Project 1

Author: Jeet Patel

Task 1:

Question 1

plot(x, y) abline(v = 2) -> The correct formatting of abline is (v=2) and not (x=2)

Question 2

I am going to knit to an .html file. In order to convert it to a pdf, I will open it in the browser, print it, and then save it as a pdf.

Question 3

The difference between a .pdf file and .rmd file is that .rmd is a script-like document with rcode while a pdf is the final output that is not editable.

```
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(readr)
library(tidyr)
library(moments)
```

Task 2: Converting Workout_Type, Experience_Level and Gender to Factors

```
gym_data <- read.csv("/Users/jeetpatel/Desktop/gym.csv")

gym_data$Workout_Type <- as.factor(gym_data$Workout_Type)
gym_data$Experience_Level <- factor(gym_data$Experience_Level, levels = c(1, 2, 3), o
rdered = TRUE)
gym_data$Gender <- as.factor(gym_data$Gender)

summary(gym_data)
```

```
##       Age            Weight           Height         Max_BPM
##  Min.   :18.00   Min.   : 40.00   Min.   :1.500   Min.   :160.0
##  1st Qu.:28.00   1st Qu.: 58.10   1st Qu.:1.620   1st Qu.:170.0
##  Median :40.00   Median : 70.00   Median :1.710   Median :180.0
##  Mean   :38.68   Mean   : 73.85   Mean   :1.723   Mean   :179.9
##  3rd Qu.:49.00   3rd Qu.: 86.00   3rd Qu.:1.800   3rd Qu.:190.0
##  Max.   :59.00   Max.   :129.90   Max.   :2.000   Max.   :199.0
##     Avg_BPM       Resting_BPM    Session_Duration Calories_Burned
##  Min.   :120.0   Min.   :50.00   Min.   :0.500    Min.   : 303.0
##  1st Qu.:131.0   1st Qu.:56.00   1st Qu.:1.040    1st Qu.: 720.0
##  Median :143.0   Median :62.00   Median :1.260    Median : 893.0
##  Mean   :143.8   Mean   :62.22   Mean   :1.256    Mean   : 905.4
##  3rd Qu.:156.0   3rd Qu.:68.00   3rd Qu.:1.460    3rd Qu.:1076.0
##  Max.   :169.0   Max.   :74.00   Max.   :2.000    Max.   :1783.0
##    Workout_Type Fat_Percentage   Water_Intake   Workout_Frequency
##  Cardio  :255   Min.   :10.00   Min.   :1.500   Min.   :2.000
##  HIIT    :221   1st Qu.:21.30   1st Qu.:2.200   1st Qu.:3.000
##  Strength:258   Median :26.20   Median :2.600   Median :3.000
##  Yoga    :239   Mean   :24.98   Mean   :2.627   Mean   :3.322
##                 3rd Qu.:29.30   3rd Qu.:3.100   3rd Qu.:4.000
##                 Max.   :35.00   Max.   :3.700   Max.   :5.000
##  Experience_Level      BMI        Gender
##  1:376            Min.   :12.32   0:462
##  2:406            1st Qu.:20.11   1:511
##  3:191            Median :24.16
##                   Mean   :24.91
##                   3rd Qu.:28.56
##                   Max.   :49.84
```

