Stat Final Project

Sheng-Lien Lee, Sijia yu, Simao Luis

2023-12-14

Q1 Data Discription:
Dataset:
Our group project aims to analyze the impact of various factors on mental health outcomes, specifically focusing on depression among adolescents. For this, we've selected the 2017-2018 National Health and Nutrition Examination Survey (NHANES) dataset, which comprehensively views health-related variables within the U.S. population. Our target demographic is adolescents aged above 18. There are mainly five parts which consists of the final datasets: 1. Demographic variables dataset, including respondent sequence number, age, ethnicity, etc. 2. Depression related dataset, including levels of depression, levels of poor appetite, etc. 3. Physical activity dataset, including vigorously working or not, physical activities like walking, etc. 4. Alcohol consumption dataset, which we only use the past 12 months alcohol intake variable. 5. Sleep condition dataset, which we mainly want a further analyzation of the explicit sleep hours for the population. After choosing the five datasets from NHANES, we combined them into one by respondent sequence number and chose the specific variables that are related with our research question.

```
#Data import
library(haven)
DEMO <- read_xpt('DEMO_J.XPT')
DPQ <- read_xpt('DPQ_J.XPT')
SLQ <- read_xpt('SLQ_J.XPT')
PAQ <- read_xpt('PAQ_J.XPT')
ALQ <- read_xpt('ALQ_J.XPT')
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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 become errors
```

```
join_1 <- left_join(DEMO, DPQ, by='SEQN')
join_2 <- left_join(join_1, SLQ, by='SEQN')
join_3 <- left_join(join_2, PAQ, by='SEQN')
join_4 <- left_join(join_3, ALQ, by='SEQN')
concise_data <- join_4[,c('SEQN', 'RIDAGEYR', 'RIAGENDR', 'RIDRETH1', 'INDFMIN2', 'DPQ020',
                          'DPQ030', 'DPQ050', 'PAQ605', 'PAQ635', 'ALQ121', 'SLD012')]
```

Data Cleaning

For data cleaning, we firstly filtered the entire datasets for the condition that we only need respondents that are aged 18 and above, which results in a reduction from 9254 respondents to 5856 respondents. Next, since this is a large dataset, it makes no sense to replace the NAs for different levels of the variables (such as depression level, sleep condition, appetite condition, etc. It makes no sense to replace them with the mean value or median value). We decide to delete all rows with no response values for the variables, which results in a reduction from 5856 respondents to 4292 respondents. Lastly, we add the explanatory variables for each of the indexing variables before, including family income level, alcohol usage condition, appetite condition, etc.

```
#Data cleaning
step_1 <- concise_data[concise_data$RIDAGEYR>=18,]
step_1$race <- factor(step_1$RIDRETH1, levels = c(1,2,3,4,5), labels = c('Mexican American',
                                                                         'Other Hispanic',
                                                                         'White',
                                                                         'Black',
                                                                         'Other Race'))
step_2 <- step_1[!is.na(step_1$INDFMIN2),]
step_3 <- step_2[!is.na(step_2$DPQ020),]
step_4 <- step_3[!is.na(step_3$DPQ030),]
step_5 <- step_4[!is.na(step_4$DPQ050),]
step_6 <- step_5[!is.na(step_5$PAQ605),]
step_7 <- step_6[!is.na(step_6$PAQ635),]
step_8 <- step_7[!is.na(step_7$ALQ121),]
step_9 <- step_8[!is.na(step_8$SLD012),]
step_9$gender <- factor(step_9$RIAGENDR, levels = c(1,2), labels = c('Male', 'Female'))
step_9$fmaily_income <- factor(step_9$INDFMIN2, levels = c(1,2,3,4,5,6,7,8,9,10,12,13,14,15,77,99),
                               labels = c('$0 to $4,999', '$5,000 to $9,999', '$10,000 to $14,999', '$15
                                          '$25,000 to $34,999', '$35,000 to $44,999', '$45,000 to $54,99
                                          '$20,000 and Over', 'Under $20,000', '$75,000 to $99,999', '$
step_9$poor_appetite <- factor(step_9$DPQ050, levels = c(0,1,2,3,7,9), labels = c('Not at all', 'Several
                                                                                  'Refused', 'Dont know'))
step_9$sleep_trouble <- factor(step_9$DPQ030, levels = c(0,1,2,3,7,9), labels = c('Not at all', 'Several
                                                                                  'Refused', 'Dont know
step_9$working_vigorously <- factor(step_9$PAQ605, levels = c(1,2,7,9), labels = c('Yes', 'No', 'Refused
step_9$alcohol_use <- factor(step_9$ALQ121, levels = c(0,1,2,3,4,5,6,7,8,9,10,77,99), labels = c('Never
                                                                                                 '2 time
                                                                                                 '3 to
step_9$physical_exercise <- factor(step_9$PAQ635, levels = c(1,2,7,9), labels = c('Yes', 'No', 'Refused
Final_data <- step_9[!(step_9$DPQ020 %in% c(7, 9)), ]
```

Descriptive Statistics

For descriptive statistics, we firstly show the summary statistics for some important factors that we want to learn about that might correlated depression levels, including the mean level of depression, median of the sleep hours, and the mode of the family income level and the sleep hours, to give audiences some knowledge about the chosen respondents of our dataset. As the result suggested, the mean level of depression of the sample population is fairly low, which explains that depression is not usual for most of the population. Specifically for depression level, we also do the frequency table and the coreesponding bar chart ("Depression Frequencybar chart") to it as visualization of the data for audiences. The income level of the sample population that appears the most $100,000 and above, which explains that the income level of the sample population is fairly wealthy. The median of the sleep hours of the sample population is 7.5 hours/day, which is a normal number of sleep hours during our daily lives.

