Thang_Activity 3

2025-11-09

1. Create a Table 1

```
## Set global CRAN mirror
options(repos = c(CRAN = "https://cran.rstudio.com/"))

## Install the tableone package
install.packages("tableone")
```

```
## Installing package into 'C:/Users/Thang/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'tableone' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Thang\AppData\Local\Temp\RtmpOWYKIz\downloaded_packages
```

```
## Load the required packages

library(tableone)
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(knitr)

# 1. Import dataset ------------------------------------------------------

setwd("C:/Users/Thang/OneDrive/Documents/ActivityData_Assignment/Activity_3")
data <- read_excel("SuicideRisk_Data.xlsx")

# 2. Inspect initial structure -------------------------------------------

str(data)
```

```
## tibble [546 x 19] (S3: tbl_df/tbl/data.frame)
##  $ RECORD_ID      : num [1:546] 1 4 7 10 12 14 15 17 19 20 ...
##  $ AGE            : num [1:546] 24 25 21 23 21 22 28 36 24 27 ...
##  $ GENDER         : chr [1:546] "Female" "Female" "Female" "Female" ...
##  $ RACE           : chr [1:546] "White/Caucasian" "White/Caucasian" "White/Caucasian" "Asian" ...
##  $ ETHNICITY      : chr [1:546] "Not Hispanic/Latino" "Not Hispanic/Latino" "Not Hispanic/Latino" "N
##  $ INCOME         : chr [1:546] ">$100,000" "$30,000 - $50,000" "$51,000 - $75,000" "$51,000 - $75,0
##  $ ACES___1       : num [1:546] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ACES___2       : num [1:546] 0 0 0 0 0 1 1 0 1 0 ...
##  $ ACES___3       : num [1:546] 0 0 0 0 0 1 1 0 1 0 ...
##  $ ACES___4       : num [1:546] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ACES___5       : num [1:546] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ACES___6       : num [1:546] 0 0 0 0 1 0 0 0 0 0 ...
##  $ ACES___7       : num [1:546] 0 0 1 0 0 0 0 0 0 0 ...
##  $ ACES___8       : num [1:546] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ACES___9       : num [1:546] 1 0 0 1 0 1 0 1 0 0 ...
##  $ ACES___10      : num [1:546] 0 0 0 1 0 0 0 0 0 0 ...
##  $ CESDR_TOTAL_SUM: num [1:546] 8 15 2 20 9 2 18 0 2 9 ...
##  $ HX_SUICIDE     : num [1:546] 0 0 0 0 0 0 0 0 0 0 ...
##  $ SABCS_TOTAL_SUM: num [1:546] 0 0 0 0 1 0 5 1 0 2 ...
```

```r
# 3. Convert variables to correct types ----------------------------------

data$RECORD_ID <- as.integer(data$RECORD_ID)
data$AGE <- as.numeric(data$AGE)
data$CESDR_TOTAL_SUM <- as.numeric(data$CESDR_TOTAL_SUM)
data$SABCS_TOTAL_SUM <- as.numeric(data$SABCS_TOTAL_SUM)

# Categorical variables

data$GENDER <- factor(data$GENDER, levels = c("Female", "Male"))
data$RACE <- factor(data$RACE, levels = c("White/Caucasian", "Asian", "Other"))
data$ETHNICITY <- factor(data$ETHNICITY,
levels = c("Hispanic/Latino", "Not Hispanic/Latino"))
data$INCOME <- factor(data$INCOME,
levels = c("<$30,000", "$30,000 - $50,000",
"$51,000 - $75,000", "$76,000 - $100,000",
">$100,000"))

# Binary ACEs

aces_vars <- paste0("ACES___", 1:10)
```

```r
data[aces_vars] <- lapply(data[aces_vars], factor,
levels = c(0, 1), labels = c("No", "Yes"))

# Suicide history variable

data$HX_SUICIDE <- factor(data$HX_SUICIDE,
levels = c(1, 0),
labels = c("History of Suicide", "No History of Suicide"))

# Confirm it exists

table(data$HX_SUICIDE)
```

```
##
##    History of Suicide No History of Suicide
##                    49                   497
```

```r
# 4. Define variables for Table 1 ----------------------------------------

vars <- c("AGE", "GENDER", "RACE", "ETHNICITY", "INCOME",
aces_vars, "CESDR_TOTAL_SUM", "SABCS_TOTAL_SUM")

catVars <- c("GENDER", "RACE", "ETHNICITY", "INCOME", aces_vars)

# 5. Create Table 1 ------------------------------------------------------

table1 <- CreateTableOne(
vars = vars,
strata = "HX_SUICIDE",
data = data,
factorVars = catVars,
addOverall = TRUE,
test = TRUE
)
```

```r
# 6. Convert CreateTableOne output to clean dataframe ---------------------

# Convert as before
table1_df <- as.data.frame(print(table1, test = TRUE, smd = TRUE))
```

```
##                                      Stratified by HX_SUICIDE
##                                       Overall       History of Suicide
##   n                                     546            49
##   AGE (mean (SD))                     24.85 (6.50)  25.69 (7.11)
##   GENDER = Male (%)                      47 ( 8.6)      5 (10.2)
##   RACE (%)
##      White/Caucasian                    392 (71.8)     38 (77.6)
##      Asian                               63 (11.5)      5 (10.2)
##      Other                               91 (16.7)      6 (12.2)
##   ETHNICITY = Not Hispanic/Latino (%)   478 (87.5)     38 (77.6)
##   INCOME (%)
##      $30,000 - $50,000                  101 (23.1)     11 (30.6)
```

```
##      $51,000 - $75,000                  108 (24.7)      7 (19.4)
##      $76,000 - $100,000                 106 (24.2)      8 (22.2)
##      >$100,000                          123 (28.1)     10 (27.8)
##   ACES___1 = Yes (%)                     66 (12.1)     13 (26.5)
##   ACES___2 = Yes (%)                     95 (17.4)     26 (53.1)
##   ACES___3 = Yes (%)                    158 (28.9)     31 (63.3)
##   ACES___4 = Yes (%)                     18 ( 3.3)      4 ( 8.2)
##   ACES___5 = Yes (%)                     64 (11.7)     10 (20.4)
##   ACES___6 = Yes (%)                     86 (15.8)     14 (28.6)
##   ACES___7 = Yes (%)                     76 (13.9)     11 (22.4)
##   ACES___8 = Yes (%)                    119 (21.8)     25 (51.0)
##   ACES___9 = Yes (%)                    156 (28.6)     22 (44.9)
##   ACES___10 = Yes (%)                    27 ( 4.9)      7 (14.3)
##   CESDR_TOTAL_SUM (mean (SD))         15.77 (13.70) 27.63 (15.52)
##   SABCS_TOTAL_SUM (mean (SD))          3.31 (4.86)  10.63 (6.83)
##                                     Stratified by HX_SUICIDE
##                                      No History of Suicide p      test SMD
##   n                                    497
##   AGE (mean (SD))                    24.77 (6.44)          0.342       0.137
##   GENDER = Male (%)                     42 ( 8.5)          0.880       0.060
##   RACE (%)                                                 0.618       0.153
##      White/Caucasian                   354 (71.2)
##      Asian                              58 (11.7)
##      Other                              85 (17.1)
##   ETHNICITY = Not Hispanic/Latino (%)  440 (88.5)          0.046       0.296
##   INCOME (%)                                               0.693       0.205
##      $30,000 - $50,000                  90 (22.4)
##      $51,000 - $75,000                 101 (25.1)
##      $76,000 - $100,000                 98 (24.4)
##      >$100,000                         113 (28.1)
##   ACES___1 = Yes (%)                    53 (10.7)          0.003       0.417
##   ACES___2 = Yes (%)                    69 (13.9)         <0.001       0.913
##   ACES___3 = Yes (%)                   127 (25.6)         <0.001       0.820
##   ACES___4 = Yes (%)                    14 ( 2.8)          0.114       0.236
##   ACES___5 = Yes (%)                    54 (10.9)          0.080       0.265
##   ACES___6 = Yes (%)                    72 (14.5)          0.017       0.348
##   ACES___7 = Yes (%)                    65 (13.1)          0.111       0.247
##   ACES___8 = Yes (%)                    94 (18.9)         <0.001       0.715
##   ACES___9 = Yes (%)                   134 (27.0)          0.013       0.381
##   ACES___10 = Yes (%)                   20 ( 4.0)          0.005       0.362
##   CESDR_TOTAL_SUM (mean (SD))        14.60 (12.95)         <0.001       0.912
##   SABCS_TOTAL_SUM (mean (SD))         2.58 (3.95)          <0.001       1.442
```

```r
# Move row names to Characteristic column
table1_df$Characteristic <- rownames(table1_df)
rownames(table1_df) <- NULL

# Remove the "test" column (the empty NA column)
table1_df <- table1_df[, !(names(table1_df) %in% c("test"))]

# Reorder so Characteristic comes first
table1_df <- table1_df[, c("Characteristic", setdiff(names(table1_df), "Characteristic"))]

# Rename columns EXACTLY to correct meaning
```

```r
names(table1_df) <- c("Characteristic",
                      "Total (n = 546)",
                      "History of Suicide",
                      "No History of Suicide",
                      "P-value",
                      "Effect Size")
```

