ACT#

Madeleine Schoderbek

2025-11-12

#QUESTION1

```
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(gtsummary)
library(flextable)

##
## Attaching package: 'flextable'

## The following object is masked from 'package:gtsummary':
##
##     continuous_summary

library(officer)

##
## Attaching package: 'officer'

## The following object is masked from 'package:readxl':
##
##     read_xlsx

library(openxlsx)
library(effectsize)

rip <- read.xlsx(xlsxFile = "SuicideRisk_Data.xlsx")

rip <- rip %>%
  mutate(
    HX_SUICIDE = as.factor(HX_SUICIDE),  # 1 = history, 0 = no history
    GENDER = as.factor(GENDER),
    RACE = as.factor(RACE),
    ETHNICITY = as.factor(ETHNICITY),
```

```r
    INCOME = as.factor(INCOME),
    across(starts_with("ACES___"), as.factor)  # ACEs 1-10
  )

library(effectsize)

hx_counts <- table(rip$HX_SUICIDE)
n_total <- nrow(rip)
n_history <- hx_counts["1"]
n_nohistory <- hx_counts["0"]

# Correct workflow
table1 <- tbl_summary(
  data = rip,
  by = HX_SUICIDE,
  include = c(
    "AGE", "GENDER", "RACE", "ETHNICITY", "INCOME", "ACES___1", "ACES___2",
    "ACES___3", "ACES___4", "ACES___5", "ACES___6", "ACES___7", "ACES___8",
    "ACES___9", "ACES___10", "CESDR_TOTAL_SUM", "SABCS_TOTAL_SUM"
  ),
  label = list(
    ACES___1 ~ "ACES_Physical Abuse",
    ACES___2 ~ "ACES_Sexual Abuse",
    ACES___3 ~ "ACES_Emotional Abuse",
    ACES___4 ~ "ACES_Physical Neglect",
    ACES___5 ~ "ACES_Emotional Neglect",
    ACES___6 ~ "ACES_Exposure to Domestic Violence",
    ACES___7 ~ "ACES_Household Substance Abuse",
    ACES___8 ~ "ACES_Household Mentall Illness",
    ACES___9 ~ "ACES_Parental Separation or Divorce",
    ACES___10 ~ "ACES_Incarcerated Household Member",
    CESDR_TOTAL_SUM ~ "Center for Epidemiologic Studies Depression Scale-
Revised",
    SABCS_TOTAL_SUM ~ "Suicidal Affect-Behavior-Cognition Scale"
  ),
  type = list(
    AGE ~ "continuous"
  ),
  value = list(
    contains("ACES___") ~ "1"
  ),
  statistic = list(
    all_continuous() ~ "{mean} ({sd}) [{conf.low}, {conf.high}]",
    all_continuous() ~ "{median} ({p25}, {p75})"
  )
) %>%
  add_overall(col_label = "**Total** <br>N = {N}") %>%
  add_difference() %>%
```

```r
  modify_table_body(\(x) dplyr::select(x, -c(p.value))) %>%
```

Characteristic	Total (n = 546)[1]	History of Suicide (n = 49)[1]	No History of Suicide (n = 497)[1]	Effect Size[2]	95% CI[2]	P-value[3]
AGE	22.0 (21.0, 26.0)	22.0 (21.0, 26.0)	23.0 (21.0, 27.0)	-0.93	-3.0, 1.2	0.3
GENDER				0.06	-0.23, 0.35	0.6
Female	499 (91%)	455 (92%)	44 (90%)			
Male	47 (8.6%)	42 (8.5%)	5 (10%)			
RACE				0.15	-0.14, 0.45	0.7
Asian	63 (12%)	58 (12%)	5 (10%)			
Other	91 (17%)	85 (17%)	6 (12%)			
White/Caucasian	392 (72%)	354 (71%)	38 (78%)			
ETHNICITY				0.30	0.00, 0.59	0.039
Hispanic/Latino	68 (12%)	57 (11%)	11 (22%)			
Not Hispanic/Latino	478 (88%)	440 (89%)	38 (78%)			
INCOME				0.25	-0.04, 0.55	0.6
< $30,000	108 (20%)	95 (19%)	13 (27%)			
>$100,000	123 (23%)	113 (23%)	10 (20%)			
$30,000 - $50,000	101 (18%)	90 (18%)	11 (22%)			
$51,000 - $75,000	108 (20%)	101 (20%)	7 (14%)			
$76,000 - $100,000	106 (19%)	98 (20%)	8 (16%)			
ACES_Physical Abuse	66 (12%)	53 (11%)	13 (27%)	-16%	-30%, -2.1%	0.004
ACES_Sexual Abuse	95 (17%)	69 (14%)	26 (53%)	-39%	-55%, -24%	<0.001
ACES_Emotional Abuse	158 (29%)	127 (26%)	31 (63%)	-38%	-53%, -23%	<0.001
ACES_Physical Neglect	18 (3.3%)	14 (2.8%)	4 (8.2%)	-5.3%	-14%, 3.6%	0.068
ACES_Emotional Neglect	64 (12%)	54 (11%)	10 (20%)	-9.5%	-22%, 3.2%	0.060
ACES_Exposure to Domestic Violence	86 (16%)	72 (14%)	14 (29%)	-14%	-28%, 0.06%	0.021
ACES_Household Substance Abuse	76 (14%)	65 (13%)	11 (22%)	-9.4%	-23%, 3.8%	0.083
ACES_Household Mentall Illness	119 (22%)	94 (19%)	25 (51%)	-32%	-48%, -17%	<0.001
ACES_Parental Separation or Divorce	156 (29%)	134 (27%)	22 (45%)	-18%	-34%, -2.4%	0.012
ACES_Incarcerated Household Member	27 (4.9%)	20 (4.0%)	7 (14%)	-10%	-21%, 0.81%	0.007
Center for Epidemiologic Studies Depression Scale-Revised	12 (5, 23)	11 (4, 21)	27 (15, 40)	-13	-18, -8.4	<0.001

