Final Project

Dahye Chung, Donguk Yoo, Hanseung Jang, Sanghyun Lee, Jungyoon Choi,
Seokyeong Park, Semin Seo, Boyeon Kim

2023-07-20

```r
library(tidyverse)
library(broom)
library(tidyr)
library(dplyr)
library(modelr)
library(boot)
library(tidyr)
library(ggplot2)
library(ggmosaic)
library(dplyr)
library(readr)
library(class)
library(caret)
```

Load the dataset

```r
library(tidyr)
library(ggplot2)
library(ggmosaic)
library(dplyr)
Sleep_health_and_lifestyle_dataset <- read_csv("Sleep_health_and_lifestyle_dataset.csv")
```

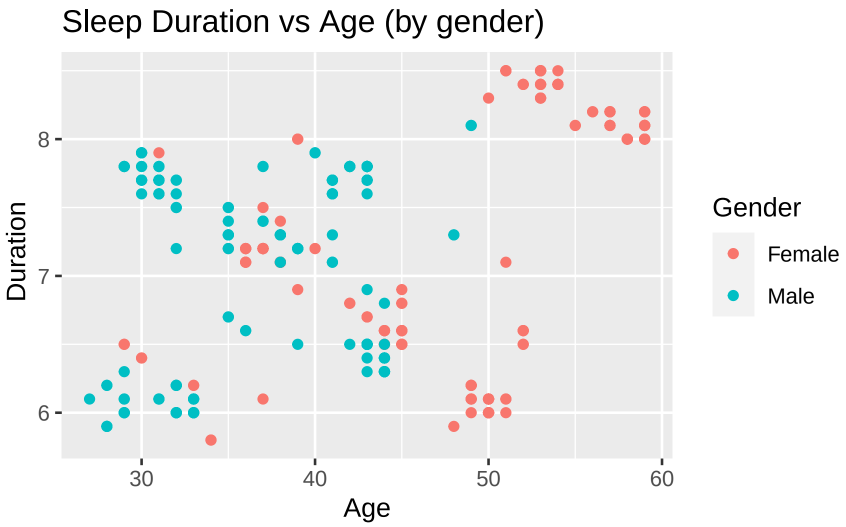
Part1

```r
Sleep_health_and_lifestyle_dataset_renamed <- Sleep_health_and_lifestyle_dataset %>%
  rename( Duration = 'Sleep Duration',
          Stress = 'Stress Level',
          Physical = 'Physical Activity Level' ,
          Quality = 'Quality of Sleep' ,
          BMI= 'BMI Category' ,
          BPressure = 'Blood Pressure' ,
          HRate = 'Heart Rate' ,
          DSteps = 'Daily Steps' ,
          Disorder = 'Sleep Disorder' )
```

Part 2

```
Sleep_health_and_lifestyle_dataset_renamed <- Sleep_health_and_lifestyle_dataset%>%
  rename( ID = "Person ID",
          Duration = 'Sleep Duration',
          Stress = 'Stress Level',
          Physical = 'Physical Activity Level' ,
          Quality = 'Quality of Sleep' ,
          BMI= 'BMI Category' ,
          BPressure = 'Blood Pressure' ,
          HRate = 'Heart Rate' ,
          DSteps = 'Daily Steps' ,
          Disorder = 'Sleep Disorder' )
```

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot()+
  geom_point( mapping = aes( x = Age , y = Duration, color = Gender))+
  labs(
   title = "Sleep Duration vs Age (by gender)",
   x= "Age", y = " Duration")
```

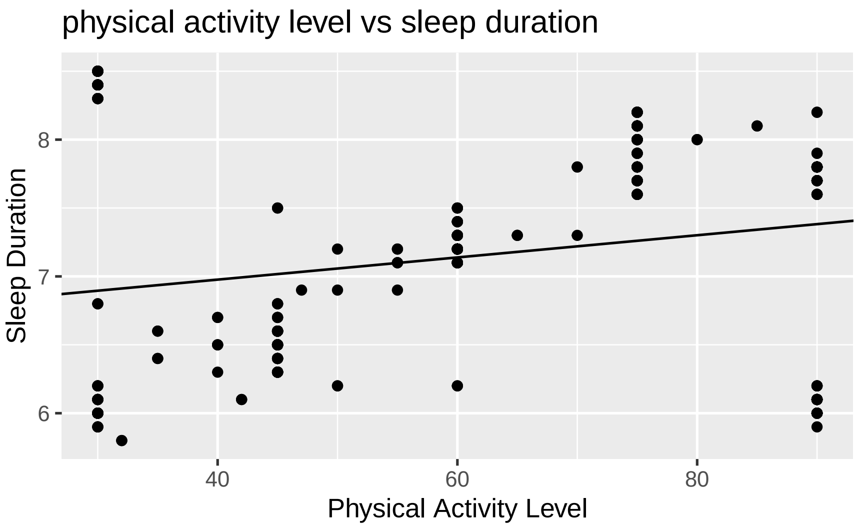