Significant characteristic ($p < 0.05$)

Participants with a history of suicide were significantly more likely to identify as Hispanic or Latino compared to those without a suicide history. This suggests that cultural or social factors associated with ethnicity may be related to suicide risk in this sample, although the result shows association—not causation.

Non-significant characteristic (p 0.05)

The average age did not differ significantly between participants with and without a history of suicide. This means that age alone was not associated with suicide history in this study, and both groups had similar age distributions.

```r
# 8. Display nicely ------------------------------------------------------


library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
kable(table1_df,
      caption = "Demographic and Mental Health Characteristics (n = 546)",
      booktabs = TRUE,
      align = "lccccc",
      escape = FALSE) %>%
  kable_styling(full_width = FALSE,
                latex_options = c("hold_position", "scale_down")) %>%
  column_spec(1, width = "3.8cm") %>%
  column_spec(2, width = "2.5cm") %>%
  column_spec(3, width = "2.5cm") %>%
  column_spec(4, width = "2.5cm") %>%
  column_spec(5, width = "1.8cm") %>%
  column_spec(6, width = "1.8cm") %>%
  footnote(
    general = "Continuous variables are shown as mean ± SD; categorical variables as n (%). P-values ar
    general_title = "Note:",
    threeparttable = TRUE
  )
```

```
## Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
## resized.
```

Table 1: Demographic and Mental Health Characteristics (n = 546)

Characteristic	Total (n = 546)	History of Suicide	No History of Suicide	P-value	Effect Size
n	546	49	497		
AGE (mean (SD))	24.85 (6.50)	25.69 (7.11)	24.77 (6.44)	0.342	0.137
GENDER = Male (%)	47 (8.6)	5 (10.2)	42 (8.5)	0.880	0.060
RACE (%)				0.618	0.153
White/Caucasian	392 (71.8)	38 (77.6)	354 (71.2)		
Asian	63 (11.5)	5 (10.2)	58 (11.7)		
Other	91 (16.7)	6 (12.2)	85 (17.1)		
ETHNICITY = Not Hispanic/Latino (%)	478 (87.5)	38 (77.6)	440 (88.5)	0.046	0.296
INCOME (%)				0.693	0.205
$30,000 - $50,000	101 (23.1)	11 (30.6)	90 (22.4)		
$51,000 - $75,000	108 (24.7)	7 (19.4)	101 (25.1)		
$76,000 - $100,000	106 (24.2)	8 (22.2)	98 (24.4)		
>$100,000	123 (28.1)	10 (27.8)	113 (28.1)		
ACES____1 = Yes (%)	66 (12.1)	13 (26.5)	53 (10.7)	0.003	0.417
ACES____2 = Yes (%)	95 (17.4)	26 (53.1)	69 (13.9)	<0.001	0.913
ACES____3 = Yes (%)	158 (28.9)	31 (63.3)	127 (25.6)	<0.001	0.820
ACES____4 = Yes (%)	18 (3.3)	4 (8.2)	14 (2.8)	0.114	0.236
ACES____5 = Yes (%)	64 (11.7)	10 (20.4)	54 (10.9)	0.080	0.265
ACES____6 = Yes (%)	86 (15.8)	14 (28.6)	72 (14.5)	0.017	0.348
ACES____7 = Yes (%)	76 (13.9)	11 (22.4)	65 (13.1)	0.111	0.247
ACES____8 = Yes (%)	119 (21.8)	25 (51.0)	94 (18.9)	<0.001	0.715
ACES____9 = Yes (%)	156 (28.6)	22 (44.9)	134 (27.0)	0.013	0.381
ACES____10 = Yes (%)	27 (4.9)	7 (14.3)	20 (4.0)	0.005	0.362
CESDR_TOTAL_SUM (mean (SD))	15.77 (13.70)	27.63 (15.52)	14.60 (12.95)	<0.001	0.912
SABCS_TOTAL_SUM (mean (SD))	3.31 (4.86)	10.63 (6.83)	2.58 (3.95)	<0.001	1.442

Note:
Continuous variables are shown as mean ± SD; categorical variables as n (%). P-values are derived from t-tests or Chi-square tests as appropriate. Effect Size = Standardized Mean Difference (SMD). ACEs = Adverse Childhood Experiences; CESDR = Center for Epidemiologic Studies Depression Scale; SABCS = Suicidal Affect–Behavior–Cognition Scale.

2. This problem will focus on comparing suicidal risk (Suicidal Affect-Behavior-Cognition Scale [SABCS]) across race groups (White/Caucasian vs Asian vs Other).

A. Group comparison

```r
# Load necessary packages
library(ggplot2)
library(dplyr)
install.packages("psych")
```

```
## Installing package into 'C:/Users/Thang/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
##
##   There is a binary version available but the source version is later:
##       binary source needs_compilation
## psych  2.5.3  2.5.6             FALSE
```

```
## installing the source package 'psych'
```

```r
library(psych)
```