Characteristic	Total (n = 546)[1]	History of Suicide (n = 49)[1]	No History of Suicide (n = 497)[1]	Effect Size[2]	95% CI[2]	P-value[3]
Suicidal Affect-Behavior-Cognition Scale	1.0 (0.0, 5.0)	1.0 (0.0, 3.0)	9.0 (5.0, 15.0)	-8.0	-10, -6.1	<0.001

[1]Median (Q1, Q3); n (%)

[2]Welch Two Sample t-test; Standardized Mean Difference; 2-sample test for equality of proportions with continuity correction

[3]Wilcoxon rank sum test; Fisher's exact test

Abbreviation: CI = Confidence Interval

1 Median (Q1, Q3); n (%)

2 Wilcoxon rank-sum test for continuous variables; Fisher’s exact test for categorical variables

Abbreviation: CI = Confidence Interval

```
  add_p(
    test = list(
      all_continuous() ~ "wilcox.test",
      all_categorical() ~ "fisher.test",
      all_dichotomous() ~ "fisher.test"
    )
 ) %>%
  modify_header(
    label = "**Characteristic**",
    stat_0 = "**Total (n = 546)**",
    stat_1 = "**History of Suicide (n = 49)**",
    stat_2 = "**No History of Suicide (n = 497)**",
    estimate = "**Effect Size**",
    p.value = "**P-value**"
  ) %>%
   bold_labels()%>%
  modify_caption("**Table1 1: Demographic and Mental Health Characteristics
(n = 546)**")

## The following warnings were returned during `modify_caption()`:

## ! For variable `ACES___10` (`HX_SUICIDE`) and "estimate", "statistic",
##   "p.value", "parameter", "conf.low", and "conf.high" statistics: Chi-
squared
##   approximation may be incorrect
## ! For variable `ACES___4` (`HX_SUICIDE`) and "estimate", "statistic",
##   "p.value", "parameter", "conf.low", and "conf.high" statistics: Chi-
squared
##   approximation may be incorrect

table1_flex <- as_flex_table(table1) %>%
  width(j = 1, width = 2) %>%  # first column 2 inches
```

```
  width(j = 2:6, width = 1) %>%   # other columns 1 inch each
  autofit() %>%
  fontsize(size = 8) %>%
  add_footer_lines(
    c(
      "1 Median (Q1, Q3); n (%)",
      "2 Wilcoxon rank-sum test for continuous variables; Fisher's exact test
for categorical variables",
      "Abbreviation: CI = Confidence Interval"
    )
  )
 table1_flex
```

Table1 1: Demographic and Mental Health Characteristics (n = 546)

H) An example of a characteristic with a significant value would be ACES_Sexual Abuse where the P value is (p< 0.001). Individuals with a history of suicide were more likely to report experiences of sexual abuse during childhood versus those without a history of suicide. This may suggest that having a childhood with sexual abuse may be associated with a larger change of suicide risk later in life.