```
model_2 <- lm(Duration ~ Physical,Sleep_health_and_lifestyle_dataset_renamed)
```

```
model_2$coefficients
```

```
## (Intercept)    Physical
## 6.652127945 0.008111349
```

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
  geom_point(mapping = aes(x = Physical, y = Duration), bin = 10) +
  geom_abline(slope = model_2$coefficients[2],
              intercept = model_2$coefficients[1])+
   labs(x = "Physical Activity Level", y = "Sleep Duration",
                    title = "physical activity level vs sleep duration" )
```



Part 3

```
Sleep_health_and_lifestyle_dataset_renamed$BMI[Sleep_health_and_lifestyle_dataset_renamed$BMI
Sleep_health_and_lifestyle_dataset_renamed$BMI[Sleep_health_and_lifestyle_dataset_renamed$BMI
Sleep_health_and_lifestyle_dataset_renamed$BMI[Sleep_health_and_lifestyle_dataset_renamed$BMI
```

```
head(Sleep_health_and_lifestyle_dataset_renamed) %>%
  select(ID, HRate, Duration, Gender, Age, Occupation, Physical, BMI, Quality) %>%
  arrange(Duration)
```

ID	HRate	Duration	Gender	Age	Occupation	Physical	BMI	Quality
4	85	5.9	Male	28	Sales Representative	30	Fat	4
5	85	5.9	Male	28	Sales Representative	30	Fat	4

ID	HRate	Duration	Gender	Age	Occupation	Physical	BMI	Quality
6	85	5.9	Male	28	Software Engineer	30	Fat	4
1	77	6.1	Male	27	Software Engineer	42	Fat	6
2	75	6.2	Male	28	Doctor	60	Normal	6
3	75	6.2	Male	28	Doctor	60	Normal	6

```
tail(Sleep_health_and_lifestyle_dataset_renamed) %>%
 select(ID, HRate, Duration, Gender, Age, Occupation, Physical, BMI, Quality) %>%
  arrange(Duration) %>%
  filter(Gender == 'Female')
```

ID	HRate	Duration	Gender	Age	Occupation	Physical	BMI	Quality
371	68	8.0	Female	59	Nurse	75	Fat	9
369	68	8.1	Female	59	Nurse	75	Fat	9
370	68	8.1	Female	59	Nurse	75	Fat	9
372	68	8.1	Female	59	Nurse	75	Fat	9
373	68	8.1	Female	59	Nurse	75	Fat	9
374	68	8.1	Female	59	Nurse	75	Fat	9

```
head(Sleep_health_and_lifestyle_dataset_renamed) %>%
 select(ID, HRate, Duration, Gender, Age, Occupation, Physical, BMI, Quality) %>%
  arrange(Duration) %>%
  filter(Gender == 'Male')
```

ID	HRate	Duration	Gender	Age	Occupation	Physical	BMI	Quality
4	85	5.9	Male	28	Sales Representative	30	Fat	4
5	85	5.9	Male	28	Sales Representative	30	Fat	4
6	85	5.9	Male	28	Software Engineer	30	Fat	4
1	77	6.1	Male	27	Software Engineer	42	Fat	6
2	75	6.2	Male	28	Doctor	60	Normal	6
3	75	6.2	Male	28	Doctor	60	Normal	6

Part4

Explore dataset

```
head(Sleep_health_and_lifestyle_dataset)
```

Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
6	Male	28	Software Engineer	5.9	4	30	8	Obese	140/90	85	3000	Insomnia

```
tail(Sleep_health_and_lifestyle_dataset)
```

Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
369	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
370	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
371	Female	59	Nurse	8.0	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
372	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
373	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
374	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea

Check summary

```
summary(Sleep_health_and_lifestyle_dataset_renamed)
```

```
##        ID            Gender               Age         Occupation
##  Min.   :  1.00   Length:374         Min.   :27.00   Length:374
##  1st Qu.: 94.25   Class :character   1st Qu.:35.25   Class :character
##  Median :187.50   Mode  :character   Median :43.00   Mode  :character
##  Mean   :187.50                      Mean   :42.18
##  3rd Qu.:280.75                      3rd Qu.:50.00
##  Max.   :374.00                      Max.   :59.00
##     Duration        Quality         Physical         Stress
##  Min.   :5.800   Min.   :4.000   Min.   :30.00   Min.   :3.000
##  1st Qu.:6.400   1st Qu.:6.000   1st Qu.:45.00   1st Qu.:4.000
##  Median :7.200   Median :7.000   Median :60.00   Median :5.000
##  Mean   :7.132   Mean   :7.313   Mean   :59.17   Mean   :5.385
##  3rd Qu.:7.800   3rd Qu.:8.000   3rd Qu.:75.00   3rd Qu.:7.000
##  Max.   :8.500   Max.   :9.000   Max.   :90.00   Max.   :8.000
##      BMI              BPressure             HRate           DSteps
##  Length:374         Length:374         Min.   :65.00   Min.   : 3000
##  Class :character   Class :character   1st Qu.:68.00   1st Qu.: 5600
##  Mode  :character   Mode  :character   Median :70.00   Median : 7000
##                                        Mean   :70.17   Mean   : 6817
##                                        3rd Qu.:72.00   3rd Qu.: 8000
##                                        Max.   :86.00   Max.   :10000
##    Disorder
##  Length:374
##  Class :character
##  Mode  :character
##
##
##
```