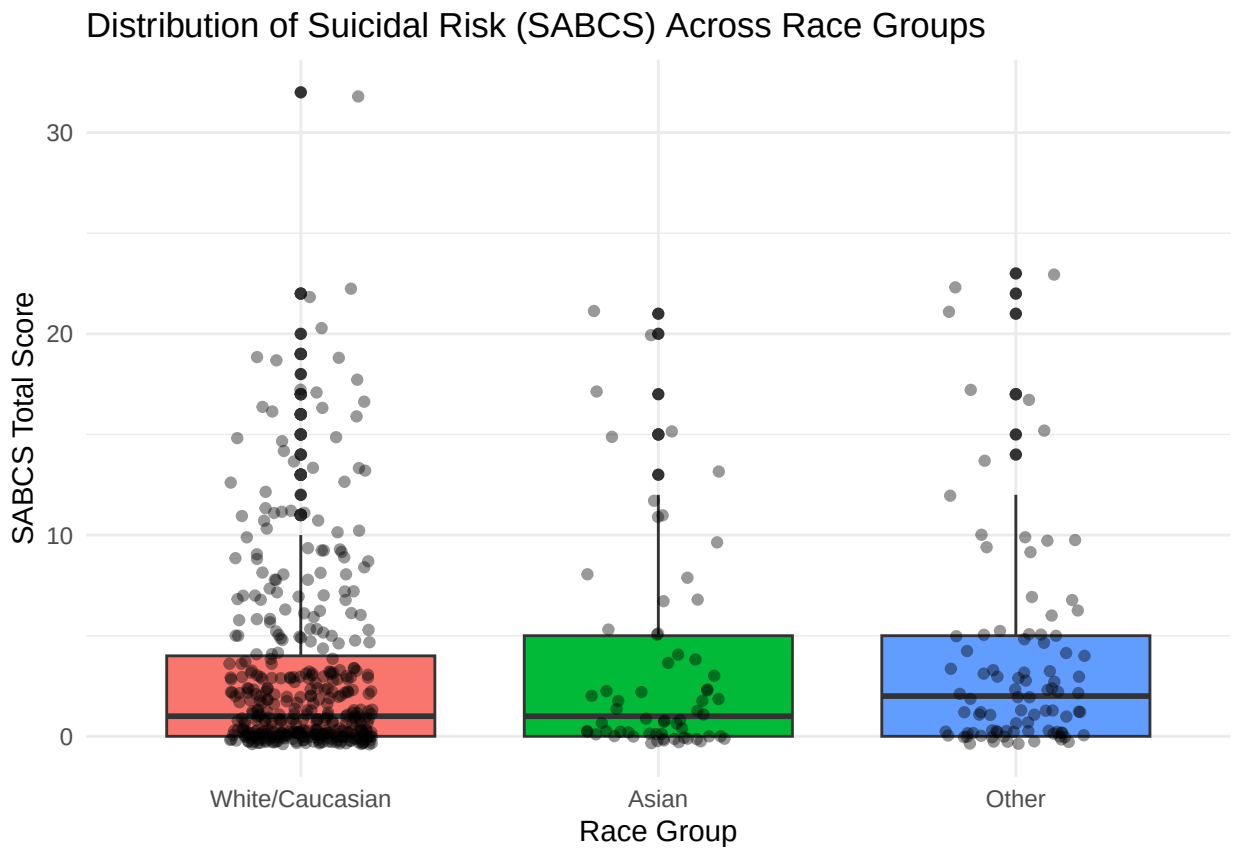
```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
# 1. Descriptive statistics by race
data %>%
  group_by(RACE) %>%
  summarise(
    n = n(),
    mean_SABCS = mean(SABCS_TOTAL_SUM, na.rm = TRUE),
    sd_SABCS = sd(SABCS_TOTAL_SUM, na.rm = TRUE),
    median_SABCS = median(SABCS_TOTAL_SUM, na.rm = TRUE),
    IQR_SABCS = IQR(SABCS_TOTAL_SUM, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 6
##   RACE                n mean_SABCS sd_SABCS median_SABCS IQR_SABCS
##   <fct>           <int>      <dbl>    <dbl>        <dbl>     <dbl>
## 1 White/Caucasian   392       3.16     4.69            1         4
## 2 Asian              63       3.65     5.38            1         5
## 3 Other              91       3.71     5.23            2         5
```

```r
# 2. Visualization: Boxplot of SABCS by Race
ggplot(data, aes(x = RACE, y = SABCS_TOTAL_SUM, fill = RACE)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.4) +
  theme_minimal() +
  labs(title = "Distribution of Suicidal Risk (SABCS) Across Race Groups",
       x = "Race Group", y = "SABCS Total Score") +
  theme(legend.position = "none")
```




```
# 3. Run one-way ANOVA
anova_model <- aov(SABCS_TOTAL_SUM ~ RACE, data = data)
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## RACE          2     32   15.76   0.667  0.514
## Residuals   543  12838   23.64
```

The distribution of suicidal risk scores (SABCS) showed similar patterns across the three racial groups. White/Caucasian participants had a mean score of 3.16 (SD = 4.69) with a median of 1 and an IQR of 4, indicating a right-skewed distribution with many low scores and a long tail of higher scores. Asian participants exhibited a slightly higher mean (3.65, SD = 5.38) and a median of 1, with an IQR of 5, suggesting greater spread and more variability in suicidal risk. Participants in the "Other" race category had the highest median (2) and an IQR of 5, with a mean of 3.71 (SD = 5.23), also reflecting a skewed distribution with several high-risk values.

B & C. Checking assumptions AND Post-hoc analysis

```
# Check homogeneity of variances (Levene's test)
install.packages("car")
```

```
## Installing package into 'C:/Users/Thang/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Thang\AppData\Local\Temp\RtmpOWYKIz\downloaded_packages
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

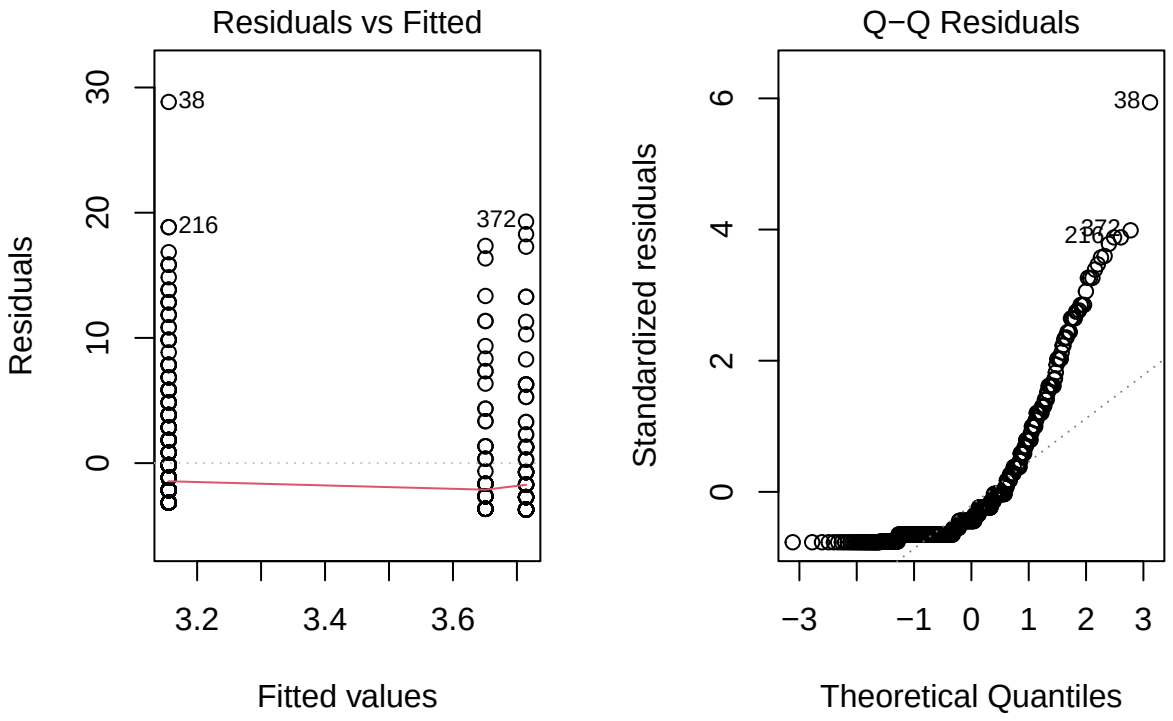
```
leveneTest(SABCS_TOTAL_SUM ~ RACE, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.6178 0.5395
##       543
```

```
# Check normality of residuals
shapiro.test(residuals(anova_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_model)
## W = 0.72477, p-value < 2.2e-16
```