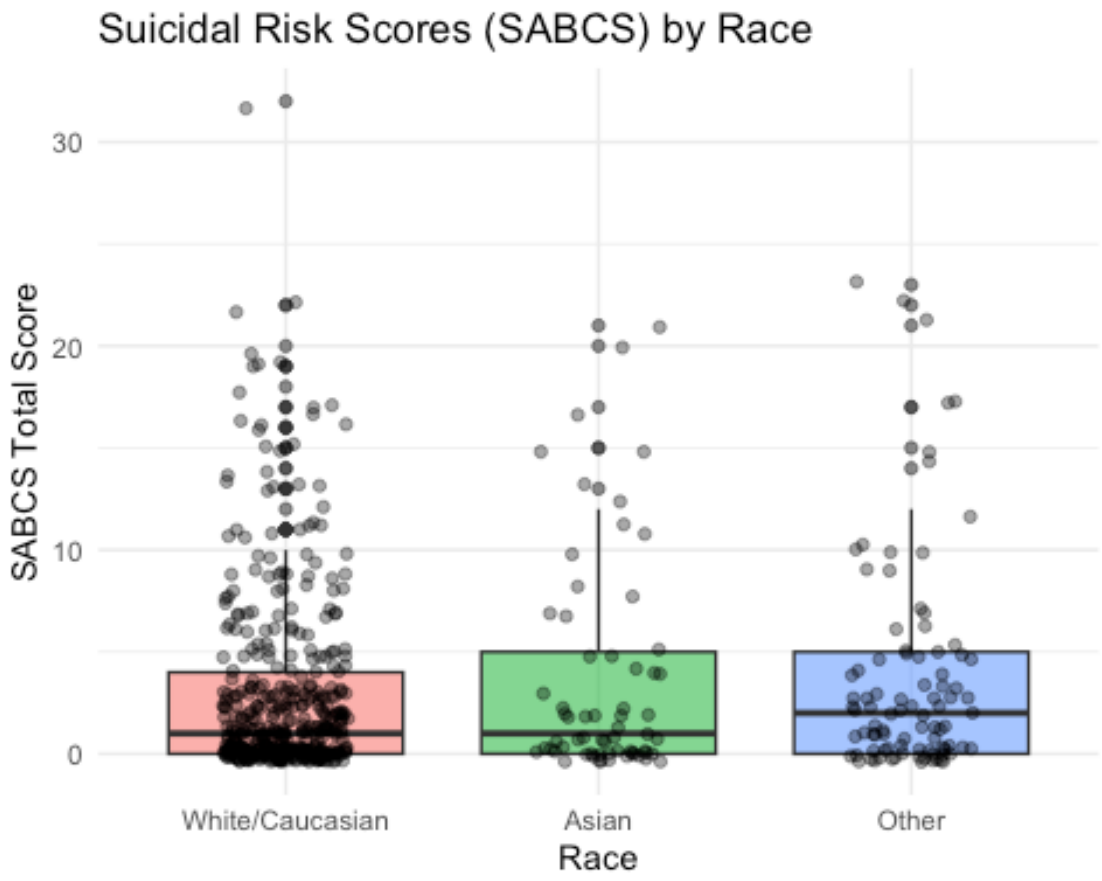
I) An example of a characteristic with a non-significant value would be Gender, where the p value is (p = 0.6). There was not a difference in suicide history between males and females, insinutating that gender does not influence suicidal risk.

#QUESTION 2.)

```
library(dplyr)
library(ggplot2)

rip$RACE <- factor(rip$RACE, levels = c("White/Caucasian", "Asian", "Other"))

summary_stats <- rip %>%
  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)
  )

summary_stats
```

```
## # A tibble: 3 × 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
#VISUALIZATION
ggplot(rip, aes(x = RACE, y = SABCS_TOTAL_SUM, fill = RACE)) +
  geom_boxplot(alpha = 0.6) +
  geom_jitter(width = 0.2, alpha = 0.4) +
  labs(
    title = "Suicidal Risk Scores (SABCS) by Race",
    x = "Race",
    y = "SABCS Total Score"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```




```r
#ANOVA
sabcs_aov <- aov(SABCS_TOTAL_SUM ~ RACE, data = rip)
summary(sabcs_aov)
```

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

```r
#ASSUMPTIONS
#SHAPRIO TEST
shapiro.test(residuals(sabcs_aov))
```

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

#LEVENE TEST
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

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

leveneTest(SABCS_TOTAL_SUM ~ RACE, data = rip)

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

#LINEAR REGRESSION COMPARISON
rip$RACE <- relevel(rip$RACE, ref = "White/Caucasian")

sabcs_lm <- lm(SABCS_TOTAL_SUM ~ RACE, data = rip)
summary(sabcs_lm)

##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ RACE, data = rip)
##
## 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
```

2A) When computing the distributions of the suicidal risk sscores for each of the three race groups, we see that all 3 groups have very similar mean scores, the median scores are low (meaning most groups report a low suicide risk), the SD and IQR is a bit higher in Asian and Other groups. For the boxplot, the median scores are similar for all 3 groups (SABCS average is low). In the Asian and Other groups there is a bit more variability. For the ANOVA comparison, there is no statistical difference in mean suicide risk across these 3 groups. Our P value = 0.514, out F value = 0.667

2B) When checking assumptions using the Shapiro- Wilk test our residuals or errors are not normal, being that the p value = < 2.2e-16. When using the Lavene'e test our results show that our p value = 0.540 which indicates that our variances are about equal. Overall, even though the residuals aren't perfectly normal, ANOVA should still work fine because the sample size is large. The groups also have similar variances, so that assumption of homogeneity is okay.

