PHCA9523 INTRODUCTORY BIOSTATISTICS

Assessment 4 Statistical Analysis in R Software

Installing Packages

```
# install.packages("readxl")
```

```
# install.packages("gtsummary")
# install.packages("broom")
# install.packages(c("ggplot2","dplyr","corrplot","RColorBrewer"))
#install.packages("car")
# install.packages("pROC")
```

Loading package

```
library(readxl)
```

```
library(gtsummary)
library(ggplot2)
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(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(RColorBrewer)
library(broom)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Importing data from Sheet1

```
# Importing data from Sheet1
sage <- read_excel("SAGE_updated.xlsx", sheet = "Sheet1")

# Displaying first few rows
head(sage)
```

```
## # A tibble: 6 × 11
##      ID centre gender   age ideation attempt health sleeplength
sleepqual
##   <dbl> <chr>   <dbl> <dbl>    <dbl>   <dbl>  <dbl>       <dbl>
<dbl>
## 1     1 China       1    67        0       0      1          22
0
## 2     2 China       2    53        0       0      2           8
0
```

```
## 3      3 China       1    60         1        1       2            7
0
## 4      4 China       2    50         0        0       1            9
0
## 5      5 China       1    57         0        0       3           10
0
## 6      6 China       2    55         1        0       2            5
1
## # ℹ 2 more variables: married <dbl>, cognitive_zscore <dbl>
```

Checking dataset structure

```
# Variable names
```

names(sage)

```
##  [1] "ID"               "centre"           "gender"           "age"
##  [5] "ideation"         "attempt"          "health"
"sleeplength"
##  [9] "sleepqual"        "married"          "cognitive_zscore"
```

```
# Structure of dataset
```

str(sage)

```
## tibble [2,356 × 11] (S3: tbl_df/tbl/data.frame)
##  $ ID              : num [1:2356] 1 2 3 4 5 6 7 8 9 10 ...
##  $ centre          : chr [1:2356] "China" "China" "China" "China"
...
##  $ gender          : num [1:2356] 1 2 1 2 1 2 2 1 2 2 ...
##  $ age             : num [1:2356] 67 53 60 50 57 55 65 73 61 77 ...
##  $ ideation        : num [1:2356] 0 0 1 0 0 1 0 0 1 0 ...
##  $ attempt         : num [1:2356] 0 0 1 0 0 0 0 0 0 0 ...
##  $ health          : num [1:2356] 1 2 2 1 3 2 3 1 1 1 ...
##  $ sleeplength     : num [1:2356] 22 8 7 9 10 5 8 7 7.5 7 ...
##  $ sleepqual       : num [1:2356] 0 0 0 0 0 1 0 0 1 0 ...
##  $ married         : num [1:2356] 2 2 1 2 2 2 2 2 2 1 ...
##  $ cognitive_zscore: num [1:2356] -1.46 1.04 -4.96 -3.53 1.99 ...
```

```
# Summary statistics
```

summary(sage)

```
##        ID            centre               gender            age
```

```
##  Min.   :   1.0   Length:2356        Min.   :1.000   Min.   : 50.00
##  1st Qu.: 589.8   Class :character   1st Qu.:1.000   1st Qu.: 55.00
##  Median :1178.5   Mode  :character   Median :2.000   Median : 63.00
##  Mean   :1178.5                      Mean   :1.598   Mean   : 64.07
##  3rd Qu.:1767.2                      3rd Qu.:2.000   3rd Qu.: 70.00
##  Max.   :2356.0                      Max.   :2.000   Max.   :106.00
##
##     ideation          attempt             health       sleeplength
##  Min.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.: 6.000
##  Median :0.0000   Median :0.00000   Median :2.000   Median : 7.000
##  Mean   :0.3663   Mean   :0.08319   Mean   :1.812   Mean   : 7.097
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.: 8.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :3.000   Max.   :24.000
##                                     NA's   :1       NA's   :83
##    sleepqual          married      cognitive_zscore
##  Min.   :0.0000   Min.   :1.000   Min.   :-11.15085
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: -2.53640
##  Median :0.0000   Median :2.000   Median :  0.01237
##  Mean   :0.1543   Mean   :1.614   Mean   :  0.02694
##  3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:  2.63819
##  Max.   :1.0000   Max.   :2.000   Max.   : 13.80052
##  NA's   :23       NA's   :5       NA's   :308
```

```
# Dimensions (rows and columns)
```

dim(sage)

```
## [1] 2356   11
```

Checking for missing values

```
# Number of missing values per variable
```

colSums(is.na(sage))

```
##               ID           centre           gender              age
##                0                0                0                0
##         ideation          attempt           health      sleeplength
##                0                0                1               83
##        sleepqual          married cognitive_zscore
##               23                5              308
```

```
# Total missing values
```

```
sum(is.na(sage))
```

```
## [1] 420
```

Examining distributions

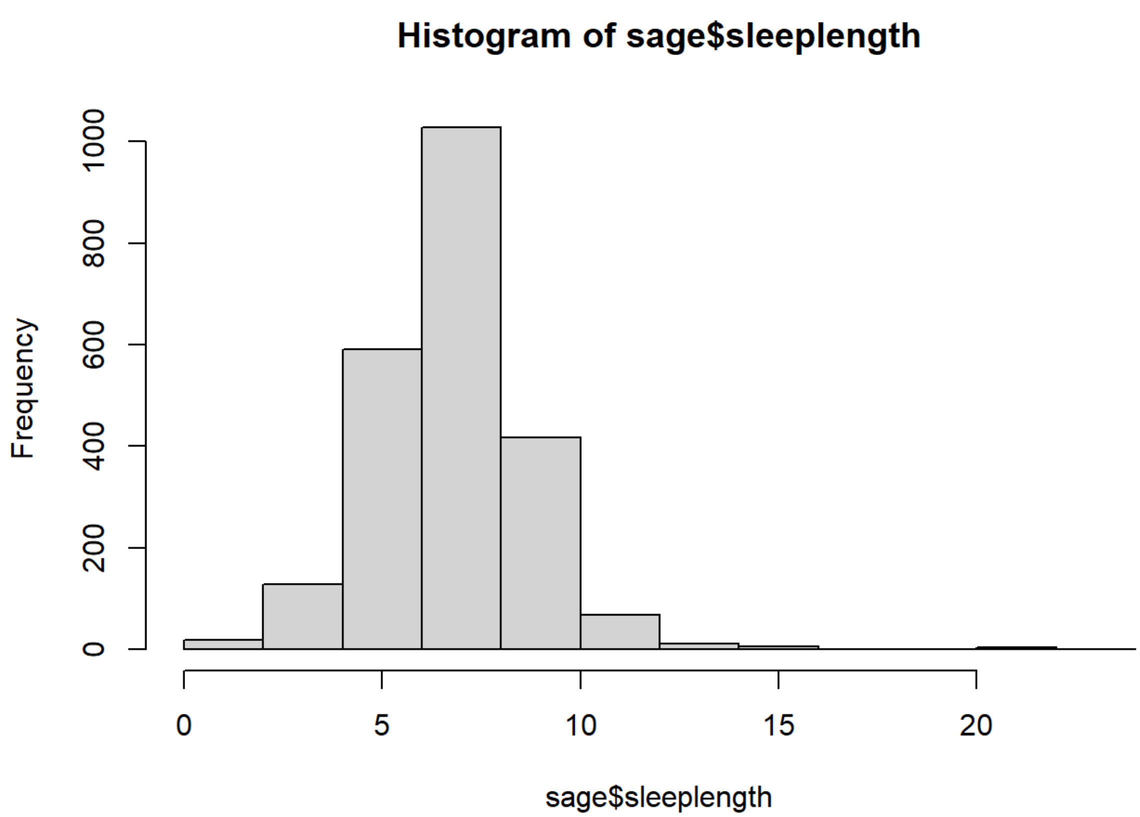
```
summary(sage$sleeplength)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   6.000   7.000   7.097   8.000  24.000      83
```

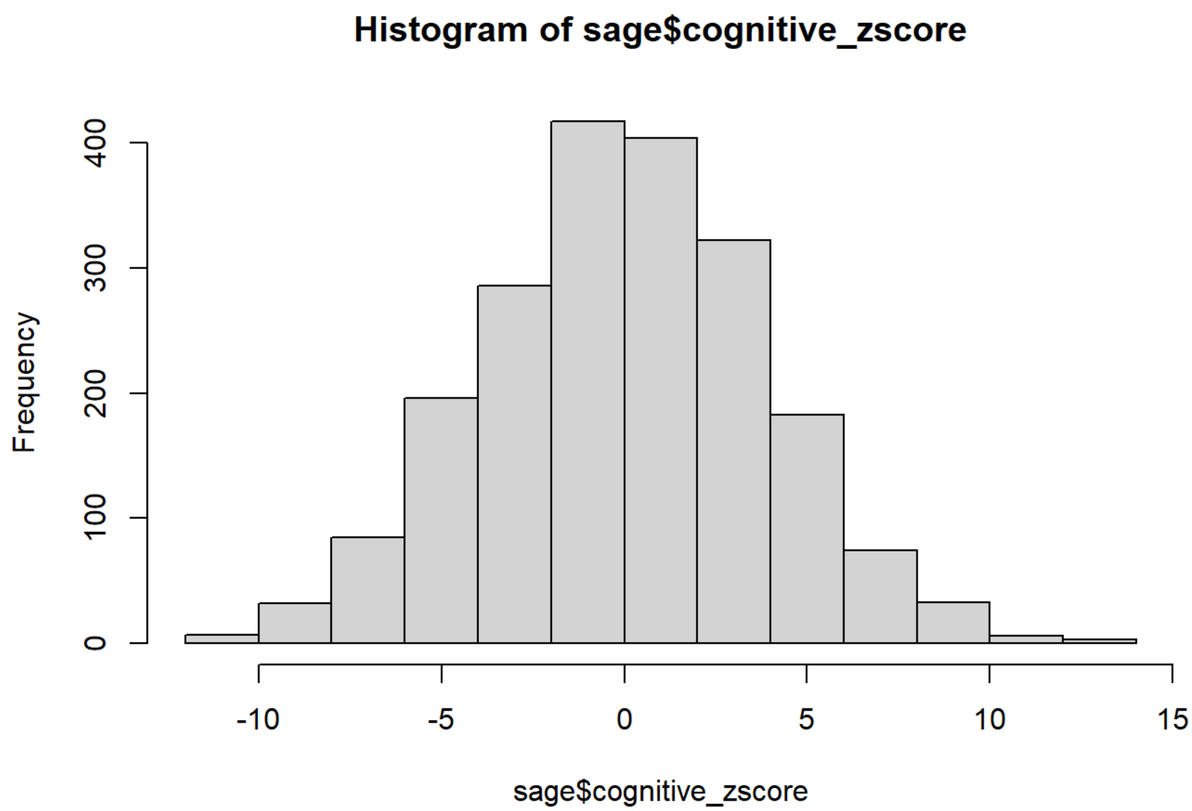
```
summary(sage$cognitive_zscore)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
## -11.15085  -2.53640   0.01237   0.02694   2.63819  13.80052
308
```

