX3     DX4
  <dbl> <dbl> <dbl+lbl>  <dbl+lbl> <dbl+l> <dbl> <chr+lbl> <chr+lbl> <chr> <dbl>
1     4    32 1 [Female] 2 [Marri~ 0 [Whi~    25 ETD [Eat~ <NA>      <NA>     NA
2     6    44 2 [Male]   1 [Divor~ 1 [Bla~    47 SUB [Sub~ ANX [Anx~ <NA>     NA
3    11    38 2 [Male]   4 [Separ~ 0 [Whi~    40 BIP [Bip~ <NA>      <NA>     NA
4    12    45 1 [Female] 2 [Marri~ 0 [Whi~    29 BIP [Bip~ OTH [Oth~ <NA>     NA
5    14    44 1 [Female] 2 [Marri~ 0 [Whi~    36 BIP [Bip~ <NA>      <NA>     NA
6    16    45 1 [Female] 2 [Marri~ 0 [Whi~    11 <NA>      <NA>      <NA>     NA
# i 49 more variables: ANXDX <dbl+lbl>, DISDX <dbl+lbl>, EATDX <dbl+lbl>,
#   MOODDX <dbl+lbl>, PSYCDX <dbl+lbl>, SUBDX <dbl+lbl>, BPDDX <dbl+lbl>,
#   OTHERDX <dbl+lbl>, UNIT <dbl+lbl>, CARE_DYS <dbl>, PPANAM <dbl+lbl>,
#   SPANAM <dbl+lbl>, PDUR <dbl>, SDUR <dbl>, NEGLECT <dbl>, SSC <dbl>,
#   SSAB <dbl+lbl>, PHYS <dbl>, SEXAB <dbl>, PAGE <dbl>, SAGE <dbl>,
#   ABUSE <dbl+lbl>, AGGR <dbl>, DES <dbl>, POSAFF1 <dbl+lbl>,
#   POSAFF2 <dbl+lbl>, PASUM <dbl>, SCL_ANX <dbl>, SCL_OBS <dbl>, ...
```

##FREQ

```r
freq(as.ordered(df$sex), plot = FALSE)
```

```
Warning: Unknown or uninitialised column: `sex`.
```

```
as.ordered(df$sex)
      Frequency Percent Cum Percent
Total         0       0
```

This table shows that more female participants went through abuse more than that of the men amount of men whpo participated.

```r
freq(as.ordered(df$mooddx), plot = FALSE)
```

```
Warning: Unknown or uninitialised column: `mooddx`.
```

```
as.ordered(df$mooddx)
      Frequency Percent Cum Percent
Total         0       0
```

This table shows the amount of people with mood disorders and there are more people with than without.

```
freq(as.ordered(df$race), plot = FALSE)
```

```
Warning: Unknown or uninitialised column: `race`.
```

```
as.ordered(df$race)
      Frequency Percent Cum Percent
Total         0       0
```

This table shows the amount of white, black and other races that went through abuse. White participants were shown to be the ones to go through abuse.

##Data Management

```
#df$race[df$race == 1 | df$race == 2 | df$race == 3 | df$race == 5 | df$race == 6] <- 1
#df$race[df$race == 4 | df$race == 7 | df$race == 8 | df$race == 9]
```