Task 3: Summarize and Describe Avg_BPM

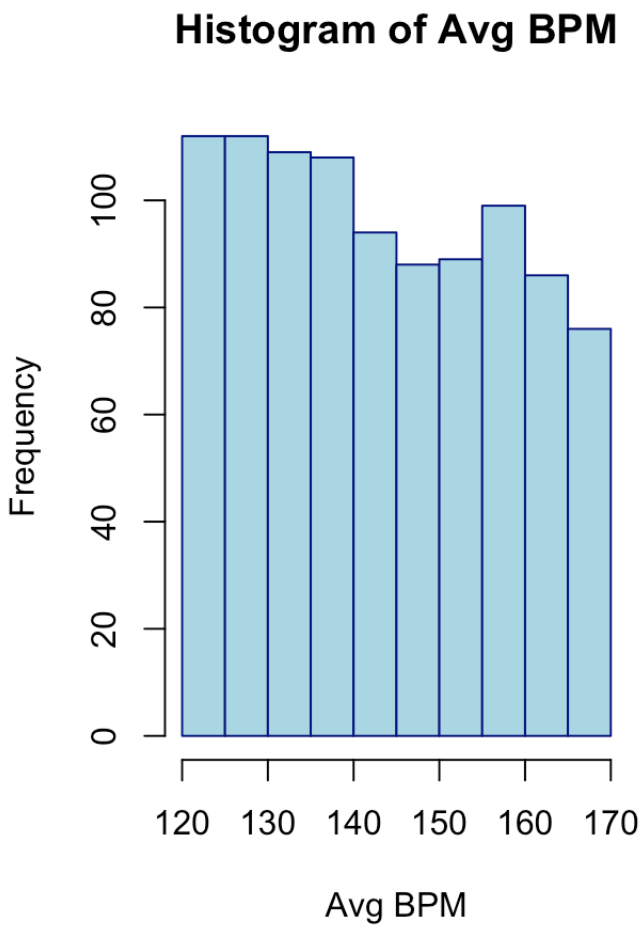
```
summary(gym_data$Avg_BPM)
```

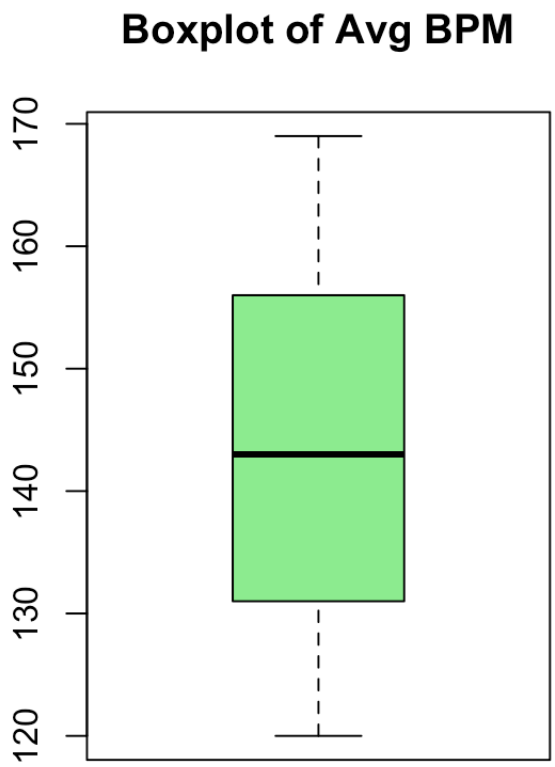
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   120.0   131.0   143.0   143.8   156.0   169.0
```

```
#NA
sum(is.na(gym_data$Avg_BPM))
```

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

```
#Histogram and Boxplot
par(mfrow = c(1,2))
hist(gym_data$Avg_BPM, main = "Histogram of Avg BPM", xlab = "Avg BPM", col = "lightb
lue", border = "navy")
boxplot(gym_data$Avg_BPM, main ="Boxplot of Avg BPM", col = "lightgreen")
```