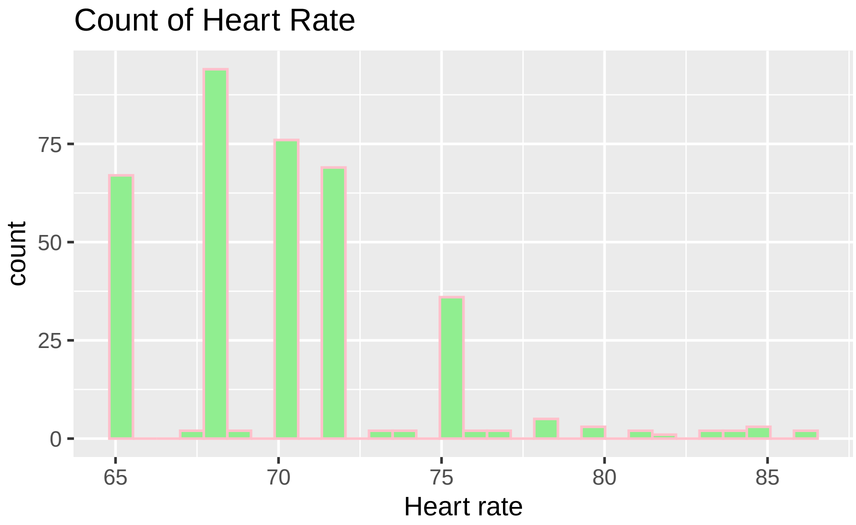
```
Sleep_health_and_lifestyle_dataset_renamed %>%
  summarize(
    standard_deviation = sd(HRate)

  )
```

standard_deviation
4.135675

Visualizing data

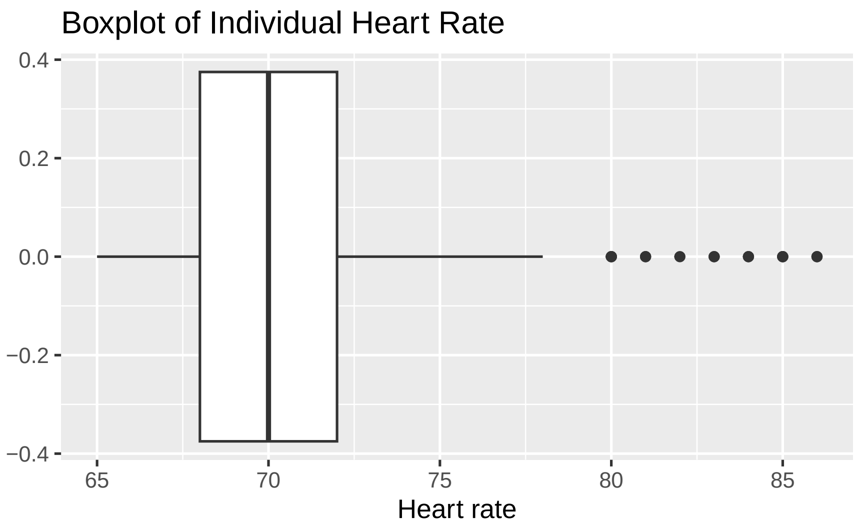
Histogram

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
    geom_histogram(mapping = aes(x = HRate), color = "pink", fill = "lightgreen") +
    labs(title = "Count of Heart Rate", x = "Heart rate")
```



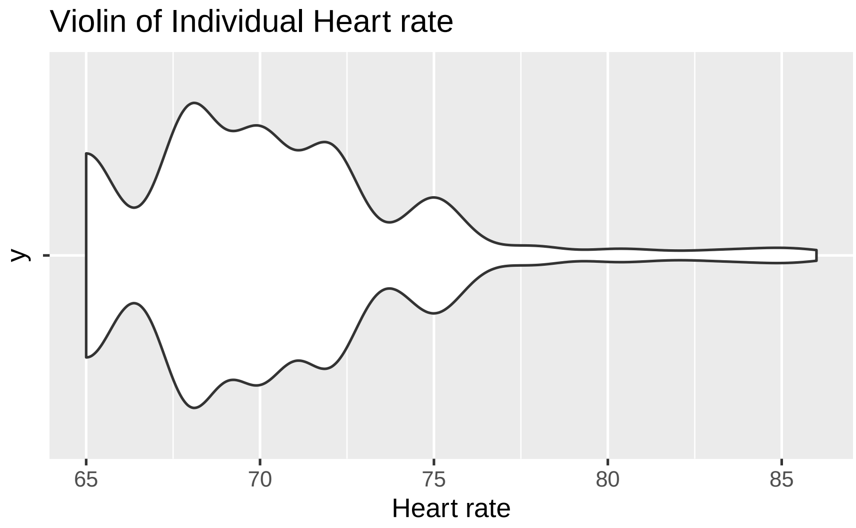
Box plot

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
    geom_boxplot(mapping = aes(x = HRate)) +
    labs(title = "Boxplot of Individual Heart Rate", x = "Heart rate")
```



