Final Project

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

2023-07-21

```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)
library(infer)
```

#Intro

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

```r
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' )
```

EDA

###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

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
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

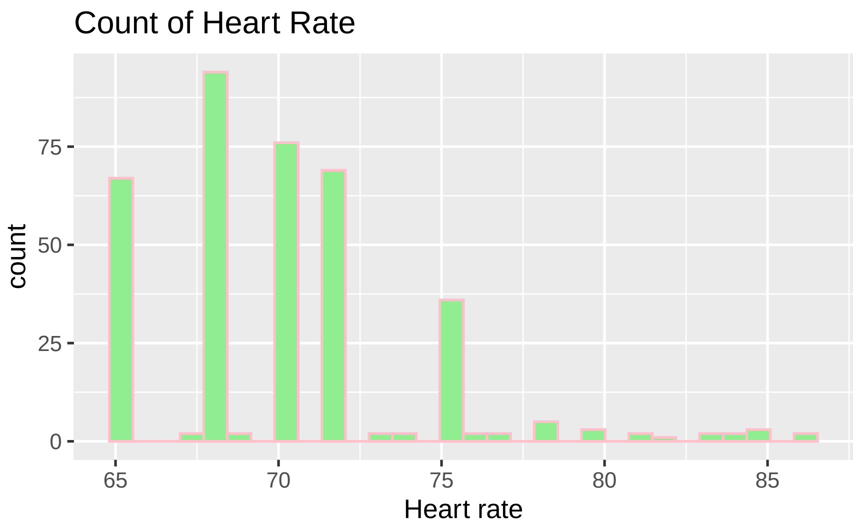
```
Sleep_health_and_lifestyle_dataset_renamed %>%
  summarize(
    standard_deviation = sd(HRate)

  )
```

standard_deviation
4.135675

Visualizing data

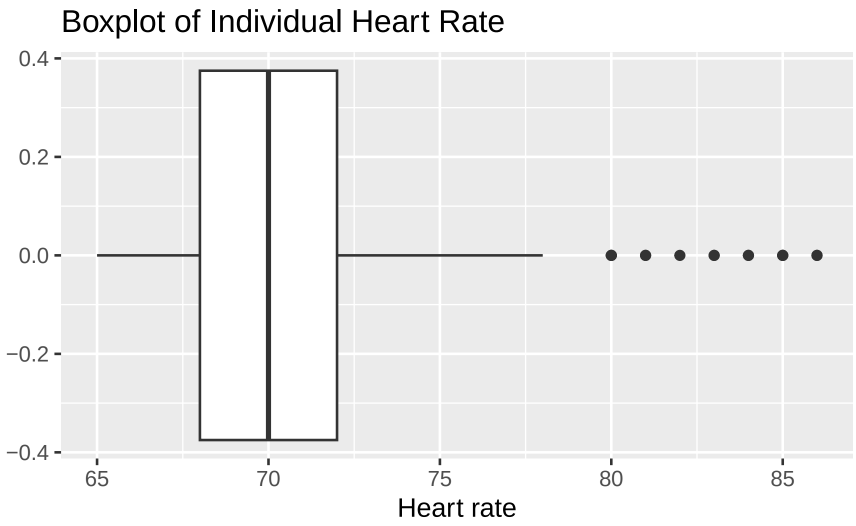
###Histogram

```
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")
```