```
#Descriptive Statistcs
mean(gym_data$Avg_BPM, na.rm = TRUE)
```

```
## [1] 143.7667
```

```
median(gym_data$Avg_BPM, na.rm = TRUE)
```

```
## [1] 143
```

```
sd(gym_data$Avg_BPM, na.rm = TRUE)
```

```
## [1] 14.3451
```

Interpretation:

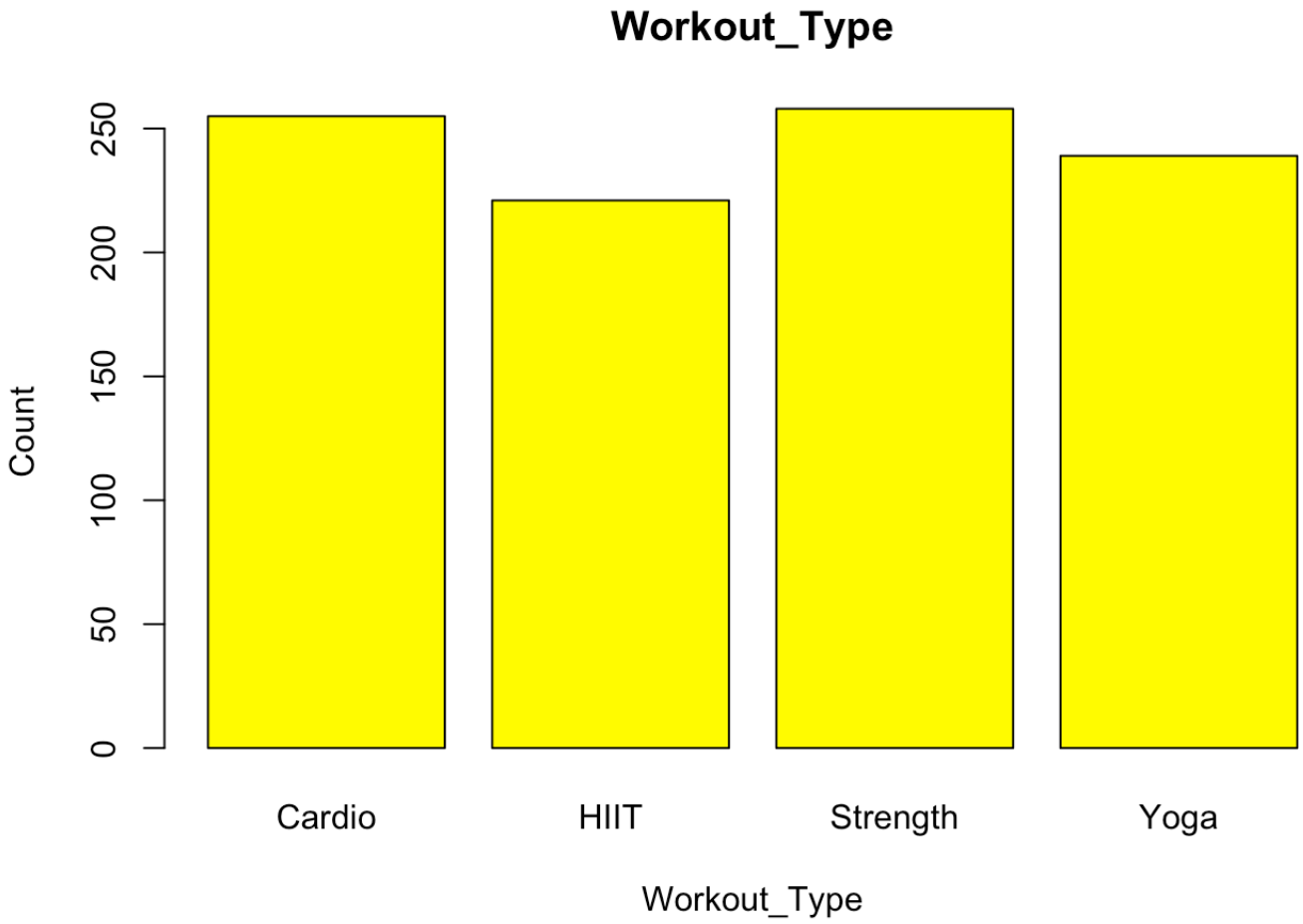
We will look at the data on average bpm and analysis it. The average bpm ranges from 120 to 169 looking at the min and max. The histogram is fairly evenly distributed and the box plot includes no major outliers. The median is 143 and the mean is 143.8 indicating a balanced distribution/symmetry since there are alike. The standard deviation is 14.35 implying variability.

Task 4: Summarize and Describe Workout_Type