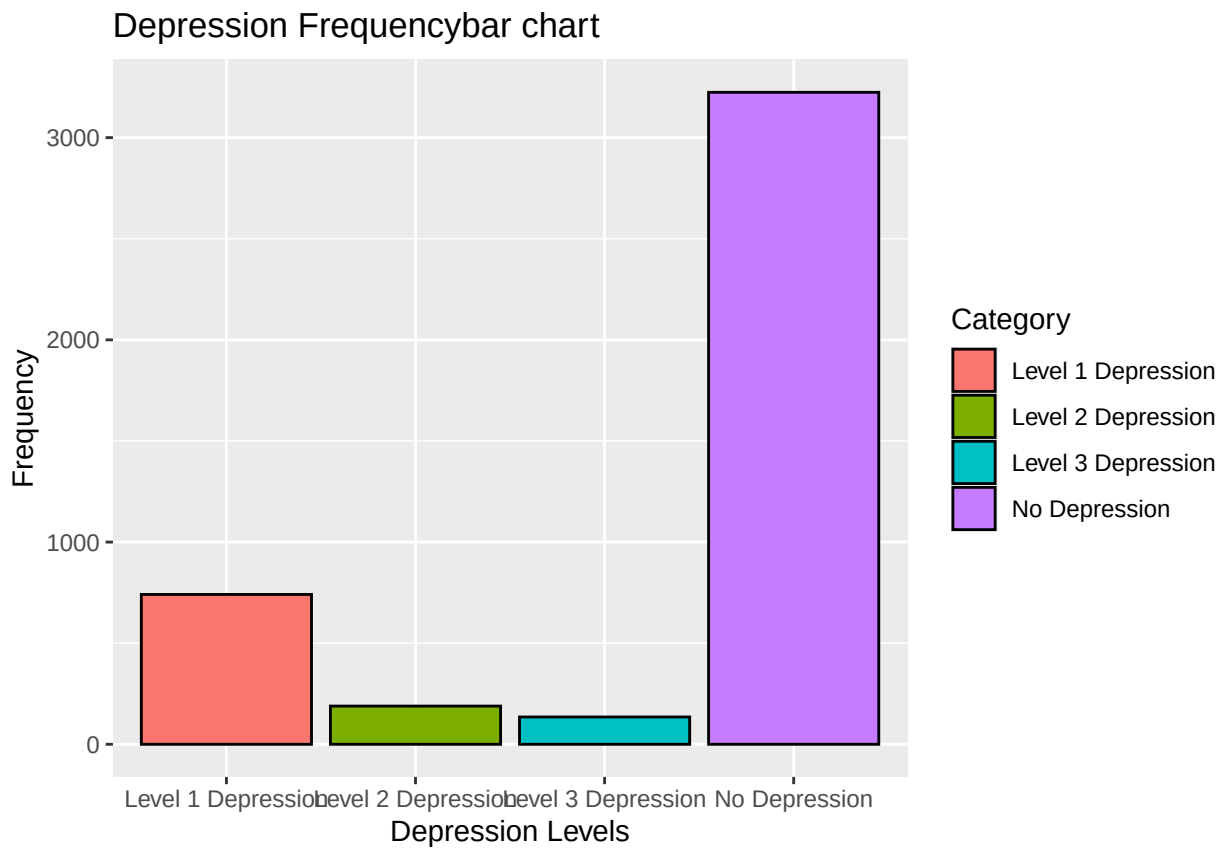
```r
#!!!!!!
Final_data = read_csv("Final_data.csv")
```

```
## New names:
## Rows: 4289 Columns: 21
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (8): race, gender, fmaily_income, poor_appetite, sleep_trouble, working... dbl
## (13): ...1, SEQN, RIDAGEYR, RIAGENDR, RIDRETH1, INDFMIN2, DPQ020, DPQ030...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
df_frequency <- table(Final_data$DPQ020)
relative_freq <- prop.table(df_frequency) * 100
freq_table <- data.frame(Category = c("No Depression", "Level 1 Depression", "Level 2 Depression", "Leve
                         Frequency = as.numeric(df_frequency),
                         Percentage = as.numeric(relative_freq))
print(freq_table)
```

```
##             Category Frequency Percentage
## 1      No Depression      3224  75.169037
## 2 Level 1 Depression       741  17.276754
## 3 Level 2 Depression       189   4.406622
## 4 Level 3 Depression       135   3.147587
```

```r
library(ggplot2)
ggplot(freq_table, aes(x = Category, y = Frequency, fill = Category)) + geom_bar(stat = "identity", posi
  labs(title = "Depression Frequencybar chart", x = "Depression Levels", y = "Frequency")
```




```
library(DescTools)
a <- mean(Final_data$DPQ020)
b <- sd(Final_data$DPQ020)
print(paste("Mean of depression level is:", a, "with a standard deviation of:", b))
```

```
## [1] "Mean of depression level is: 0.35532758218699 with a standard deviation of: 0.711460154527443"
```

```
c <- median(Final_data$SLD012)
print(paste("Median of sleep hours is:", c))
```

```
## [1] "Median of sleep hours is: 7.5"
```

```
d <- mean(Final_data$SLD012, trim=0.1)
print(paste("Trimed mean of depression level is:", d))
```

```
## [1] "Trimed mean of depression level is: 7.58141567142443"
```

```
e <- Mode(Final_data$INDFMIN2)
print(paste("Mode of family income level is:", e))
```

```
## [1] "Mode of family income level is: 15"
```

```r
f <- Mode(Final_data$SLD012)
print(paste("Mode of sleep hour is:", f))
```

```
## [1] "Mode of sleep hour is: 8"
```

Next, we do the 5 quantiles and the calculated IQR for the sleep hours variable. Since this is a critical variable that may have some influence on depression level, we want to make sure that the descriptive statistics for this variable makes sense. The IQR for sleep hours is 2, which results in a range of 6.5 hours to 8.5 hours of sleep as Q2 and Q4, which explains that the middle 50% of the observations corresponds to the common sleep hour range. Then, we also calculate the variance of the sleep hours variable and its corresponding coefficient of variation, which is 0.35. This illustrates that the dataset sleep hours is a little bit dispersed. In order to determine whether we need to do transformation of the data, we also calculate the skewness of the sleep hours data, which is -0.073. This explains that the data is a little bit left skewed, but approximately normal. Hence we decided not to transform this data. To further prove this, we add a histogram for the sleep hours dataset with its mean and median ("Histogram of Sleep Hours"). As can be seen from the distribution, it is approximately normally distributed (quantile plot "Quantile Plot of Sleep Hours").

```r
quantile(Final_data$SLD012)
```

```
##   0%  25%  50%  75% 100%
##  2.0  6.5  7.5  8.5 14.0
```

```r
IQR(Final_data$SLD012)
```

```
## [1] 2
```

```r
var(Final_data$SLD012)
```

```
## [1] 2.665194
```

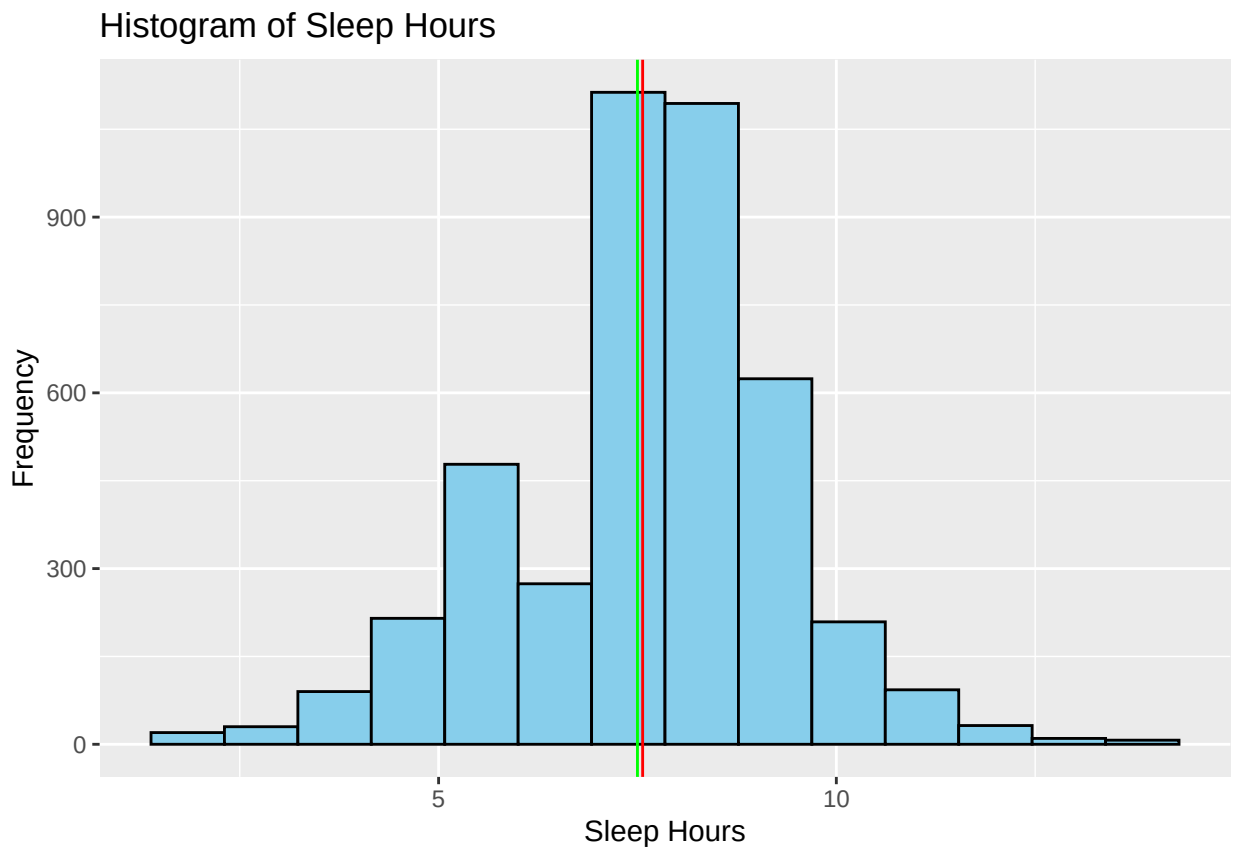
```r
CV_sleep_hours <- var(Final_data$SLD012)/mean(Final_data$SLD012)
print(CV_sleep_hours)
```

```
## [1] 0.3523035
```