```
df$RACE[is.na(df$RACE)] <- 0
```

```
df <- janitor::clean_names(df)
```

```
names (df)
```

```
 [1] "id"              "age"             "sex"             "mar"
 [5] "race"            "ses"             "dx"              "dx2"
 [9] "dx3"             "dx4"             "anxdx"           "disdx"
[13] "eatdx"           "mooddx"          "psycdx"          "subdx"
[17] "bpddx"           "otherdx"         "unit"            "care_dys"
[21] "ppanam"          "spanam"          "pdur"            "sdur"
[25] "neglect"         "ssc"             "ssab"            "phys"
[29] "sexab"           "page"            "sage"            "abuse"
[33] "aggr"            "des"             "posaff1"         "posaff2"
[37] "pasum"           "scl_anx"         "scl_obs"         "scl_dep"
[41] "scl_hostility"   "scl_int"         "scl_par"         "scl_pho"
[45] "scl_psy"         "scl_som"         "scl_add"         "scl_gsi"
[49] "traums"          "pastpt"          "pres_pt"         "tr_time"
[53] "sisdb_tot"       "sisdb_sub"       "sisdb_eat"       "sisdb_seximp"
[57] "sisdb_sharm"     "sisdb_suic"      "sptss"
```

```r
#df <- df %>%
  #age_group = case_when()
      #age < 30 ~ "Under 30",
      #age >= 30 & age < 60 ~ "30-59",
      #age >= 60 ~ "60+"
    #sex = recode(as.character(sex
                 "1" = "Male"
                 "2" = "Female"
```

```r
#df_summary <- df %>%
  #group_by(sex, age_group) %>%
  #summarise(
    #mean_depression = mean(scl_dep, na.rm = TRUE),
    #count = n()
```

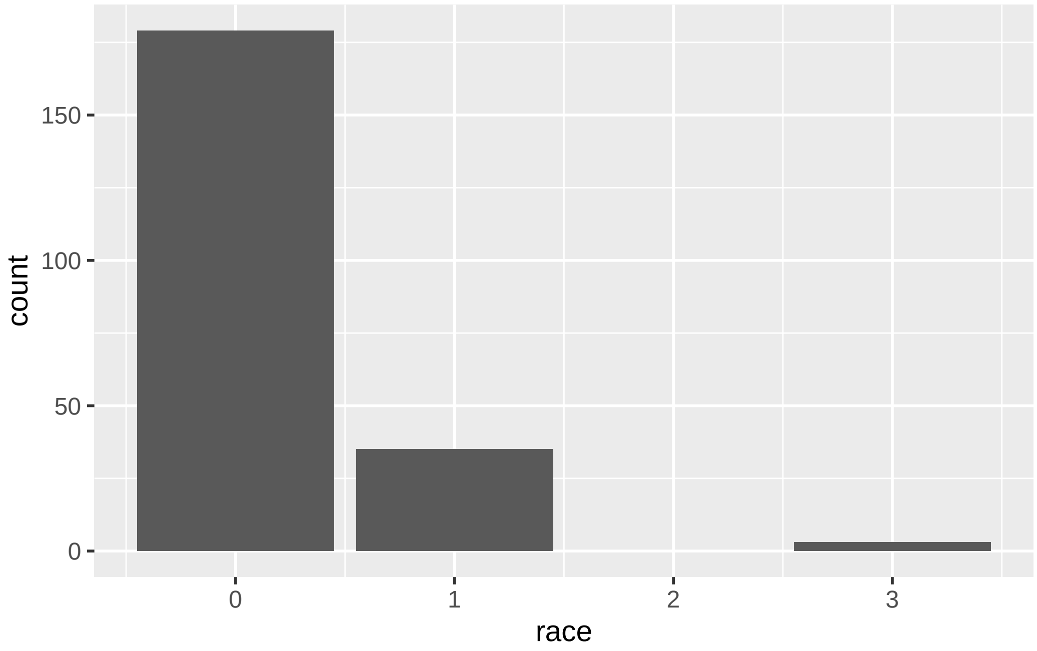
```r
#df_summary <- df %>%
  #group_by(sex, age_group) %>%
  #summarise(
    #mean_depression = mean(scl_dep, na.rm = TRUE),
    #count = n(),
    #.groups = "drop"
```

```r
#ggplot(df_summary, aes(x = age_group, y = mean_depression, fill = sex)) +
  #geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Mean Depression Scores by Age Group and Sex",
    x = "Age Group",
    y = "Mean Depression Score"
  ) +
  theme_minimal()
```