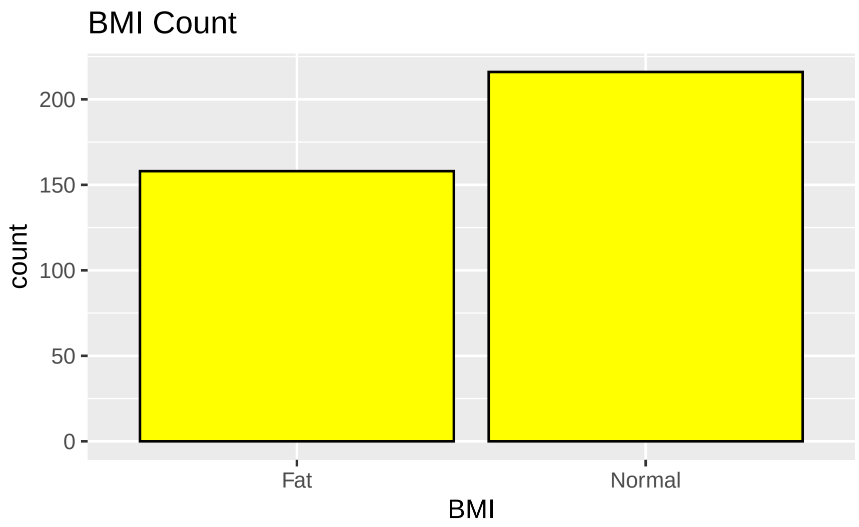
Violin plot

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
    geom_violin(mapping = aes(x = HRate, y ="")) +
    labs(title = "Violin of Individual Heart rate", x = "Heart rate", y = "y")
```



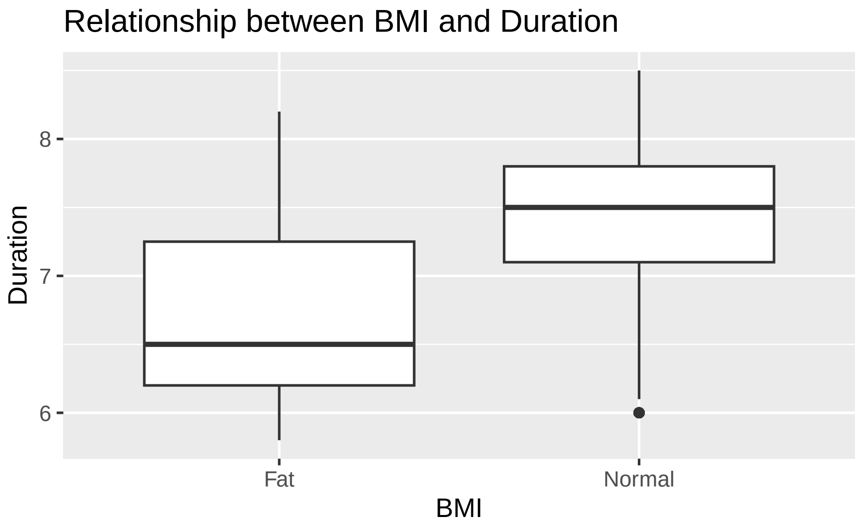
Bar Graph

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
    geom_bar(mapping = aes(x = BMI), color = "black", fill = "yellow") +
    labs(title = "BMI Count", x = "BMI")
```