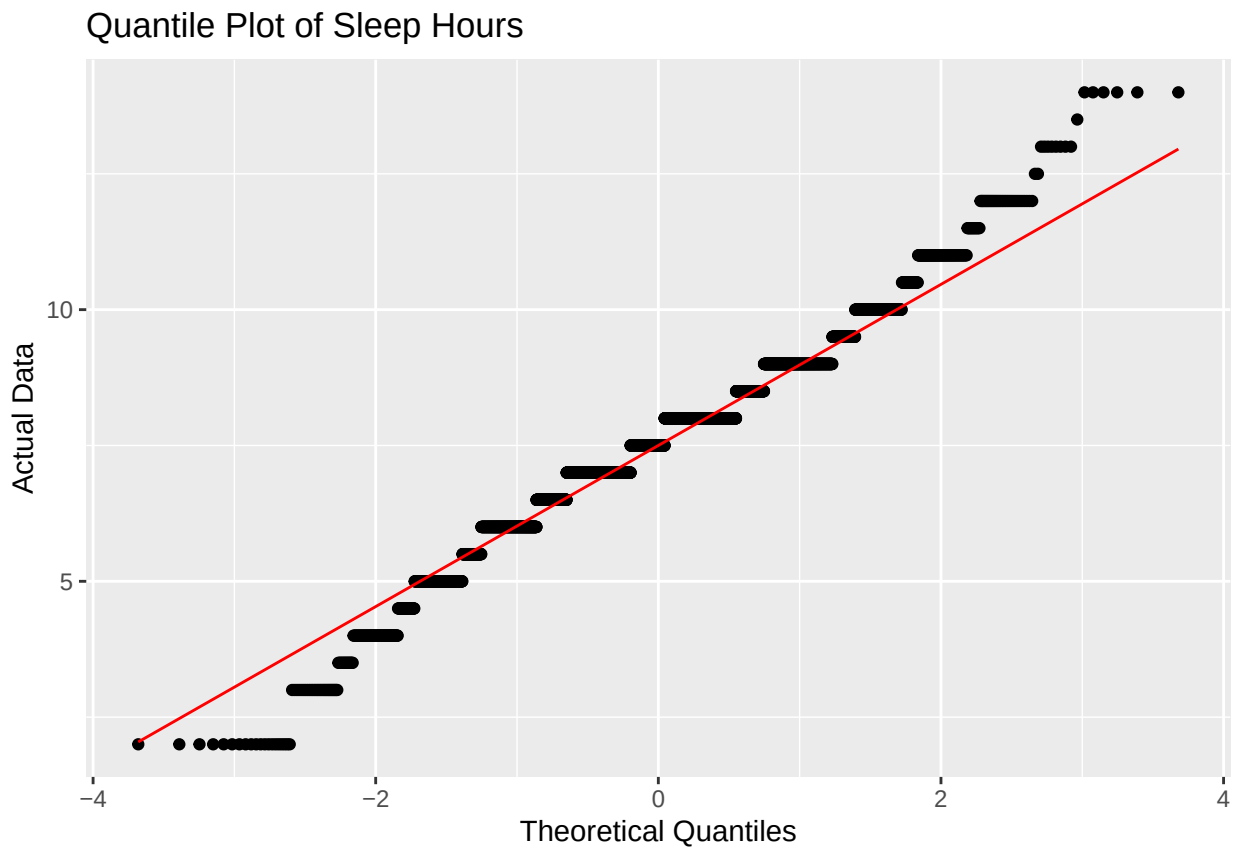
```r
library(moments)
skewness(Final_data$SLD012)
```

```
## [1] -0.07330052
```

```r
num_bins <- (ceiling(log2(nrow(Final_data)))) + 1
ggplot(Final_data, aes(x = SLD012)) +
  geom_histogram(bins = num_bins, fill = "skyblue", color = "black") +
  geom_vline(aes(xintercept = mean(SLD012)), color = "red", show.legend = F) +
  geom_vline(aes(xintercept = median(SLD012)), color = "green", show.legend = F) +
  labs(title = "Histogram of Sleep Hours", x = "Sleep Hours", y = "Frequency")
```