```
NULL
```

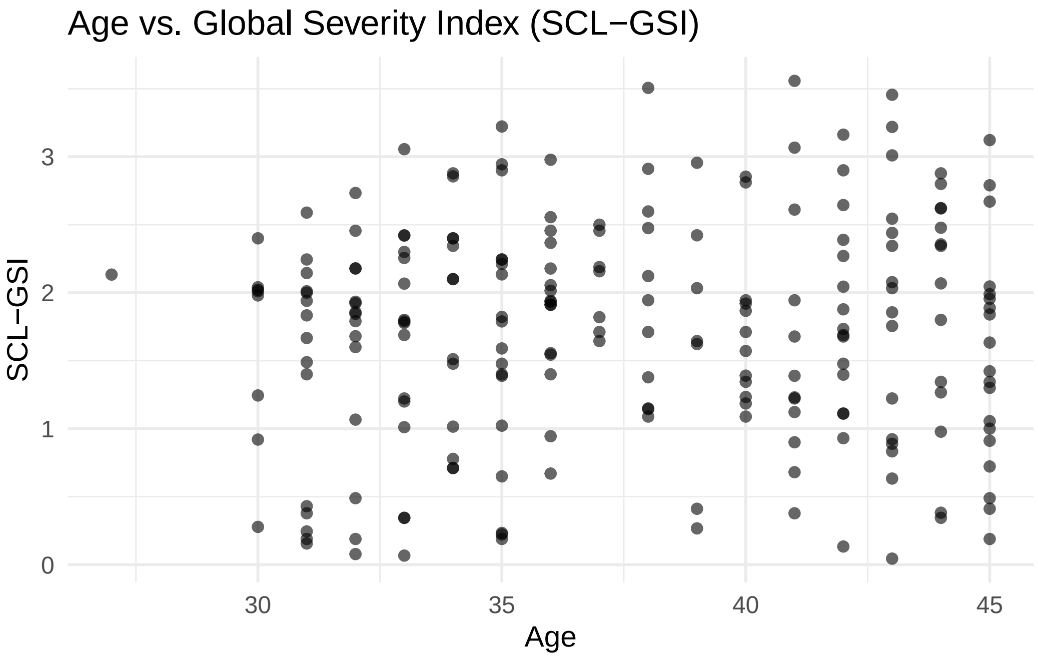
##Univariate First Graph representing race

```r
ggplot(df, aes(x = race)) + geom_bar()
```



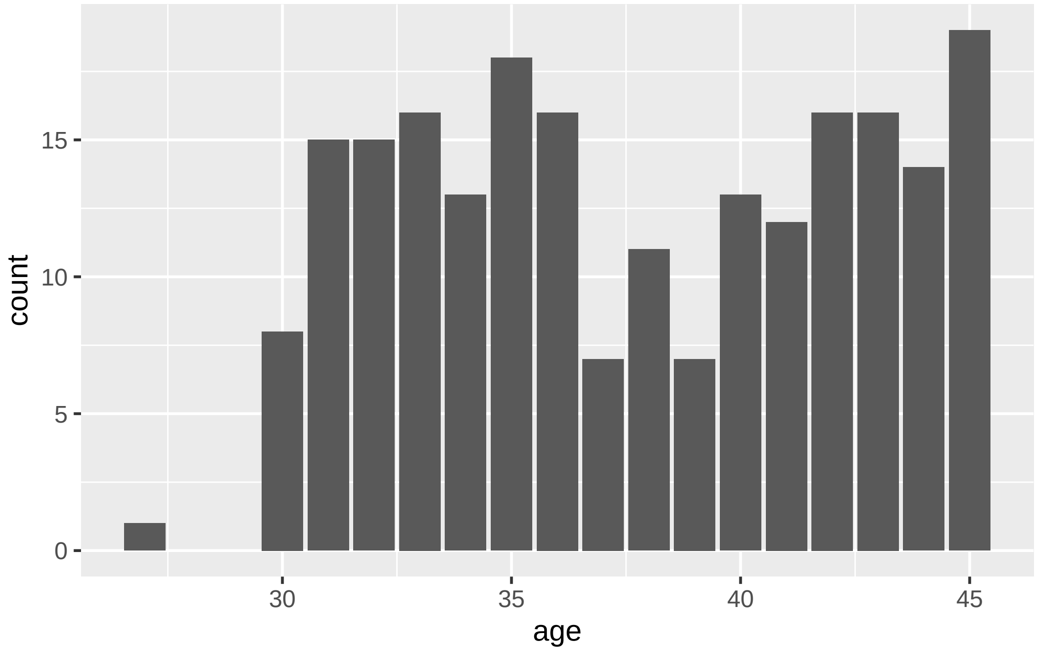
```
ggplot(df, aes(x = age, y = scl_gsi)) +
  geom_point(alpha = 0.6) +
  labs(title = "Age vs. Global Severity Index (SCL-GSI)", x = "Age", y = "SCL-GSI") +
  theme_minimal()
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