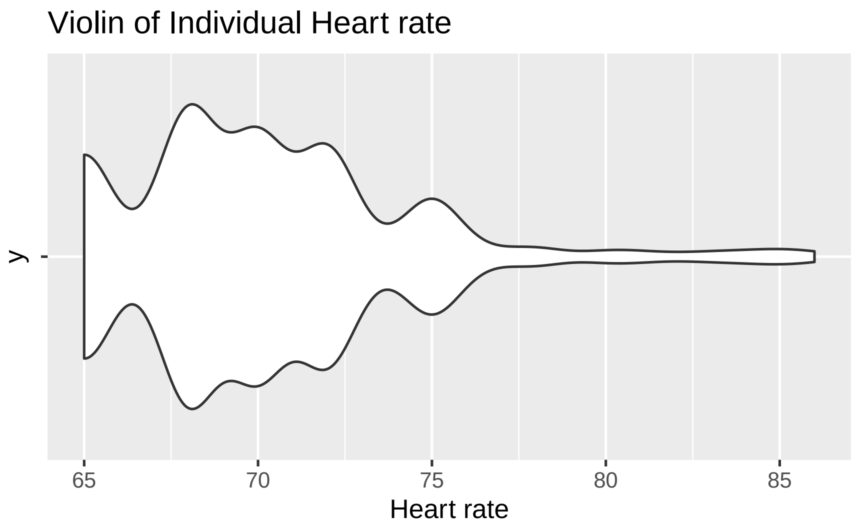
###Box plot

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



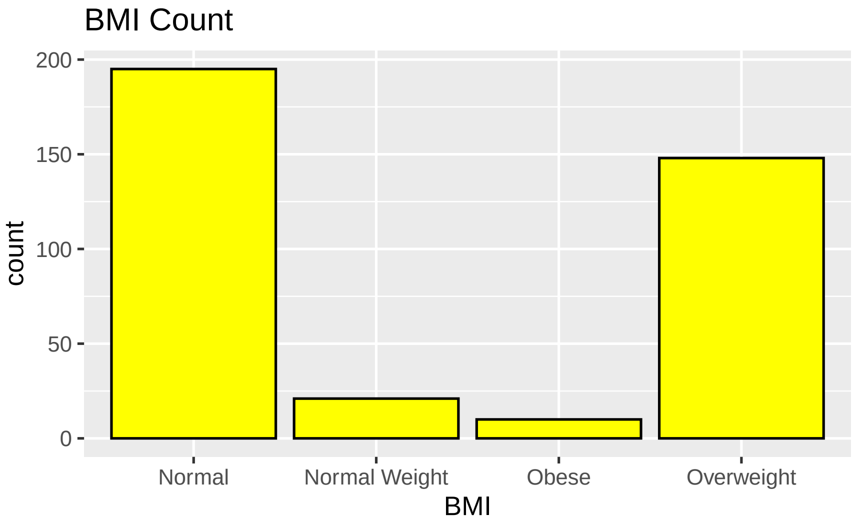
Violin plot

```
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")
```



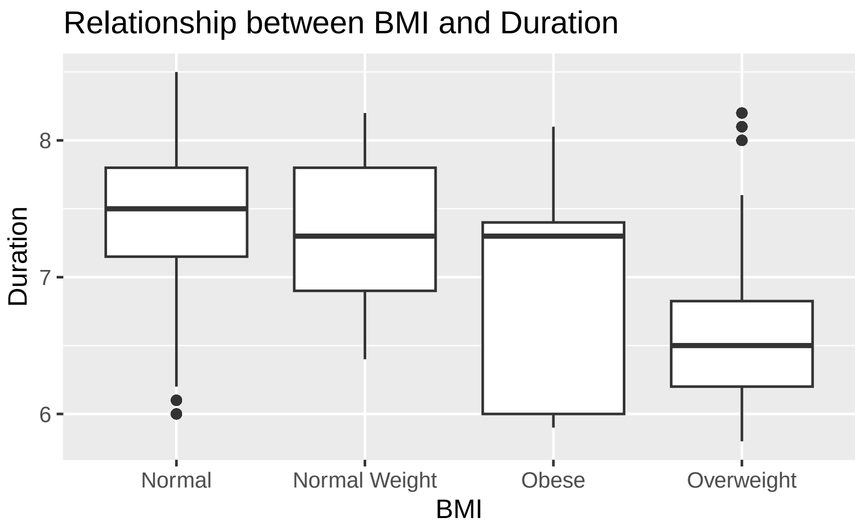
###Bar Graph

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



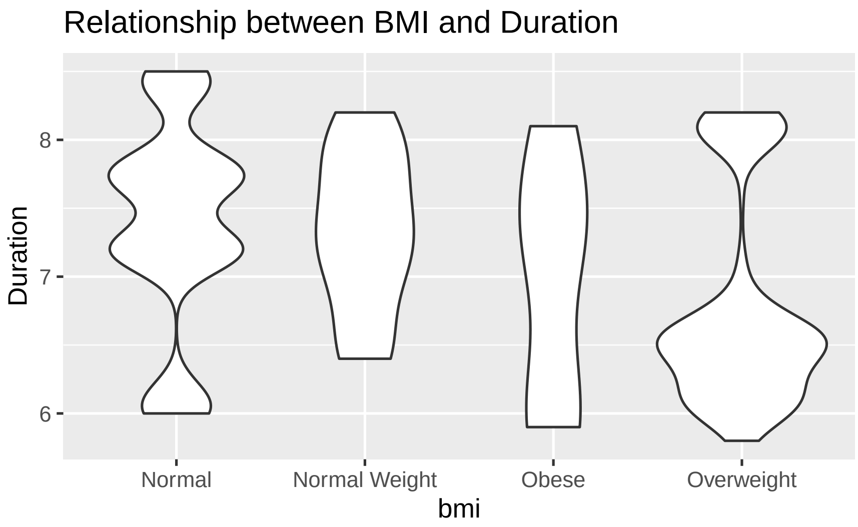
###Box plot

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



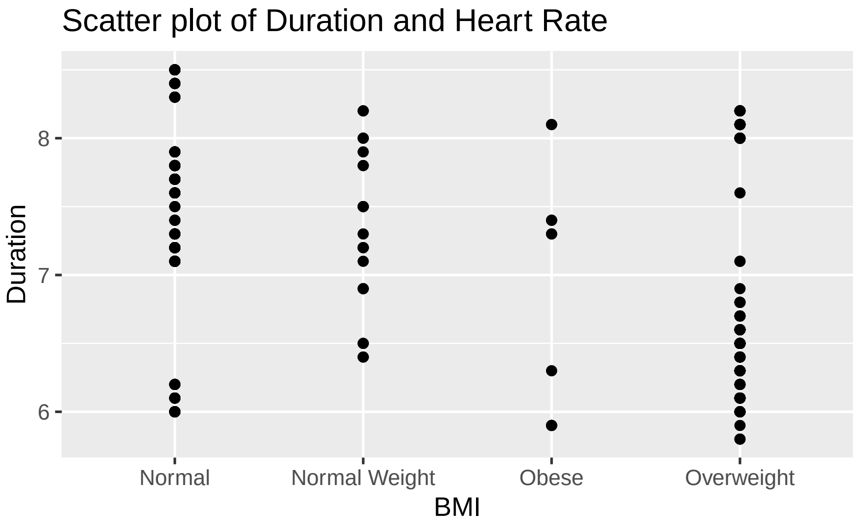
###Violin plot

```
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"
)
```