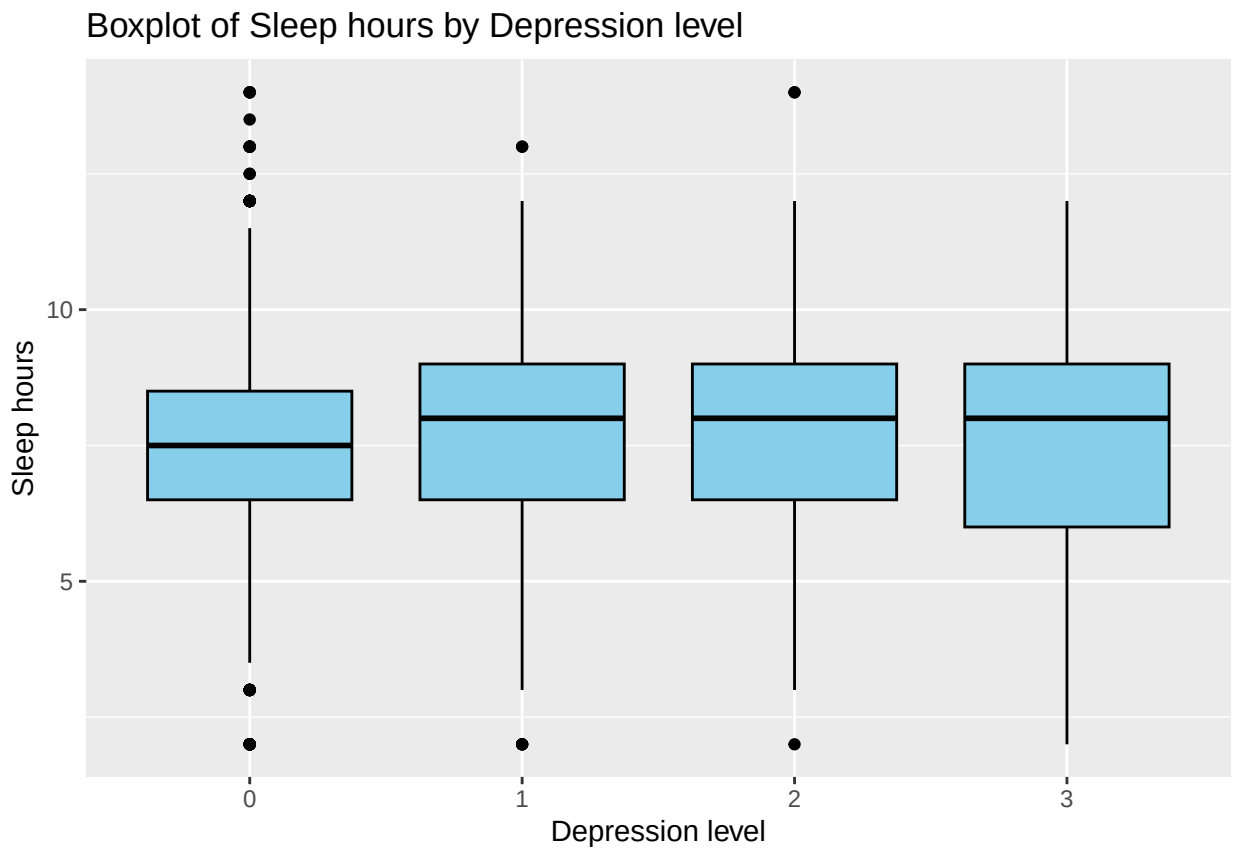



```
ggplot(Final_data, aes(sample = SLD012)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Quantile Plot of Sleep Hours", x = "Theoretical Quantiles", y = "Actual Data")
```

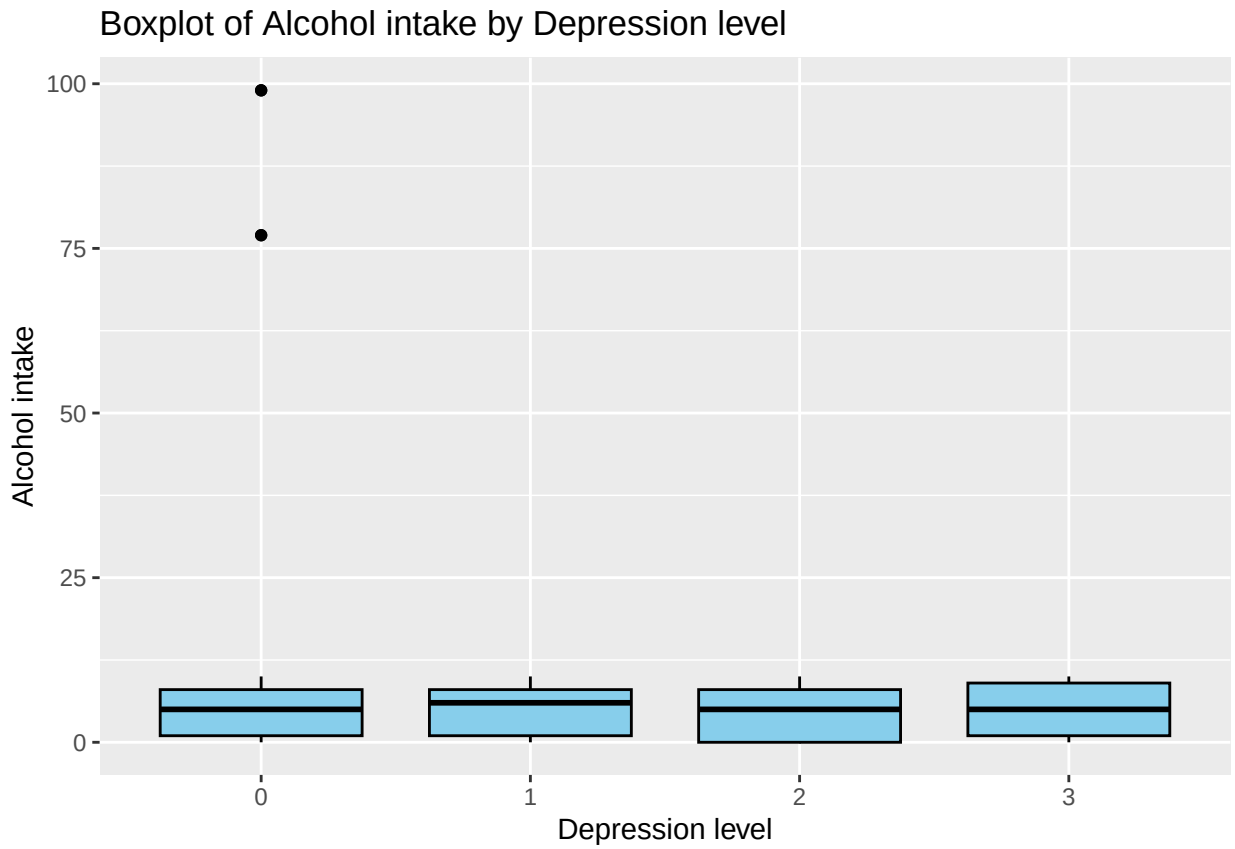


For the visualizations for some other critical variables, we do a side-by-side boxplot for the sleep hours variable ("Boxplot of Sleep hours by Depression level") and the alcohol intake variable ("Boxplot of Alcohol intake by Depression level") for audiences to see the distribution of respondents given their depression level. We also do a scatter plot for age and depression level ("Scatter Plot of Age and Sleep Hours"), with a covariance level of 0.65, which illustrates a positively correlated relation. We also do a histogram for the age dataset ("Histogram of Ages") to see whether we want to do the transformation, and we can see from the distribution that it is not seemingly normally distributed. For further confirmation, we do a quantile plot ("Quantile Plot of Ages") for the variable ages, and we can see from the graph that the points at two ends do not lie totally on the line y = x, which infers that the data is not normally distributed. After confirmation, we decide to do the box-cox transformation for the ages data.

```
ggplot(Final_data, aes(x = as.factor(DPQ020), y = SLD012)) + geom_boxplot(fill = "skyblue", color = "bla
  labs(title = "Boxplot of Sleep hours by Depression level", x = "Depression level", y = "Sleep hours")
```