```
hist(sage$sleeplength)
```



```
hist(sage$cognitive_zscore)
```



Replacing Missing Values

```
# Mode function
getmode <- function(v) {
  uniqv <- unique(v[!is.na(v)])
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Health (mode)
sage$health[is.na(sage$health)] <- getmode(sage$health)

# Sleep quality (mode)
sage$sleepqual[is.na(sage$sleepqual)] <- getmode(sage$sleepqual)

# Married (mode)
sage$married[is.na(sage$married)] <- getmode(sage$married)
```

```
# Sleep length (median)
sage$sleeplength[is.na(sage$sleeplength)] <- median(sage$sleeplength, na.rm = TRUE)

# Cognitive score (median)
sage$cognitive_zscore[is.na(sage$cognitive_zscore)] <- median(sage$cognitive_zscore,
na.rm = TRUE)
```

Further Checking for Missing Values

```
colSums(is.na(sage))
```

```
##               ID           centre           gender              age
##                0                0                0                0
##         ideation          attempt           health      sleeplength
##                0                0                0                0
##        sleepqual          married cognitive_zscore
##                0                0                0
```

```
sum(is.na(sage))
```

```
## [1] 0
```

Converting variables to appropriate types

```
sage$centre <- as.factor(sage$centre)
```

```
sage$gender <- factor(
  sage$gender,
  levels = c(1,2),
  labels = c("Male","Female")
)

sage$ideation <- factor(
  sage$ideation,
  levels = c(0,1),
  labels = c("No","Yes")
)

sage$attempt <- factor(
  sage$attempt,
  levels = c(0,1),
  labels = c("No","Yes")
)
```

```
sage$health <- factor(
  sage$health,
  levels = c(1,2,3),
  labels = c("Bad/Very Bad",
             "Moderate",
             "Good/Very Good")
)

sage$sleepqual <- factor(
  sage$sleepqual,
  levels = c(0,1),
  labels = c("No","Yes")
)

sage$married <- factor(
  sage$married,
  levels = c(1,2),
  labels = c("Not currently married",
             "Married/cohabitating")
)
```

```
nrow(sage)
```

```
## [1] 2356
```

```
table(sage$gender)
```

```
##
##   Male Female
##    948   1408
```

```
table(sage$ideation)
```

```
##
##   No  Yes
## 1493  863
```

```
table(sage$sleepqual)
```

```
##
##   No  Yes
## 1996  360
```

```
summary(sage$age)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    50.00   55.00   63.00   64.07   70.00  106.00
```

```
summary(sage$sleeplength)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.000   6.000   7.000   7.094   8.000  24.000
```

```
summary(sage$cognitive_zscore)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -11.15085  -2.08809   0.01237   0.02503   2.20198  13.80052
```

The data is loaded into R Studio and appropriately processed for analysis. Missing value are found in the dataset and flagged and correctly replaced by employing the mode or the median values (Palanivinayagam and Damaševičius, 2023). Variables are coded accordingly to fit the data structure, and a data quality assessment is carried out. The entire process of analysis involved the use of R codes, and it involves to generate proper analytical results.

Exploratory Data Analysis

Descriptive Characteristics of Study Participants

```
tbl1 <- sage %>%
```

```
select(age,
       gender,
       centre,
       health,
       sleeplength,
       sleepqual,
       married,
       ideation,
       cognitive_zscore) %>%
tbl_summary(
  statistic = list(
    all_continuous() ~ "{mean} ({sd})",
    all_categorical() ~ "{n} ({p}%)"
  )
)
```

tbl1

Characteristic	N = 2,356[1]
age	64 (10)
gender	
Male	948 (40%)
Female	1,408 (60%)
centre	
China	327 (14%)
Ghana	406 (17%)
India	1,156 (49%)
Russian Federation	308 (13%)
South Africa	159 (6.7%)
health	
Bad/Very Bad	851 (36%)
Moderate	1,096 (47%)
Good/Very Good	409 (17%)
sleeplength	7.09 (1.97)
sleepqual	360 (15%)
married	
Not currently married	907 (38%)
Married/cohabitating	1,449 (62%)
ideation	863 (37%)

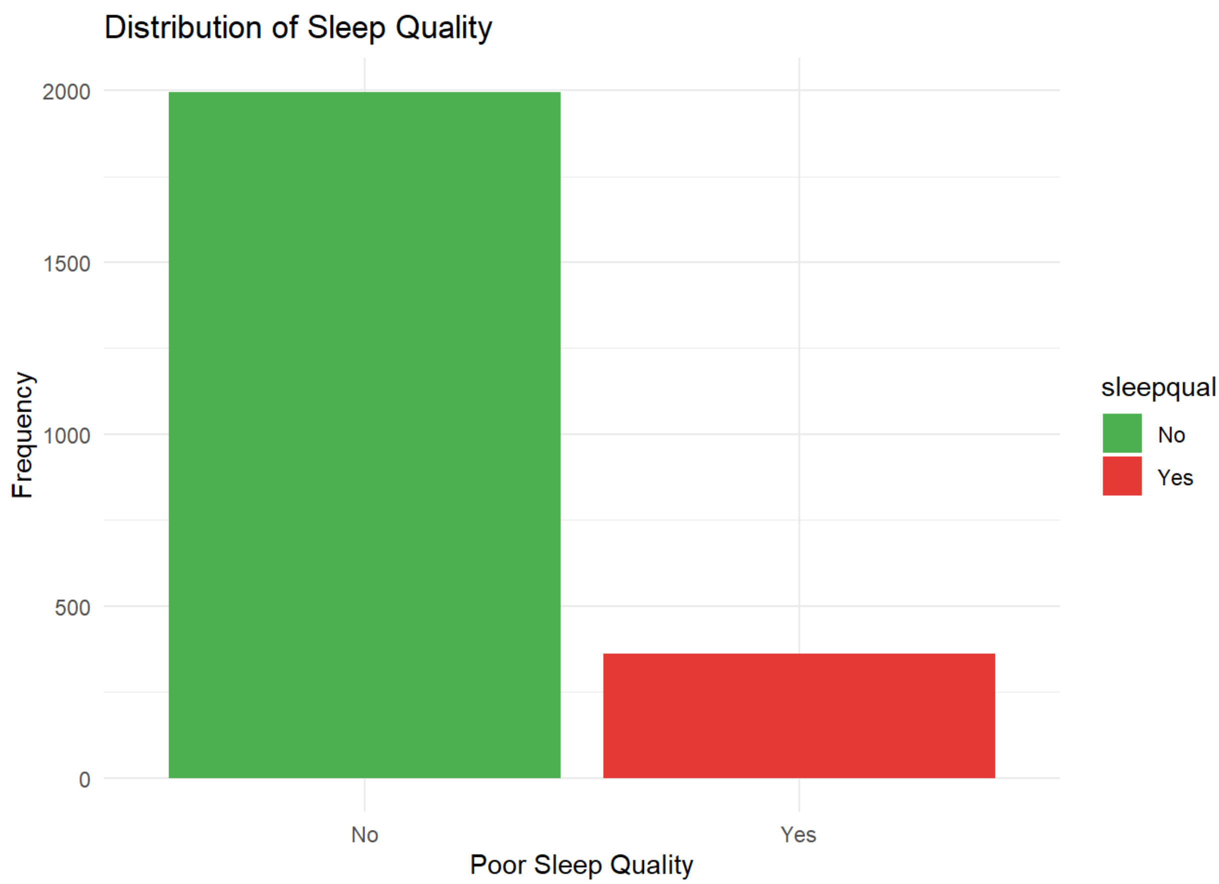
Characteristic	N = 2,356[1]
cognitive_zscore	0.0 (3.6)

[1] Mean (SD); n (%)

The participants comprised 2,356 individuals with an average age of 64 years, and 60% of the participants are female. The average time for sleep is 7.1 hours, and 15% has poor sleep quality. Suicidal thoughts are experienced by 37% of the participants, and the average cognitive z-score is calculated as zero.

Distribution of Sleep Quality