Data Wrangling

```
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
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

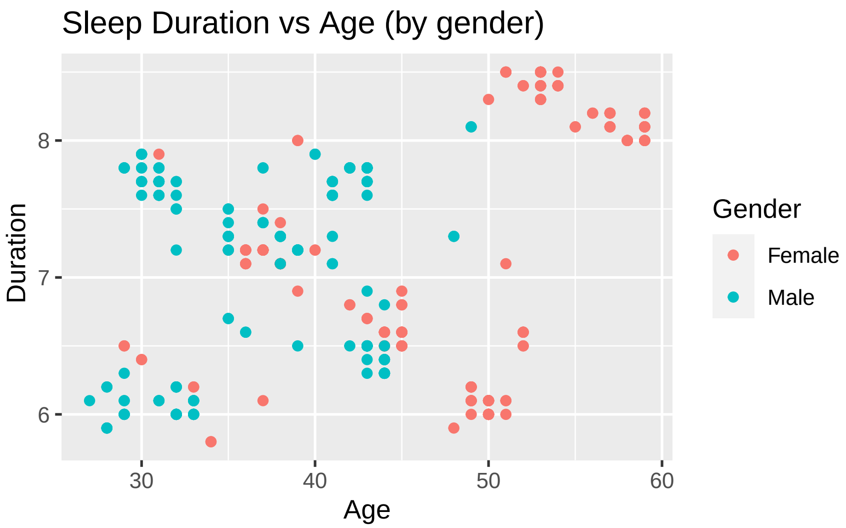
```
head(Sleep_health_and_lifestyle_dataset_renamed) %>%
 select(ID, HRate, Duration, Gender, Age, Occupation, Physical, BMI, Quality) %>%
  arrange(Duration) %>%
  filter(Gender == 'Male')
```

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

Data Visualization

```
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")
```

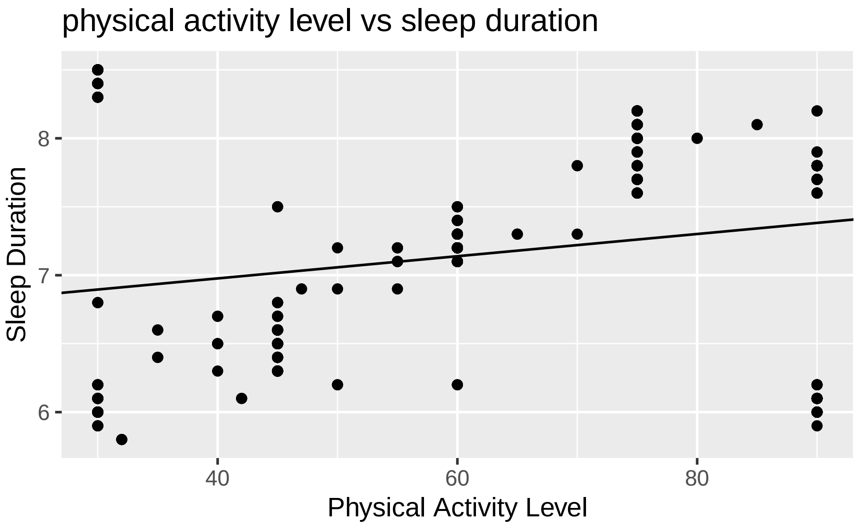


```
model_2 <- lm(Duration ~ Physical,Sleep_health_and_lifestyle_dataset_renamed)
```

```
model_2$coefficients
```

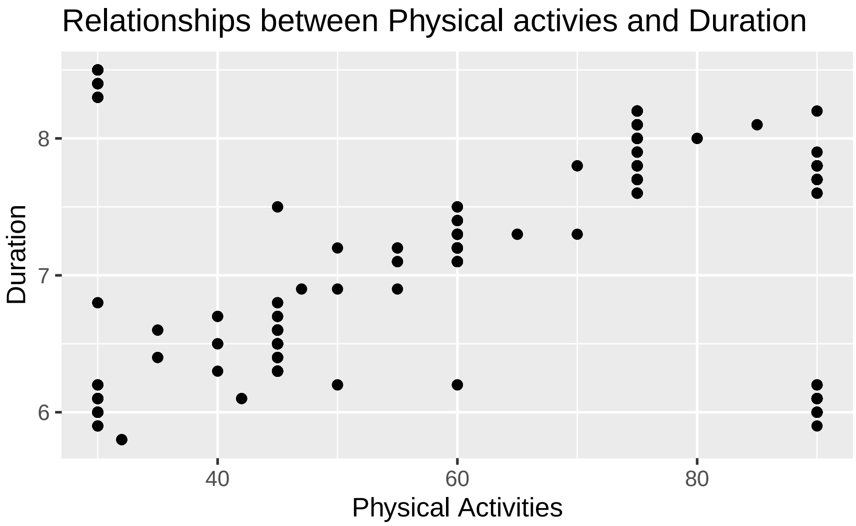
```
## (Intercept)    Physical
## 6.652127945 0.008111349
```

