Data602GroupProject

Group 4

2024-10-11

Introduction

As graduate students, balancing academic responsibilities and maintaining a healthy lifestyle can become challenging, especially when convenience means grabbing a quick bite to eat for lunch at fast food restaurants like A&W, Diary Queen, McDonald's, Subway or Tim Hortons. They offer easily accessible meals but they come with concerns about their nutritional content, high calorie and processed fat content.

Beyond graduate life, in today's world, we need to have an important discussion about the potential long-term impacts of these foods on our health. With this in mind, our project aims to address the issue of calorie and nutrient awareness among students, and make it easier for us to navigate the complexities of nutrition and fast food choices.

We initially sourced our data from Kaggle using a dataset that provides a breakdown of the nutrient and calorie value of items on the menu of selected fast-food restaurants. This dataset also contained Weight Watcher Points for each menu item. The Weight Watcher Points are a point system program that enables people to easily track their diet and understand the nutrition implications of their fast food choices. People who join the Weight Watcher Program are given a maximum daily and weekly allowance of points.

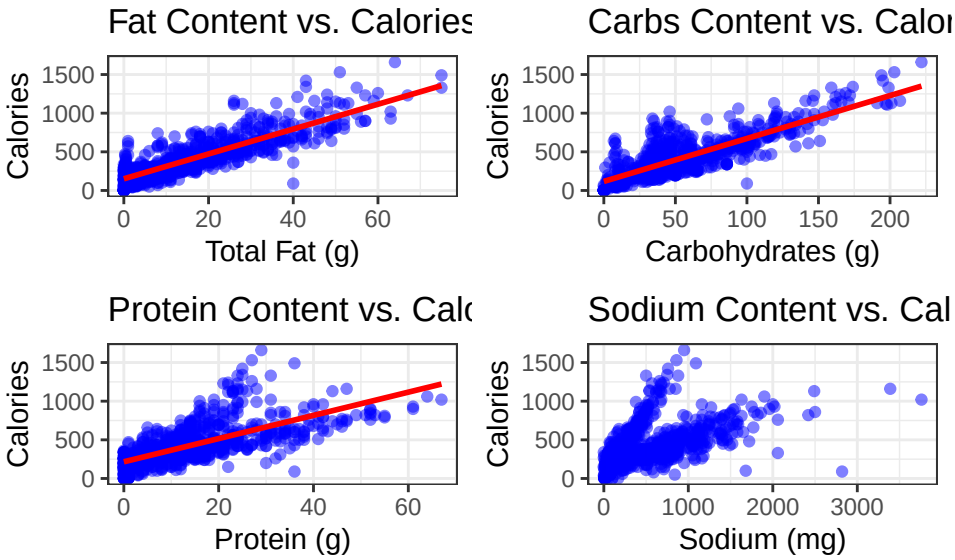
Within the Kaggle dataset, we noticed that the Weight Watcher Points in this data were more than the actual Weight Watcher Points obtainable. We also noticed that the Weight Watcher Points were oddly similar to the value of the calories for each menu item.

To solvee this, we went to the original website where this information was hosted which is Fast Food Nutrition's website. We selected five Canadian fast food restaurants (A&W, Diary Queen, McDonald's, Subway, and Tim Hortons) from this site and webscraped their nutrition and weight watcher points information using Python. The webscrapping codes are available here. We then updated our code to reflect this new data.

Exploratory Data Analysis (EDA)

Relationship between Calories and Other Nutrients

```
grid.arrange(cal_fat, cal_carb, cal_prot, cal_sodium, ncol=2)
```



Calories is highly correlated to three major nutrients: Fat, Carbohydrates, and Protein. The more of these nutrients your food contains, the more calories you are eating. For these major nutrients, there are menu items that contain 0g of a single nutrient but has high caloric value. This can imply that there is a multilinear relationship between calories and all these nutrients, not just a single one which we will address with a multilinear regression analysis later on.

There seems to be two linear trends in the relationship between Calories and Sodium, which almost mimics the trend in Calories and Protein. Sodium is primarily contained in salt and spices. This could imply that the salted items on the menu are generally proteins.