```
ggplot(sage, aes(x = sleepqual, fill = sleepqual)) +
  geom_bar() +
  scale_fill_manual(values = c("#4CAF50", "#E53935")) +
  labs(
    title = "Distribution of Sleep Quality",
    x = "Poor Sleep Quality",
    y = "Frequency"
  ) +
  theme_minimal()
```



It is visulized that most of the subjects had better quality sleep, as 85% has no poor quality of sleep, while only 15% have poor quality sleep. It indicates that sleep problems were rare among the test subjects, but there are those who had poor quality sleep.

Suicidal Ideation Distribution

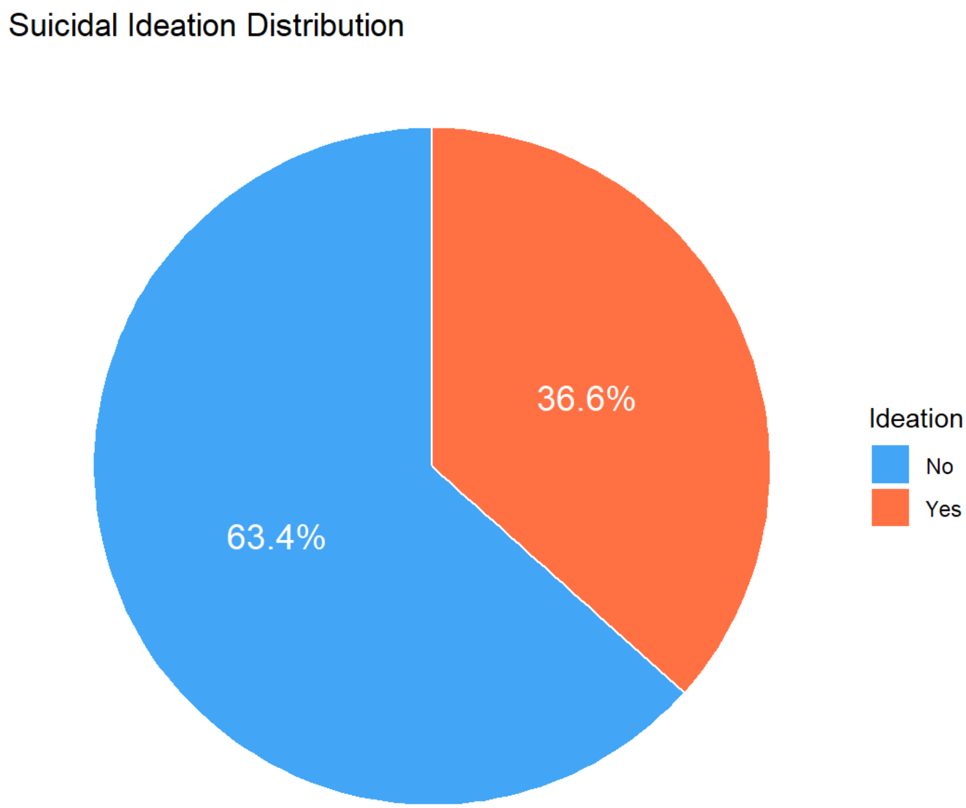
```
ideation_counts <- as.data.frame(table(sage$ideation))
```

```
colnames(ideation_counts) <- c("Ideation", "Count")

ideation_counts <- ideation_counts %>%
  mutate(
    Percentage = round(Count / sum(Count) * 100, 1),
    Label = paste0(Percentage, "%")
  )

ggplot(ideation_counts,
       aes(x = "", y = Count, fill = Ideation)) +
  geom_col(width = 1, color = "white") +
  geom_text(
    aes(label = Label),
```

```
  position = position_stack(vjust = 0.5),
  size = 5,
  color = "white"
) +
coord_polar("y") +
scale_fill_manual(values = c("#42A5F5", "#FF7043")) +
labs(
  title = "Suicidal Ideation Distribution",
  fill = "Ideation"
) +
theme_void()
```



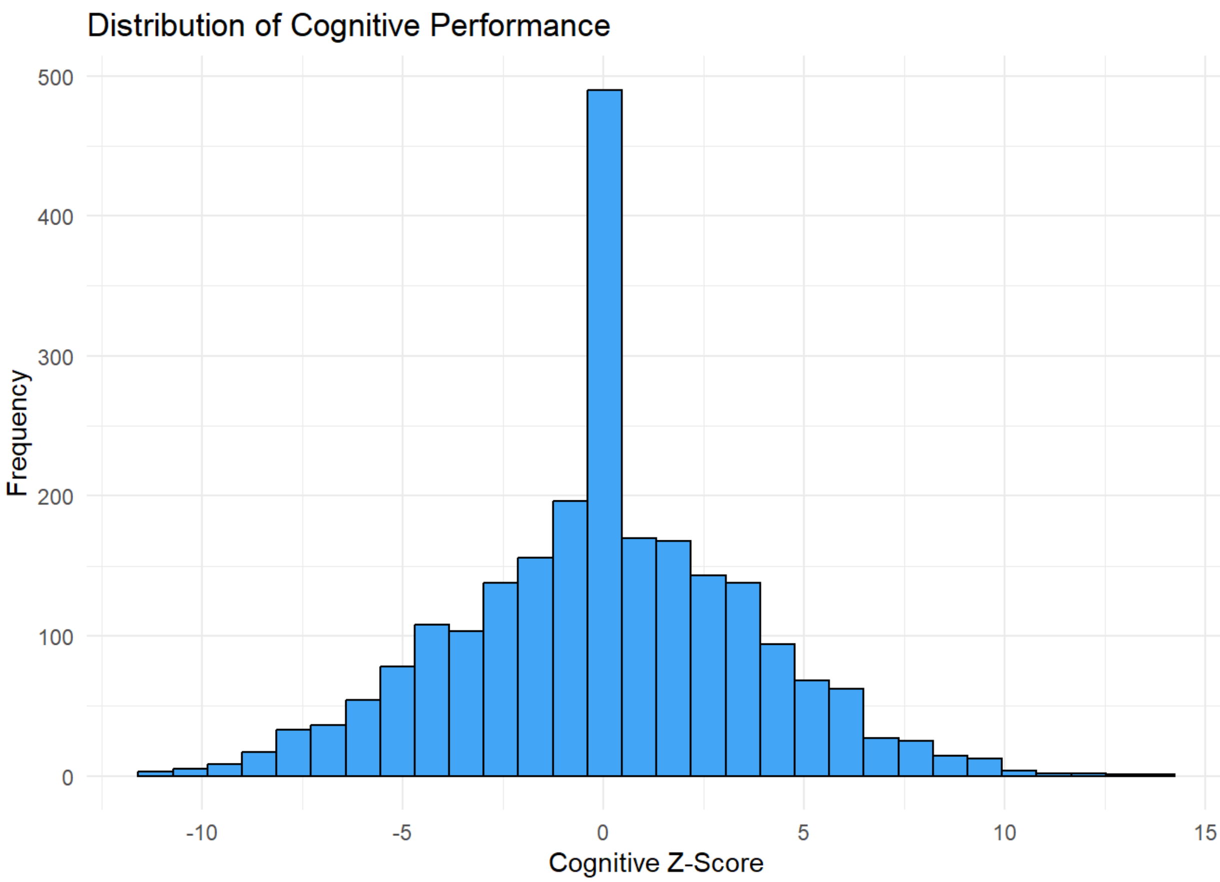
About 37% of subjects has suicidal ideations, where 63% of participants has no suicidal ideations. The high incidence level makes the topic of suicidal ideation very important from a public health perspective, thus necessitating the need to identify the associated risks.

Distribution of Cognitive Performance