```
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" )
```



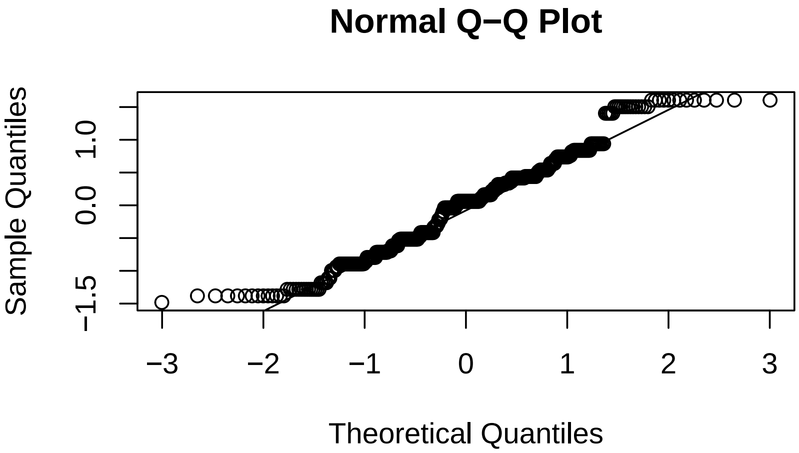
Modeling

```
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)
```

```
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
```

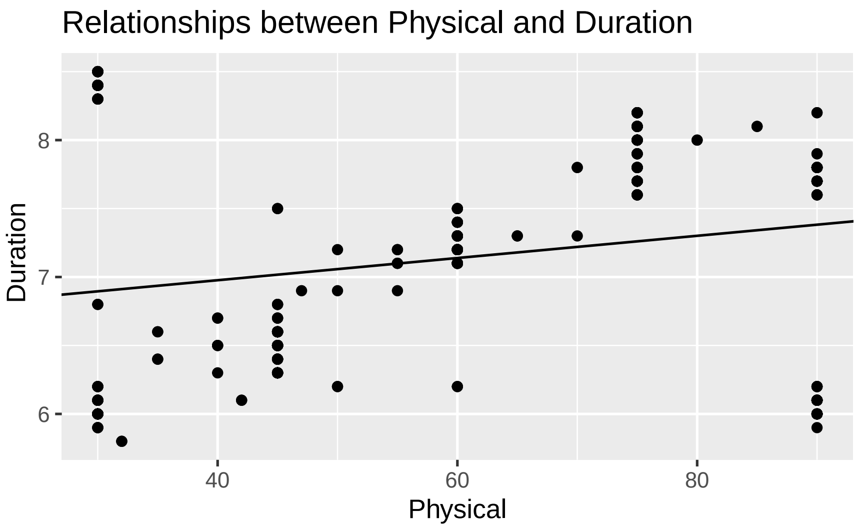
```
Renamed_other_model%>%
  tidy()
```

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" )
```



#Advanced Modeling

```
continuous_model <- lm(Duration ~ Gender + Age + Occupation + DSteps + BMI + Physical, data = S
coefficients <- tidy (continuous_model)
coefficients
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.1538671	0.2889932	17.8338716	0.0000000
GenderMale	-0.2383578	0.1312962	-1.8154203	0.0703006
Age	0.0632686	0.0063798	9.9170335	0.0000000
OccupationDoctor	0.4771137	0.1436900	3.3204376	0.0009918

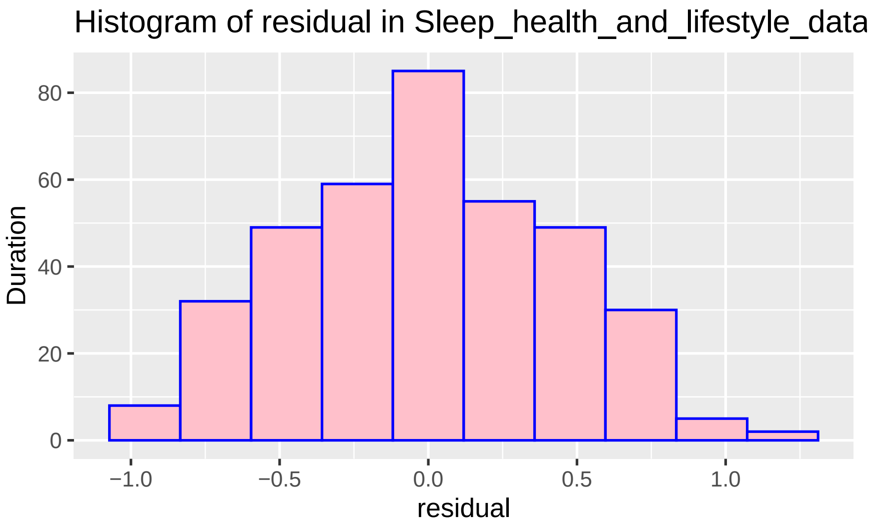
term	estimate	std.error	statistic	p.value
OccupationEngineer	0.3329909	0.1416952	2.3500512	0.0193162
OccupationLawyer	0.3442380	0.1574134	2.1868409	0.0294044
OccupationManager	0.0481333	0.4644330	0.1036388	0.9175143
OccupationNurse	-0.2414602	0.1234123	-1.9565332	0.0511838
OccupationSales Representative	0.6880740	0.3890627	1.7685429	0.0778265
OccupationSalesperson	0.2949398	0.1766485	1.6696421	0.0958692
OccupationScientist	0.2136997	0.2697812	0.7921220	0.4288171
OccupationSoftware Engineer	0.8351780	0.2669096	3.1290667	0.0018984
OccupationTeacher	0.3032818	0.1249970	2.4263134	0.0157491
DSteps	-0.0002816	0.0000290	-9.7161550	0.0000000
BMINormal Weight	-0.0266284	0.1150426	-0.2314658	0.8170860
BMIObese	-1.4449954	0.2024347	-7.1380799	0.0000000
BMIOverweight	-1.0938883	0.1247308	-8.7699953	0.0000000
Physical	0.0271587	0.0023259	11.6767906	0.0000000