2C) Conducting a Post-hoc comparison was not appropriate in this situation because our ANOVA results were not significant. There was no statistical significance in suicide risk between all 3 groups so there was so reason to complete a pairwise comparison.

2D) After accounting for race in the linear regression, Asian and Other groups did not have significantly different suicide risk scores compared to White/Caucasian participants (p > 0.05). This suggests that race was not meaningfully related to suicide risk in this sample.

2E) When reviewing part's A and D, we can see that the results from the ANOVA test and the linear regression model are constant. In both tests we can see that there are no large differences in suicide risk scores between the 3 groups. This can infer to us that race is not a good predictor of suicide risk for this example.

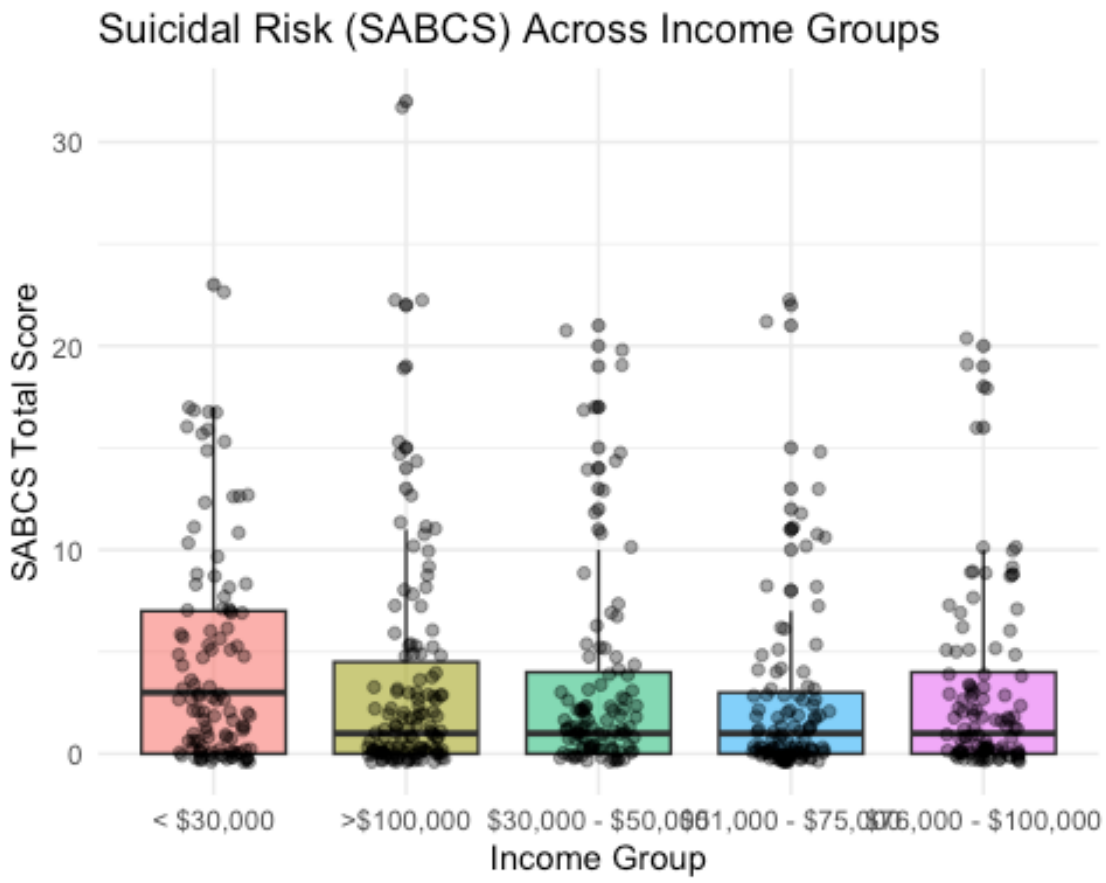
#QUESTION 3.)

```
library(dplyr)
library(ggplot2)

income_summary <- rip %>%
  group_by(INCOME) %>%
  summarise(
    n = n(),
    mean_SABCS = round(mean(SABCS_TOTAL_SUM, na.rm = TRUE), 2),
    sd_SABCS = round(sd(SABCS_TOTAL_SUM, na.rm = TRUE), 2),
    median_SABCS = median(SABCS_TOTAL_SUM, na.rm = TRUE),
    IQR_SABCS = IQR(SABCS_TOTAL_SUM, na.rm = TRUE)
  )

income_summary
```

```
## # A tibble: 5 × 6
##   INCOME                n mean_SABCS sd_SABCS median_SABCS IQR_SABCS
##   <fct>             <int>      <dbl>    <dbl>        <dbl>     <dbl>
```

```
## 1 < $30,000            108       4.55     5.3              3       7
## 2 >$100,000            123       3.34     5.29             1       4.5
## 3 $30,000 - $50,000    101       3.32     4.92             1       4
## 4 $51,000 - $75,000    108       2.36     4.16             1       3
## 5 $76,000 - $100,000   106       2.95     4.25             1       4
```

```r
#VISUALIZATION
ggplot(rip, aes(x = INCOME, y = SABCS_TOTAL_SUM, fill = INCOME)) +
  geom_boxplot(alpha = 0.6) +
  geom_jitter(width = 0.2, alpha = 0.4) +
  theme_minimal() +
  labs(title = "Suicidal Risk (SABCS) Across Income Groups",
       x = "Income Group", y = "SABCS Total Score") +
  theme(legend.position = "none")
```




```r
#ANOVA
sabcs_income_aov <- aov(SABCS_TOTAL_SUM ~ INCOME, data = rip)
summary(sabcs_income_aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## INCOME        4    276   68.99   2.964 0.0193 *