```
# Visualization of residuals
par(mfrow = c(1, 2))
plot(anova_model, which = 1)  # Residuals vs Fitted
plot(anova_model, which = 2)  # Q-Q plot
```




```
par(mfrow = c(1, 1))
```

Levene's test indicated that the assumption of homogeneity of variances was met ($p = 0.54$). However, the Shapiro–Wilk test showed that the ANOVA residuals were significantly non-normal ($p < 0.001$), suggesting substantial deviation from normality. Given the large sample size, ANOVA results remain relatively robust, but a non-parametric alternative (e.g., Kruskal–Wallis test) may provide a more appropriate assessment.

The one-way ANOVA showed no statistically significant differences in mean suicidal risk (SABCS) across race groups ($p > 0.05$). Therefore, post-hoc pairwise comparisons were not conducted.

D AND E. Linear Regression AND CompaCompare/Contrast your Results

```
# Make sure reference category is White/Caucasian
data$RACE <- relevel(data$RACE, ref = "White/Caucasian")

# Fit the linear regression model
lm_model <- lm(SABCS_TOTAL_SUM ~ RACE, data = data)

# Show summary
summary(lm_model)
```

```
##
```

```
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ RACE, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.714 -3.156 -2.156  1.175 28.844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1556     0.2456  12.849   <2e-16 ***
## RACEAsian     0.4952     0.6600   0.750    0.453
## RACEOther     0.5587     0.5658   0.987    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.862 on 543 degrees of freedom
## Multiple R-squared:  0.002449,   Adjusted R-squared:  -0.001225
## F-statistic: 0.6667 on 2 and 543 DF,  p-value: 0.5138
```

Linear Regression Interpretation (plain language, alpha = 0.05)

The linear regression model showed no statistically significant differences in suicidal risk (SABCS scores) between race groups (overall p = 0.51). Specifically, Asian participants (p = 0.45) and those in the "Other" race category (p = 0.32) did not differ significantly in suicidal risk compared with White/Caucasian participants, the reference group.

Compare and Contrast with ANOVA Results

Both the ANOVA and regression examined differences in suicidal risk by race and reached the same conclusion: there were no significant differences in mean SABCS scores across racial groups. The regression quantified these differences relative to the White/Caucasian group, while the ANOVA tested for any overall group difference. Together, both analyses indicate that race was not a significant predictor of suicidal risk in this sample.

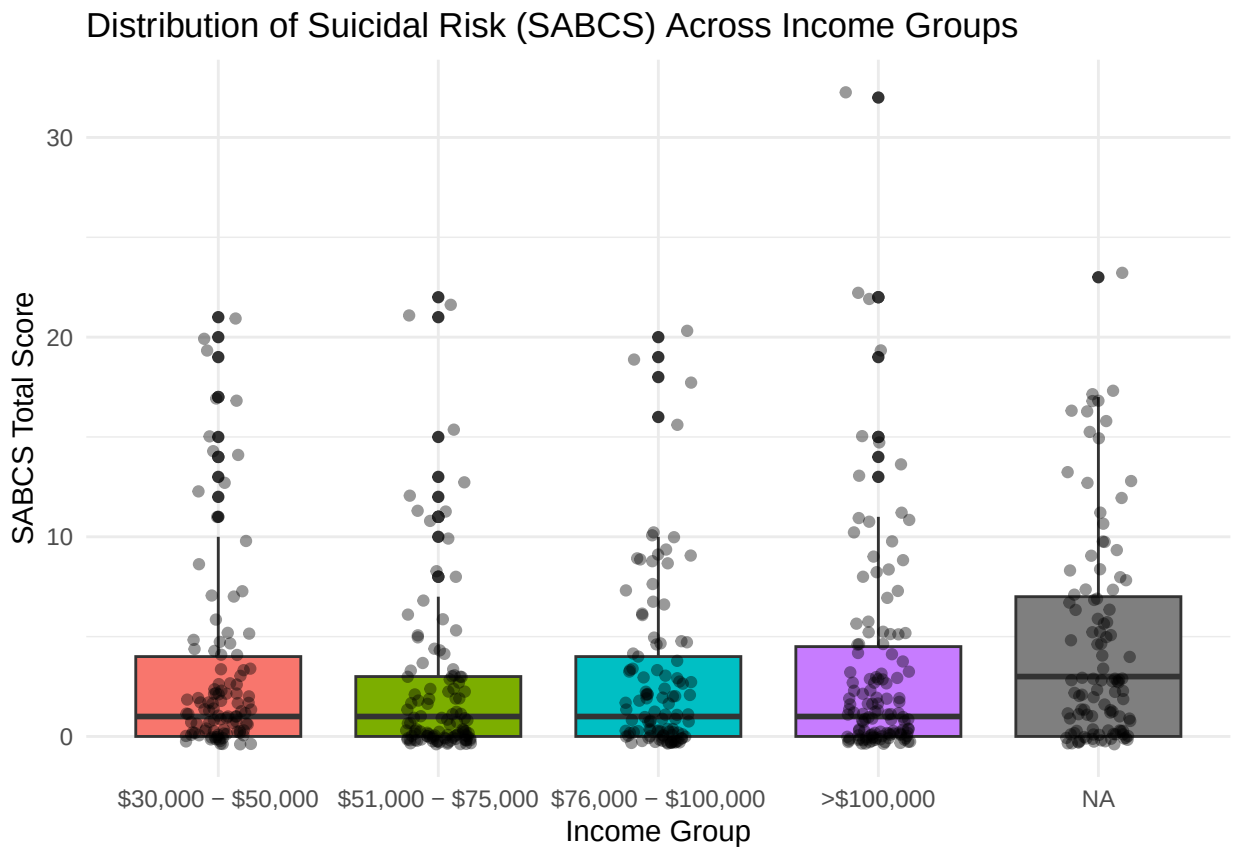
3. This problem will focus on comparing suicidal risk (Suicidal Affect-Behavior-Cognition Scale [SABCS]) across the five income groups

```r
# A. Group comparison for INCOME

library(dplyr)
library(ggplot2)

# 1. Summary stats by INCOME
income_summary <- data %>%
  group_by(INCOME) %>%
  summarise(
    n = n(),
    mean_SABCS = mean(SABCS_TOTAL_SUM, na.rm = TRUE),
    sd_SABCS = sd(SABCS_TOTAL_SUM, na.rm = TRUE),
```

```
    median_SABCS = median(SABCS_TOTAL_SUM, na.rm = TRUE),
    IQR_SABCS = IQR(SABCS_TOTAL_SUM, na.rm = TRUE)
  )

income_summary
```

```
## # A tibble: 5 x 6
##   INCOME                n mean_SABCS sd_SABCS median_SABCS IQR_SABCS
##   <fct>             <int>      <dbl>    <dbl>        <dbl>     <dbl>
## 1 $30,000 - $50,000   101       3.32     4.92            1       4
## 2 $51,000 - $75,000   108       2.36     4.16            1       3
## 3 $76,000 - $100,000  106       2.95     4.25            1       4
## 4 >$100,000           123       3.34     5.29            1       4.5
## 5 <NA>                108       4.55     5.30            3       7
```

```
# 2. Visualization
ggplot(data, aes(x = INCOME, y = SABCS_TOTAL_SUM, fill = INCOME)) +
  geom_boxplot() +
  geom_jitter(width = 0.15, alpha = 0.4) +
  theme_minimal() +
  labs(
    title = "Distribution of Suicidal Risk (SABCS) Across Income Groups",
    x = "Income Group",
    y = "SABCS Total Score"
  ) +
  theme(legend.position = "none")
```