```r
r_squared <- glance(continuous_model)$r.squared
```

```r
Sleep_health_and_lifestyle_dataset_df <- Sleep_health_and_lifestyle_dataset_renamed %>%
  add_predictions(continuous_model) %>%
  add_residuals(continuous_model)
```

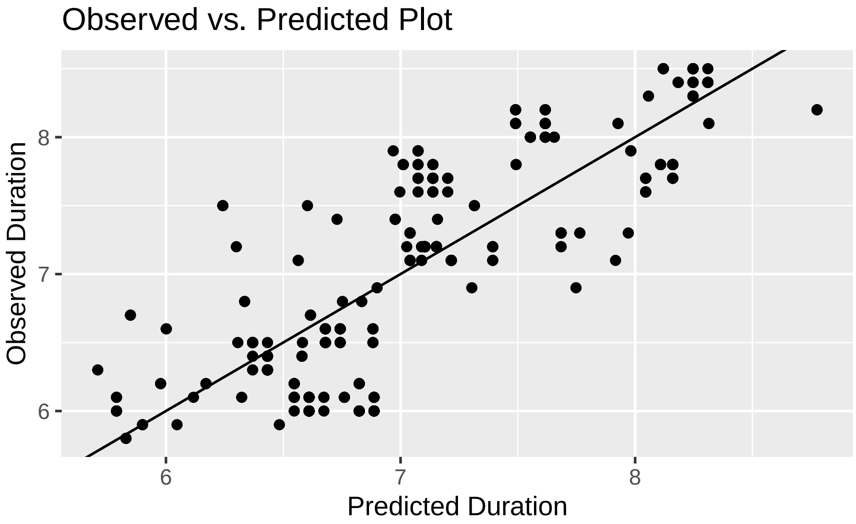
###Histogram of residual in Sleep_health_and_lifestyle_dataset_df

```r
Sleep_health_and_lifestyle_dataset_df %>%
  ggplot() +
  geom_histogram(mapping = aes(x = resid), color = "blue", fill = "pink", bins = 10) +
  labs(x = "residual", y ="Duration",
title = "Histogram of residual in Sleep_health_and_lifestyle_dataset_df")
```



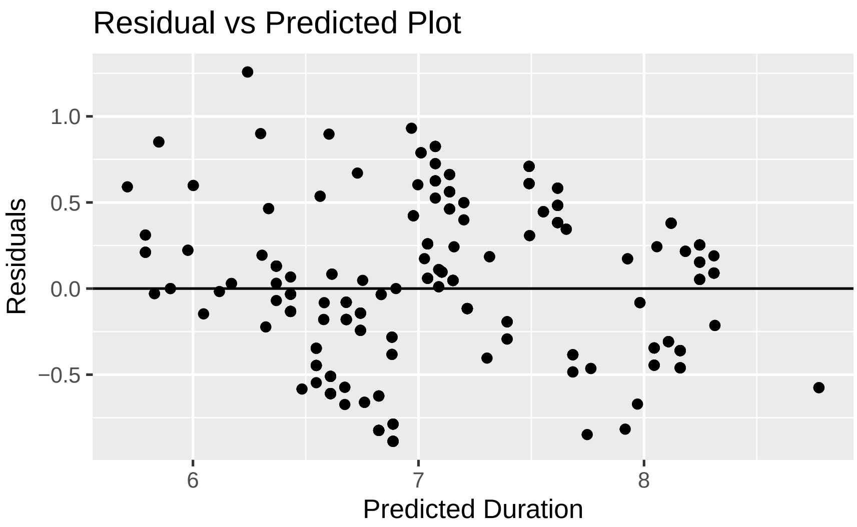
###Observed vs. Predicted Plot