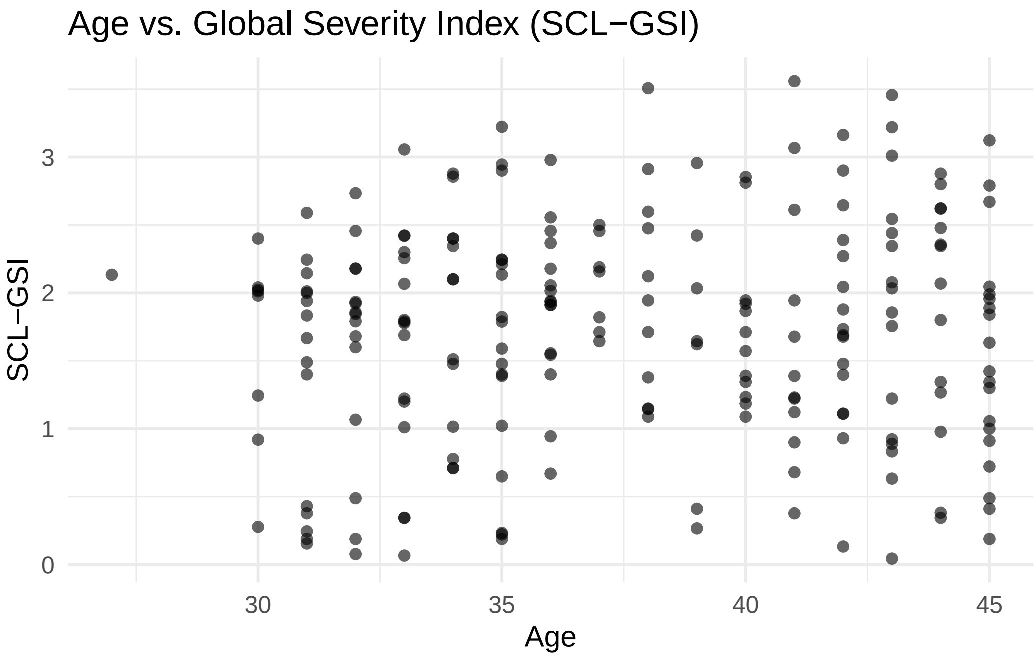
Second graph representing age

```
ggplot(df, aes(x = age)) + geom_bar()
```



```
ggplot(df, aes(x = age, y = scl_gsi)) +
  geom_point(alpha = 0.6) +
  labs(title = "Age vs. Global Severity Index (SCL-GSI)", x = "Age", y = "SCL-GSI") +
  theme_minimal()
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



1 Chi-Squared

