5301-Final Report

Soham Neeraj Agarkar (1002157894) & Utkarsh Pant (1002170893

2023-12-12

Contents

Effect of Gender and Diet on Weight Loss

General Declaration

```
library(WRS2)
```

```
## Warning: package 'WRS2' was built under R version 4.3.2
```

```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.3.2
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
library(outliers)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

Introduction

In today's world there is a very heavy focus on weight loss. Scores of corporations are constantly locked in a struggle to make products that target factors such as weight gain, weight loss and general beauty products and market them to a wide spectrum of audiences. In such a market, data analysis could play a crucial role and allow corporations to make an informed decision on what their product needs versus what their customers want.

Today we set out to understand which attributes are the most likely to affect a person's weight loss. We will work with the diet data set that is already a present in the [WRS2] package.

Problem Statement

Perform Two-Way ANOVA on the features *gender* and *diet.type* to determine which is more influential to *weight.loss* feature.

Descriptive Analysis

Below is the distribution of our data's descriptive analysis :

```r
str(diet)
```

```
## 'data.frame':    76 obs. of  7 variables:
##  $ gender        : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age           : int  22 46 55 33 50 50 37 28 28 45 ...
##  $ height        : int  159 192 170 171 170 201 174 176 165 165 ...
##  $ diet.type     : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ initial.weight: int  58 60 64 64 65 66 67 69 70 70 ...
##  $ final.weight  : num  54.2 54 63.3 61.1 62.2 64 65 60.5 68.1 66.9 ...
##  $ weight.loss   : num  3.8 6 0.7 2.9 2.8 ...
```

We can observe from the above the analysis the structure of our data set. The data possesses two categorical variables and the rest are numerical.

```r
summary(diet$gender)
```

```
## Female   Male
##     43     33
```

```r
summary(diet$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   32.50   39.00   39.22   47.25   60.00
```

```r
summary(diet$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   141.0   163.8   169.0   170.8   175.2   201.0
```

```r
summary(diet$diet.type)
```

```
##  A  B  C
## 24 25 27
```

```r
summary(diet$initial.weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   58.00   66.00   72.00   72.29   78.00   88.00
```

```r
summary(diet$final.weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.00   61.95   68.95   68.34   73.67   84.50
```

```r
summary(diet$weight.loss)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.100   2.300   3.700   3.946   5.650   9.200
```

Here we see the feature wise summary statistics.

Below we see the features statistics while being grouped by ***gender :***

```r
diet %>%
  group_by(diet$gender) %>%
  summarise(
    Mean = mean(diet$age),
    Median = median(diet$age),
    SD = sd(diet$age),
    Min = min(diet$age),
    Max = max(diet$age)
  )
```

```
## # A tibble: 2 x 6
##   `diet$gender`  Mean Median    SD   Min   Max
##   <fct>         <dbl>  <dbl> <dbl> <int> <int>
## 1 Female         39.2     39  9.91    16    60
## 2 Male           39.2     39  9.91    16    60
```

```r
diet %>%
  group_by(diet$gender) %>%
  summarise(
    Mean = mean(diet$height),
    Median = median(diet$height),
    SD = sd(diet$height),
    Min = min(diet$height),
    Max = max(diet$height)
  )
```

```
## # A tibble: 2 x 6
##   `diet$gender`  Mean Median    SD   Min   Max
##   <fct>         <dbl>  <dbl> <dbl> <int> <int>
## 1 Female         171.    169  11.4   141   201
## 2 Male           171.    169  11.4   141   201
```

```r
diet %>%
  group_by(diet$gender) %>%
  summarise(
    Mean = mean(diet$initial.weight),
    Median = median(diet$initial.weight),
    SD = sd(diet$initial.weight),
    Min = min(diet$initial.weight),
    Max = max(diet$initial.weight)
  )
```

```
## # A tibble: 2 x 6
##   `diet$gender` Mean Median    SD   Min   Max
##   <fct>         <dbl>  <dbl> <dbl> <int> <int>
## 1 Female         72.3     72  7.97    58    88
## 2 Male           72.3     72  7.97    58    88
```