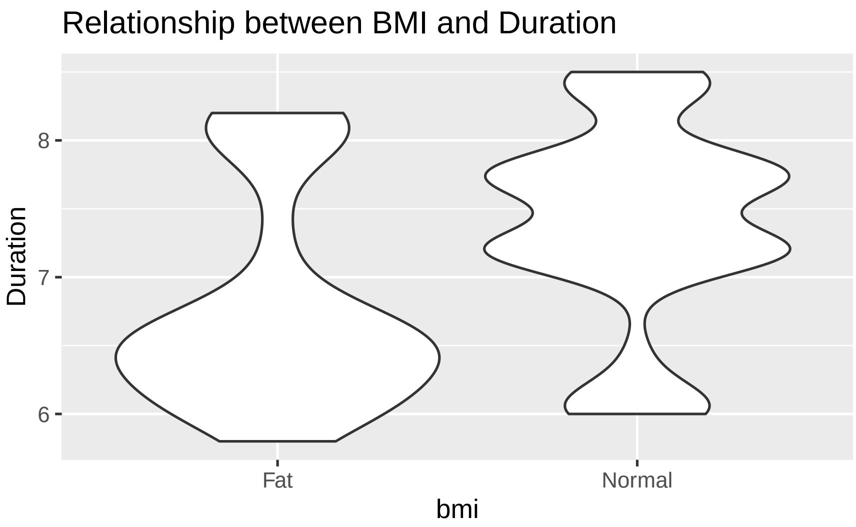
Box plot

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
    geom_boxplot(mapping = aes(x = BMI, y = Duration)) +
    labs(title = "Relationship between BMI and Duration", x = "BMI")
```



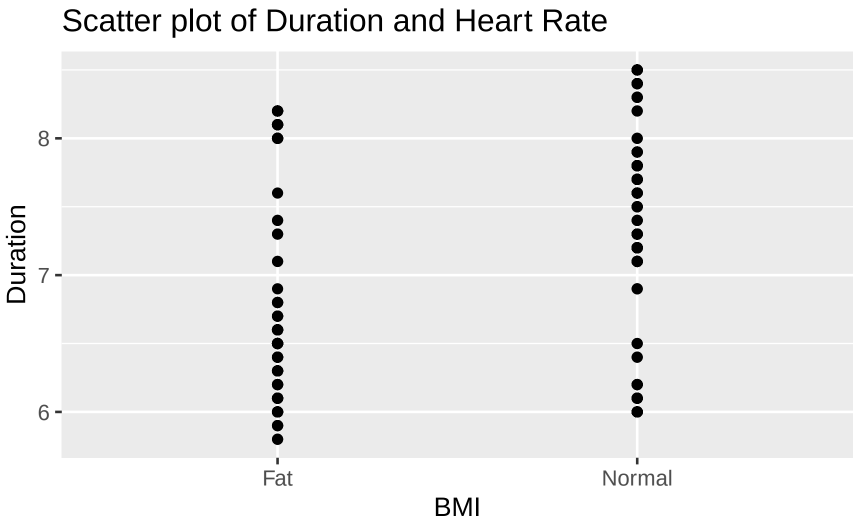
Violin plot

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  ggplot() +
    geom_violin(mapping = aes(x = BMI, y = Duration)) +
    labs(title = "Relationship between BMI and Duration", x = "bmi", y = "Duration")
```



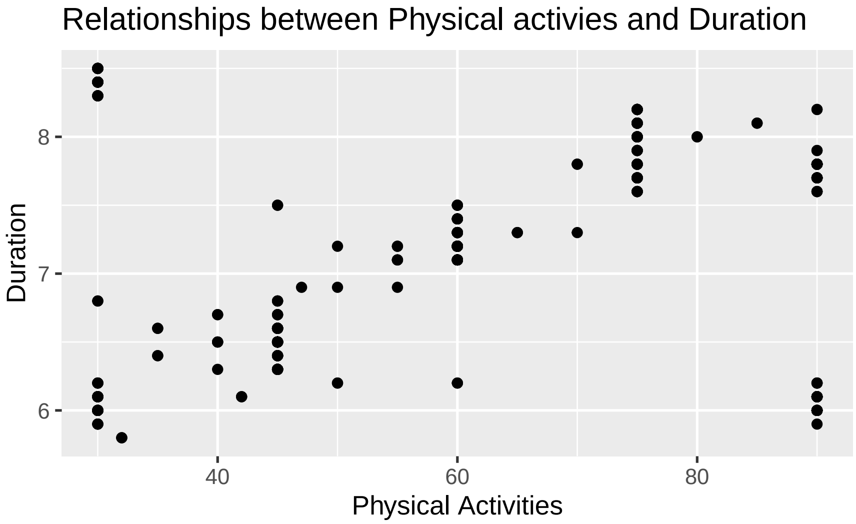
#Scatter plot_Duration and Heart Rate

```
Sleep_health_and_lifestyle_dataset_renamed    %>%
ggplot()    +
geom_point(mapping = aes(x = BMI, y = Duration))    +
labs(
title = "Scatter plot of Duration and Heart Rate",
x = "BMI",
y = "Duration"
)
```