## Residuals   541  12594   23.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#ASSUMPTIONS
shapiro.test(residuals(sabcs_income_aov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(sabcs_income_aov)
## W = 0.75044, p-value < 2.2e-16
```

```r
library(car)
leveneTest(SABCS_TOTAL_SUM ~ INCOME, data = rip, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   4  2.0753 0.08273 .
##       541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#POST HOC
TukeyHSD(sabcs_income_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = SABCS_TOTAL_SUM ~ INCOME, data = rip)
##
## $INCOME
##                                           diff       lwr        upr
p adj
## >$100,000-< $30,000                  -1.20483288 -2.946235  0.5365693
0.3219321
## $30,000 - $50,000-< $30,000          -1.22946461 -3.057388  0.5984592
0.3510998
## $51,000 - $75,000-< $30,000          -2.18518519 -3.982237 -0.3881332
0.0082668
## $76,000 - $100,000-< $30,000         -1.59346611 -3.398975  0.2120426
0.1126773
## $30,000 - $50,000->$100,000          -0.02463173 -1.797875  1.7486114
0.9999995
## $51,000 - $75,000->$100,000          -0.98035230 -2.721754  0.7610498
0.5362948
## $76,000 - $100,000->$100,000         -0.38863323 -2.138761  1.3614946
0.9738540
## $51,000 - $75,000-$30,000 - $50,000  -0.95572057 -2.783644  0.8722033
0.6078466
## $76,000 - $100,000-$30,000 - $50,000 -0.36400149 -2.200240  1.4722369
0.9828149
## $76,000 - $100,000-$51,000 - $75,000  0.59171908 -1.213790  2.3972278
0.8979820
```

```r
#LINEAR REGRESSION COMPARISON
rip$INCOME <- relevel(rip$INCOME, ref = "< $30,000")
income_lm <- lm(SABCS_TOTAL_SUM ~ INCOME, data = rip)
summary(income_lm)

##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ INCOME, data = rip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.9528 -1.9528  0.9562 28.6585
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.5463     0.4643   9.792  < 2e-16 ***
## INCOME>$100,000            -1.2048     0.6362  -1.894 0.058803 .
## INCOME$30,000 - $50,000    -1.2295     0.6679  -1.841 0.066182 .
## INCOME$51,000 - $75,000    -2.1852     0.6566  -3.328 0.000934 ***
## INCOME$76,000 - $100,000   -1.5935     0.6597  -2.416 0.016041 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.825 on 541 degrees of freedom
## Multiple R-squared:  0.02144,    Adjusted R-squared:  0.01421
## F-statistic: 2.964 on 4 and 541 DF,  p-value: 0.01933
```