```
diet %>%
  group_by(diet$gender) %>%
  summarise(
    Mean = mean(diet$final.weight),
    Median = median(diet$final.weight),
    SD = sd(diet$final.weight),
    Min = min(diet$final.weight),
    Max = max(diet$final.weight)
  )
```

```
## # A tibble: 2 x 6
##   `diet$gender` Mean Median    SD   Min   Max
##   <fct>         <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Female         68.3   69.0  8.06    53  84.5
## 2 Male           68.3   69.0  8.06    53  84.5
```

```
diet %>%
  group_by(diet$gender) %>%
  summarise(
    Mean = mean(diet$weight.loss),
    Median = median(diet$weight.loss),
    SD = sd(diet$weight.loss),
    Min = min(diet$weight.loss),
    Max = max(diet$weight.loss)
  )
```

```
## # A tibble: 2 x 6
##   `diet$gender` Mean Median    SD   Min   Max
##   <fct>         <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Female         3.95    3.7  2.51 -2.10   9.2
## 2 Male           3.95    3.7  2.51 -2.10   9.2
```

Below we see the feature statistics while grouped by ***diet.type*** :

```
diet %>%
  group_by(diet$diet.type) %>%
  summarise(
    Mean = mean(diet$age),
    Median = median(diet$age),
    SD = sd(diet$age),
    Min = min(diet$age),
    Max = max(diet$age)
  )
```

```
## # A tibble: 3 x 6
##   `diet$diet.type`  Mean Median    SD   Min   Max
```

```
##   <fct>             <dbl>  <dbl> <dbl> <int> <int>
## 1 A                  39.2     39  9.91    16    60
## 2 B                  39.2     39  9.91    16    60
## 3 C                  39.2     39  9.91    16    60
```

```
diet %>%
  group_by(diet$diet.type) %>%
  summarise(
    Mean = mean(diet$height),
    Median = median(diet$height),
    SD = sd(diet$height),
    Min = min(diet$height),
    Max = max(diet$height)
  )
```

```
## # A tibble: 3 x 6
##   `diet$diet.type`  Mean Median    SD   Min   Max
##   <fct>             <dbl>  <dbl> <dbl> <int> <int>
## 1 A                  171.    169  11.4   141   201
## 2 B                  171.    169  11.4   141   201
## 3 C                  171.    169  11.4   141   201
```

```
diet %>%
  group_by(diet$diet.type) %>%
  summarise(
    Mean = mean(diet$initial.weight),
    Median = median(diet$initial.weight),
    SD = sd(diet$initial.weight),
    Min = min(diet$initial.weight),
    Max = max(diet$initial.weight)
  )
```

```
## # A tibble: 3 x 6
##   `diet$diet.type`  Mean Median    SD   Min   Max
##   <fct>             <dbl>  <dbl> <dbl> <int> <int>
## 1 A                  72.3     72  7.97    58    88
## 2 B                  72.3     72  7.97    58    88
## 3 C                  72.3     72  7.97    58    88
```

```
diet %>%
  group_by(diet$diet.type) %>%
  summarise(
    Mean = mean(diet$final.weight),
    Median = median(diet$final.weight),
    SD = sd(diet$final.weight),
    Min = min(diet$final.weight),
    Max = max(diet$final.weight)
  )
```

```
## # A tibble: 3 x 6
##   `diet$diet.type`  Mean Median    SD   Min   Max
##   <fct>             <dbl>  <dbl> <dbl> <dbl> <dbl>
```

```
## 1 A                     68.3   69.0  8.06    53  84.5
## 2 B                     68.3   69.0  8.06    53  84.5
## 3 C                     68.3   69.0  8.06    53  84.5
```

```
diet %>%
  group_by(diet$diet.type) %>%
  summarise(
    Mean = mean(diet$weight.loss),
    Median = median(diet$weight.loss),
    SD = sd(diet$weight.loss),
    Min = min(diet$weight.loss),
    Max = max(diet$weight.loss)
  )
```

```
## # A tibble: 3 x 6
##   `diet$diet.type`  Mean Median    SD   Min   Max
##   <fct>            <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 A                 3.95    3.7  2.51 -2.10   9.2
## 2 B                 3.95    3.7  2.51 -2.10   9.2
## 3 C                 3.95    3.7  2.51 -2.10   9.2
```

Frequency of Categorical variables :