Processed (Bad) Fats Distributions in Each Restaurant

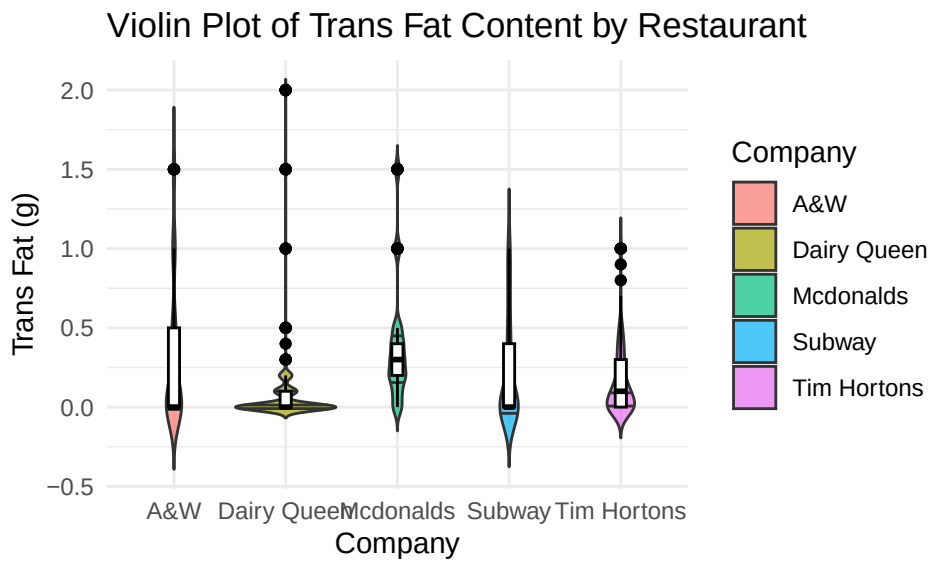
Unsaturated Fat: These are the good fats, and doctors say they should be the majority of fat that people eat. For cooking, they usually come in the form of liquid oils, not solid fats. Unsaturated fats are listed on food packages as polyunsaturated fats and monounsaturated fats.

Saturated Fat: These fats are often derived from animals and generally take a more solid form. They raise "bad" cholesterol and can contribute to heart disease. The government recommends that saturated fats make up less than 10 per cent of daily calories. Common sources include: high-fat cheeses, high-fat cuts of meat, whole-fat milk and cream and ice cream, butter, palm and coconut oils etc.

TRANS FAT: These are the worst fats, and the FDA is forcing food companies to phase them out. They are made when hydrogen is added to vegetable oil, usually to create a certain consistency or increase shelf life, and they are also called partially hydrogenated oils. Many of them have already been phased out, but foods that are more likely to contain trans fats are: fried items, pie crusts, stick margarine, ready-to-use frosting, coffee creamers, some microwave popcorn and frozen pizza, and some cakes, crackers and cookies.

What then is the trans fat content of menu items in these different restaurants?

```
print(trans_fat_violin_plot)
```




```
companies = unique(fast_food_df$Company)
for (company in companies) {
  cat("Favstats for", company, ":\n")
  stats = favstats(filter(fast_food_df, Company == company)$Trans.Fat..g., na.rm = TRUE)
  print(stats)
}
```

```
## Favstats for A&W :
##  min Q1 median  Q3 max      mean        sd   n missing
##    0  0      0 0.5 1.5 0.2873874 0.4257864 111       0
## Favstats for Dairy Queen :
##  min Q1 median  Q3 max      mean        sd   n missing
##    0  0      0 0.1   2 0.1391111 0.3809986 225       0
## Favstats for Mcdonalds :
##  min  Q1 median  Q3 max      mean      sd   n missing
##    0 0.2    0.3 0.4 1.5 0.3841727 0.37169 139       0
## Favstats for Subway :
##  min Q1 median  Q3 max      mean        sd  n missing
##    0  0      0 0.4   1 0.2044444 0.3274943 45       0
## Favstats for Tim Hortons :
##  min Q1 median  Q3 max      mean        sd   n missing
##    0  0    0.1 0.3   1 0.1564444 0.2122704 225       0
```