PART 5 __ Modeling

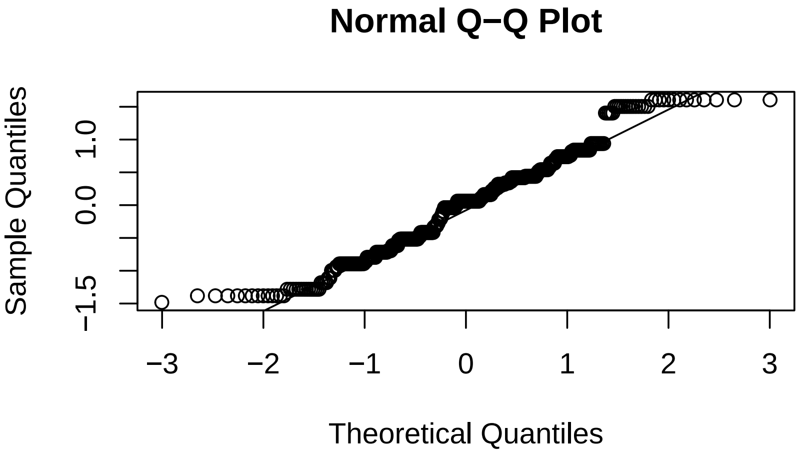
```
Sleep_health_and_lifestyle_dataset_renamed%>%
  ggplot()+
  geom_point( mapping = aes( x  = Physical , y = Duration)) +
  labs(title = "Relationships between Physical activies and Duration",
       x = "Physical Activities" , y = "Duration")
```



```
data <- Sleep_health_and_lifestyle_dataset_renamed

model <- lm(Duration ~ Physical, data = Sleep_health_and_lifestyle_dataset_renamed)


summary(model)
```

```
##
## Call:
## lm(formula = Duration ~ Physical, data = Sleep_health_and_lifestyle_dataset_renamed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48215 -0.59686  0.06119  0.43952  1.60453
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.652128   0.121379  54.805  < 2e-16 ***
## Physical    0.008111   0.001935   4.191 3.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7786 on 372 degrees of freedom
## Multiple R-squared:  0.0451, Adjusted R-squared:  0.04253
## F-statistic: 17.57 on 1 and 372 DF,  p-value: 3.467e-05
```

```
residuals <- residuals(model)

qqnorm(residuals)
qqline(residuals)
```



```
labs( title  = "QQplot" , x = "Theoretical" , y = "Quantaties")
```

```
## $x
## [1] "Theoretical"
##
## $y
## [1] "Quantaties"
##
## $title
## [1] "QQplot"
##
## attr(,"class")
## [1] "labels"
```

```
Renamed_other_model <- lm(Duration ~ Physical, data = Sleep_health_and_lifestyle_dataset_renam
```

```
Renamed_other_model$coefficients
```

```
## (Intercept)    Physical
## 6.652127945 0.008111349
```

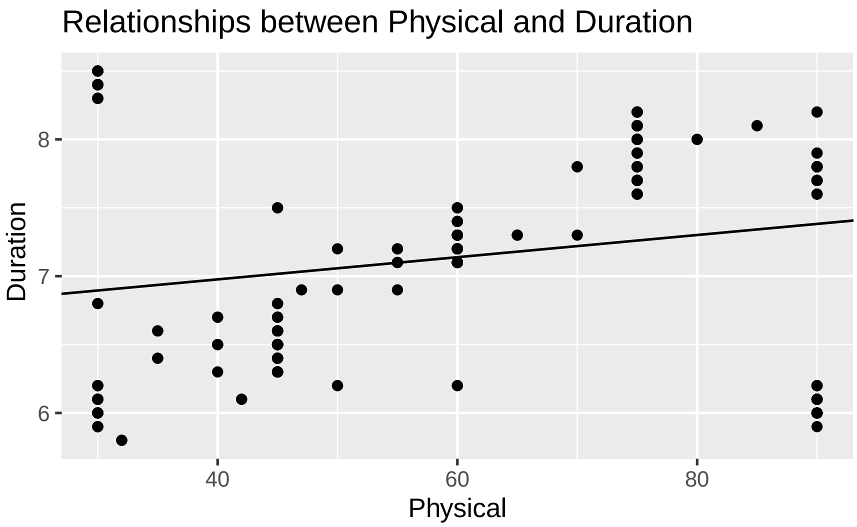
```
Renamed_other_model%>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	6.6521279	0.1213792	54.804523	0.00e+00
Physical	0.0081113	0.0019352	4.191459	3.47e-05

```
Renamed_other_model%>%
  glance()%>%
  select(r.squared)
```

r.squared
0.0450969

```
Sleep_health_and_lifestyle_dataset_renamed%>%
  ggplot()+
  geom_point(mapping = aes( x  = Physical , y = Duration) )+
  geom_abline(slope = Renamed_other_model$coefficients[2]    ,
              intercept = Renamed_other_model$coefficients[1]   )+
  labs( title = "Relationships between Physical and Duration",
        x = " Physical ",
        y = " Duration" )
```



Load the dataset