```
#Frequency Table
table(gym_data$Workout_Type)
```

```
##
##   Cardio     HIIT Strength     Yoga
##      255      221      258      239
```

```
#Barplot
barplot(table(gym_data$Workout_Type), main = "Workout_Type", col ="yellow", xlab = "W
orkout_Type", ylab = "Count")
```

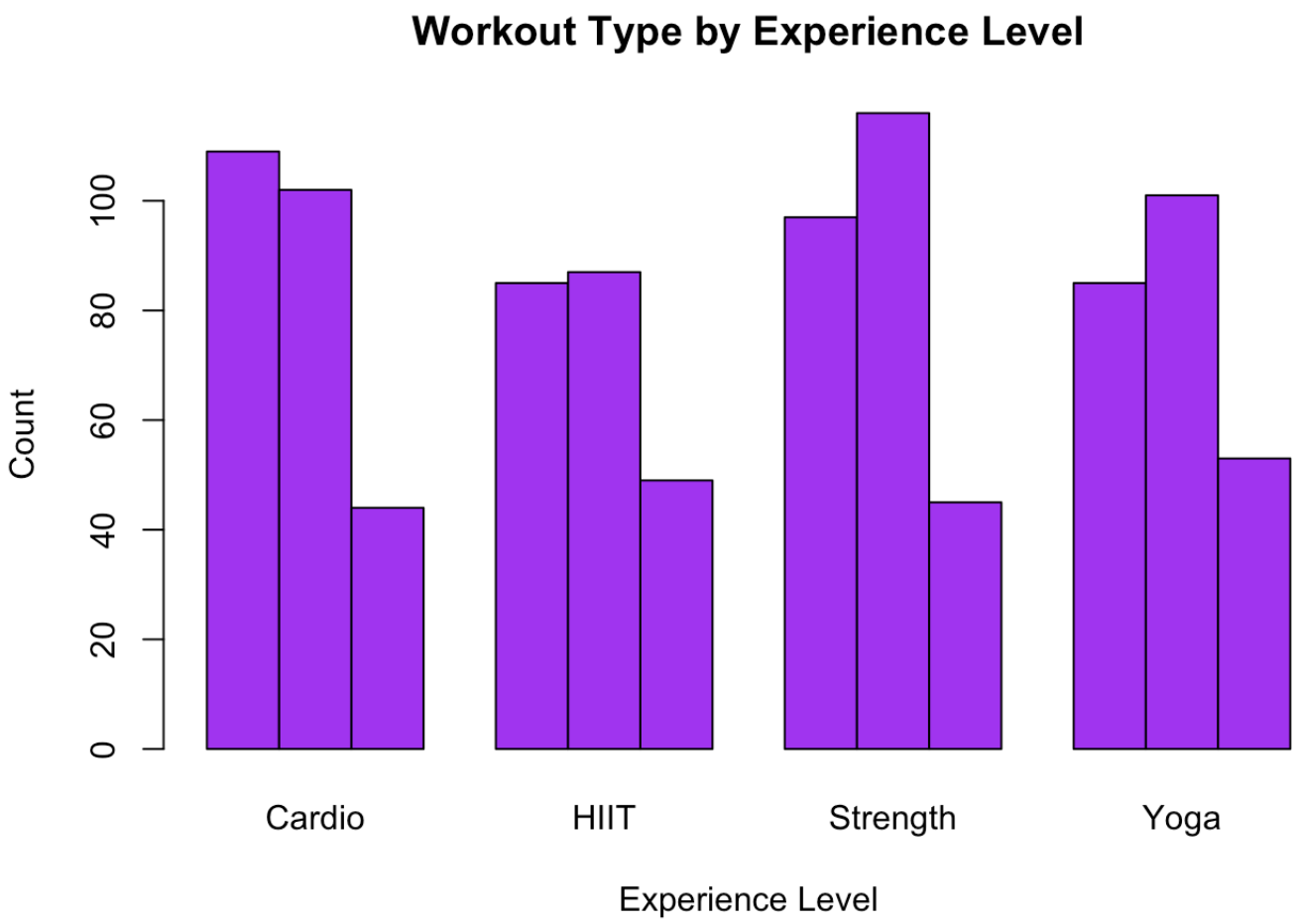


Interpretation:

We will look at the data on workout type and analysis it. Looking at the frequency table, we see that Strength and Cardio (255 and 258) are more common and Yoga and HIIT (221 and 239) are only slightly less common since there frequencies are smaller. In addition, the histogram is uniformly distributed but the bars for Cardio and Strength are higher than HIIT and Yoga. Therefore we conclude the workout types are all equally common with no major difference.

Task 5: Workout_Type Varying by Experience_Level