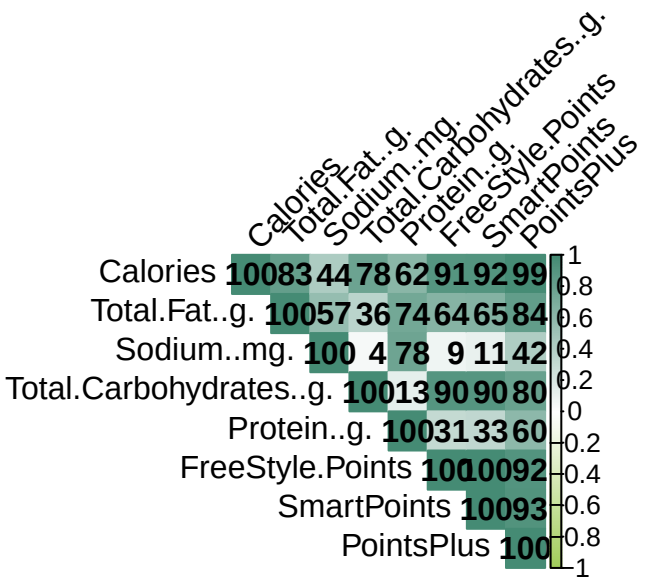
It looks like these companies are making efforts to phase out Trans Fat since their central tendencies are less than 0.5g. All of these companies have outliers with items that have more than 1g of trans fat.

A&W and Diary Queen are closest to 0g, generally. We can test this hypothesis later on. McDonald's has the highest median trans fat than the others. Perhaps consider menu items from A&W as a "healthier" alternative to McDonald's.

Correlations between Selected Nutrients, Calories, and Weight Watcher Points

There are three categories of Weight Watcher Points: Freestyle Points, Smart Points, and Plus Points.

```
# Heat map of the correlation matrix
corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black",
         tl.srt = 45,
         col = colorRampPalette(c("darkolivegreen3", "white",
                                  "aquamarine4"))(200),
         addCoef.col = "black", addCoefasPercent = TRUE)
```



From the correlation plot, we see that Calories are highly correlated with Fat and Carbs, with medium correlation to Proteins.

For the Weight Watcher Points categories: 1. Freestyle Points: Highly correlated with calories and carbohydrates. 2. SmartPoints: Highly correlated with Calories and Carbohydrates. 3. PointsPlus: Most correlated with Calories. Good correlations with carbs and fat. All the weight watcher point categories are correlatable with calories.

Relationships between Weight Watcher Points and Calories.

```
# calories vs freestyle
fsp = ggplot(fast_food_df,
             mapping = aes(x = Calories, y = FreeStyle.Points)) +
  geom_point() +
  geom_smooth(mapping = aes(group = Company, color = Company),
              show.legend = TRUE, method = 'loess', formula = 'y~x')  +
  labs(title = "Scatter Plot of Calories vs. Weight Watchers Freestyle Points",
       x = "Calories", y = "Freestyle Points", color = "Company")

# calories vs points plus
pp = ggplot(fast_food_df,
             mapping = aes(x = Calories, y = PointsPlus)) +
  geom_point() +
  geom_smooth(mapping = aes(group = Company, color = Company),
              show.legend = TRUE, method = 'loess', formula = 'y~x')  +
  labs(title = "Scatter Plot of Calories vs. Weight Watchers Plus Points",
```