```
# Frequency table for a gender
table(diet$gender)
```

```
##
## Female   Male
##     43     33
```

```
# Percentage table for a gender
prop.table(table(diet$gender)) * 100
```

```
##
##   Female     Male
## 56.57895 43.42105
```

```
# Frequency table for a diet.type
table(diet$diet.type)
```

```
##
##  A  B  C
## 24 25 27
```

```
# Percentage table for a categorical variable
prop.table(table(diet$diet.type)) * 100
```

```
##
##        A        B        C
## 31.57895 32.89474 35.52632
```

Checking for missing values :

```r
colSums(is.na(diet))
```

```
##         gender            age         height      diet.type initial.weight 
##              0              0              0              0              0 
##   final.weight    weight.loss 
##              0              0
```

As evidenced by the result of the code above, there are **no missing values** in our data set.

Plots and Visualizations

Plots for Gender

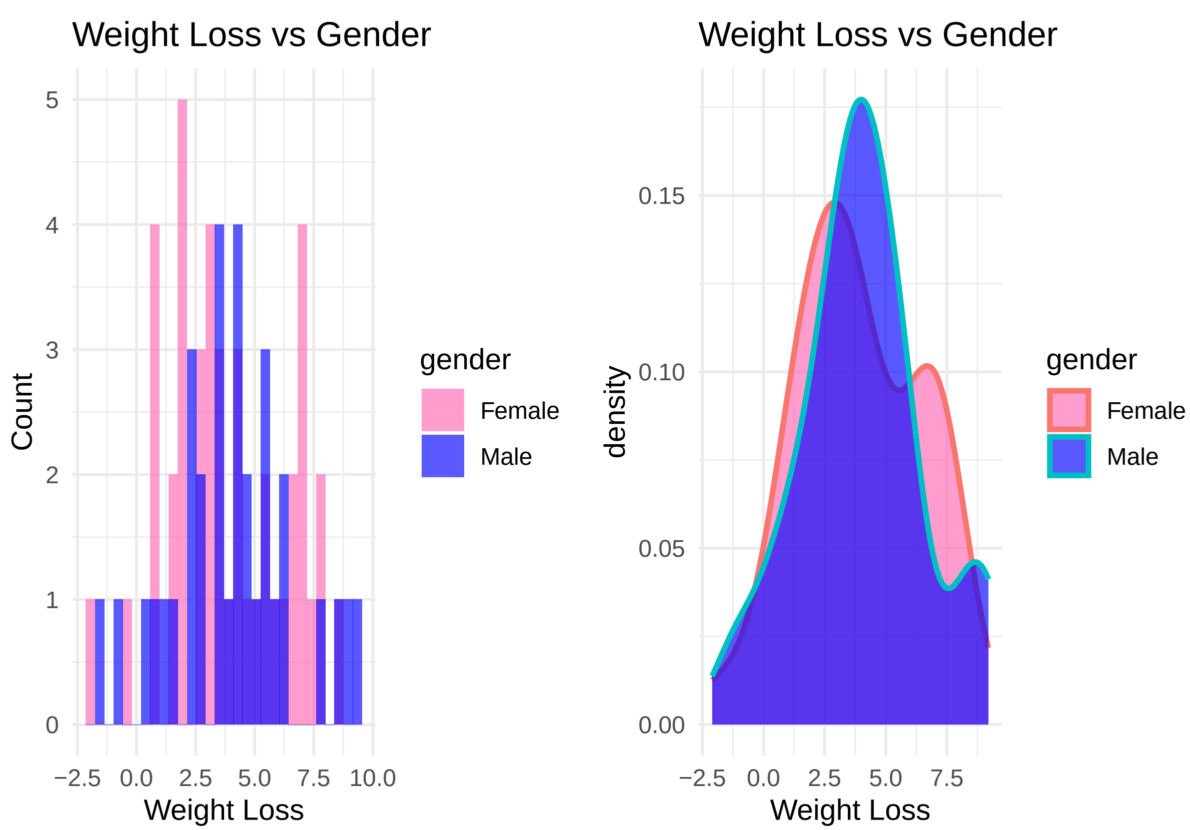
```
#specifying colors for the plot
gender_colors <- c("hotpink", "blue")