```
table(df$sex)
```

```

  1   2 
125  92 
```

```
table(df$dx)
```

```
AFF ANX BIP DEP DIS ETD MPD OTH PTS SCZ SUB
 13   1  27  55   7   4  41  10   6  13  18
```

```r
table(df$sex, df$dx)
```

```
    AFF ANX BIP DEP DIS ETD MPD OTH PTS SCZ SUB
  1   8   1  12  29   6   4  34   2   6   1   6
  2   5   0  15  26   1   0   7   8   0  12  12
```

```r
chisq.test(table(df$sex, df$dx))
```

```
Warning in chisq.test(table(df$sex, df$dx)): Chi-squared approximation may be
incorrect
```

```

    Pearson's Chi-squared test

data:  table(df$sex, df$dx)
X-squared = 46.381, df = 10, p-value = 1.223e-06
```

```r
table(df$sex, df$abuse)
```

```

     0  1  2  3
  1 15 22 15 71
  2 28 42  4 16
```

```r
chisq.test(table(df$sex, df$abuse))
```

```

    Pearson's Chi-squared test

data:  table(df$sex, df$abuse)
X-squared = 47.342, df = 3, p-value = 2.939e-10
```

```
df <- df %>%
  mutate(abuse = as.factor(abuse))
```

##T-Test

```
t.test(age ~ sex, data = df)
```

```

	Welch Two Sample t-test

data:  age by sex
t = 0.012603, df = 195.4, p-value = 0.99
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to
95 percent confidence interval:
 -1.297984  1.314680
sample estimates:
mean in group 1 mean in group 2 
       37.70400        37.69565 
```

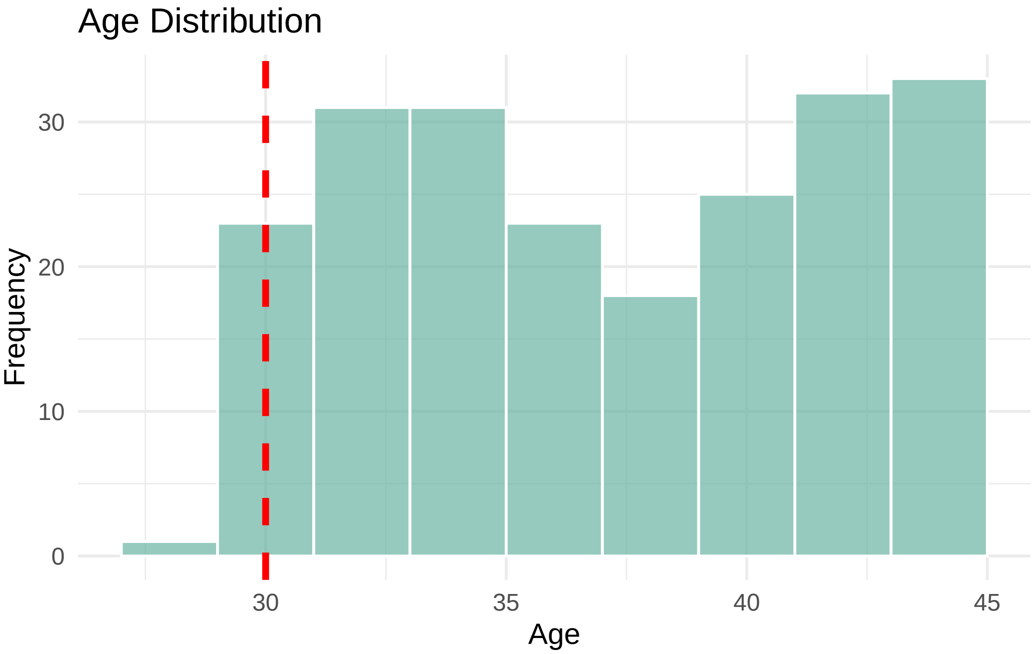
```
df <- df %>%
  mutate(
    sex = as.factor(sex),
    abuse = as.factor(abuse)
  )
```