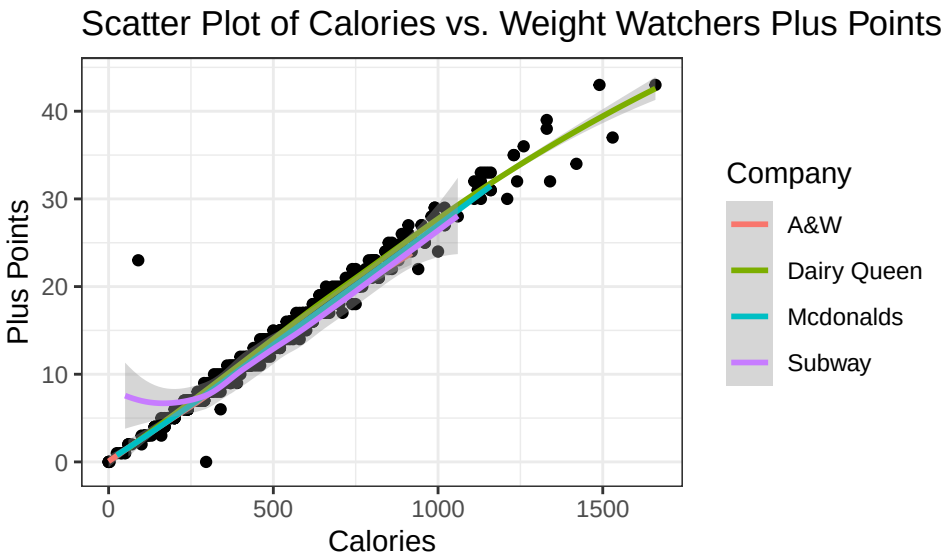
```
       x = "Calories", y = "Plus Points", color = "Company")

# calories vs smart points
sp = ggplot(fast_food_df,
            mapping = aes(x = Calories, y = SmartPoints)) +
  geom_point() +
  geom_smooth(mapping = aes(group = Company, color = Company),
              show.legend = TRUE, method = 'loess', formula = 'y~x')  +
  labs(title = "Scatter Plot of Calories vs. Weight Watchers Smart Points",
       x = "Calories", y = "Smart Points", color = "Company")
```

```
print(fsp)
```

The calories are highly correlatable with the Freestyle Points but McDonald's trend deviate slightly from that of A&W, Dairy Queen, and Subway.

```
print(pp)
```



The calories are highly correlatable with the Plus Points and all listed restaurants follow the same trend.

```
print(sp)
```

The calories are highly correlatable with the Smart Points but McDonald's trend deviate slightly from that of A&W, KFC, and Subway.

Hypothesis Testing

State the hypothesis

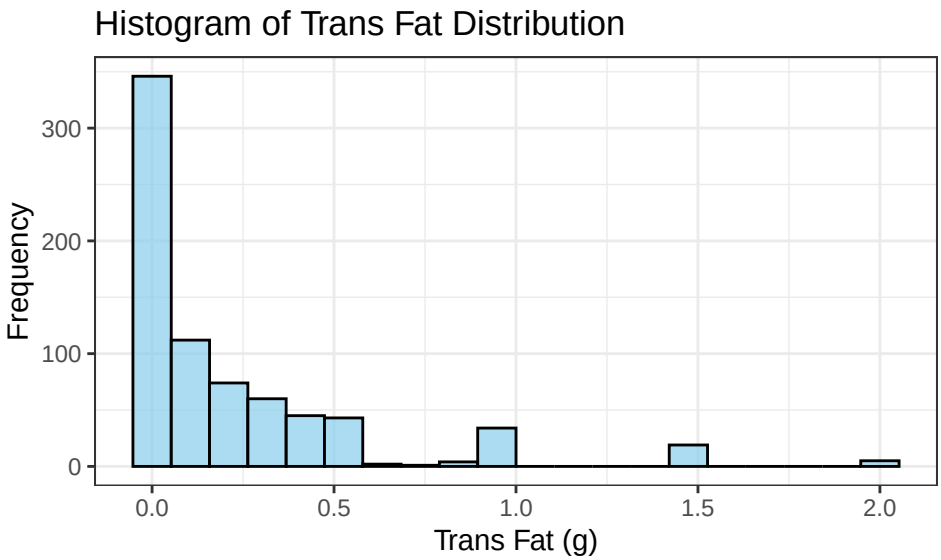
We identified in the EDA that the trans fat content of menu items in A&W and Dairy Queen were closer to 0g than other restaurants. We want to test and see whether the trans fat in menu items of both companies are equal, on average.