```
#Barplot
barplot(table(gym_data$Experience_Level, gym_data$Workout_Type), beside = TRUE, col =
"purple", main="Workout Type by Experience Level", xlab="Experience Level", ylab="Cou
nt")
```




```
#Table
# Compute the conditional distributions (proportions within each experience level)
workout_table <- prop.table(table(gym_data$Experience_Level, gym_data$Workout_Type))

# Display the table
workout_table
```

```
##
##         Cardio       HIIT   Strength       Yoga
##   1 0.11202467 0.08735868 0.09969168 0.08735868
##   2 0.10483042 0.08941418 0.11921891 0.10380267
##   3 0.04522097 0.05035971 0.04624872 0.05447071
```

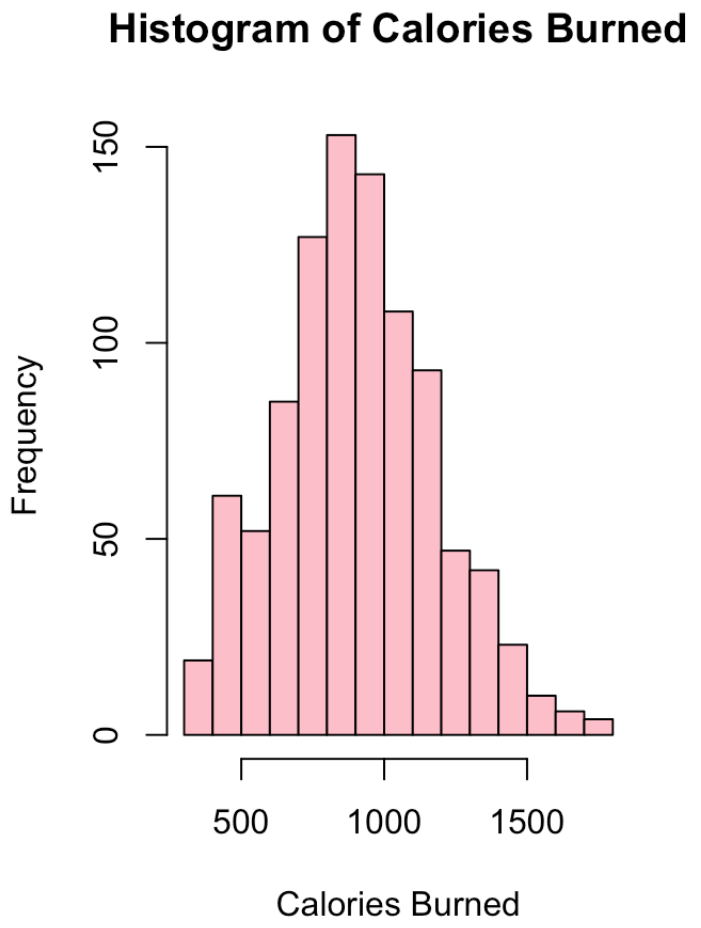
Interpretation:

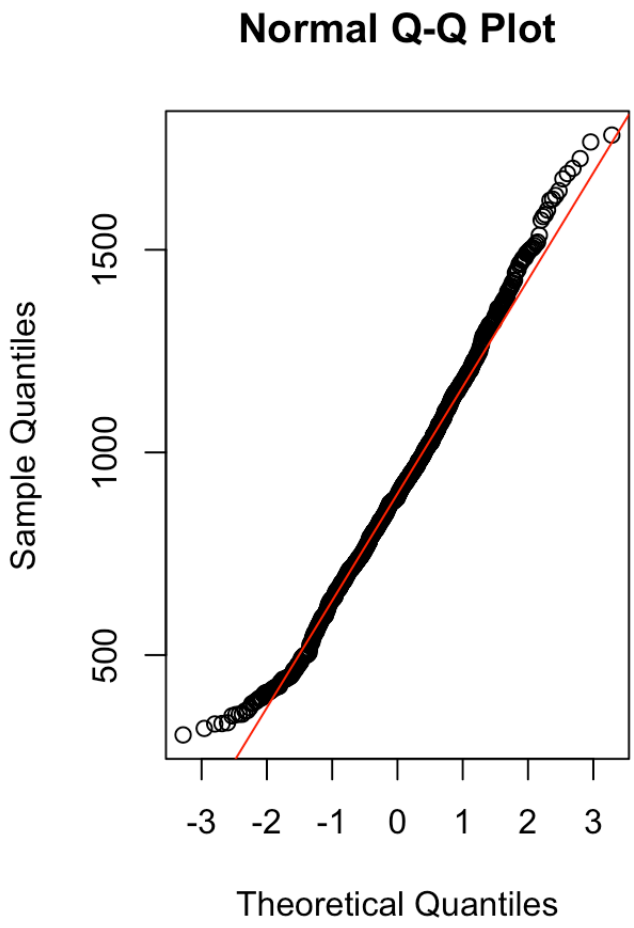
The data does not imply the workout type varies by experience level. When we look at the bar plot, the heights of the bars are nearly identical and differ only slightly between the different types of workouts. The table support my statements in that the conditional probabilities are very similar. A key example is the

beginner statistics including Cardio with the conditional probabilities of 11.2% for Cardio, 8.7% for HIIT, 9.9% for Strength, and 8.7% for Yoga. Therefore, we conclude workout type does not significantly vary by experience level.

Task 6: Normality Test for Calories_Burned