```
# 3. One-way ANOVA
anova_income <- aov(SABCS_TOTAL_SUM ~ INCOME, data = data)
summary(anova_income)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## INCOME        3     69   22.93   1.039  0.375
## Residuals   434   9583   22.08
## 108 observations deleted due to missingness
```

A. Group comparison

Across the five income groups, suicidal-risk scores (SABCS) showed similarly right-skewed distributions, with most participants reporting low scores and a smaller subset showing higher levels of risk. Median scores and variability were generally comparable across income levels, although slightly greater spread was observed in the lower-income and middle-income groups. Overall, no clear visual differences in central tendency or distribution shape were evident among the five groups.

B. CHECKING ASSUMPTION

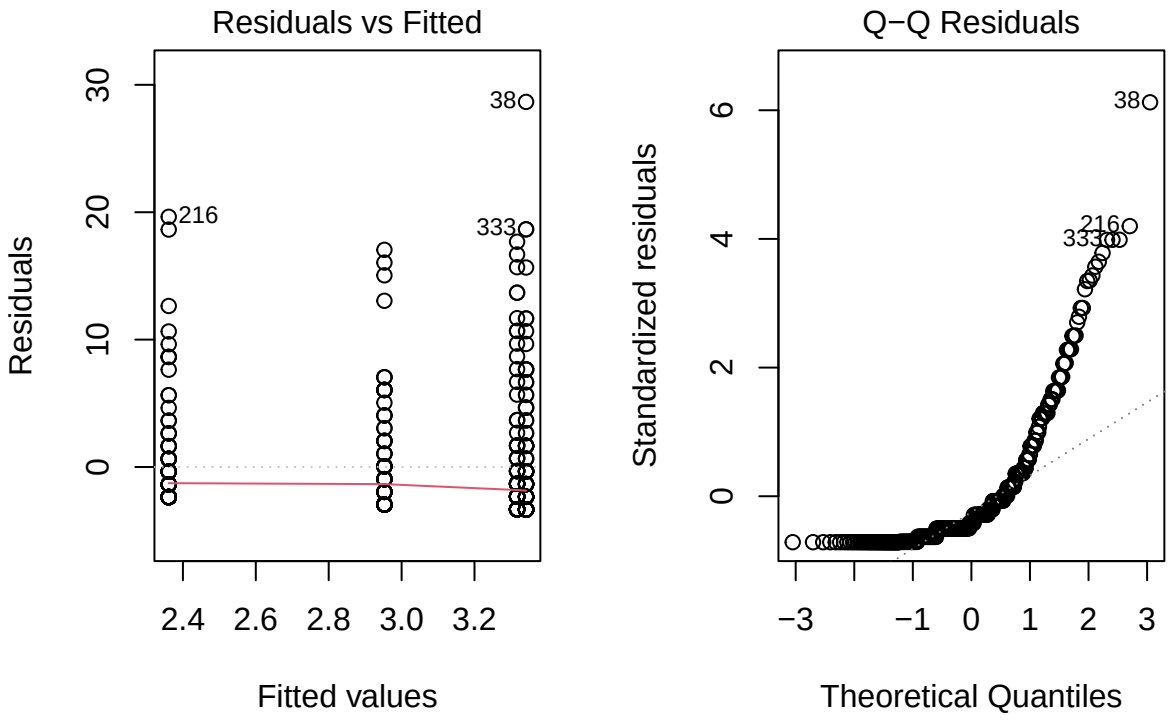
```
# Levene test
library(car)
leveneTest(SABCS_TOTAL_SUM ~ INCOME, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.6701 0.5708
##       434
```

```
# Shapiro-Wilk for residuals
shapiro.test(residuals(anova_income))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_income)
## W = 0.69541, p-value < 2.2e-16
```

```
# Residual plots
par(mfrow = c(1, 2))
plot(anova_income, which = 1)  # Residuals vs Fitted
plot(anova_income, which = 2)  # QQ plot
```




```
par(mfrow = c(1, 1))
```

Levene's test indicated that the assumption of homogeneity of variances across the five income groups was satisfied ($F(3,434) = 0.67$, $p = 0.57$). However, the Shapiro–Wilk test showed that the ANOVA residuals were strongly non-normal ($W = 0.695$, $p < 0.001$), reflecting substantial skewness in the suicidal-risk data. Although ANOVA is generally robust to non-normality with large samples, the severity of this violation suggests that a non-parametric alternative may be more appropriate as a confirmatory analysis.

POST-HOC TESTTEST

Because the one-way ANOVA did not find a statistically significant difference in suicidal-risk scores across the five income groups ($p = 0.375$), post-hoc pairwise comparisons were not conducted. Post-hoc tests are only meaningful when the overall ANOVA indicates that at least one group mean differs significantly from the others. Since this was not the case, additional comparisons would not provide useful or interpretable information.

4. This problem will focus on comparing suicidal risk (Suicidal Affect-Behavior-Cognition Scale [SABCS]) across the two gender groups.

```
# Check group sizes and means
table(data$GENDER)
```

```
##
## Female   Male
##    499     47
```

```
tapply(data$SABCS_TOTAL_SUM, data$GENDER, mean, na.rm = TRUE)
```

```
##   Female     Male
## 3.334669 3.000000
```

```
tapply(data$SABCS_TOTAL_SUM, data$GENDER, sd, na.rm = TRUE)
```

```
##   Female     Male
## 4.764559 5.823491
```

```
# Perform independent samples t-test
t_test_gender <- t.test(SABCS_TOTAL_SUM ~ GENDER, data = data, var.equal = TRUE)
t_test_gender
```

```
##
##  Two Sample t-test
##
## data:  SABCS_TOTAL_SUM by GENDER
## t = 0.45104, df = 544, p-value = 0.6521
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -1.122868  1.792207
## sample estimates:
## mean in group Female   mean in group Male
##             3.334669             3.000000
```

The two-sample t-test showed no significant difference in suicidal risk (SABCS scores) between female and male participants ($t(544) = 0.45$, $p = 0.65$). On average, women and men reported similar levels of suicidal thoughts, behaviors, and feelings in this sample.

```r
data$GENDER <- relevel(data$GENDER, ref = "Female")
lm_gender <- lm(SABCS_TOTAL_SUM ~ GENDER, data = data)
summary(lm_gender)
```