#histogram weightloss by gender
p1 <- ggplot(data = diet, aes(x = weight.loss, fill = gender)) +
  geom_histogram(position = "identity", alpha = 0.65) +
  labs(title = "Weight Loss vs Gender", x = "Weight Loss", y = "Count") +
  scale_fill_manual(values = gender_colors) +
  theme_minimal()

#density plot for weightloss by diet.type
p2 <- ggplot(data = diet, aes(x=weight.loss, fill = gender)) +
  geom_density(aes(x=weight.loss, color = gender), position  = "identity", alpha = 0.65, linewidth = 1)
  labs(title = "Weight Loss vs Gender", x= "Weight Loss") +
  scale_fill_manual(values = gender_colors) +
  theme_minimal()

#Gender plots
p1 + p2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Box plot for Gender:

```
#boxplot for weightloss by gender
ggplot(data =diet, aes(x=weight.loss, fill = gender)) +
  geom_boxplot(aes(x = gender, y=weight.loss, group = gender)) +
  labs(title = "Weight Loss grouped by Gender", x="Gender", y = "Weight Loss") +
  scale_fill_manual(values = gender_colors) +
  theme_minimal()
```

Plots for Diet Type:

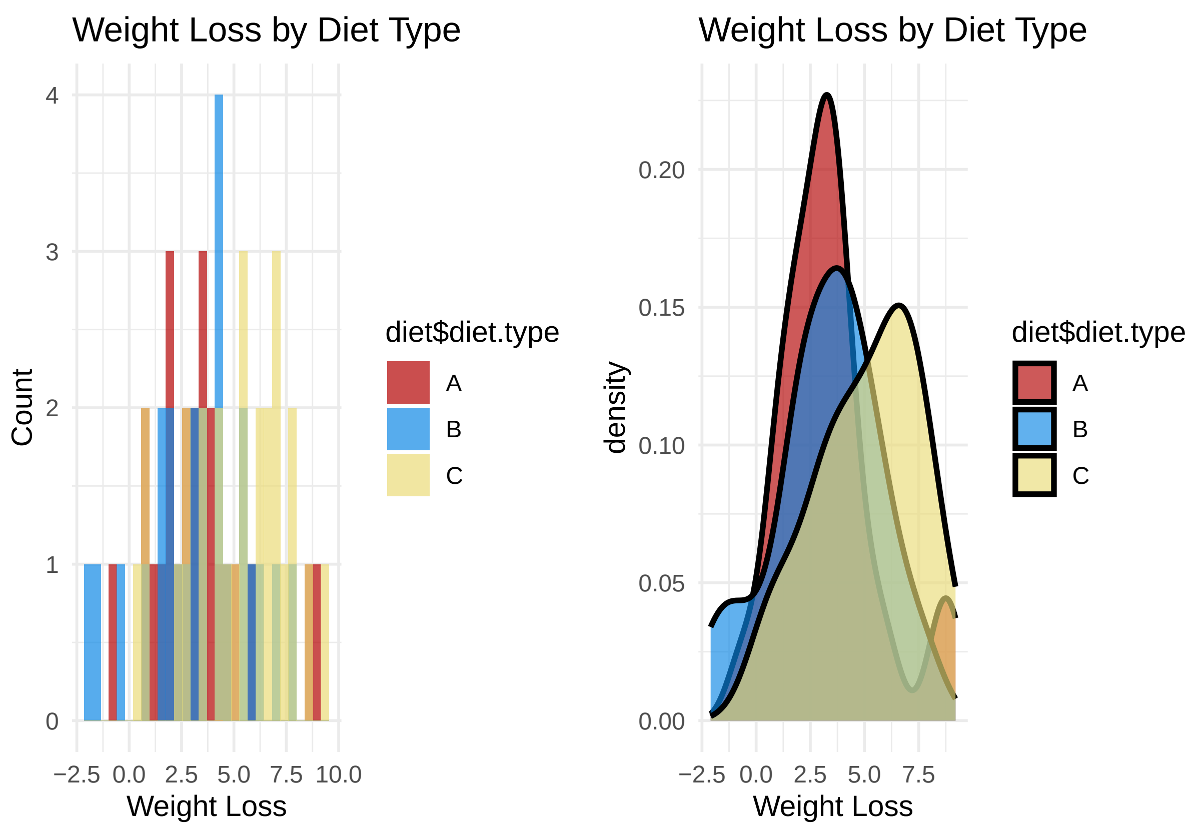
```
#diet colors
diet_colors <- c("#b30000", "#0d88e6", "#ebdc78")

#Histogram plots for weight loss by diet type
p3 <- ggplot(data= diet, aes(x = diet$weight.loss, fill = diet$diet.type)) +
  geom_histogram(position = "identity", alpha = 0.7) +
  labs(title = "Weight Loss by Diet Type", x="Weight Loss", y="Count") +
  scale_fill_manual(values = diet_colors) +
  theme_minimal()

p4 <- ggplot(data = diet, aes(x = diet$weight.loss, fill = diet$diet.type)) +
  geom_density(aes(x=weight.loss), position = "identity", alpha =0.65, linewidth =1) +
  labs(title = "Weight Loss by Diet Type", x= "Weight Loss") +
  scale_fill_manual(values = diet_colors) +
  theme_minimal()
```