```
t.test(df$age, mu = 30)
```

```

	One Sample t-test

data:  df$age
t = 23.606, df = 216, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 37.05752 38.34341
sample estimates:
mean of x 
 37.70046 
```

```
ggplot(df, aes(x = age)) +
  geom_histogram(binwidth = 2, fill = "#69b3a2", alpha = 0.7, color = "white") +
  geom_vline(xintercept = 30, linetype = "dashed", color = "red", size = 1.2) +
  labs(
    title = "Age Distribution",
    x = "Age",
    y = "Frequency"
  ) +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



People affected by abuse mostly have effects ages 30 and 40.

##ANOVA

```
set.seed(123)
```

```
aov_age_sex <- aov(age ~ sex, data = df)
```

```r
summary(aov_age_sex)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
sex           1      0   0.004       0   0.99
Residuals   215   4988  23.198
```

This shows that there is no significant correlation between sex and age of people affected by abuse.

##Correlation

```r
cor.test(df$age, df$scl_gsi, use = "complete.obs")
```

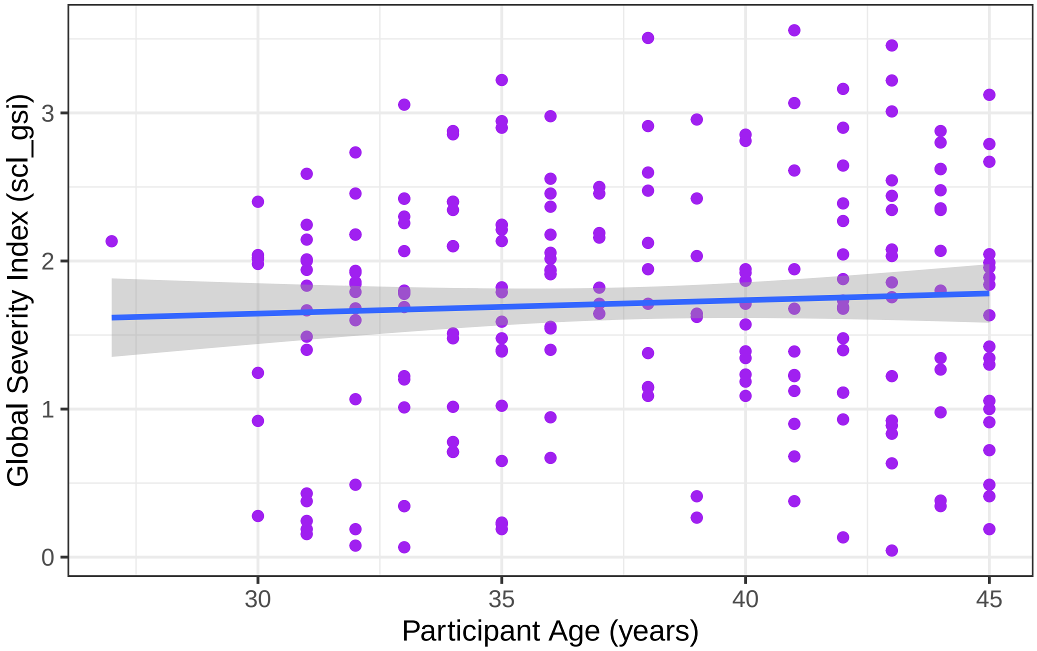
```

	Pearson's product-moment correlation

data:  df$age and df$scl_gsi
t = 0.79039, df = 214, p-value = 0.4302
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08011896  0.18610366
sample estimates:
       cor 
0.05395103 
```

```r
ggplot(data = df, aes(x = age, y = scl_gsi)) +
  geom_point(color = "purple") +
  theme_bw() +
  labs(x = "Participant Age (years)", y = "Global Severity Index (scl_gsi)") +
  stat_smooth(method = lm)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



There’s no real link between age and psychological distress