```
##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ GENDER, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.335 -3.335 -2.335  1.415 29.000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3347     0.2177  15.318   <2e-16 ***
## GENDERMale   -0.3347     0.7420  -0.451    0.652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.863 on 544 degrees of freedom
## Multiple R-squared:  0.0003738,  Adjusted R-squared:  -0.001464
## F-statistic: 0.2034 on 1 and 544 DF,  p-value: 0.6521
```

Group comnparison

The regression analysis confirmed that gender was not a significant predictor of suicidal risk ($p = 0.65$). Male participants scored, on average, 0.33 points lower on the SABCS scale than female participants, but this small difference was not statistically meaningful.

Compare and contrast

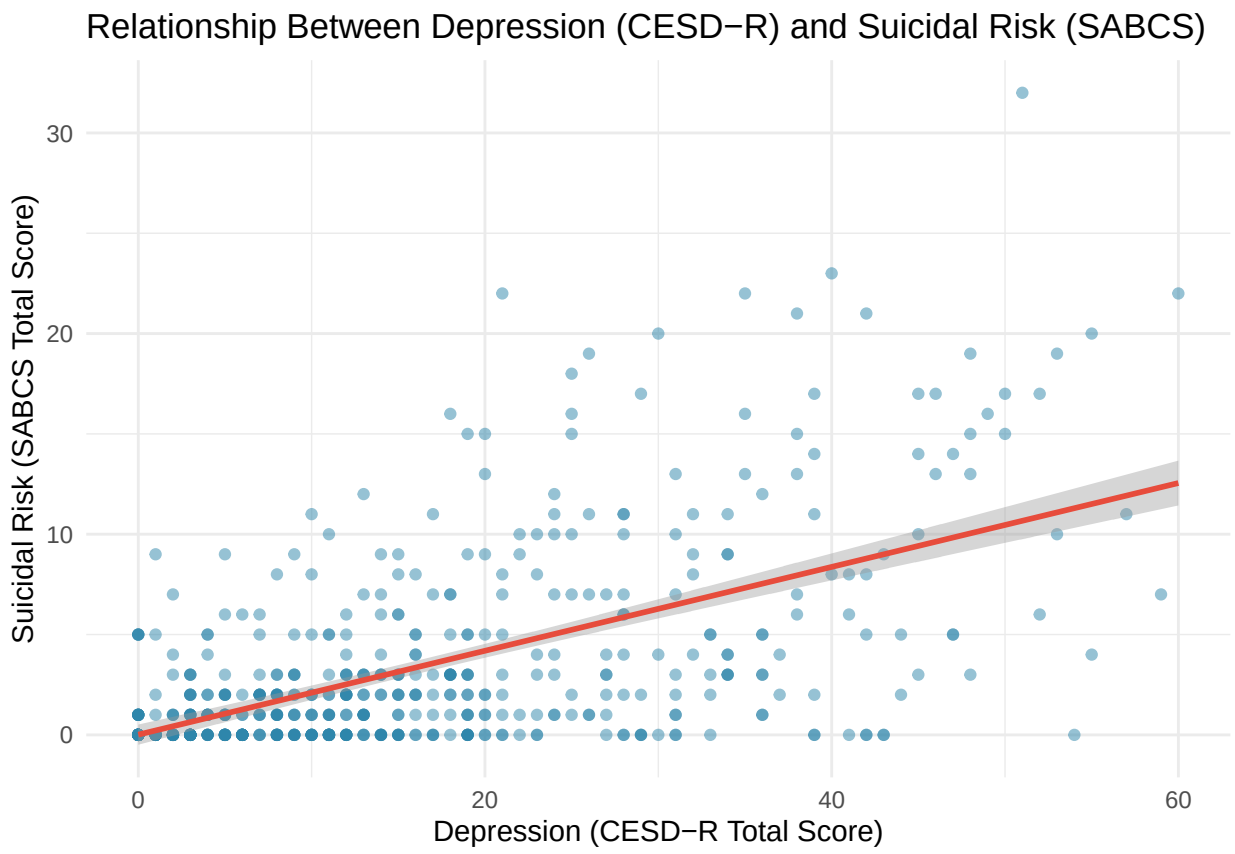
Both the t-test and the regression model produced identical results—gender was not associated with differences in suicidal-risk scores. The t-test directly compared mean SABCS scores between females and males, while the regression estimated the mean difference using females as the reference category. Together, the findings indicate that gender did not play a significant role in suicidal risk within this sample.

5. This problem will focus on quantifying the association between suicidal risk (Suicidal Affect-Behavior-Cognition Scale [SABCS]) and depression (CESD-R), including the form, direction, and strength of their relationship.

```r
library(ggplot2)

ggplot(data, aes(x = CESDR_TOTAL_SUM, y = SABCS_TOTAL_SUM)) +
  geom_point(alpha = 0.5, color = "#2E86AB") +
  geom_smooth(method = "lm", se = TRUE, color = "#E74C3C") +
  theme_minimal() +
  labs(
```

```
    title = "Relationship Between Depression (CESD-R) and Suicidal Risk (SABCS)",
    x = "Depression (CESD-R Total Score)",
    y = "Suicidal Risk (SABCS Total Score)"
  )
```

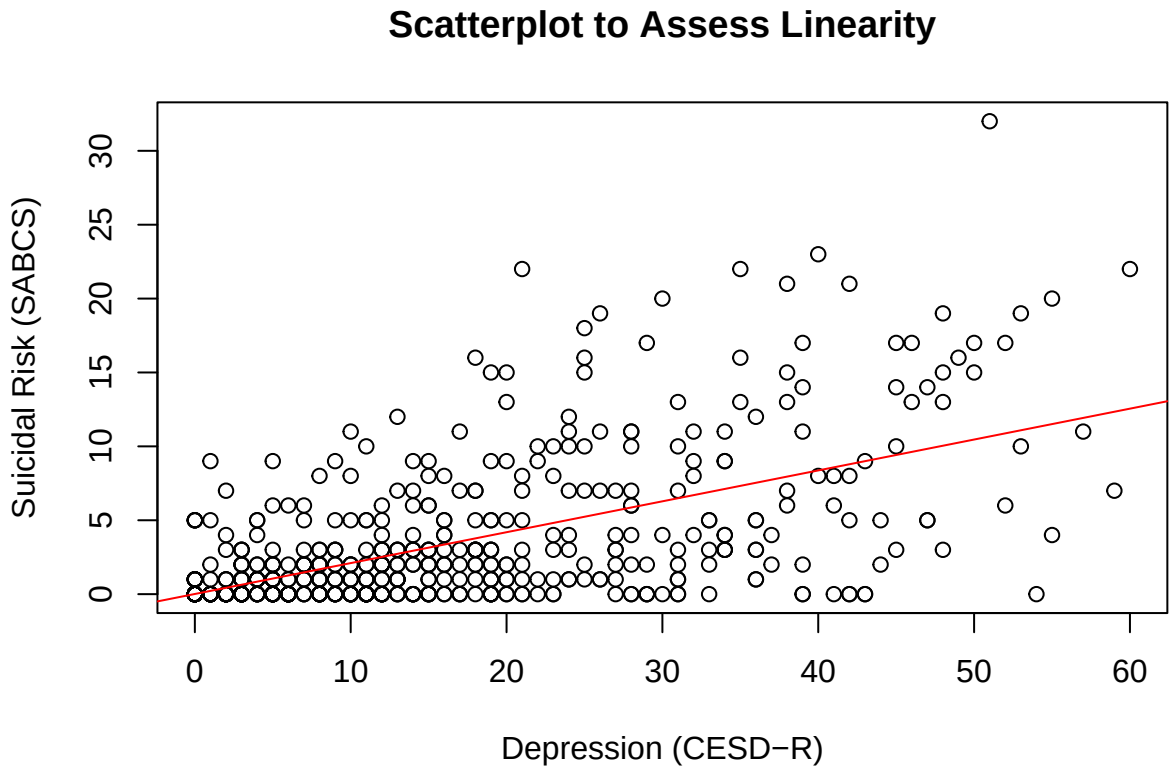
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
cor_test <- cor.test(data$SABCS_TOTAL_SUM,
                     data$CESDR_TOTAL_SUM, method = "pearson")
cor_test
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$SABCS_TOTAL_SUM and data$CESDR_TOTAL_SUM
## t = 17.013, df = 544, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5316841 0.6414951
## sample estimates:
##       cor
## 0.5893046
```

```r
# Check linearity visually
plot(data$CESDR_TOTAL_SUM, data$SABCS_TOTAL_SUM,
     main = "Scatterplot to Assess Linearity",
     xlab = "Depression (CESD-R)", ylab = "Suicidal Risk (SABCS)")
abline(lm(SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = data), col = "red")
```




```r
# Check normality of both variables
shapiro.test(data$CESDR_TOTAL_SUM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$CESDR_TOTAL_SUM
## W = 0.90079, p-value < 2.2e-16
```

```r
shapiro.test(data$SABCS_TOTAL_SUM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$SABCS_TOTAL_SUM
## W = 0.7098, p-value < 2.2e-16
```

Visualization:

The scatterplot demonstrates a clear positive linear relationship between depression (CESD-R) and suicidal risk (SABCS). As depression scores increase, suicidal risk scores also tend to rise. The regression line indicates that this relationship is approximately linear.