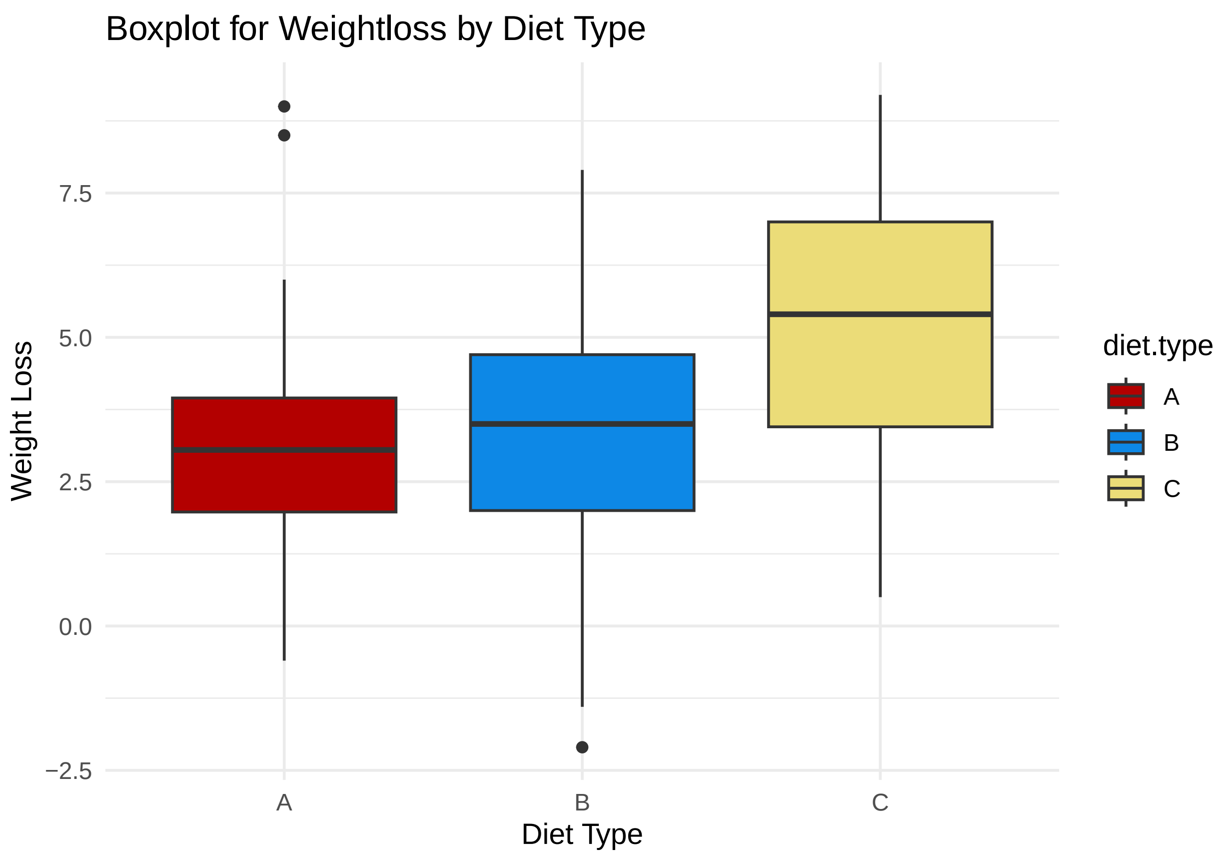
```r
#Plots for weight loss by diet type
p3 + p4
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Box plot for Diet Type:

```r
#boxplot for weight loss by diet type
ggplot(data = diet, aes(x=weight.los, fill = diet.type)) +
  geom_boxplot(aes(x=diet.type, y=weight.loss, group = diet.type)) +
  labs(title = "Boxplot for Weightloss by Diet Type", x = "Diet Type", y = "Weight Loss") +
  scale_fill_manual(values = diet_colors) +
  theme_minimal()
```



The above box plot for the *diet.type* feature shows some outliers in diet type 'A' and 'B'. Simply eyeballing it we can assume that while 'A' appears to have outliers, it is important to keep in mind that those values are only outliers among instances that follow diet type 'A'. If we consider the whole feature, those values aren't considered outliers as the values for diet type 'C' stretches all the way up to include those values.

Diet type 'B', on the other hand, seems to have one outlier that falls far below the lowest value in the entire feature among all categories ('A', 'B', 'C').

Outlier Detection:

Now that we know that the feature has one outlier, we can try to determine what outlier it is. Since we only have to detect one outlier, we can use the grubbs single outlier detection test.

```
#outlier detection
grubbs.test(diet$weight.loss)
```

```
##
##  Grubbs test for one outlier
##
## data:  diet$weight.loss
## G = 2.41282, U = 0.92134, p-value = 0.5369
## alternative hypothesis: lowest value -2.1 is an outlier
```

The above code displays the outlier in the data. Since there is only outlier in the data, we chose to leave it in as it doesn't affect the overall outcome of the analysis.

ANOVA

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences among group means in a sample. It is an extension of the t-test, allowing for the comparison of means across more than two groups. ANOVA assesses whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

Null and Alternate Hypotheses

Null Hypothesis (H_O) : Means of groups by *gender* are similar, means of groups by *diet.type* are similar and there is no significant interaction between *gender* and *diet.type* features.

Alternate Hypothesis (H_a) : There is a significant difference betweeen the means of atleast one factor or there is significant interaction between *gender* and *diet.type.*

Assumptions for ANOVA

ANOVA makes the following assumptions :

Independance of of observations Since we know that the data has been collected at random and without any bias we can surmise that that the observations are independent.

Normality ANOVA assumes that the data within each group are normally distributed.

Homogeneity of Variance This means that the variability within each group is roughly the same.

Pre-Analysis Testing for Weight Loss grouped by Gender

```
#Normality Test for Gender
shapiro.test(diet$weight.loss[diet$gender == "Female"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet$weight.loss[diet$gender == "Female"]
## W = 0.96956, p-value = 0.3054
```

```
shapiro.test(diet$weight.loss[diet$gender == "Male"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet$weight.loss[diet$gender == "Male"]
## W = 0.97529, p-value = 0.6383
```

If we observe the p-values acquired from the test, we can conclude that even when grouped the data is still somewhat normally distributed. We say this because our $\alpha = 0.05$ and both p-values are greater than 0.05.

```r
#Levene's Test for Homogeneity of Variance
leveneTest(data = diet, group = diet$gender, y= diet$weight.loss)
```

```
## Levene's Test for Homogeneity of Variance (center = median: diet)
##       Df F value Pr(>F)
## group  1   0.202 0.6544
##       74
```

Since, our p-value is greater than α (0.05), we can conclude that there exists Homogeneity of Variances in the *weight.loss* feature when grouped by *gender*.

In the case of *weight.loss* grouped by *gender*, all our tests have yielded posisitve.

Pre-Analysis Testing for Weight Loss grouped by Diet Type

```r
#Normality Test for Diet Type
shapiro.test(diet$weight.loss[diet$diet.type == "A"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet$weight.loss[diet$diet.type == "A"]
## W = 0.92553, p-value = 0.07749
```

```r
shapiro.test(diet$weight.loss[diet$diet.type == "B"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet$weight.loss[diet$diet.type == "B"]
## W = 0.97936, p-value = 0.8722
```

```r
shapiro.test(diet$weight.loss[diet$diet.type == "C"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet$weight.loss[diet$diet.type == "C"]
## W = 0.96013, p-value = 0.372
```

In all cases, the p-value is greater than α (0.05). Hence we can say that the data is normally distributed even when split into groups.