```
ggplot(sage,
```

```
    aes(x = cognitive_zscore)) +
```

```
geom_histogram(
  bins = 30,
  fill = "#42A5F5",
  color = "black"
) +
labs(
  title = "Distribution of Cognitive Performance",
  x = "Cognitive Z-Score",
  y = "Frequency"
) +
theme_minimal()
```



The distribution of cognitive z-scores is normal and centered around zero, with a majority of respondents falling close to the mean cognitive performance. Some outliers with exceptionally low and high cognitive scores are observed.

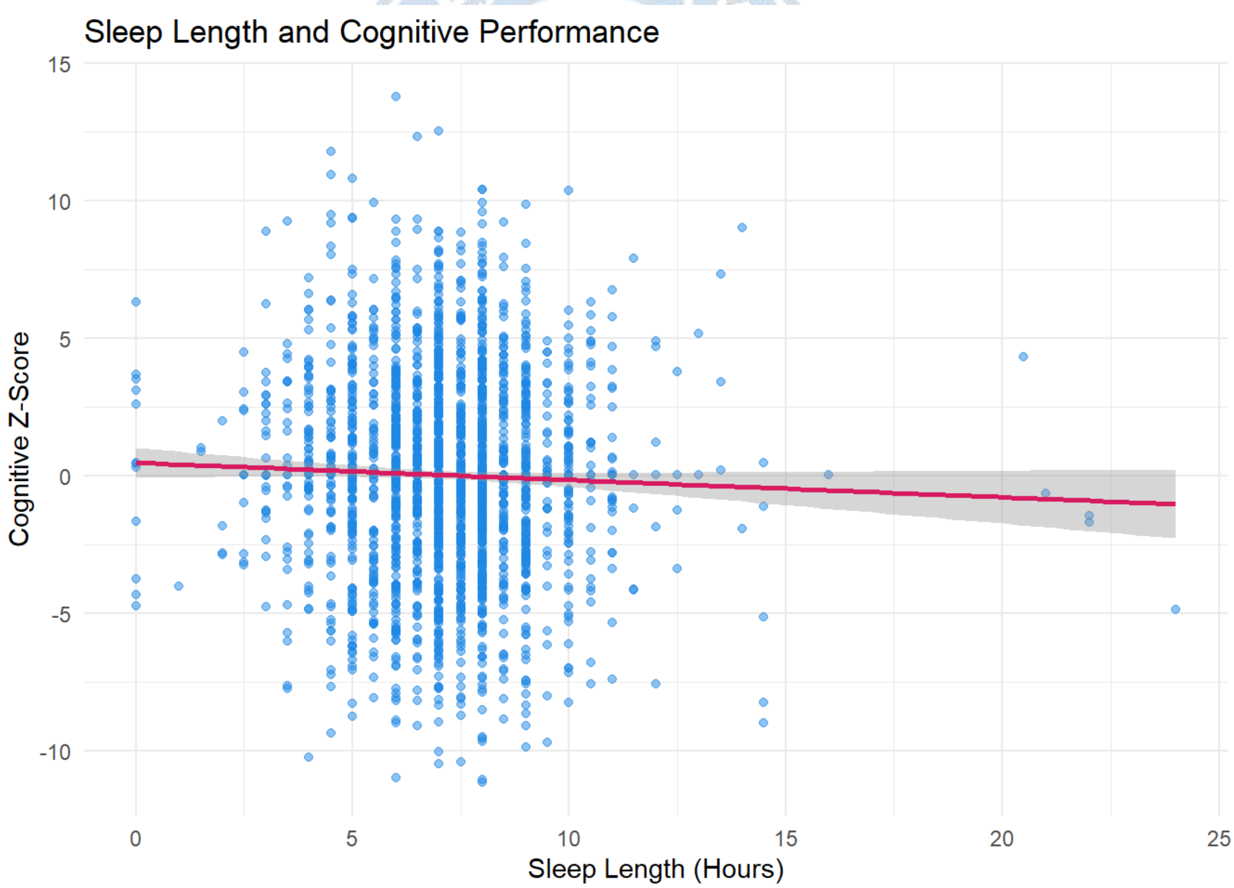
Sleep Length and Cognitive Performance

```
ggplot(sage,
```

```
    aes(x = sleeplength,
```

```
         y = cognitive_zscore)) +
geom_point(
  color = "#1E88E5",
  alpha = 0.5
) +
geom_smooth(
  method = "lm",
  color = "#D81B60",
  se = TRUE
) +
labs(
  title = "Sleep Length and Cognitive Performance",
  x = "Sleep Length (Hours)",
  y = "Cognitive Z-Score"
) +
theme_minimal()
```

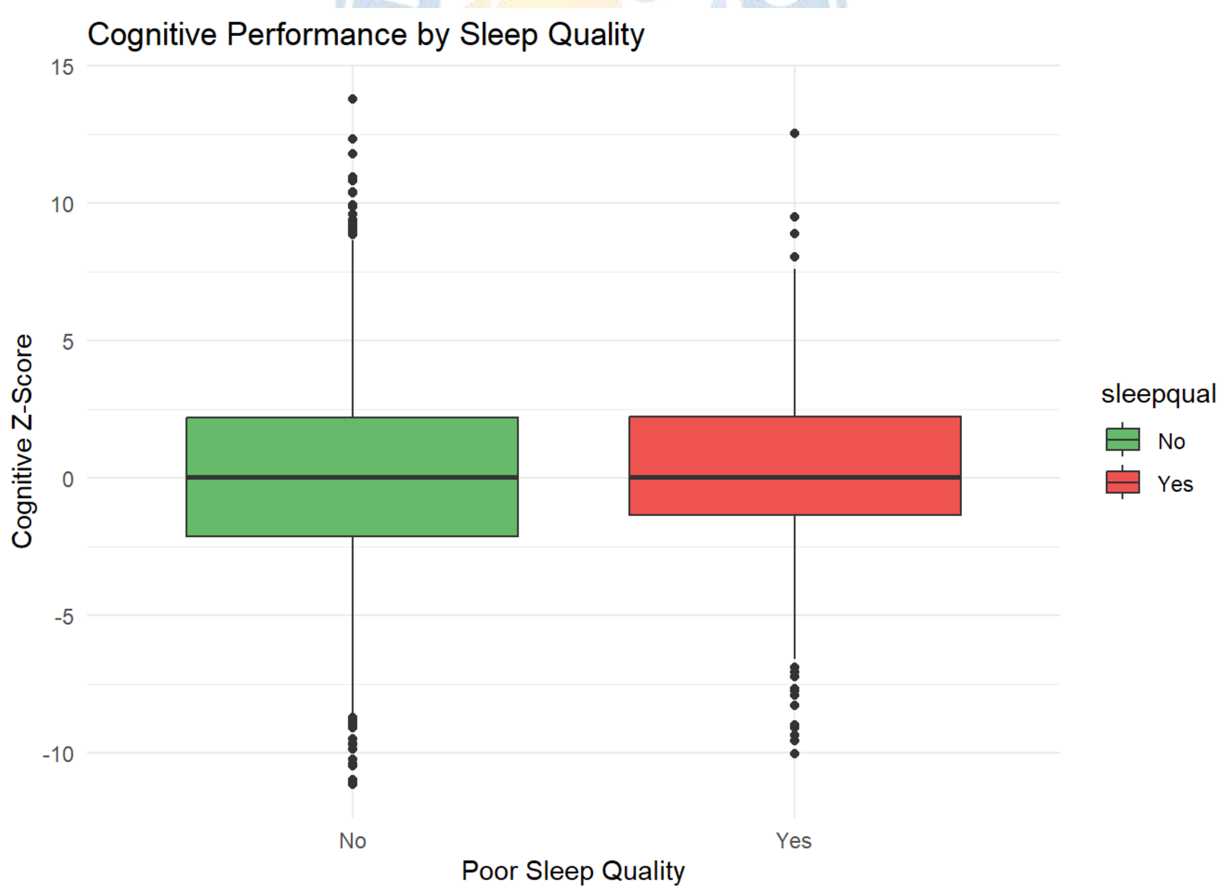
```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplot revealed large variability in cognitive performance among different sleeping durations. The regression line has an evident slight negative trend, implying some weak correlation between cognitive performance and sleep duration.