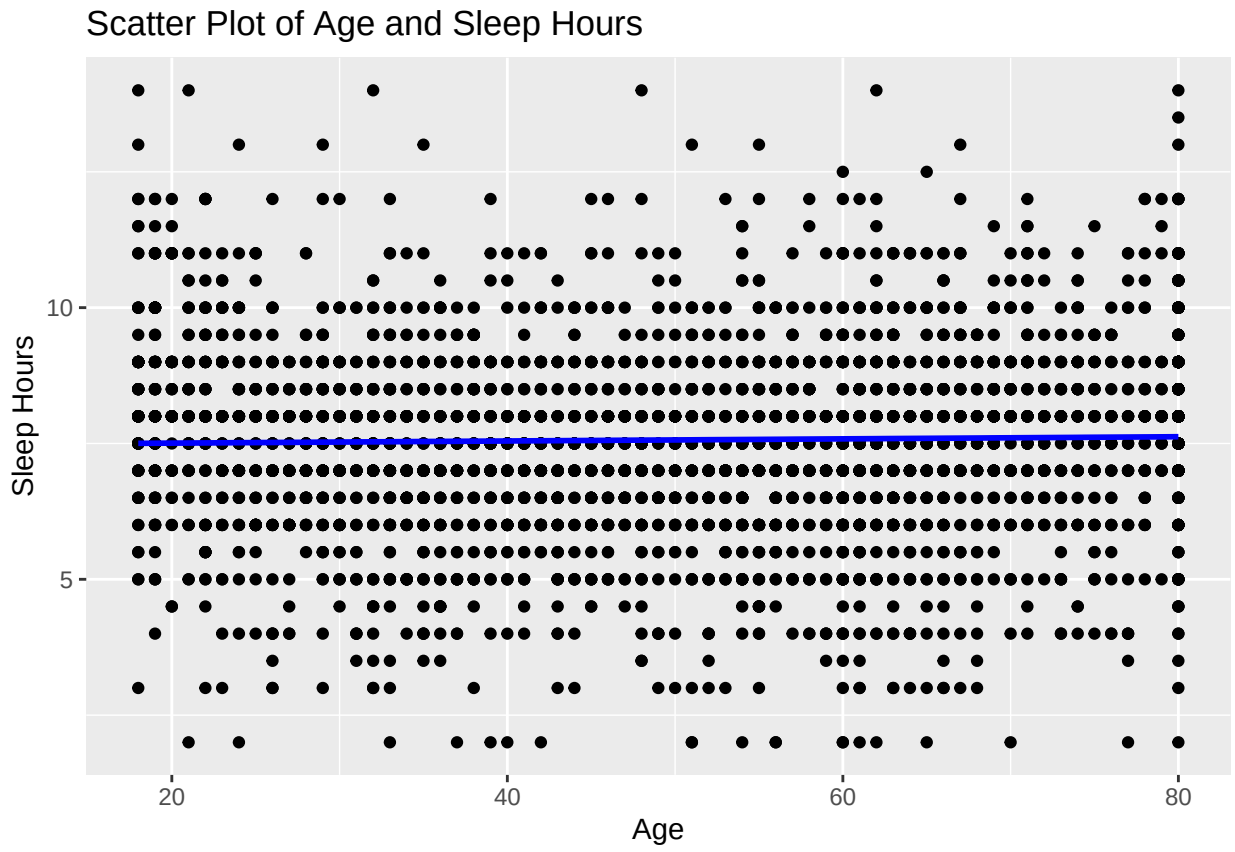



```
ggplot(Final_data, aes(x = as.factor(DPQ020), y = ALQ121)) + geom_boxplot(fill = "skyblue", color = "bla
  labs(title = "Boxplot of Alcohol intake by Depression level", x = "Depression level", y = "Alcohol int
```




```
ggplot(Final_data, aes(x = RIDAGEYR, y = SLD012)) + geom_point() + geom_smooth(method = "lm", se = FALSE
  labs(title = "Scatter Plot of Age and Sleep Hours", x = "Age", y = "Sleep Hours")
```

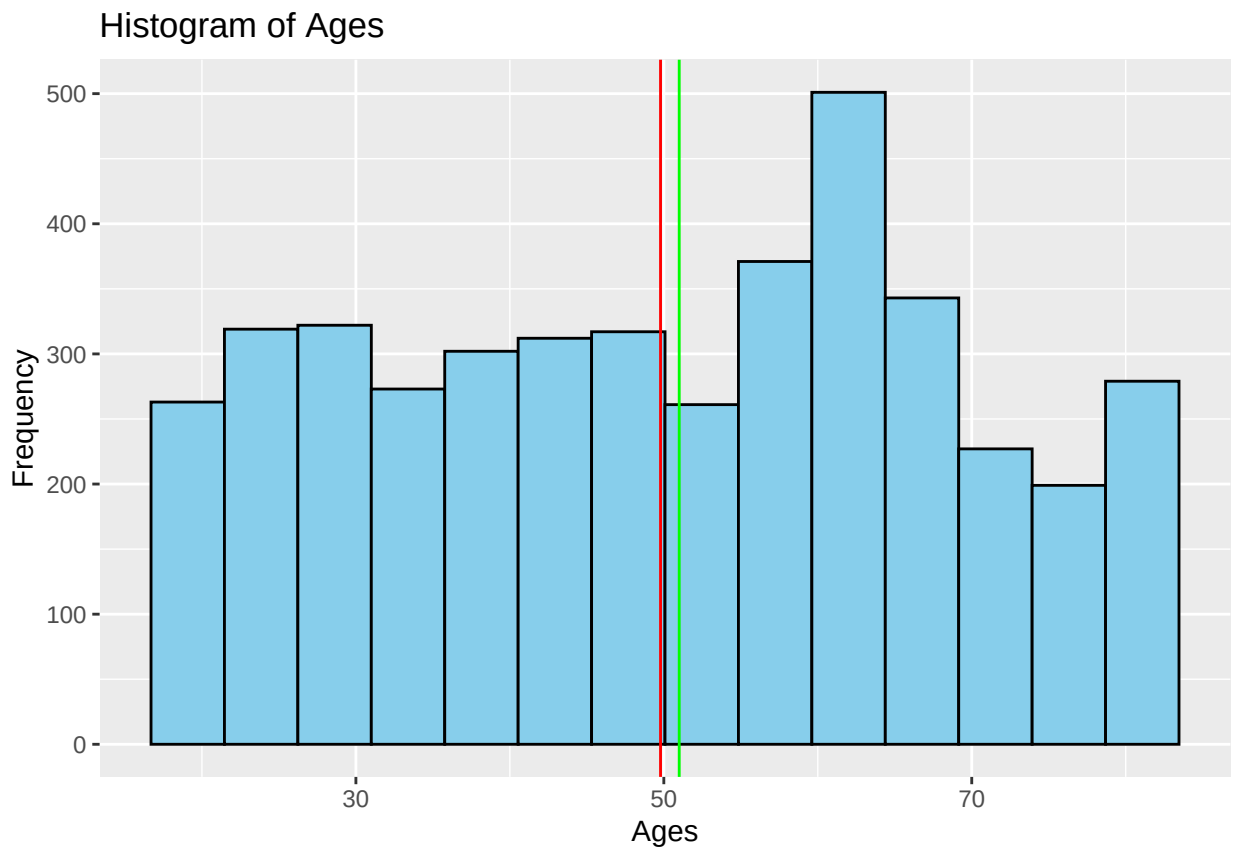
```
## `geom_smooth()` using formula = 'y ~ x'
```