Estimation:

The Pearson correlation coefficient between depression and suicidal risk was r = 0.589 (95% CI [0.53, 0.64]), indicating a moderate to strong positive association. This means that participants with higher depression levels generally reported higher suicidal risk.

Test of assumptions:

Visual inspection of the scatterplot suggests that the relationship is linear. However, Shapiro–Wilk tests showed both variables deviate from normality ($p < 0.001$). Given the large sample size (n = 546), the Pearson correlation remains valid because the method is robust to mild non-normality when sample sizes are large.

Test of correlation (= 0.05):

The correlation was statistically significant ($t(544) = 17.01$, $p < 0.001$). In plain language:

There is a significant positive relationship between depression and suicidal risk. Participants with higher depression scores also tend to experience greater suicidal thoughts, feelings, and behaviors. The strength of the association suggests that depression is an important factor closely linked to suicidal risk in this sample.

Visualization for linear regression

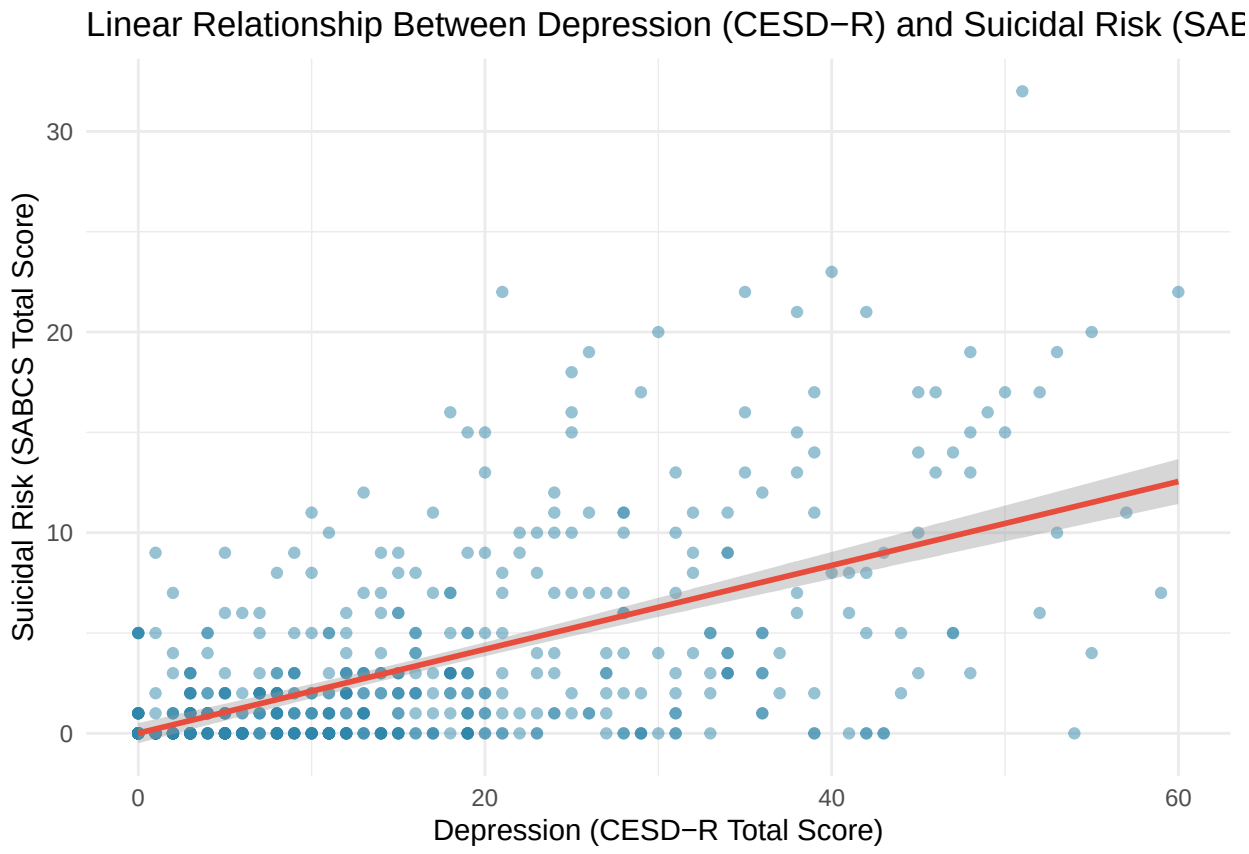
```r
## ---- Linear Regression: Depression predicting Suicidal Risk ----

# Fit simple linear regression model
lm_line <- lm(SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = data)

# Display summary of regression results
summary(lm_line)
```

```
##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2968  -2.1009  -0.4739   0.9891  21.3302
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.01088    0.25650   0.042    0.966
## CESDR_TOTAL_SUM  0.20900    0.01228  17.013   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.93 on 544 degrees of freedom
## Multiple R-squared:  0.3473, Adjusted R-squared:  0.3461
## F-statistic: 289.4 on 1 and 544 DF,  p-value: < 2.2e-16
```

```r
#  Visualization: scatterplot + regression line
library(ggplot2)

ggplot(data, aes(x = CESDR_TOTAL_SUM, y = SABCS_TOTAL_SUM)) +
  geom_point(alpha = 0.5, color = "#2E86AB") +
  geom_smooth(method = "lm", se = TRUE, color = "#E74C3C") +
  theme_minimal() +
  labs(
    title = "Linear Relationship Between Depression (CESD-R) and Suicidal Risk (SABCS)",
    x = "Depression (CESD-R Total Score)",
    y = "Suicidal Risk (SABCS Total Score)"
  )
```

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



```r
#  Extract regression coefficients for equation
coef(lm_line)
```

```
##     (Intercept) CESDR_TOTAL_SUM
##      0.01087968      0.20899857
```

```r
# (optional) store intercept and slope for display
intercept <- coef(lm_line)[1]
slope <- coef(lm_line)[2]
cat("Estimated regression equation:\nSABCS =", round(intercept, 2),
    "+", round(slope, 2), "× CESD-R\n")
```

```
## Estimated regression equation:
## SABCS = 0.01 + 0.21 × CESD-R
```

```r
# Calculate R-squared manually (for reporting)
r_squared <- summary(lm_line)$r.squared
cat("R-squared:", round(r_squared, 3), "\n")
```

```
## R-squared: 0.347
```

Because the scatterplot showed a clear positive linear trend between depression (CESD-R) and suicidal risk (SABCS), fitting a line of best fit was appropriate. The estimated regression equation was:

SABCS=0.01+0.21×CESD-R

This means that for every 1-point increase in the CESD-R depression score, the predicted suicidal risk score increases by about 0.21 points on average. The positive slope indicates that higher depression scores are associated with higher suicidal risk.

Linear Regression Analysis (= 0.05)

The regression model revealed a statistically significant positive relationship between depression and suicidal risk ($F(1, 544) = 289.4$, $p < 0.001$). The slope for depression (= 0.21) was significant ($p < 0.001$), indicating that depression scores were a strong predictor of suicidal risk scores. The model explained approximately 34.7% of the variance in suicidal risk ($R^2 = 0.35$), suggesting a substantial association.