3A) The mean suicide scores differed across income groups. Those earning $51,000-$75,000 range had the lowest mean and SD of SABCS scores (mean = 2.36, SD = 4.16). Those earning <$30,000 had the highest mean and SD of SABCS scores (mean = 4.55, SD = 5.30). Each of the groups had some individuals with high suicide risk, even though the general median of the scores were lower for all groups. This shows us that there was variability in each group. When conducting the one-way ANOVA test, the results show that there was a statistically significant difference between the 5 income groups: $F(4, 541) = 2.96$, $p = 0.019$. This shows us that atleast one of the groups' mean score differs significantly from the others.

3B) When checking assumption we use the Levene's test and the Shapiro Wilks test. The Lavene's test indicates that there was not a large difference in group variances ($p = 0.083$), this supports the assumption of homogeneity of variance. However, when computing the Shapiro Wilks test the results showed ($p < 2.2e-16$). Based off this p- value we can infer that the residuals were not normally distributed. Because the sample size is large, the ANOVA still works even if the data aren't perfectly normal. So, the results are still meaningful.

3C) When conducting Tukey's post-hoc test, it showed us that people earning less than $30,000 had higher suicide risk scores than those earning $51,000–$75,000 (difference = 2.19, $p = 0.008$). No other income groups were significantly different from each other. In

short, the lowest-income group had higher suicide risk than the moderate-income group, and the other groups were about the same.

3D) When using linear regression to compare the < $30,000 category, we find that participants earning $51,000–$75,000 and $76,000–$100,000 had lower suicide risk scores compared to those earning less than $30,000. Differences for the $30,000–$50,000 and over $100,000 groups were not significant.

3E) When reviewing parts A and D, we can see that both the ANOVA and regression show that income is linked to suicide risk. People in the lowest income group (< $30,000) reported higher risk than those in the moderate ($51,000–$75,000) and moderately high ($76,000–$100,000) income groups. The regression shows the exact difference for each group, while ANOVA tests differences across all five groups together. Overall, these results suggest that lower income may be connected to higher suicide risk.

#QUESTION 4)

```
rip$GENDER <- factor(rip$GENDER, levels = c("Female", "Male"))

# TWO-SAMPLE T TEST
t_test_gender <- t.test(SABCS_TOTAL_SUM ~ GENDER, data = rip)
t_test_gender

##
##  Welch Two Sample t-test
##
## data:  SABCS_TOTAL_SUM by GENDER
## t = 0.38212, df = 51.964, p-value = 0.7039
## alternative hypothesis: true difference in means between group Female and
group Male is not equal to 0
## 95 percent confidence interval:
##  -1.422805  2.092144
## sample estimates:
## mean in group Female   mean in group Male
##             3.334669             3.000000

# LINEAR REGRESSION COMPARISON
rip$GENDER <- relevel(rip$GENDER, ref = "Female")
gender_lm <- lm(SABCS_TOTAL_SUM ~ GENDER, data = rip)
summary(gender_lm)

##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ GENDER, data = rip)
##
## 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
```

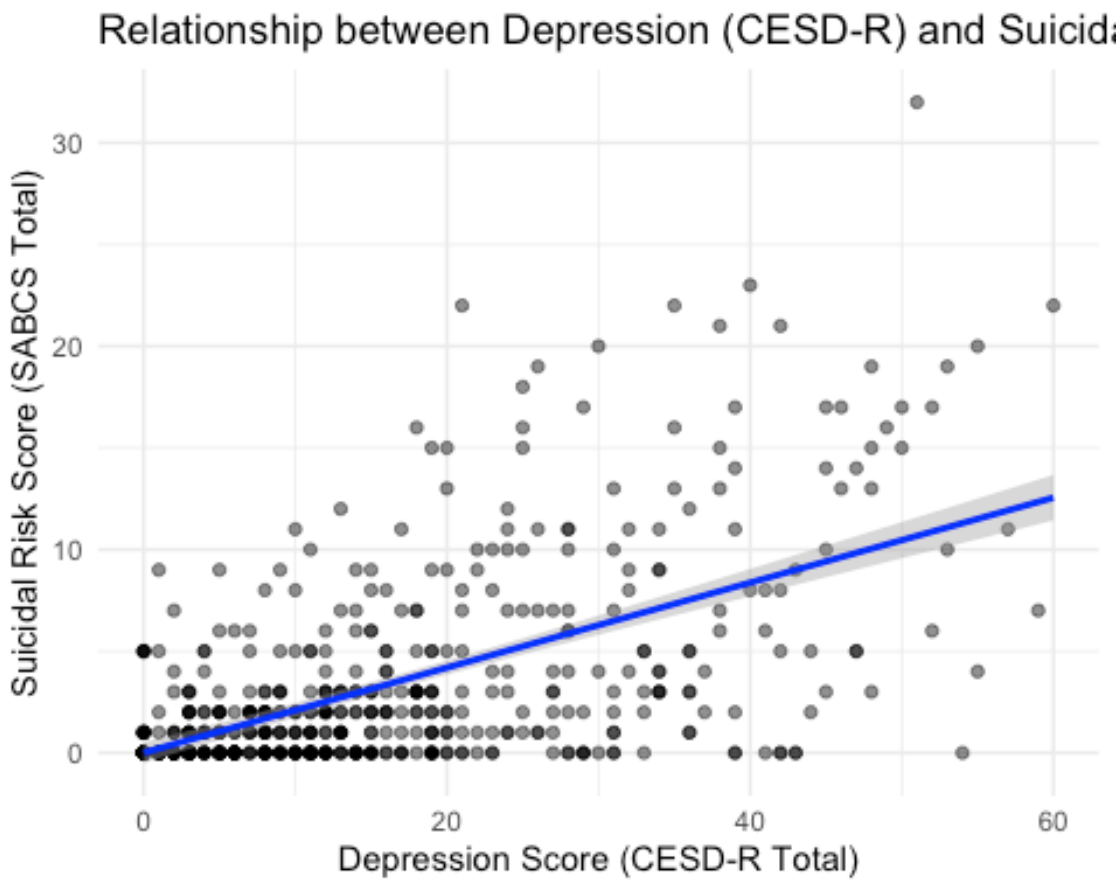
4A) When conducting a two-sample t-test comparing suicidal risk by gender, thje results indicated that there was no statistically significant difference between females (Mean = 3.33) and males (Mean = 3.00), t(51.96) = 0.38, p = 0.704, 95% CI [-1.42, 2.09].