```
#Histogram and Q-Q Plot
par(mfrow = c(1,2))
hist(gym_data$Calories_Burned, main="Histogram of Calories Burned", xlab ="Calories B
urned", col = "pink")
qqnorm(gym_data$Calories_Burned)
qqline(gym_data$Calories_Burned, col = "red")
```






```
#Skew and Kurtosis
skewness(gym_data$Calories_Burned)
```

```
## [1] 0.2778918
```

```
kurtosis(gym_data$Calories_Burned)
```

```
## [1] 2.938077
```

```
#Shaprio-Wilk Test
shapiro.test(gym_data$Calories_Burned)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gym_data$Calories_Burned
## W = 0.99176, p-value = 2.982e-05
```

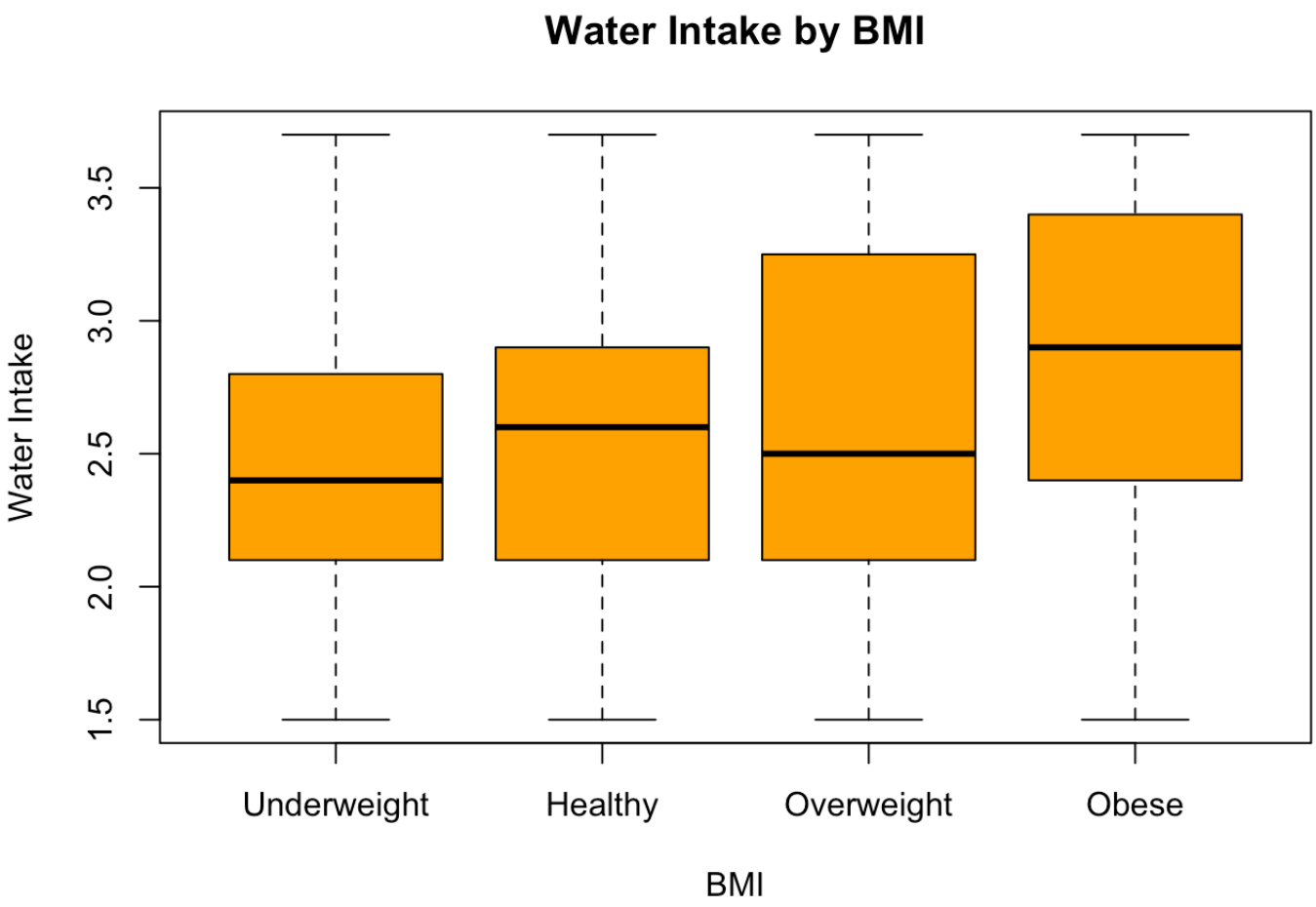
Analysis:

There are multiple reasons to assume Calories_Burned is normally distributed. First we look at the graphs. The histogram is symmetric and bell-shaped and the points on the normal q-q plot align up on the reference line implying normality. Next, we look at the skewness and kurtosis. The skewness is .278 which is small close to 0 meaning there are very few outliers. In addition our kurtosis is 2.94 which is close to 3 which is the kurtosis of a normal distribution. Therefore we imply normality once again. Lastly, we look at the Shapiro-Wilk test. Our null hypothesis is the calories burned follows a normal distribution. Our alternative hypothesis is the calories burned does not follow a normal distribution. Our pvalue is less the significance level (.00002982 < .01) and we reject the null hypothesis. Therefore the Shapiro-Wilk test concludes the calories burned is not normally distribution. To summarize, the histogram, normal q-q plot, skewness, and kurtosis imply a normal distribution but the Shapiro-Wilk test does not imply normality.

Task 7: BMI

```
gym_data$bmi_class <- cut(gym_data$BMI,
                          breaks = c(-Inf, 18.5, 25, 30, Inf),
                          labels = c("Underweight", "Healthy", "Overweight", "Obes
e"),
                          ordered_result = TRUE)