Cognitive Performance by Sleep Quality

```
ggplot(sage,
     aes(x = sleepqual,
         y = cognitive_zscore,
         fill = sleepqual)) +
  geom_boxplot() +
  scale_fill_manual(
    values = c("#66BB6A", "#EF5350")
  ) +
  labs(
    title = "Cognitive Performance by Sleep Quality",
    x = "Poor Sleep Quality",
    y = "Cognitive Z-Score"
  ) +
  theme_minimal()
```

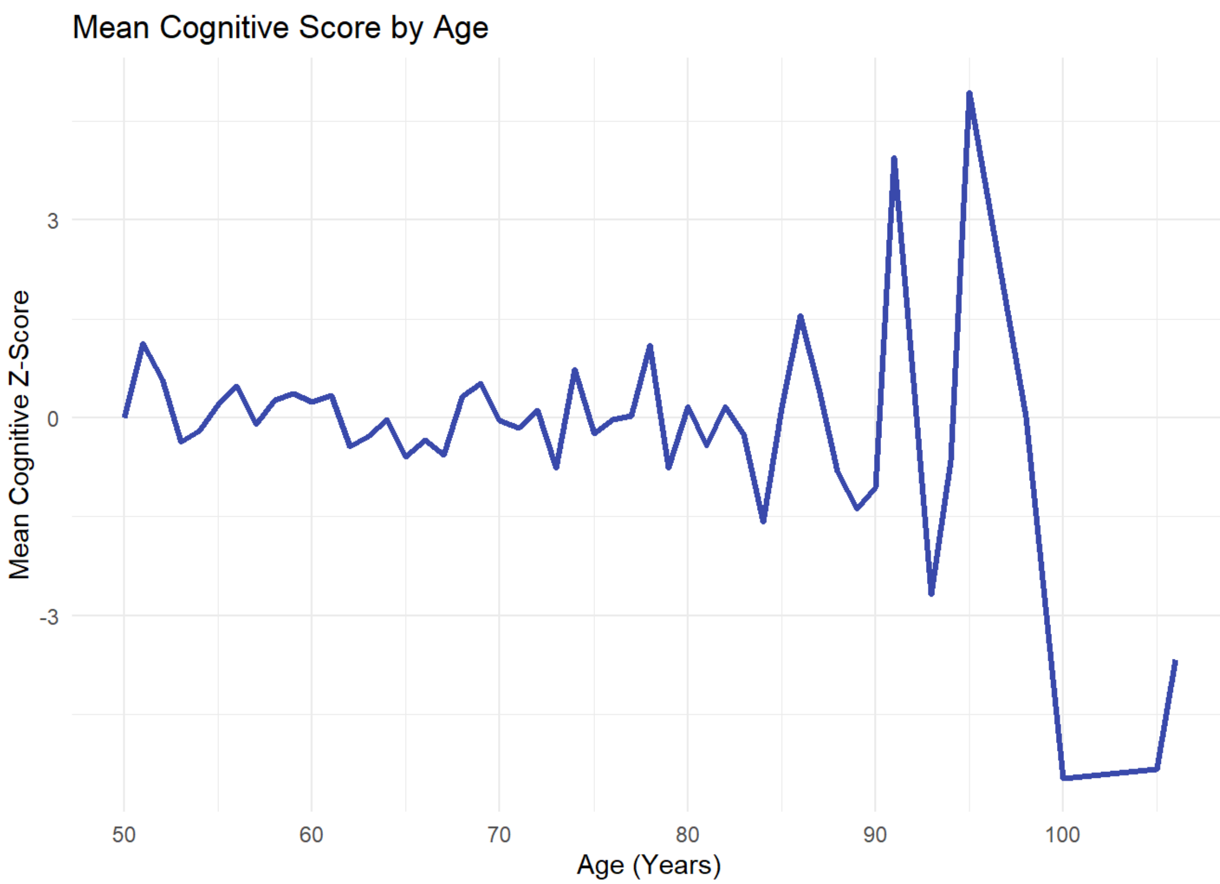


Respondents that has poor sleep quality and those with good sleep quality has roughly equivalent performance in terms of their cognitive performance. Small variations can be observed between the two groups, as significant overlap implies that sleep quality does not have much effect on cognitive performance.

Mean Cognitive Score by Age

```
age_summary <- sage %>%
  group_by(age) %>%
  summarise(
    mean_cognitive = mean(cognitive_zscore)
  )

ggplot(age_summary,
       aes(x = age,
           y = mean_cognitive)) +
  geom_line(color = "#3949AB", linewidth = 1.2) +
  labs(
    title = "Mean Cognitive Score by Age",
    x = "Age (Years)",
    y = "Mean Cognitive Z-Score"
  ) +
  theme_minimal()
```



Mean cognitive scores showed variations according to age groups, which are more pronounced amongst the eldest subjects. The observed variation can be attributed to the relatively small number of samples in the older age groups. It seems that there is no age-related trend based on the descriptive statistics.

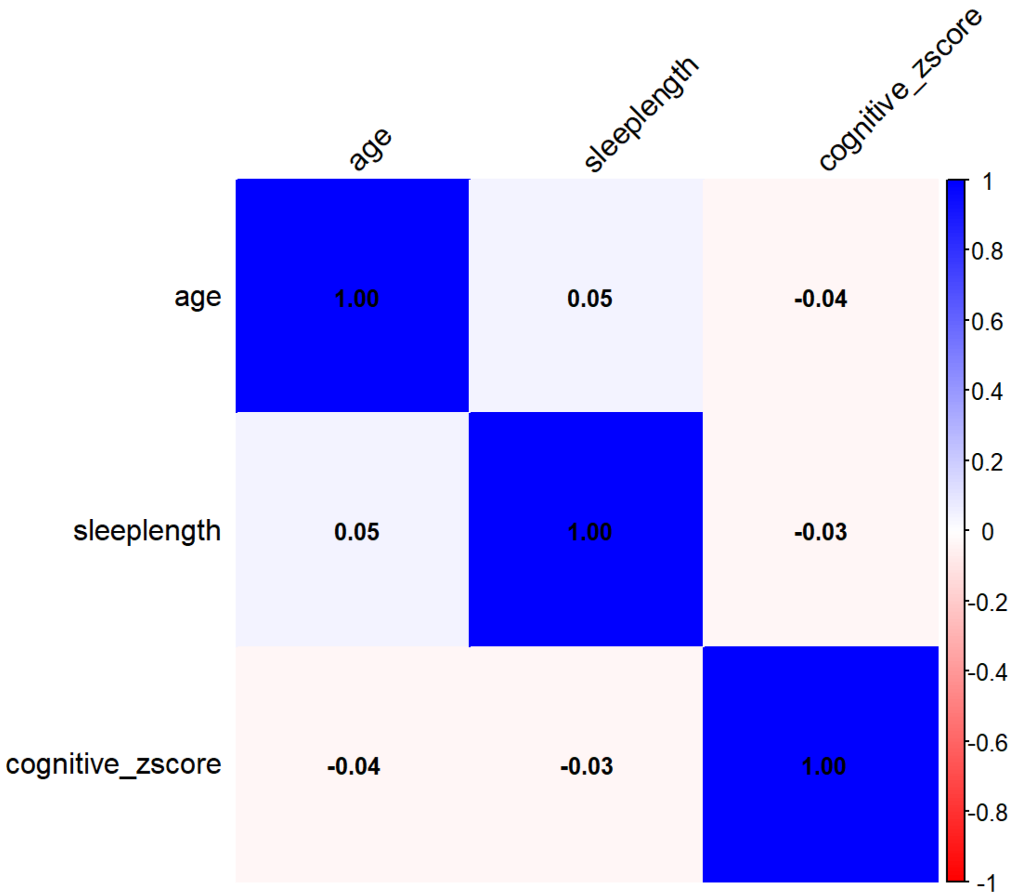
Correlation Matrix

```
numeric_data <- sage %>%
  select(age,
         sleeplength,
         cognitive_zscore)

cor_matrix <- cor(numeric_data)

corrplot(
  cor_matrix,
  method = "color",
  col = colorRampPalette(
    c("red", "white", "blue")
  )(200),
```

```
  addCoef.col = "black",
  number.cex = 0.8,
  tl.col = "black",
  tl.srt = 45
)
```