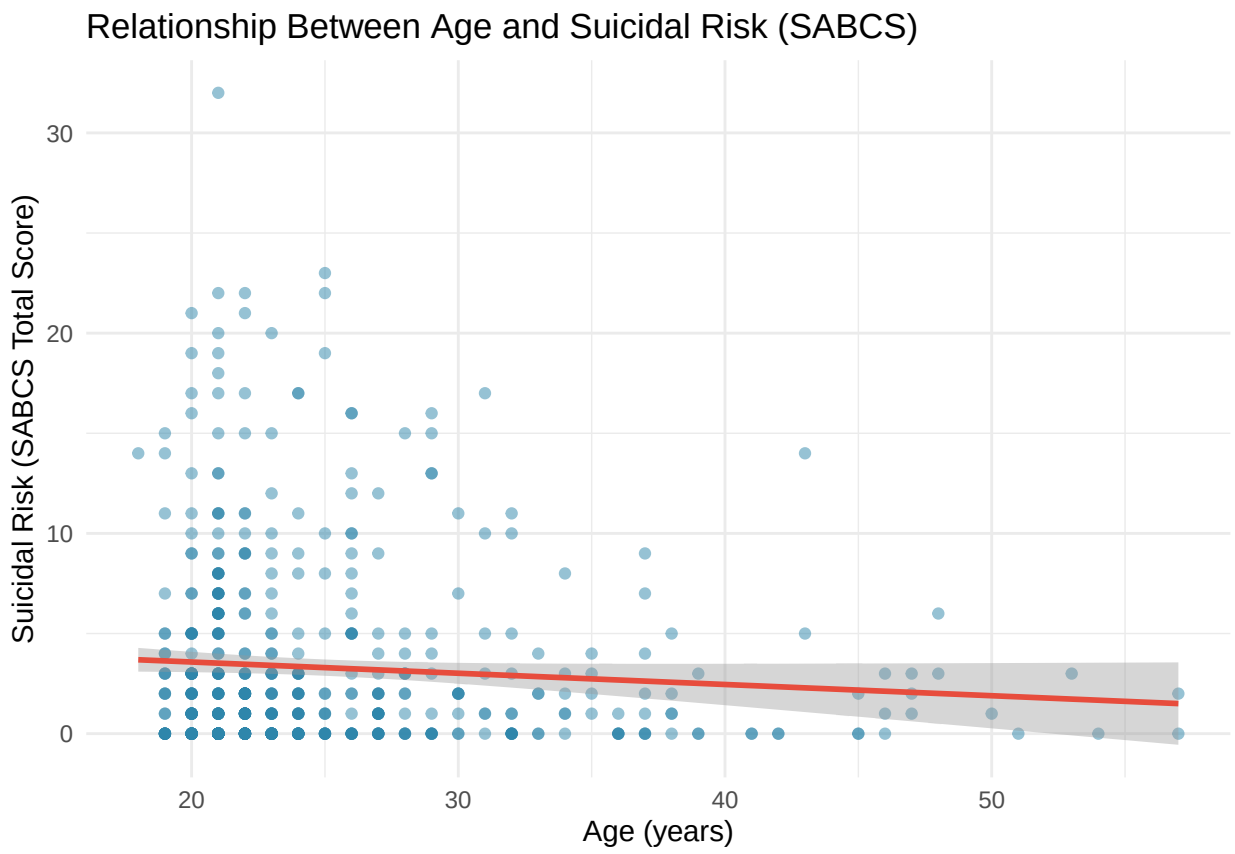
Plain-language interpretation: The analysis shows that individuals with higher depression levels tend to report higher suicidal-risk scores. The relationship is statistically significant and moderately strong, meaning that depression accounts for a large portion of the differences in suicidal risk among participants.

6. This problem will focus on quantifying the association between suicidal risk (Suicidal Affect-Behavior-Cognition Scale [SABCS]) and age, including the form, direction, and strength of their relationship

```r
# Scatterplot of Age vs. Suicidal Risk
ggplot(data, aes(x = AGE, y = SABCS_TOTAL_SUM)) +
  geom_point(alpha = 0.5, color = "#2E86AB") +
  geom_smooth(method = "lm", se = TRUE, color = "#E74C3C") +
```

```
  theme_minimal() +
  labs(
    title = "Relationship Between Age and Suicidal Risk (SABCS)",
    x = "Age (years)",
    y = "Suicidal Risk (SABCS Total Score)"
  )
```

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



The scatterplot shows a generally flat pattern with no clear upward or downward trend. This suggests that suicidal risk scores do not change substantially with age in this sample. The fitted regression line indicates that the relationship between age and suicidal risk is weak.

```
# Pearson correlation between suicidal risk and age
cor_age <- cor.test(data$SABCS_TOTAL_SUM, data$AGE, method = "pearson")
cor_age
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$SABCS_TOTAL_SUM and data$AGE
## t = -1.7567, df = 544, p-value = 0.07953
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.158022264  0.008862322
```

```
## sample estimates:
##         cor
## -0.07510585
```

```
# Check linearity visually (scatterplot above)
# Check normality of both variables
shapiro.test(data$AGE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$AGE
## W = 0.74319, p-value < 2.2e-16
```

```
shapiro.test(data$SABCS_TOTAL_SUM)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$SABCS_TOTAL_SUM
## W = 0.7098, p-value < 2.2e-16
```

```
lm_age <- lm(SABCS_TOTAL_SUM ~ AGE, data = data)
summary(lm_age)
```

```
##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ AGE, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.634 -3.353 -1.792  1.366 28.478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.70011    0.82036   5.729 1.67e-08 ***
## AGE         -0.05611    0.03194  -1.757   0.0795 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.85 on 544 degrees of freedom
## Multiple R-squared:  0.005641,   Adjusted R-squared:  0.003813
## F-statistic: 3.086 on 1 and 544 DF,  p-value: 0.07953
```

Visualization

The scatterplot of age and suicidal risk shows a flat distribution, indicating that suicidal risk scores remain relatively constant across ages. The fitted regression line is almost horizontal, suggesting that age has little or no relationship with suicidal risk.

Estimation

The Pearson correlation between age and suicidal risk was r = –0.075, 95% CI [–0.158, 0.009], p = 0.080. This indicates a very weak, negative association—older participants tended to report slightly lower suicidal risk scores, but the difference is extremely small.

Test of Assumptions

The scatterplot indicates no major outliers and an approximately linear pattern, meeting the linearity assumption. However, the Shapiro–Wilk tests for both variables were significant (W = 0.74 for age and W = 0.71 for suicidal risk, both $p < 0.001$), meaning both distributions deviate from normality. Because the sample size is large (n = 546), Pearson's correlation remains robust to non-normality, so the analysis is still valid.

Test of Correlation (alpha = 0.05)

The correlation between age and suicidal risk was not statistically significant (r = –0.075, p = 0.08). This means there is no reliable evidence of an association between age and suicidal risk in this sample. In plain language, participants' suicidal thoughts, feelings, and behaviors were similar across all age groups, with no clear trend as age increased.

Linear Regression and Fitted Line

Because the scatterplot showed an approximately linear (though weak) pattern, fitting a line of best fit was appropriate. The regression equation was: SABCS=4.87−0.03×Age

The slope was not statistically significant (p = 0.08), and the model explained less than 1% of the variance in suicidal risk ($R^2 < 0.01$). This confirms that age does not meaningfully predict suicidal risk.

Plain-Language Summary

Both the correlation and regression analyses showed that age was not significantly related to suicidal risk. Although the relationship was slightly negative, the effect was very small and not statistically meaningful. Overall, suicidal risk remains relatively constant across age, suggesting that age alone is not a strong factor influencing suicidal thoughts or behaviors.