1. Null Hypothesis(Ho) - Mean Trans Fat of menu items in A&W and Dairy Queen is equal. i.e.$\mu_{aw} = \mu_{dq}$

2. Alternative Hypothesis(H1) - Mean Trans fat of menu items in A&W and Dairy Queen is not equal i.e.$\mu_{aw} \neq \mu_{dq}$

First, check the distribution of Trans Fat.

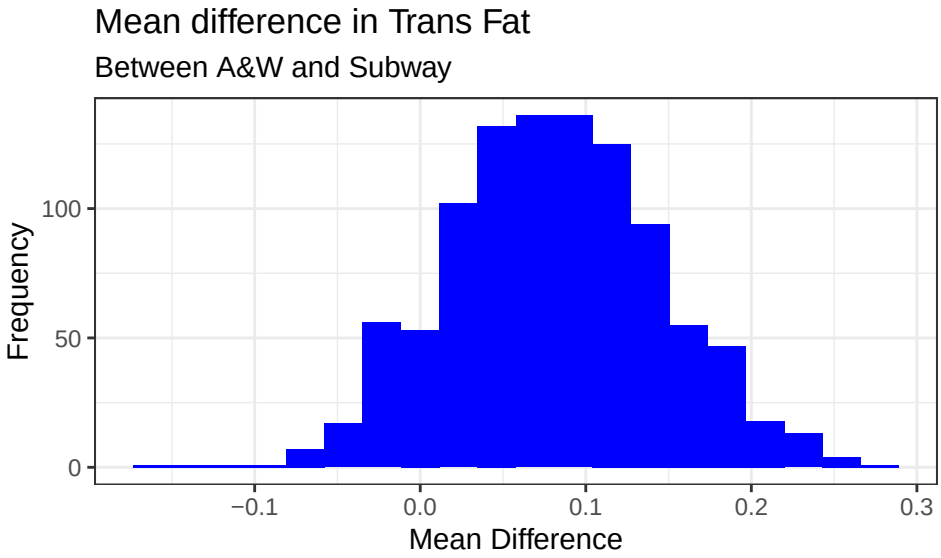
```
print(hist_tf)
```



AS the data is right skewed and not normally distributed, we will use bootstrapping as a non-parametric approach and permutation test as a parametric approach to test the hypothesis.

Bootstrapping

```
ggplot(b, aes(x=result)) + geom_histogram(fill='blue', bins=20) +
  labs(title = 'Mean difference in Trans Fat',
       subtitle = 'Between A&W and Subway',
       x = 'Mean Difference', y='Frequency')
```




```
# 95% confidence interval
quantile(b$result, c(0.025,0.975))
```

```
##        2.5%       97.5%
## -0.03935586  0.20992342
```

Permutation test

```
# Display the results from permutation test
cat("Permutation Test p-value:", p_value)
```

```
## Permutation Test p-value: 0.133
```

Result

Based on the values of the confidence interval and the p-value from the permutation test, we have 95% confidence in rejecting the null hypothesis, and we can accept the alternative hypothesis with 5% level of significance.

This means that we are 95% confident that the average trans fat content of items on the menu of A&W and Dairy Queen are not the same.

Based on the fact that the confidence intervals contains negative values, we can also conclude with 95% confidence that the average trans fat content in Subway is greater than that of A&W.

Regression Analysis

Carbohydrates, Fat, Protein and Calories

We identified that calories were highly correlated with fat, protein, and carbohydrates. We can create a multilinear regression analysis between these nutrients and the calories.