The values of the correlation coefficients between “age”, “sleeping hours”, and “cognitive scores” are very low, almost near zero, according to the correlation matrix results.

Linear Regression Modelling

Checking Correlation Between Sleep Variables

```
# Check coding
```

```
table(sage$sleepqual)
```

```
##
```

```
##   No  Yes
## 1996  360
```

```
# Correlation between sleep variables
cor.test(
  sage$sleeplength,
  as.numeric(sage$sleepqual) - 1
)
```

```
##
##  Pearson's product-moment correlation
##
## data:  sage$sleeplength and as.numeric(sage$sleepqual) - 1
## t = -14.042, df = 2354, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3148517 -0.2403180
## sample estimates:
##        cor
## -0.2780032
```

Fitting Linear Regression Model

```
lm1 <- lm(
  cognitive_zscore ~ sleeplength + sleepqual,
  data = sage
)

summary(lm1)
```

```
##
## Call:
## lm(formula = cognitive_zscore ~ sleeplength + sleepqual, data =
sage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.114  -2.126  -0.011   2.165  13.717
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.44525    0.29660   1.501    0.133
```

```
## sleeplength  -0.06027    0.03897  -1.547    0.122
## sleepqualYes  0.04794    0.21384   0.224    0.823
##
## Residual standard error: 3.587 on 2353 degrees of freedom
## Multiple R-squared:  0.001212,   Adjusted R-squared:  0.0003632
## F-statistic: 1.428 on 2 and 2353 DF,  p-value: 0.24
```

Presenting Results in a Table

```
tidy(lm1)
```

```
## # A tibble: 3 × 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    0.445     0.297      1.50    0.133
## 2 sleeplength   -0.0603    0.0390    -1.55    0.122
## 3 sleepqualYes   0.0479    0.214      0.224   0.823
```

Check Multicollinearity

```
vif(lm1)
```

```
## sleeplength   sleepqual
##    1.083759    1.083759
```

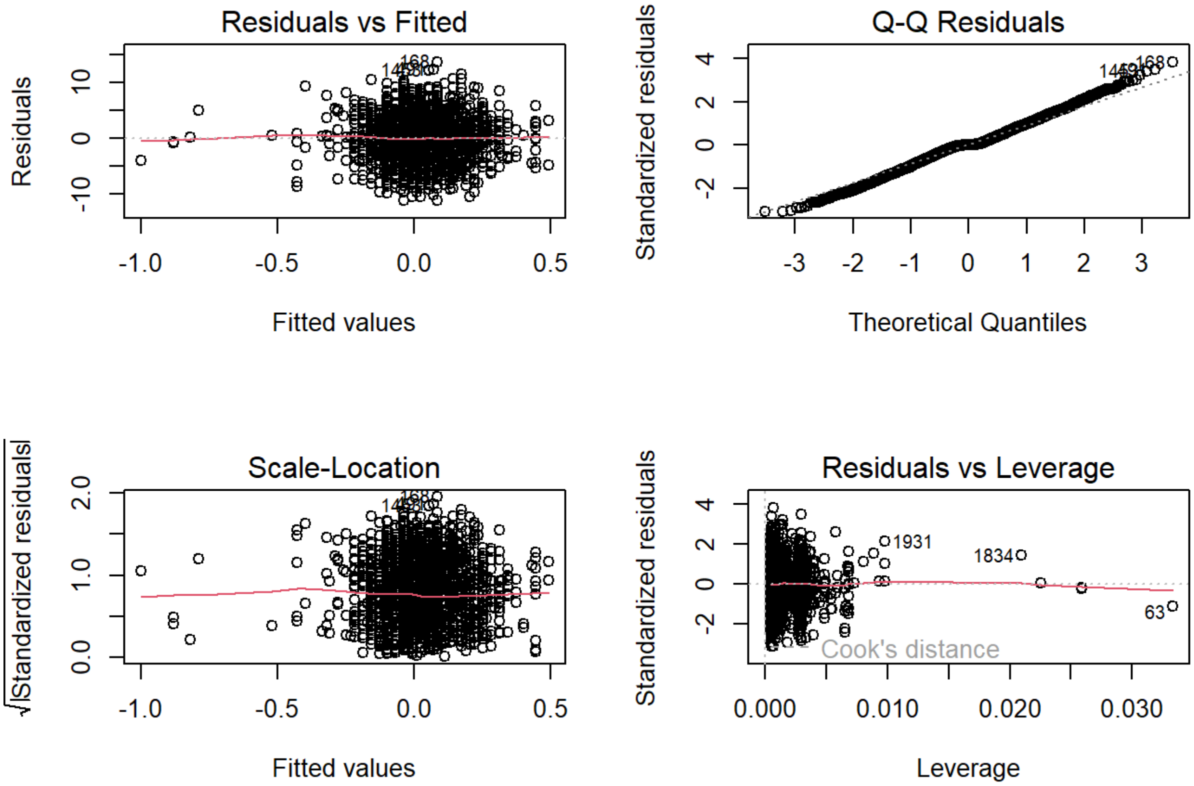
Finding Adjusted R^2

```
summary(lm1)$adj.r.squared
```

```
## [1] 0.0003632134
```

Model Diagnostics

```
par(mfrow = c(2,2))
plot(lm1)
```



“Linear Regression” is performed to analyze the relationship between sleep patterns and cognitive ability. Before performing regression, there is a very weak negative correlation found between sleep length and sleep quality “($r = -0.278$, $p<0.001$)”. There is no issue with multicollinearity as both independent variables has “VIFs” below 1.1 (1.08). Sleep length is negatively related to cognitive ability (“$\beta = -0.060$”, “$p = 0.122$”), meaning that for every one unit increase in hours of sleep, the value of cognitive “z-score” reduced by 0.06 units. People that reported poor sleep quality has marginally higher cognitive “z-score” (“$\beta = 0.048$”, “$p = 0.823$”) than people with good sleep quality; as the relationship do not reach statistical significance. The model is not significant (“$F = 1.43$”, “$p = 0.24$”) and accounted for 0.04% of the variability in cognitive scores (“Adjusted $R^2 = 0.0004$”). It can be observed from the diagnostic plots that the assumptions underlying the model are well satisfied.