# Boxplot: Water Intake by BMI Class
boxplot(Water_Intake ~ bmi_class, data=gym_data, main="Water Intake by BMI", xlab="BM
I", ylab="Water Intake", col="orange")
```




```
#Summary Statistics
aggregate(Water_Intake ~ bmi_class, data=gym_data, FUN=function(x) c(mean=mean(x, na.
rm=TRUE), sd=sd(x, na.rm=TRUE)))
```

```
##     bmi_class Water_Intake.mean Water_Intake.sd
## 1 Underweight         2.4732143       0.5418391
## 2     Healthy         2.5754054       0.5869082
## 3  Overweight         2.6000000       0.6393437
## 4       Obese         2.8932292       0.5430243
```

Analysis:

The mean water intake increases as BMI increaes with Obese people with the most water consumption. First, we analyze the graph, where obese people have a higher water intake on average compared to the others. In addition, its median intake is the highest along with the higher Q1 and Q3. Second, we analyze the summary statistics where Obese people have the highest mean of 2.89. On top of that, the mean is linearly increasing and all different from each other. Therefore, we conclude water intake varies by bmi class.

Task 8: Calories_Burned Hypothesis Test

```
#Mean Calories Burned > 890

test_Calorie <- t.test(gym_data$Calories_Burned, mu=890, alternative="greater", conf.
level=0.98)
test_Calorie
```

```
##
##  One Sample t-test
##
## data:  gym_data$Calories_Burned
## t = 1.7645, df = 972, p-value = 0.03898
## alternative hypothesis: true mean is greater than 890
## 98 percent confidence interval:
##  887.4475      Inf
## sample estimates:
## mean of x
##  905.4224
```

Conclusion:

The test we are conducting is whether the population mean Calories_Burned is greater than 890. Our null hypothesis is the mean calories burned is less than of equal to 890. Our alternative hypothesis is the mean calories burned is more than 890. From the test, our pvalue is .03898. Since our pvalue is greater than our significance level(.03898 > .02), we ail to reject the null hypothesis. Therefore we conclude there is not enough statistical evidence that the mean calories burned is more than 890. In addition, our confidence interval is (887.45, Inf) and there are integers below 890 meaning there is a probability the mean is not greater than 890. Thus, we fail to reject the hypothesis.

Task 9: Sesssion_Duration Hypothesis Test

```
# Normality
shapiro.test(gym_data$Session_Duration[gym_data$Gender == 1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gym_data$Session_Duration[gym_data$Gender == 1]
## W = 0.98419, p-value = 2.36e-05
```

```
shapiro.test(gym_data$Session_Duration[gym_data$Gender == 0])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gym_data$Session_Duration[gym_data$Gender == 0]
## W = 0.98552, p-value = 0.0001461
```

```
# Independent t-test
test_Session <- t.test(Session_Duration ~ Gender, data=gym_data, var.equal=TRUE, con
f.level=.97)
test_Session
```

```
##
##  Two Sample t-test
##
## data:  Session_Duration by Gender
## t = 0.38019, df = 971, p-value = 0.7039
## alternative hypothesis: true difference in means between group 0 and group 1 is no
t equal to 0
## 97 percent confidence interval:
##  -0.03950535  0.05625800
## sample estimates:
## mean in group 0 mean in group 1
##        1.260823        1.252446
```

Conclusion:

The test we are conducting is whether there is enough evidence the mean of Session_Duration is different for those who identify as gender 1 and those who do not. Our null hypothesis is there is no difference in session duration between genders. The alternative hypothesis is there is a difference in session duration between genders. From the test, our pvalue is .7039. Since our pvalue is greater than our significance level(.7039 > .03), we fail to reject the null hypothesis. Therefore we conclude there is no difference in session duration between genders. In addition, our confidence interval is (-.0395, .0563) and since 0 is in between the bounds we fail to reject the null hypothesis.