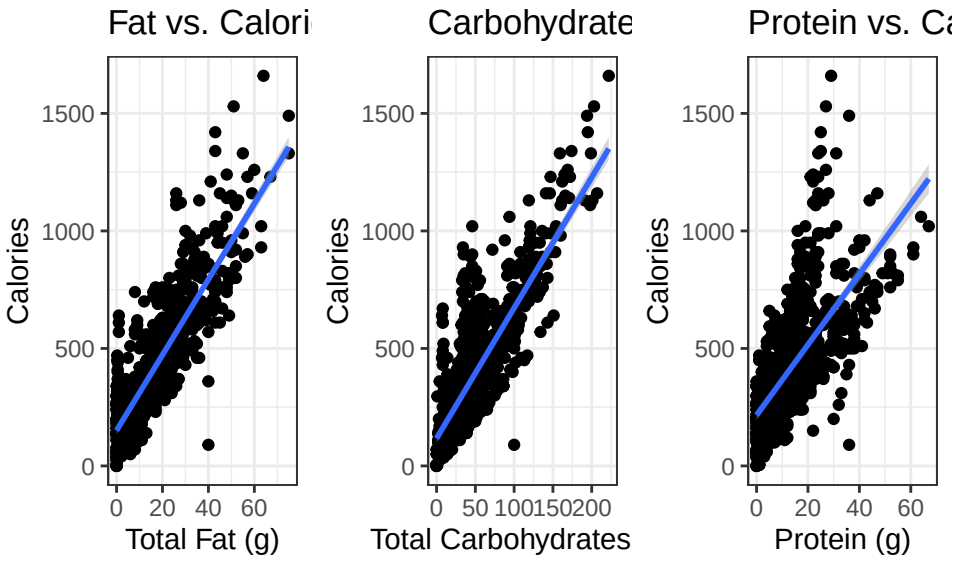
First we can recalculate the correlation coefficient between these nutrients and Calories.

```
cat("The linear correlation between calories and fat is", fat_cal_cor,
    "\nthat of calories and protein is", prot_cal_cor,
    "\nand that of calories and carbohydrates is", carb_cal_cor)
```

```
## The linear correlation between calories and fat is 0.8499438
## that of calories and protein is 0.6887784
## and that of calories and carbohydrates is 0.8133983
```

Next, we visualize and build a multi linear regression model.

```
grid.arrange(fat, carbs, protein, ncol=3)
```




```
print(ml_reg)
```

```
##
## Call:
## lm(formula = Calories ~ Total.Carbohydrates..g. + Protein..g. +
##     Total.Fat..g., data = fast_food_df)
##
## Coefficients:
##             (Intercept)  Total.Carbohydrates..g.              Protein..g.
##                   5.516                    3.847                    4.204
##           Total.Fat..g.
##                   8.643
```

Based on the multi linear regression model:

$\widehat{y} = 5.516 + 3.847C + 4.204P + 8.643F$ where $\widehat{y}$ is the predicted Calories, C is carbohydrate content, P is the protein content, and F is the fat content.

Next, we check the significance of the model coefficients.

```
summary(ml_reg)$coefficients[, "Pr(>|t|)"]
```