Logistic Regression Modelling

Fitting Logistic Regression Model

```
logit1 <- glm(
  ideation ~ sleeplength + sleepqual,
```

```
  family = binomial(link = "logit"),
  data = sage
)

summary(logit1)
```

```
##
## Call:
## glm(formula = ideation ~ sleeplength + sleepqual, family =
binomial(link = "logit"),
##     data = sage)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.78569    0.17287  -4.545 5.50e-06 ***
## sleeplength   0.01527    0.02264   0.674      0.5
## sleepqualYes  0.79721    0.12086   6.596 4.22e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3095.6  on 2355  degrees of freedom
## Residual deviance: 3050.5  on 2353  degrees of freedom
## AIC: 3056.5
##
## Number of Fisher Scoring iterations: 4
```

Obtaining Odds Ratios

```
exp(coef(logit1))
```

```
##  (Intercept)  sleeplength sleepqualYes
##    0.4558052    1.0153885    2.2193346
```

Obtaining 95% Confidence Intervals

```
exp(confint(logit1))
```

```
## Waiting for profiling to be done...
##                  2.5 %    97.5 %
## (Intercept)  0.3245470 0.6396368
## sleeplength  0.9711621 1.0614399
```

```
## sleepqualYes 1.7518345 2.8143467
```

Creating Results Table of Logistic Regression

```
tidy(logit1,
     exponentiate = TRUE,
     conf.int = TRUE)
```

```
## # A tibble: 3 × 7
##   term        estimate std.error statistic  p.value conf.low
conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
<dbl>
## 1 (Intercept)    0.456    0.173     -4.54  5.50e- 6    0.325
0.640
## 2 sleeplength    1.02     0.0226     0.674 5.00e- 1    0.971
1.06
## 3 sleepqualYes   2.22     0.121      6.60  4.22e-11    1.75
2.81
```

Logistic regression analysis is conducted to determine the relationship between sleep length and sleep quality in relation to suicidal ideation. The inclusion of the two sleep variables is justified due to the presence of low correlation between the two factors (“r = -0.278”) and lack of multicollinearity (“VIF = 1.08”). There is no statistically significant relationship between sleep length and suicidal ideation (“OR = 1.015”, “95% CI: 0.971–1.061”, “p = 0.500”). Poor sleep quality is significantly related to suicidal ideation (“OR = 2.219”, “95% CI: 1.752–2.814”, “p < 0.001”). Individuals with poor sleep quality had about 2.2 times greater likelihood of suicidal ideation than those with good sleep quality.

Extending to Multivariable Models

Adjusted Linear Regression

```
lm_adj <- lm(
  cognitive_zscore ~ sleeplength + sleepqual +
    age + gender + centre,
  data = sage
)
```

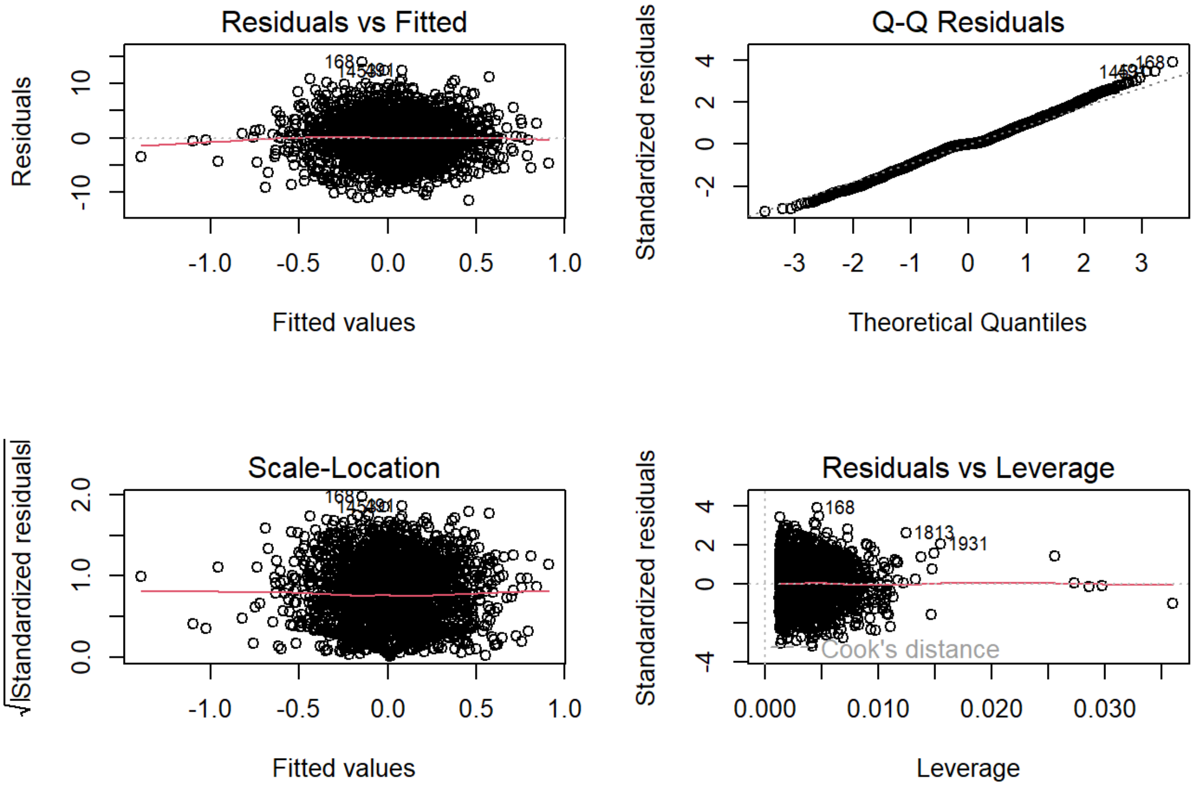
```
summary(lm_adj)
```

```
##
## Call:
## lm(formula = cognitive_zscore ~ sleeplength + sleepqual + age +
##     gender + centre, data = sage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4725  -2.1551  -0.0131   2.1735  13.9470
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.256715   0.595126   2.112   0.0348 *
## sleeplength               -0.057758   0.039774  -1.452   0.1466
## sleepqualYes               0.120935   0.221062   0.547   0.5844
## age                       -0.015095   0.007645  -1.975   0.0484 *
## genderFemale              -0.162323   0.152792  -1.062   0.2882
## centreGhana                0.482243   0.271398   1.777   0.0757 .
## centreIndia                0.166307   0.229933   0.723   0.4696
## centreRussian Federation  0.260806   0.287361   0.908   0.3642
## centreSouth Africa         0.369557   0.348533   1.060   0.2891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.586 on 2347 degrees of freedom
## Multiple R-squared:  0.004701,   Adjusted R-squared:  0.001309
## F-statistic: 1.386 on 8 and 2347 DF,  p-value: 0.1975
```

Model Diagnostics of adjusted Linear Regression

```
par(mfrow = c(2,2))
```

```
plot(lm_adj)
```



Adjusted Logistic Regression

```
logit_adj <- glm(
  ideation ~ sleeplength + sleepqual +
    age + gender + centre,
  family = binomial,
  data = sage
)

summary(logit_adj)
```

```
##
## Call:
## glm(formula = ideation ~ sleeplength + sleepqual + age + gender +
##     centre, family = binomial, data = sage)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.069652   0.356237  -5.810 6.26e-09 ***
```

```
## sleeplength              -0.005057   0.023597  -0.214 0.830303
## sleepqualYes              0.805966   0.127514   6.321 2.61e-10 ***
## age                       0.010247   0.004497   2.279 0.022684 *
## genderFemale              0.390463   0.091604   4.263 2.02e-05 ***
## centreGhana               0.808569   0.165855   4.875 1.09e-06 ***
## centreIndia               0.457150   0.144586   3.162 0.001568 **
## centreRussian Federation  0.674521   0.173483   3.888 0.000101 ***
## centreSouth Africa        1.025114   0.206524   4.964 6.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3095.6  on 2355  degrees of freedom
## Residual deviance: 2986.2  on 2347  degrees of freedom
## AIC: 3004.2
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(logit_adj))
```

```
##              (Intercept)              sleeplength
sleepqualYes
##                0.1262297                0.9949557
2.2388574
##                      age             genderFemale
centreGhana
##                1.0102997                1.4776654
2.2446940
##              centreIndia centreRussian Federation       centreSouth
Africa
##                1.5795652                1.9630924
2.7874146
```

```
exp(confint(logit_adj))
```

```
## Waiting for profiling to be done...

##                              2.5 %    97.5 %
## (Intercept)              0.0626080 0.2531372
## sleeplength              0.9497355 1.0418622
## sleepqualYes             1.7446885 2.8769745
## age                      1.0014290 1.0192461
## genderFemale             1.2354693 1.7693731
## centreGhana              1.6255018 3.1155566
## centreIndia              1.1937236 2.1050286
```

```
## centreRussian Federation 1.3994109 2.7637175
## centreSouth Africa       1.8616130 4.1862490
```