```
Sleep_health_and_lifestyle_dataset <- read_csv(file = "Sleep_health_and_lifestyle_dataset.csv"
  col_types = cols(
    'Person ID' = col_character(),
    'Age' = col_double(),
    'Sleep Duration' = col_double(),
    'Stress Level' = col_double(),
    'Physical Activity Level' = col_double(),
    'Quality of Sleep' = col_double(),
    'BMI Category' = col_character(),
    'Blood Pressure' = col_character(),
    'Heart Rate' = col_double(),
    'Daily Steps' = col_double(),
    'Sleep Disorder' = col_character()
  ))
```

Rename

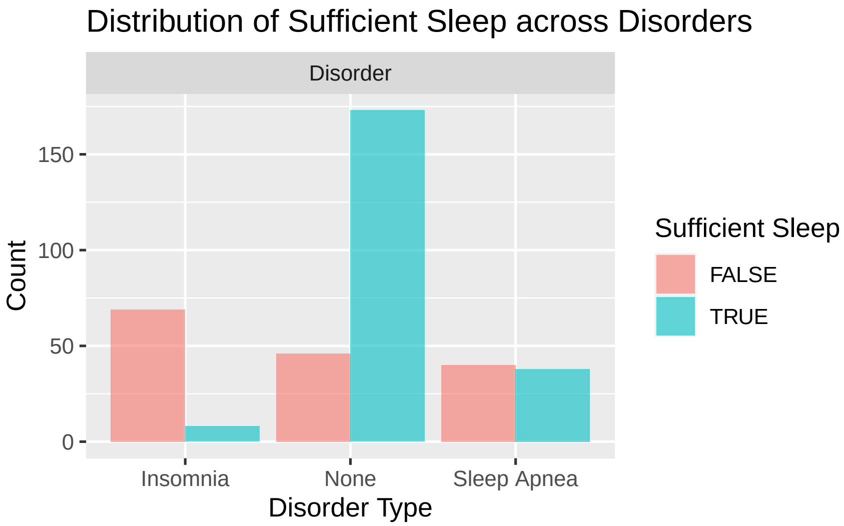
```
Sleep_health_and_lifestyle_dataset_renamed <- Sleep_health_and_lifestyle_dataset %>%
  rename(ID = 'Person ID',
         Duration = 'Sleep Duration',
         Stress = 'Stress Level',
         Physical = 'Physical Activity Level',
         Quality = 'Quality of Sleep',
         BMI = 'BMI Category',
         BPressure = 'Blood Pressure',
         HRate = 'Heart Rate',
         DSteps = 'Daily Steps',
         Disorder = 'Sleep Disorder')
```

Parse Sleep Data

```
sleep_data <- Sleep_health_and_lifestyle_dataset_renamed %>%
    mutate(sufficient_sleep = as.logical(Duration >= 7.0))
```

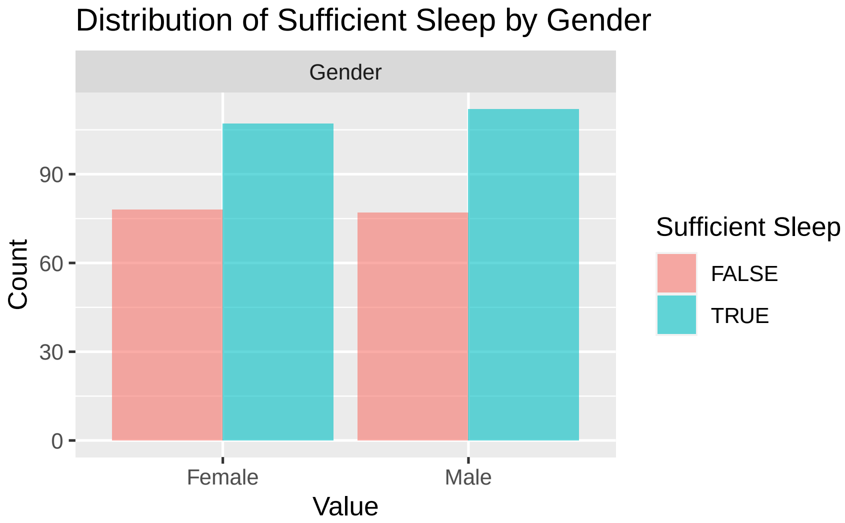
Sleep Data Disorders