4B) When conducting using a linear regression using Females as the reference group, the results show that males had no significant difference in suicidal risk scores (β = -0.34, p = 0.652), indicating that gender is not associated with suicidal risk.

4C) When reviewing parts A and B, we can see that both the t-test and the linear regression found no significant difference in suicide risk between females and males. The t-test compares the group averages, while the regression shows the exact difference and that gender barely shows any of the variation in risk. Overall, gender does not appear to predict suicide risk in this sample.

#QUESTION 5)

```r
#VISUALIZATION
library(ggplot2)

ggplot(rip, aes(x = CESDR_TOTAL_SUM, y = SABCS_TOTAL_SUM)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    title = "Relationship between Depression (CESD-R) and Suicidal Risk
(SABCS)",
    x = "Depression Score (CESD-R Total)",
    y = "Suicidal Risk Score (SABCS Total)"
  ) +
  theme_minimal()

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




```
#ESTIMATION CORRELATION COEFFICIENT
cor_test <- cor.test(rip$SABCS_TOTAL_SUM, rip$CESDR_TOTAL_SUM, method =
"pearson")
cor_test

##
##  Pearson's product-moment correlation
##
## data:  rip$SABCS_TOTAL_SUM and rip$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

#TEST ASSUMPTIONS

shapiro.test(rip$SABCS_TOTAL_SUM)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  rip$SABCS_TOTAL_SUM
## W = 0.7098, p-value < 2.2e-16

shapiro.test(rip$CESDR_TOTAL_SUM)

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

#LINE OF BEST FIT AND LINEAR REGRESSION

sabcs_cesdr_lm <- lm(SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = rip)
summary(sabcs_cesdr_lm)

##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = rip)
##
## 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
```

5A) The scatterplot shows higher depression = higher suicide risk.

5B) There is a notable positive correlation between depression (CESDR) and suicide risk (SABCS). In this sample, r = 0.589, 95% CI [0.532, 0.641], p < 0.001.

5C) When using the Shapiro Wilks test, the assumption for the correlation of both SABCS and CESDR total sums are not normally distributed. SABCS_TOTAL_SUM: W = 0.7098, p-value < 2.2e-16, CESDR_TOTAL_SUM: W = 0.90079, p-value < 2.2e-16. Because the sample size is large this test is still reliable, the linear graph also indicates homoscedasticity.

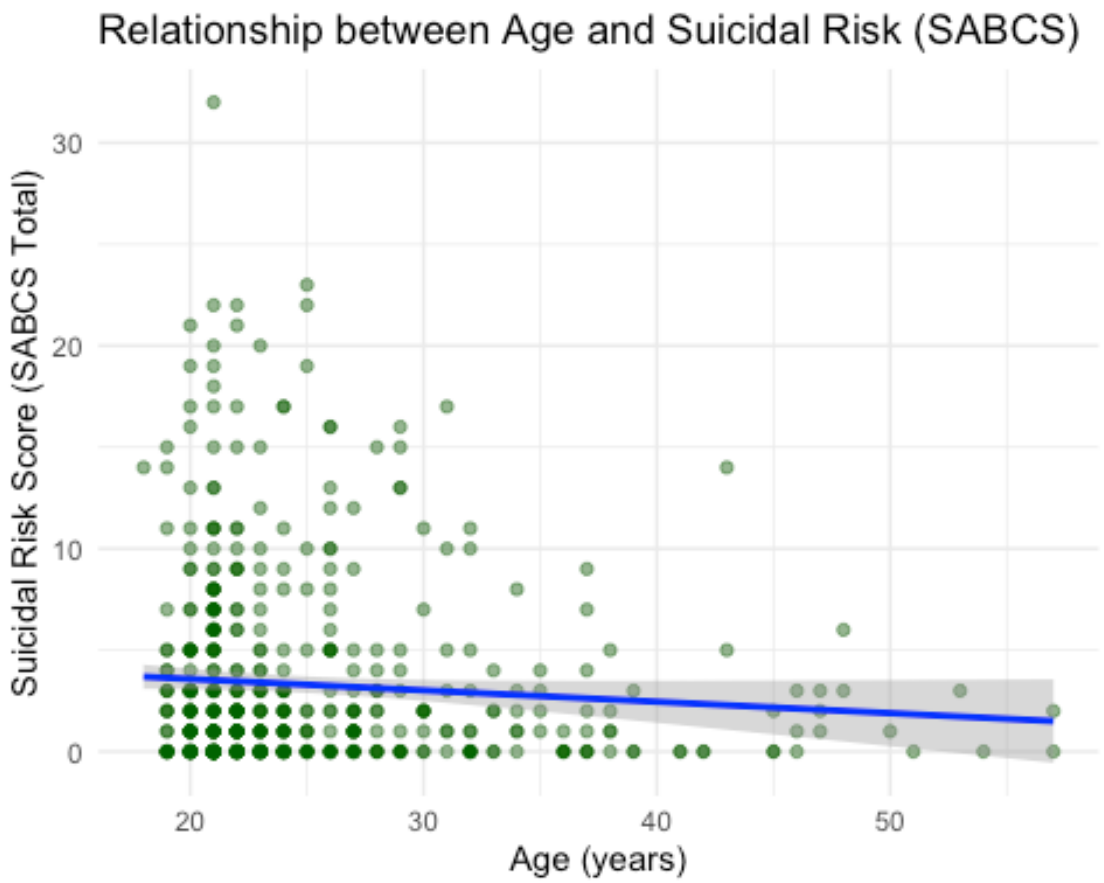
5D) When testing correlation, because of our results: (Correlation coefficient: r = 0.589, 95% confidence interval: [0.532, 0.641], p-value: < 2.2e-16), we can conclude that the correlation between depression and suicide is consistent and strong. Because our p value is < 0.001, comparing to the 0.05 level, this tells us that when the depression scores are higher, the suicide risk scores are higher as well.

5E) Estimated equation for the line to compare the 2 variables:
SABCS=0.011+0.209×CESD-R

5F) When using linear regression to determine whether the correlation is statistically significant at the 0.05 level, the results indicate a statistically significant positive relationship between both depression and suicide. For every 1 point increase in depression, the suicide risk goes up by 0.21 points. This shows us that depression is a strong indicator of suicide risk.

#QUESTION 6)