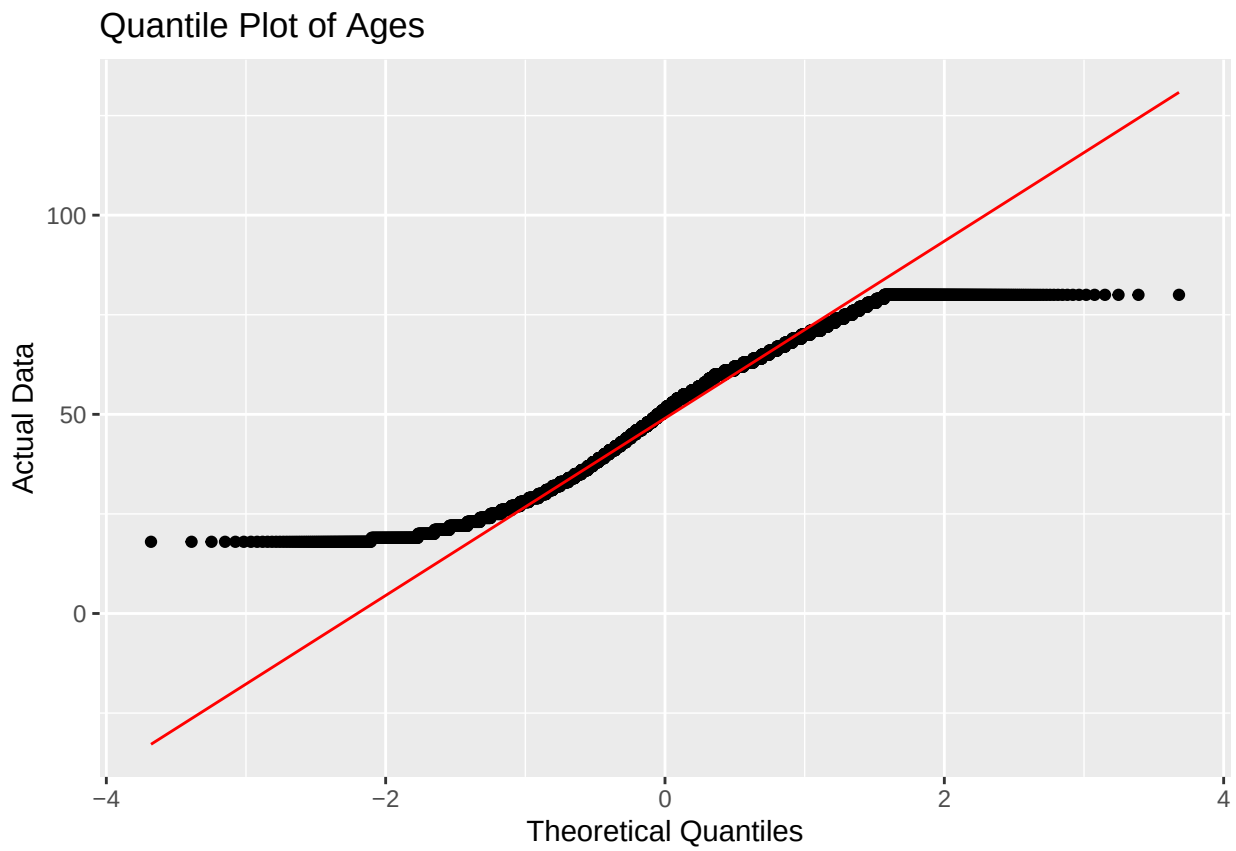

```r
covariance <- cov(Final_data$RIDAGEYR, Final_data$SLD012)
cat("Covariance:", covariance, "\n")
```

```
## Covariance: 0.6511954
```

```r
num_bins <- (ceiling(log2(nrow(Final_data)))) + 1
ggplot(Final_data, aes(x = RIDAGEYR)) +
  geom_histogram(bins = num_bins, fill = "skyblue", color = "black") +
  geom_vline(aes(xintercept = mean(RIDAGEYR)), color = "red", show.legend = F) +
  geom_vline(aes(xintercept = median(RIDAGEYR)), color = "green", show.legend = F) +
  labs(title = "Histogram of Ages", x = "Ages", y = "Frequency")
```



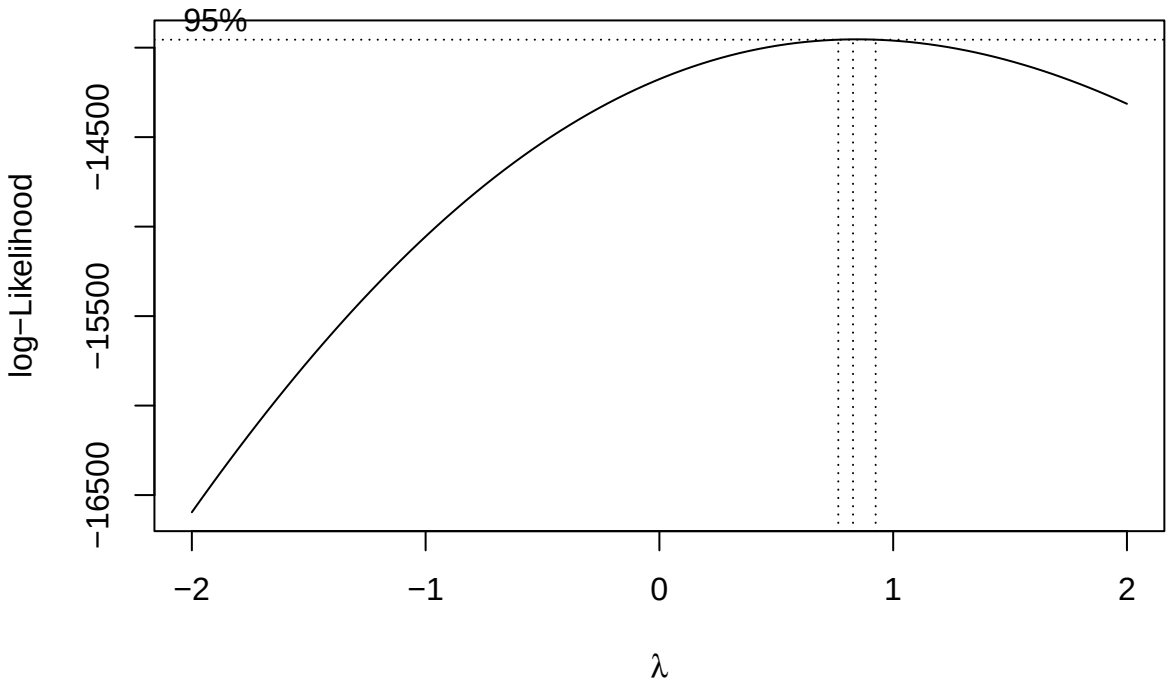

```
ggplot(Final_data, aes(sample = RIDAGEYR)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Quantile Plot of Ages", x = "Theoretical Quantiles", y = "Actual Data")
```




```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
bc <- boxcox(Final_data$RIDAGEYR ~ 1)
```




```
bc_lambda <- bc$x[which.max(bc$y)]
bc_data_age <- (Final_data$RIDAGEYR^bc_lambda - 1)/bc_lambda
```

Q2: To what extent is depression (DPQ020) related to sleep disorders (DPQ030), sleep hours (SLD012), physical activity (PAQ605), poor appetite (DPQ050), and alcohol use (ALQ121)?

Hypothesis:
Null Hypothesis (H_0): There is no correlation between depression and the variables above.
Alternative Hypothesis (H_1): There is a correlation between depression and the variables above.