```
Sleep_health_and_lifestyle_dataset_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = Duration)) +
  geom_abline(slope = 1, intercept = 0) +
  labs(title = "Observed vs. Predicted Plot", x = "Predicted Duration", y = "Observed Duration
```



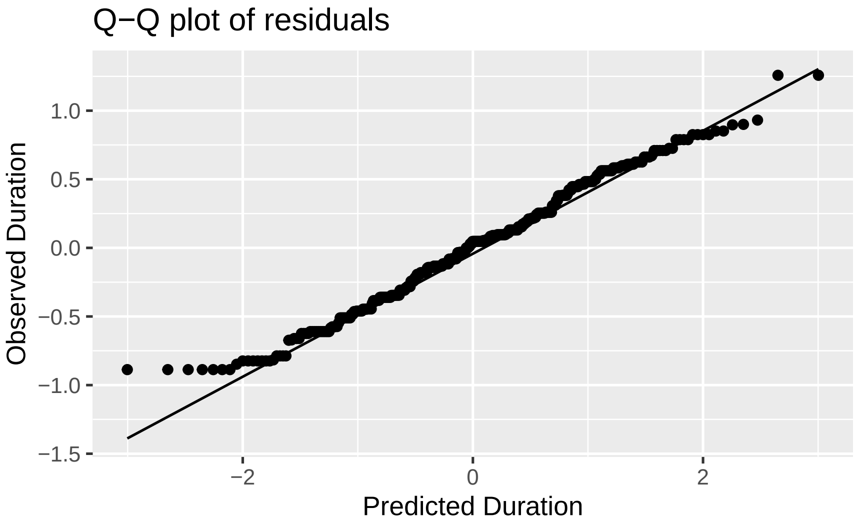
###Residual vs Predicted Plot

```
Sleep_health_and_lifestyle_dataset_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline(yintercept = 0) +
  labs( title= "Residual vs Predicted Plot",
        x = "Predicted Duration",
        y = "Residuals")
```



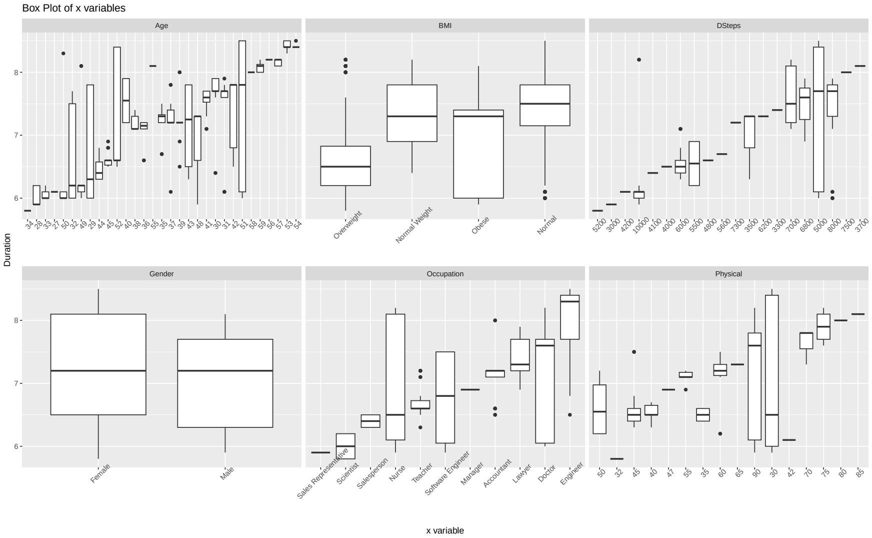
###Q-Q Plot (Obeserved vs Predicted Plot)

```
Sleep_health_and_lifestyle_dataset_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid))+
  labs(title = "Q-Q plot of residuals", x= "Predicted Duration", y= "Observed Duration")
```



###Box Plot

```
Sleep_health_and_lifestyle_dataset_df %>%
  pivot_longer(
    cols = Gender:Occupation | Physical | BMI | DSteps,
    names_to = "column",
    values_to = "value",
    values_transform = list(value = 'factor')
) %>%
ggplot() +
  geom_boxplot(aes(x = reorder(value, Duration, FUN = median), y = Duration)) +
  facet_wrap(~column, scales = "free_x") +
  labs(x = "x variable", y = "Duration", title = "Box Plot of x variables") +
  theme(axis.text.x = element_text(angle = 45))
```



#Predictive Analysis

###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

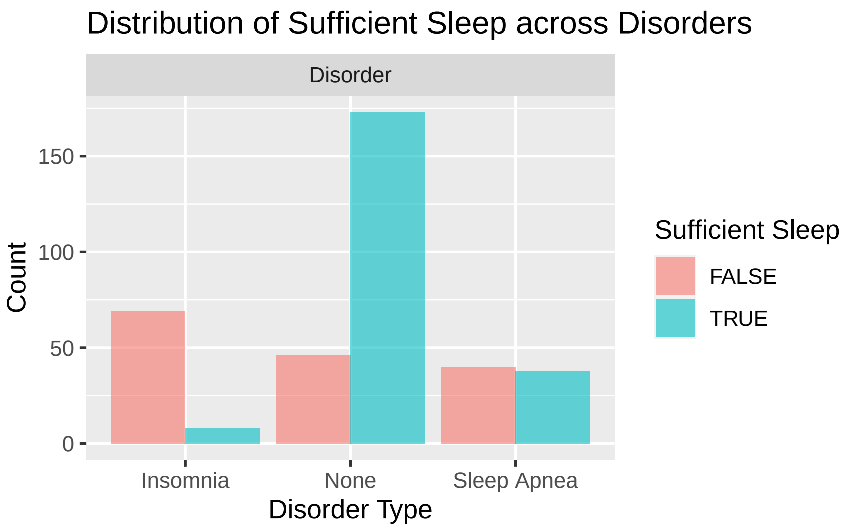
```r
Sleep_health_and_lifestyle_dataset_renamed <- Sleep_health_and_lifestyle_dataset %>%
  rename(ID = 'Person ID',
         Duration = 'Sleep Duration',
         Stress = 'Stress Level',
         Physical = 'Physical Activity Level',
         Quality = 'Quality of Sleep',
         BMI = 'BMI Category',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