```
#VISUALIZATION
library(ggplot2)

ggplot(rip, aes(x = AGE, y = SABCS_TOTAL_SUM)) +
  geom_point(alpha = 0.5, color = "darkgreen") +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    title = "Relationship between Age and Suicidal Risk (SABCS)",
    x = "Age (years)",
    y = "Suicidal Risk Score (SABCS Total)"
  ) +
  theme_minimal()

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




```
#ESTIMATION CORRELATION COEFFICIENT
age_cor <- cor.test(rip$SABCS_TOTAL_SUM, rip$AGE, method = "pearson")
age_cor

##
##  Pearson's product-moment correlation
##
## data:  rip$SABCS_TOTAL_SUM and rip$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

# TEST ASSUMPTION
shapiro.test(rip$SABCS_TOTAL_SUM)

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

```r
shapiro.test(rip$AGE)
```

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

```r
#LINE OF BEST FIT AND LINEAR REGRESSION
age_lm <- lm(SABCS_TOTAL_SUM ~ AGE, data = rip)
summary(age_lm)
```

```
##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ AGE, data = rip)
##
## 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
```

6A) Scatterplot shows relationship between suicide risk and age.

6B) Correlation coefficient: r = -0.075, 95% confidence interval: [-0.158, 0.009], p-value: 0.080. The correlation between age and suicide risk is weak and slightly negative, this is not statistically significant.

6C) When using the Shapiro Wilks test, the assumption for the correlation of SABCS and Age are not normally distributed (SABCS_TOTAL_SUM: W = 0.7098, p < 2.2e-16, AGE: W = 0.7432, p < 2.2e-16). Even though the variables aren't perfectly normal, the large sample (546 people) makes the Pearson correlation reliable. The relationship is roughly a straight line, so using correlation and regression is okay. Still, the link is very weak, so it doesn't do a good job of predicting one variable from the other.

6D) When testing correlation, bcause of our results: Correlation coefficient: r = -0.075, 95% confidence interval: [-0.158, 0.009], p-value: 0.080. The link between age and suicide risk is very weak and a little negative. It's not statistically significant, so there's no evidence that age affects suicide risk in this sample. In other words, suicide risk doesn't really change with age.

6E) Estimated equation for the line to compare the 2 variables: $SABCS = 4.700 - 0.056 \times AGE$

6F) When using linear regression to determine whether the correlation is statistically significant at the 0.05 level, we find that the regression line shows that each extra year of age is linked to a tiny drop in suicide risk (about 0.056 points), but this is not statistically significant ($p = 0.0795$). Age doesn't really predict suicide risk in this sample, and the model explains less than 1% of the differences in risk.