Significance level: $\alpha = 0.05$

Computing in R:

```
# Loading necessary library
library(readr)
data <- read_csv("Final_data.csv", show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
# Selecting relevant columns
data_selected <- data[c('DPQ020', 'DPQ030', 'DPQ050', 'PAQ605', 'ALQ121', 'SLD012')]

# Converting selected data to numeric type
data_selected <- data.frame(lapply(data_selected, function(x) as.numeric(as.character(x))))
```

```r
# Computing the correlation matrix
cor_matrix <- cor(data_selected, use="complete.obs")

# Initializing a matrix to store p-values
p_matrix <- matrix(NA, ncol=ncol(data_selected), nrow=ncol(data_selected))
rownames(p_matrix) <- colnames(data_selected)
colnames(p_matrix) <- colnames(data_selected)

# Computing p-values
for (i in 1:ncol(data_selected)) {
  for (j in 1:ncol(data_selected)) {
    if (i != j) {
      test_result <- cor.test(data_selected[,i], data_selected[,j], method="pearson")
      p_matrix[i,j] <- test_result$p.value
    }
  }
}

# Setting a significance level of 0.05
significance_level <- 0.05

# Applying flags for significant correlations
signif_flags <- ifelse(p_matrix < significance_level, "*", " ")

# Combining the correlation values and significance flags
result_matrix <- matrix(paste0(round(cor_matrix, 2), signif_flags), ncol=ncol(cor_matrix))
rownames(result_matrix) <- colnames(cor_matrix)
colnames(result_matrix) <- colnames(cor_matrix)

#!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
result<-as.data.frame(result_matrix)
cat("Corelation Metrix\n")
```

```
## Corelation Metrix
```

```r
# Printing the result as a data frame for better formatting
print(result)
```

```
##        DPQ020 DPQ030 DPQ050 PAQ605 ALQ121 SLD012
## DPQ020    1NA   0.4*  0.39* -0.03  -0.01   0.03
## DPQ030   0.4*    1NA  0.38* -0.01  -0.04* -0.07*
## DPQ050  0.39*  0.38*    1NA -0.01      0  -0.01
## PAQ605 -0.03  -0.01  -0.01     1NA -0.03   0.09*
## ALQ121 -0.01  -0.04*     0  -0.03     1NA  0.01
## SLD012  0.03  -0.07* -0.01   0.09*  0.01     1NA
```

Result:
The results show that depression has a significant positive correlation with both sleep disorders ($r = 0.4$, $p < 0.05$) and abnormal appetite ($r = 0.39$, $p < 0.05$), prompting rejection of the null hypothesis. These results suggest a two-way interaction where depression may exacerbate sleep disorders and poor appetite, while these conditions might also contribute to the severity of depression. Conversely, the correlations between depression and physical activity ($r = -0.03$), alcohol use ($r = -0.01$), and sleep duration ($r = 0.03$) were found

to be weak and not statistically significant. These findings underline the multifaceted nature of depression's interplay with different lifestyle and health factors, emphasizing the need for comprehensive approaches to understanding and managing depression.

Q3: Are there statistically significant differences in depression levels based on gender, family income, and race?
Hypothesis:
Null Hypothesis (H_0): There are no significant differences in depression levels based on gender, family income, and race.
Alternative Hypothes (H_1): There are significant differences in depression levels based on at least one of the following variables: gender, family income, and race.
Significance Level: $\alpha = 0.05$
Computing in R:

```
# Loading the data
library(dplyr)
library(readr)
data <- read_csv("Final_data.csv", show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
# Multi-way ANOVA for Depression Levels by Gender, Family Income, and Race
anova_multiway <- aov(DPQ020 ~ gender + fmaily_income + race, data = data)
summary(anova_multiway)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## gender           1   14.9  14.871  30.499 3.54e-08 ***
## fmaily_income   15   67.7   4.515   9.261  < 2e-16 ***
## race             4    6.9   1.717   3.521   0.0071 **
## Residuals     4268 2081.0   0.488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Mean comparison for gender
mean_gender <- aggregate(DPQ020 ~ gender, data = data, mean)
print(mean_gender)
```

```
##   gender    DPQ020
## 1 Female 0.4151659
## 2   Male 0.2973841
```

```
# Post-hoc test for race
posthoc_race <- TukeyHSD(anova_multiway, "race")
print(posthoc_race)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = DPQ020 ~ gender + fmaily_income + race, data = data)
##
## $race
```

```
##                                       diff         lwr        upr     p adj
## Mexican American-Black           0.04457719 -0.05404344 0.14319782 0.7317418
## Other Hispanic-Black             0.11912075  0.00510959 0.23313190 0.0354371
## Other Race-Black                 0.07726734 -0.01717279 0.17170747 0.1678110
## White-Black                      0.09121498  0.01435868 0.16807127 0.0106186
## Other Hispanic-Mexican American  0.07454355 -0.04953871 0.19862581 0.4721181
## Other Race-Mexican American      0.03269015 -0.07369031 0.13907061 0.9186005
## White-Mexican American           0.04663779 -0.04449263 0.13776821 0.6298973
## Other Race-Other Hispanic       -0.04185341 -0.16263965 0.07893284 0.8789960
## White-Other Hispanic            -0.02790577 -0.13550349 0.07969195 0.9547950
## White-Other Race                 0.01394764 -0.07264141 0.10053669 0.9922626
```

Result: The ANOVA analysis demonstrates significant effects of gender, family income, and race on depression levels (DPQ020), leading to the rejection of the null hypothesis. Gender differences are pronounced, with females showing higher mean depression scores (0.415) compared to males (0.297), indicating higher depression levels among females ($F(1, 4268) = 30.499$, $p < 0.001$). Family income significantly affects depression levels ($F(15, 4268) = 9.261$, $p < 0.001$), suggesting a link between economic status and depression. Racial disparities are also evident; mainly, Other Hispanic and White groups exhibit higher depression levels than the Black group ($F(4, 4268) = 3.521$, $p = 0.0071$). These results highlight the critical role of gender, income, and race in understanding and addressing depression, emphasizing the need for focused mental health strategies tailored to these demographic factors.

Q4: How do sleep disorders (DPQ030), sleep hours (SLD012), physical activity (PAQ605), poor appetite (DPQ050), and alcohol use (ALQ121) predict depression levels (DPQ020) among US population aged over 18 years?

Hypothesis:

Null Hypothesis (H_0): Physical exercise, sleep disorder, and alcohol use do not have a significant predictive impact on depression levels.

Alternative Hypothesis (H_1): At least one of the factors - physical exercise, sleep disorder, or alcohol use - has a significant predictive impact on depression levels.

Significance Level: $\alpha = 0.05$

Computing in R:

```r
# Reading the dataset
data <- read_csv("Final_data.csv", show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```r
# Fitting the linear regression model
model <- lm(DPQ020 ~ DPQ030 + SLD012 + PAQ605 + DPQ050 + ALQ121, data = data)

# Summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = DPQ020 ~ DPQ030 + SLD012 + PAQ605 + DPQ050 + ALQ121,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2906 -0.3099 -0.1115 -0.0448  2.9219
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.414e-03  5.645e-02   0.167 0.867556
## DPQ030       2.174e-01  1.078e-02  20.169  < 2e-16 ***
## SLD012       2.224e-02  5.884e-03   3.779 0.000159 ***
## PAQ605      -3.794e-02  1.951e-02  -1.945 0.051864 .
## DPQ050       2.494e-01  1.290e-02  19.338  < 2e-16 ***
## ALQ121       9.441e-06  2.196e-03   0.004 0.996570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.625 on 4283 degrees of freedom
## Multiple R-squared:  0.2292, Adjusted R-squared:  0.2283
## F-statistic: 254.7 on 5 and 4283 DF,  p-value: < 2.2e-16
```

Result: In the linear regression model, sleep disorders and poor appetite are the most significant contributors to depression levels (r = 0.2174 and 0.2494, respectively, both p < 2e-16). Sleep duration also showed a significant, albeit smaller, effect (r = 0.02224, p = 0.000159), indicating that longer sleep duration is associated with higher depression levels. Physical activity displayed a potential but marginally non-significant inverse relationship with depression (r = -0.03794, p = 0.051864), suggesting that increased physical activity might slightly reduce depression levels. Alcohol use, however, did not significantly impact depression levels (p = 0.996570). Overall, the model was statistically significant, accounting for approximately 22.92% of the variance in depression levels (Multiple R-squared = 0.2292, Adjusted R-squared = 0.2283), with an F-statistic of 254.7 on 5 and 4283 degrees of freedom (p-value < 2.2e-16). The residual standard error was 0.625, indicating a reasonable fit of the model, although the range of residuals (-3.2906 to 2.9219) suggests the presence of some outliers. These findings highlight the importance of sleep and appetite disturbances in predicting depression levels, with a lesser but notable role for sleep duration and less clear contributions from physical activity and alcohol use.