###Parse Sleep Data

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

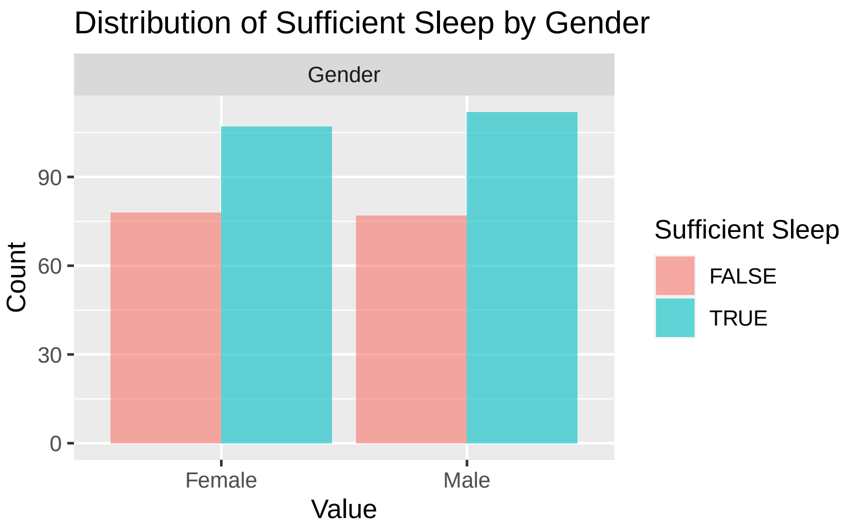
###Sleep Data Disorders

```r
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")
```



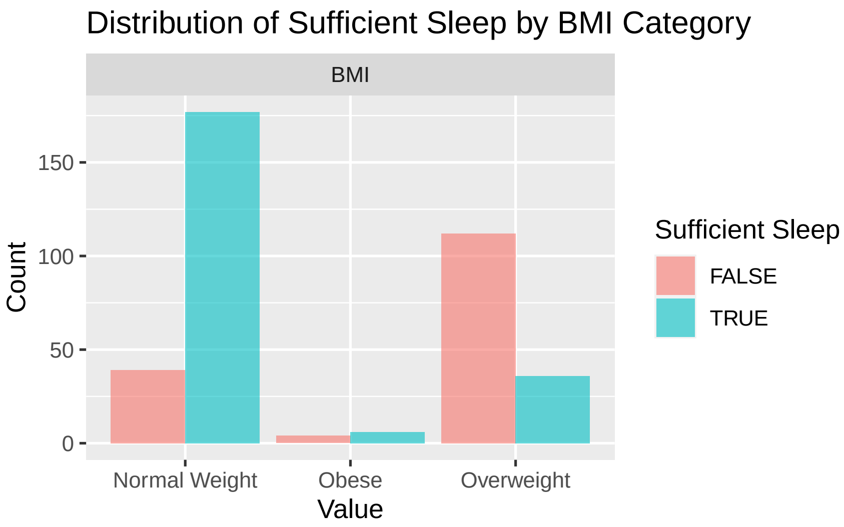
###Sleep Data Gender

```
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, ]

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

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

###Predict

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

###Test