```r
#Levene's Test for Homogeneity of Variance
leveneTest(data = diet, group = diet$diet.type, y= diet$weight.loss)
```

```
## Levene's Test for Homogeneity of Variance (center = median: diet)
##       Df F value Pr(>F)
## group  2  0.4629 0.6313
##       73
```

Since, our p-value is greater than α (0.05), we can conclude that there exists Homogeneity of Variances in the *weight.loss* feature when grouped by *diet.type*.

In the case of *weight.loss* grouped by *diet.type,* all tests have yielded positive.

Having met all assumptions, we can now move forward with main ANOVA test itself.

ANOVA Test

```
#Two-way ANOVA
model <- aov(weight.loss ~ gender * diet.type, data=diet)

summary(model)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## gender            1    0.3   0.278   0.052 0.82062
## diet.type         2   60.4  30.209   5.619 0.00546 **
## gender:diet.type  2   33.9  16.952   3.153 0.04884 *
## Residuals        70  376.3   5.376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Judging by the results of the ANOVA, we can conclude that ***diet.type*** plays an important role in the analysis. ***gender*** by itself has no importance but when taken in conjunction with ***diet.type*** it may be a facotr in weight loss.

Thus, we reject the null hypothesis H_O that the means of groups by *gender* are similar, means of groups by *diet.type* are similar and there is no significant interaction between *gender* and *diet.type* features.

Post-Hoc Analysis

```r
#post-hoc analysis
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.loss ~ gender * diet.type, data = diet)
##
## $gender
##                  diff        lwr      upr     p adj
## Male-Female 0.1221283 -0.9480861 1.192343 0.8206233
##
## $diet.type
##            diff        lwr      upr     p adj
## B-A -0.03484966 -1.6215073 1.551808 0.9984761
## C-A  1.84475570  0.2871469 3.402365 0.0162482
## C-B  1.87960536  0.3385771 3.420634 0.0128844
##
## $`gender:diet.type`
##                         diff        lwr       upr     p adj
## Male:A-Female:A    0.6000000 -2.2129628 3.4129628 0.9887997
## Female:B-Female:A -0.4428571 -3.0107291 2.1250148 0.9958151
## Male:B-Female:A    1.0590909 -1.6782698 3.7964516 0.8656520
## Female:C-Female:A  2.8300000  0.3052886 5.3547114 0.0191170
## Male:C-Female:A    1.1833333 -1.4893925 3.8560592 0.7855223
## Female:B-Male:A   -1.0428571 -3.8558199 1.7701056 0.8852416
## Male:B-Male:A      0.4590909 -2.5093998 3.4275816 0.9975014
## Female:C-Male:A    2.2300000 -0.5436187 5.0036187 0.1863470
## Male:C-Male:A      0.5833333 -2.3256625 3.4923292 0.9915569
## Male:B-Female:B    1.5019481 -1.2354126 4.2393087 0.5963201
## Female:C-Female:B  3.2728571  0.7481458 5.7975685 0.0040103
## Male:C-Female:B    1.6261905 -1.0465354 4.2989163 0.4833188
## Female:C-Male:B    1.7709091 -0.9260048 4.4678230 0.3965102
## Male:C-Male:B      0.1242424 -2.7117126 2.9601974 0.9999949
## Male:C-Female:C   -1.6466667 -4.2779524 0.9846191 0.4513580
```

Having rejected the null hypothesis H_O. We wanted to know which factor plays the biggest role in Weight Loss.

Conclusion

After conducting a through pre-testing and then a Two-Way Analysis, we can safely say that Diet Type plays the most important role in Weight Loss. Even within the Various Diet Types the best Type for Weight Loss is that of Diet 'C'.

This Analysis should allow anyone that followed it so far to understand what potential implications it could have at an industrial level.

Citations and References

Couturier, D. L., Nicholls, R., and Fernandes, M. (2018). ANOVA with R: analysis of the diet data set. Retrieved online.