Q5: According to the four-year trend recording of the sleeping duration1, the average sleeping duration for the worldwide population is 7.1 hours with a variance of 1.96. Is the variance significantly different from the sample population(variance is 2.64) aged 18 and above with depression level under 3?

Hypothesis Null Hypothesis(H_0):The world population variance of sleeping duration is same as the sleeping duration of the US population variance whose age is 18 and above.
Alternative Hypothesis(H_1):The world population variance of sleeping duration is different from the sleeping duration of the US population variance whose age is 18 above.

Significance Level: $\alpha = 0.05$

Computing in R:

```r
data <- read.csv("Final_data.csv")
#filter our potential depression samples(DPQ020 between 1~3)
non_depression_samples<- data[data$DPQ020 < 3, ]
#one population variance test
n = length(non_depression_samples$DPQ020)
sample_var = var(non_depression_samples$SLD012)
var_0 = 1.96
T_obs = (n-1)*var_0/sample_var
p_value = 2*pchisq(T_obs,n-1)
print(paste("The p-value is",p_value))
```

```
## [1] "The p-value is 8.53009925671137e-38"
```

Result:
According to our test result, the p-value is 8.530099e-38. It is smaller than our significant level alpha = 0.05. Therefore, we reject the null hypothesis and infer that the US population's variance of sleeping duration is different from the worldwide population.

Q6: For the proportion of the sample population who is bothered by depression (with depression level equal to 3), does the proportion of the sample who have abnormal appetite higher than 0.5? **Hypothesis:**
Null Hypothesis(H_0): The population bothered by depression (with depression level equal to 3) has a proportion of abnormal appetite that is equal to or less than 0.5.
Alternative Hypothesis(H_1): The alternative hypothesis is that the population bothered by depression (with a depression level equal to 3) has the proportion of abnormal appetite that is more than 0.5. **Significance Level:** $\alpha = 0.05$
Computing in R:

```
data <- read.csv("Final_data.csv")
#filter our potential depression samples(DPQ020 between 1~3)
depression3_samples<- data[data$DPQ020 == 3, ]
#one proportion test
Num_abnormal_eating <- sum(depression3_samples$DPQ050 == 1)+sum(depression3_samples$DPQ050 == 2)+sum(dep
prop.test(Num_abnormal_eating, length(depression3_samples$DPQ050), p = 0.5, alternative = 'greater')
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  Num_abnormal_eating out of length(depression3_samples$DPQ050), null probability 0.5
## X-squared = 13.067, df = 1, p-value = 0.0001503
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5858228 1.0000000
## sample estimates:
##         p
## 0.6592593
```

Result: Our test has the p-value, 0.0001503 which is less than our significant level alpha = 0.05. We reject the null hypothesis and accept that people suffering from depression have more than half have the abnormal appetite.

Q7: For the sample who reported having trouble sleeping(level = 2 and = 3), is there a significantly different rate between the sample population of those who have depression problems (depression level equal to 3) and those who don't have depression problems(depression level equal to 0)?
Hypothesis:
Null Hypothesis(H_0): The proportion of trouble sleeping in the population suffering from depression is the same as the population who doesn't have depression.
Alternative Hypothesis(H_1): The proportion of those having trouble sleeping in the population suffering from depression is the significant different from the population who doesn't have depression.
Significance Level: $\alpha = 0.05$
Computing in R:

```
data <- read.csv("Final_data.csv")
#discriminate depression3 and non depression
depression3_samples<- data[data$DPQ020 == 3, ]
non_depression_samples<- data[data$DPQ020 == 0, ]
depression3_trouble_sleeping<-sum(depression3_samples$DPQ030 == 2)+sum(depression3_samples$DPQ030 == 3)
non_depression_trouble_sleeping<-sum(non_depression_samples$DPQ030 == 2)+sum(non_depression_samples$DPQ
```

```r
x<-c(depression3_trouble_sleeping,non_depression_trouble_sleeping)
n<-c(length(depression3_samples$DPQ030),length(non_depression_samples$DPQ030))
prop.test(x, n, p = NULL, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 325.65, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.4354266 0.6080253
## sample estimates:
##    prop 1    prop 2
## 0.6222222 0.1004963
```

Result: The p-value of our test is less than 2.2e-16 which is smaller than our significant level alpha = 0.05. Therefore, we reject the null hypothesis and infer that the proportion of having trouble sleeping is different between those who have depression and those who don't have depression.

Q8: Does the variance of the income level of the sample aged 18 and above which has a depression level equal to 3 significantly different from the variance of the income level of the sample aged 18 and above which has a depression level of less than 3?
Hypothesis:
Null Hypothesis(H_0): The variance of income level is the same between those who have a depression level equal to 3(sample variance = 123.51) and those who have a depression level less than 3(sample variance = 267.8739).
Alternative Hypothesis(H_1): The alternative hypothesis is that the variance of income level is different between those who have a depression level equal to 3 and those who have a depression level less than 3.
Significance Level: $\alpha = 0.05$
Computing in R:

```r
data <- read.csv("Final_data.csv")
depression3_samples<- data[data$DPQ020 == 3, ]
less_depression_samples<- data[data$DPQ020 < 3, ]
var.test(depression3_samples$INDFMIN2, less_depression_samples$INDFMIN2, ratio=1, alt="two.sided", conf
```

```
##
##  F test to compare two variances
##
## data:  depression3_samples$INDFMIN2 and less_depression_samples$INDFMIN2
## F = 0.46107, num df = 134, denom df = 4153, p-value = 3.421e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3663190 0.5969399
## sample estimates:
## ratio of variances
##          0.4610686
```

Result: According to our result, we have the p-value = 3.421e-08. It is less than our significant level alpha = 0.05. Hence, we reject the null hypothesis and infer that the variance of income level is different between those who have a depression level equal to 3 and those who have a depression level less than 3.

Conclusion:
The final project report presents a comprehensive analysis of factors influencing mental health, US population aged 18 and above using the 2017-2018 National Health and Nutrition Examination Survey dataset. Key findings include a significant positive correlation between depression and sleep disorders, and abnormal appetite, suggesting a bidirectional relationship. However, weak or no significant correlations were found with physical activity, alcohol use, and sleep duration. Additionally, the study revealed gender, family income, and racial disparities in depression levels, with females, lower-income groups, and certain racial groups showing higher depression rates. These results underline the complexity of depression, highlighting the need for comprehensive and tailored approaches to mental health management.