```
sleep_data %>%
  pivot_longer(cols = c(Disorder), names_to = "variable", values_to = "value") %>%
  group_by(variable, value, sufficient_sleep) %>%
  summarise(count = n()) %>%
  ggplot() +
  geom_bar(
    mapping = aes(x = value, y = count, fill = sufficient_sleep),
    position = "dodge",
    alpha = 0.6,
    stat = "identity"
  ) +
  facet_wrap(~ variable, scales = "free") +
  labs(title = "Distribution of Sufficient Sleep across Disorders",
       x = "Disorder Type",
       y = "Count",
       fill = "Sufficient Sleep")
```



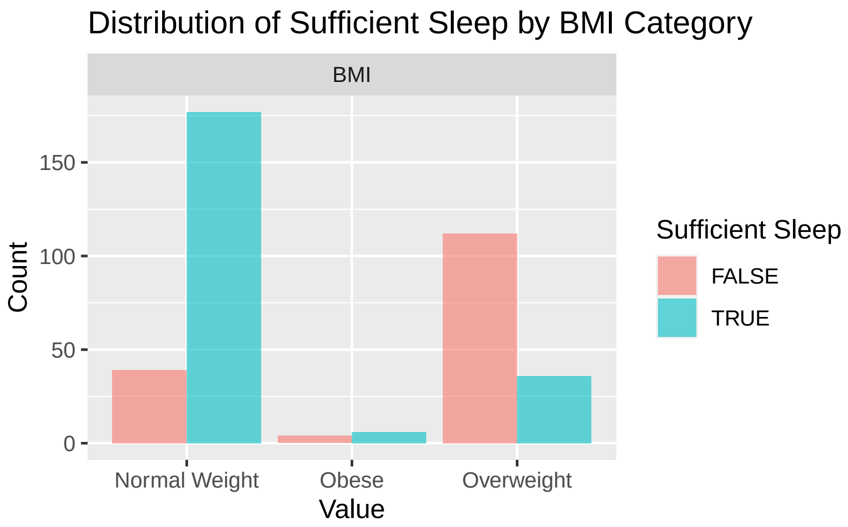
Sleep Data Gender

```
sleep_data %>%
  pivot_longer(cols = c(Gender), names_to = "variable", values_to = "value") %>%
  group_by(variable, value, sufficient_sleep) %>%
  summarise(count = n()) %>%
  ggplot() +
  geom_bar(
    mapping = aes(x = value, y = count, fill = sufficient_sleep),
    position = "dodge",
    alpha = 0.6,
    stat = "identity"
  ) +
  facet_wrap(~ variable, scales = "free") +
  labs(title = "Distribution of Sufficient Sleep by Gender",
       x = "Value",
       y = "Count",
       fill = "Sufficient Sleep")
```



Sleep Data BMI

```
sleep_data %>%
  pivot_longer(cols = c(BMI), names_to = "variable", values_to = "value") %>%
  mutate(value = ifelse(value == "Normal", "Normal Weight", value)) %>%
  group_by(variable, value, sufficient_sleep) %>%
  summarise(count = n()) %>%
  ggplot() +
  geom_bar(
    mapping = aes(x = value, y = count, fill = sufficient_sleep),
    position = "dodge",
    alpha = 0.6,
    stat = "identity"
  ) +
  facet_wrap(~ variable, scales = "free") +
  labs(title = "Distribution of Sufficient Sleep by BMI Category",
       x = "Value",
       y = "Count",
       fill = "Sufficient Sleep")
```



Mode

```
mode_gender <- as.character(names(which.max(table(sleep_data$Gender))))
mode_occupation <- as.character(names(which.max(table(sleep_data$Occupation))))
mode_bmi <- as.character(names(which.max(table(sleep_data$BMI))))

sleep_data <- sleep_data %>%
mutate(
  Gender = if_else(is.na(Gender), mode_gender, Gender),
  Occupation = if_else(is.na(Occupation), mode_occupation, Occupation),
  BMI = if_else(is.na(BMI), mode_bmi, BMI)
)
```

Sufficient Sleep

```
sleep_data$sufficient_sleep <- ifelse(sleep_data$Duration >= 7, "Sufficient", "Insufficient")
```

Saparate Train, Test Set

```
set.seed(123)
train_indices <- createDataPartition(sleep_data$sufficient_sleep, p = 0.7, list = FALSE)
trainingSet <- sleep_data[train_indices, ]
testSet <- sleep_data[-train_indices, ]
```

```r
trainingSet$sufficient_sleep <- as.factor(trainingSet$sufficient_sleep)
testSet$sufficient_sleep <- as.factor(testSet$sufficient_sleep)

training_Outcomes <- trainingSet$sufficient_sleep
test_Outcomes <- testSet$sufficient_sleep
```

Train

```r
model <- glm(sufficient_sleep ~ Age + Gender + Occupation + Physical + DSteps + BMI, data = tra
```

Predict

```r
predictions <- predict(model, newdata = testSet, type = "response")
```

Test

```r
threshold <- 0.5
predicted_classes <- as.factor(ifelse(predictions >= threshold, "Sufficient", "Insufficient"))
actual_classes <- test_Outcomes
accuracy <- sum(predicted_classes == actual_classes) / length(actual_classes)
print(paste("Accuracy:", accuracy))
```

```
## [1] "Accuracy: 0.981981981981982"
```

```r
model_1_preds <- testSet %>%
  add_predictions(model, type = "response") %>%
  mutate(
    outcome = as.factor(if_else(condition = pred > threshold,
                        "Sufficient", "Insufficient"))
  )
```