```
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))
```

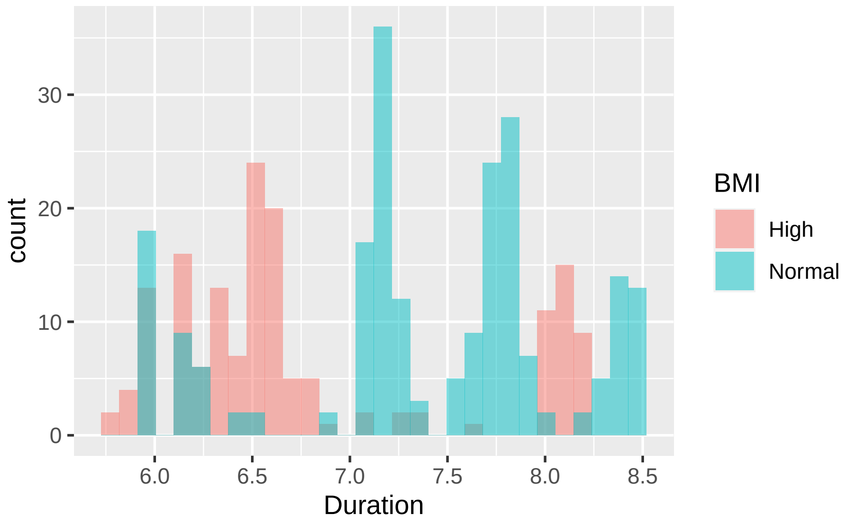
```
## [1] "Accuracy: 0.981981981981982"
```

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

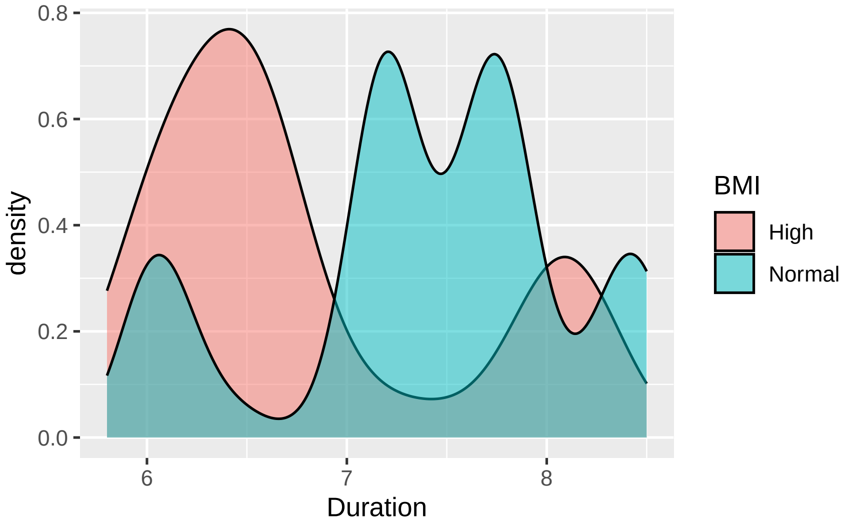
#Hypothesis Testing

```
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 =
```

```
Sleep_health_and_lifestyle_dataset_renamed %>%
  filter(BMI == "Normal" | BMI == "High") %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = Duration, fill = BMI),
    position = "identity",
alpha = 0.5
  )
```



```
Sleep_health_and_lifestyle_dataset_renamed %>%
  filter(BMI == "Normal" | BMI == "High") %>%
  ggplot() +
  geom_density(
    mapping = aes(x = Duration, fill = BMI),
    position = "identity",
alpha = 0.5
  )
```



```
Sleep_health_and_lifestyle_dataset_renamed %>%
summarize(
mean = mean(Duration),
median = median(Duration),
standard_deviation = sd(Duration),
minimum = min(Duration),
```

```r
maximum = max(Duration)
)
```

mean	median	standard_deviation	minimum	maximum
7.132086	7.2	0.7956567	5.8	8.5

```r
Model <- lm(Duration ~ BMI, data = Sleep_health_and_lifestyle_dataset_renamed)
Simulation_results <-
  Sleep_health_and_lifestyle_dataset_renamed %>%
  specify(Duration ~ BMI) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("Normal", "High"))
```

```r
Shl_obs_stat <-
  Sleep_health_and_lifestyle_dataset_renamed %>%
  specify(formula = Duration ~ BMI) %>%
  calculate(stat = "diff in means", order = c("Normal","High"))
```

```r
Shl_null <- Sleep_health_and_lifestyle_dataset_renamed %>%
  specify(Duration ~ BMI) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute")
```

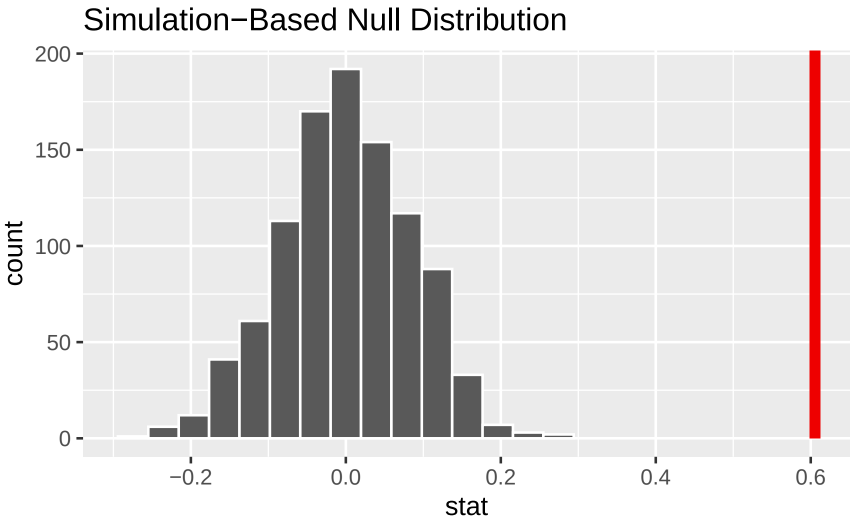
```r
Shl_null %>% get_p_value(obs_stat = Shl_obs_stat, direction = "right")
```

p_value
1

```r
p_value <- Shl_null %>% get_p_value(obs_stat = Shl_obs_stat, direction = "right")
```

```r
Simulation_results %>%
 visualize() +
 shade_p_value(obs_stat = Shl_obs_stat, direction = "right")
```

```
## Warning in min(diff(unique_loc)): min에 전달되는 인자들 중 누락이 있어 Inf를
## 반환합니다
```


Simulation−Based Null Distribution
200
150
100
50
0
count
−0.2
0.0
0.2
0.4
0.6
stat