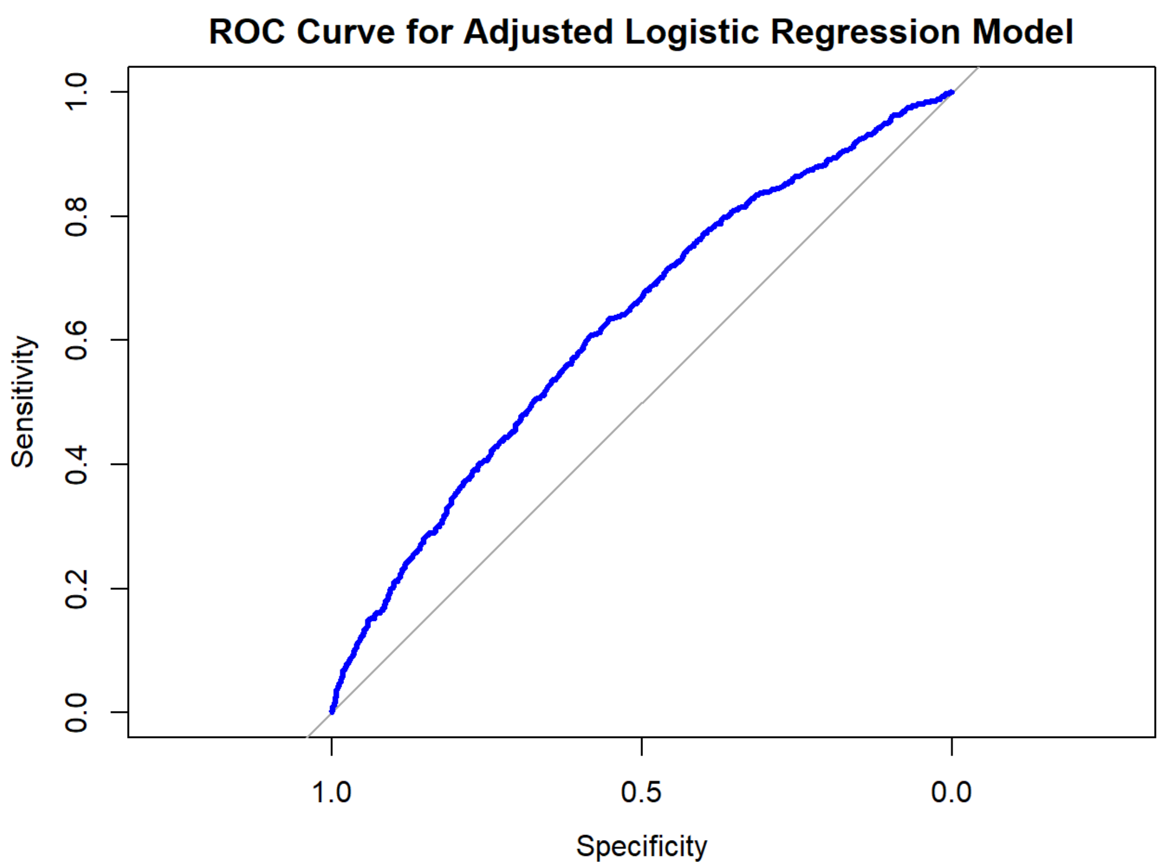
ROC Curve for Adjusted Logistic Regression Model

```
prob <- predict(logit_adj, type = "response")
roc_obj <- roc(sage$ideation, prob)
```

```
## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```

```
plot(
  roc_obj,
  col = "blue",
  lwd = 3,
  main = "ROC Curve for Adjusted Logistic Regression Model"
)
```



```
auc(roc_obj)

## Area under the curve: 0.6251
```

“Linear Regression” analysis is further expanded with control for “age”, “gender”, and “country” to test whether there was any confounding influence. The “β” value for sleep duration changed from “-0.060” in the bivariate analysis to “-0.058” in the multiple regression analysis; at the same time, the “β” value for poor sleep quality changed from 0.048 to 0.121. None of the variables showed significance “($p > 0.05$)” in the analysis and, thus, demonstrated no confounding influence. “Age” is found to be significantly related to reduced cognitive performance (“β = -0.015”; “p = 0.048”). “Logistic Regression” model, the relationship between poor quality sleep and suicidal ideation was still very significant even after adjustment (“OR = 2.24”; “95% CI: 1.74-2.88”; “$p < 0.001$”) compared to the unadjusted “OR of 2.22”. Length of sleep did not become statistically significant (“OR = 0.995”; “95% CI: 0.95-1.04”; “p = 0.830”). Very small changes in these measures of effect indicate that there is little evidence of confounding by demographics. Other variables, such as age (“OR = 1.01”), female sex (“OR = 1.48”), and some countries, are found to be related to suicidal ideation. The “AUC” value from the logistic regression model is moderate at 0.625.

Model Evaluation and Critical Reflection

The performance of the adjusted linear regression model is analyzed with the help of diagnostic graphs and indices for model fit. The plot of residuals versus fitted values revealed that the assumption of linearity is meet to some extent, where the Q-Q plot implies that residuals are normally distributed, showing only slight violations at the extreme ends. The chart of scale-location visualises the stability of residual variance, which confirms the assumption of homoscedasticity. The graph of residuals versus leverage pointed out several influential cases in the data set; there are no significant outliers. Adjusted “R^2” is 0.0013, meaning that 0.13% of the variability in cognitive performance was accounted for by the model (Catalán et al. 2025). Regarding the “ROC” analysis on the logistic regression model, the area under the curve (“AUC”) is found as 0.625, which is considered weak discrimination (Shariatnia et al. 2022). It shows that the predictive ability of the adjusted model in discriminating between those with and without suicidal thoughts is around 62.5%. Sleep quality continued to be a significant predictor of the risk of suicidal ideation (“OR = 2.24”, “95% CI = 1.74 – 2.88”; “$p < 0.001$”), while sleep duration is non-significant. Several limitations must be noted where first, the linear regression model has very little explanatory power, meaning there might have been relevant variables related to cognitive function that are not incorporated in the model (Niu et al. 2025). Second, the logistic regression model showed only average discriminant power, implying that there could be further variables that could help predict the outcome of interest. Finally, self-reports about sleep quality and suicidal thoughts might be biased, although the findings of the model are still useful to analyze (Klimes et al. 2022). Specifically, the models offer public health insight, especially the high correlation between the two outcomes.

Communication and Policy Insight

The research looked at associations between sleep behaviour, cognitive functioning, and suicidal ideation among 2,356 older persons. According to exploratory analyses, most participants experienced high-quality sleep, but 37% has suicidal ideation. Tt became clear for regression analysis that there is no significant relationship between sleep duration and cognitive functioning or suicidal ideation. Also, it is found that there is no significant association between sleep quality and cognitive functioning. Poor sleep quality is associated with suicidal ideation, as participants with poor sleep had more than two times higher odds of suicidal thoughts than those with good sleep quality (“OR = 2.24”, “95% CI: 1.74 – 2.88”). The importance of the study lies in the fact that the issue of sleep quality is identified as one of the key risk factors in determining the possibility of mental disorders in seniors. At the same time, it is necessary to be careful when analyzing the obtained data owing to its cross-sectional and self-reporting nature, which does not allow making any conclusions about causality. The study provides new ideas for studying other variables affecting cognition and mental condition.

References

- Palanivinayagam, A. and Damaševičius, R., 2023. Effective handling of missing values in datasets for classification using machine learning methods. Information, 14(2), p.92.
- Catalán, A., Aymerich, C., Rodríguez-Sánchez, J.M., Pedruzo, B., de Pablo, G.S., Gil, P., Aguayo, F., Acasuso, G., Collado-Pérez, A., Goena, J. and Ibarretxe, O., 2025. Peripheral inflammation and neurocognitive functioning in early psychosis: Specific associations of TNF-α and IL-6 with social cognition. European Psychiatry, 68(1), p.e95.
- Shariatnia, S., Ziaratban, M., Rajabi, A., Salehi, A., Abdi Zarrini, K. and Vakili, M., 2022. Modeling the diagnosis of coronary artery disease by discriminant analysis and logistic regression: a cross-sectional study. BMC medical informatics and decision making, 22(1), p.85.
- Niu, T., Liu, T., Luo, Y.T., Pang, P.C.I., Huang, S. and Xiang, A., 2025. Decoding student cognitive abilities: a comparative study of explainable AI algorithms in educational data mining. Scientific Reports, 15(1), p.26862.
- Klimes-Dougan, B., Mirza, S.A., Babkin, E. and Lanning, C., 2022. Biased reporting of past self-injurious thoughts and behaviors: A literature review. Journal of affective disorders, 308, pp.596-606.