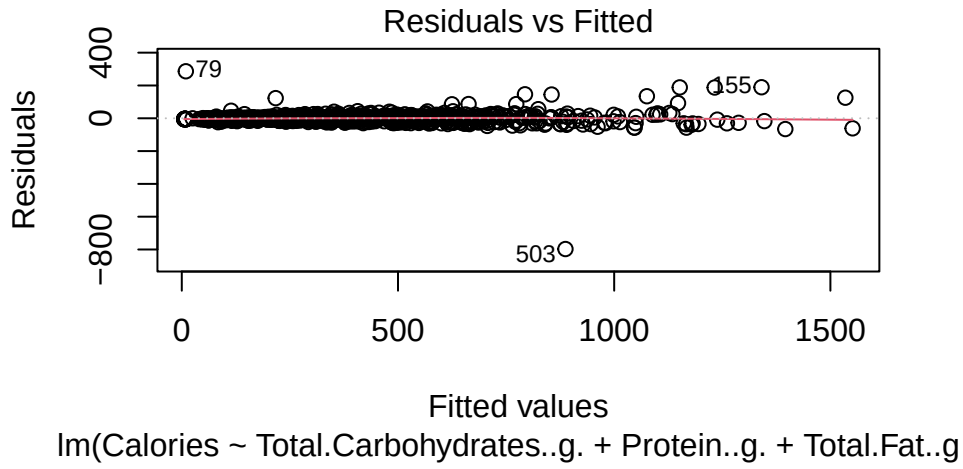
```
##
## 0.041094654360042483232096088840989978052675724029541015625000000000000000000000000000000000000000000000
##
## 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##
## 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##
## 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

The p-value of the intercept is greater than 0, so the intercept model coefficient β_0 is insignificant. The p-value of the model coefficients of all dependent variables is 0, so the dependent variable model coefficient β_1 is significant.

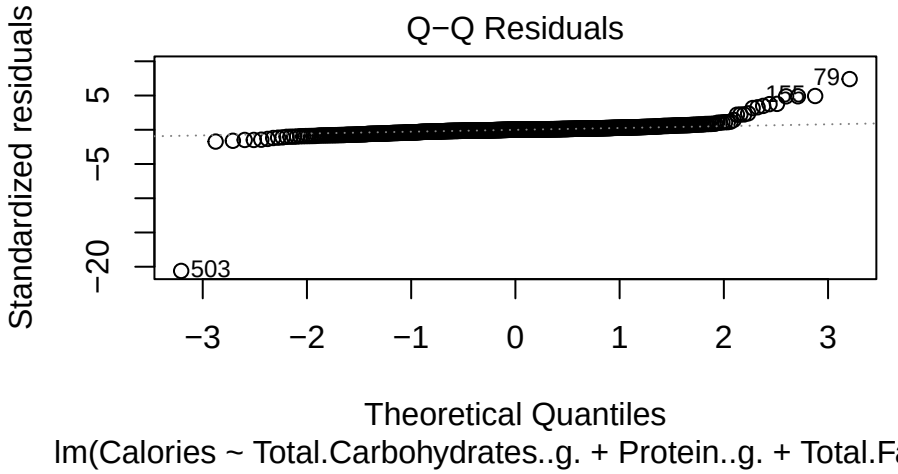
Essentially, for every 1g of carbohydrates, you can expect to add roughly 4 calories in a menu item, for every 1g of protein, you expect to add 4 calories, and for every 1g of Fat, you expect to add, you can expect to lose roughly 3 of your Weight Watchers Plus Points.

Finally, we visualize the residual plots.

```
plot(ml_reg, which=1)
```




```
plot(ml_reg, which=2)
```



The residuals pass the test of independence (Residuals vs Fitted). According to the Q-Q plot, it seems fairly normal but there are outliers at the extremes.

Calories and Weight Watchers Plus Points Calories

From the EDA, since the Calories vs. Plus Points follow the same trend for all listed companies, we can create a regression analysis between the Calories and the Plus Points.

First we can recalculate the correlation coefficient between Calories and Plus Points.

```
cor(fast_food_df$Calories, fast_food_df$PointsPlus, use='complete.obs',
    method = 'pearson')
```

```
## [1] 0.9851189
```

There is a strong positive linear correlation between Calories and Plus Points.

Next, we visualize the regression line and build a simple linear regression model.

```
ggplot(fast_food_df, mapping = aes(x = Calories, y = PointsPlus)) +
  geom_point() +
  stat_smooth(geom='smooth', method = 'lm', formula = 'y~x')  +
  labs(title = "Scatter Plot of Calories vs. Weight Watchers Plus Points",
       x = "Calories", y = "Plus Points")
```


```
reg = lm (PointsPlus ~ Calories, data=fast_food_df)
print(reg)
```

```
##
## Call:
## lm(formula = PointsPlus ~ Calories, data = fast_food_df)
##
## Coefficients:
## (Intercept)     Calories
##    0.009975     0.027144
```

Based on the simple linear regression model:

$\widehat{y} = 0.009975 + 0.027144X$ where $\widehat{y}$ is the predicted Weight Watchers Plus Points and X is the caloric value of the menu item.

Next, we check the significance of the model coefficients.

```
summary(reg)$coefficients[, "Pr(>|t|)"]
```

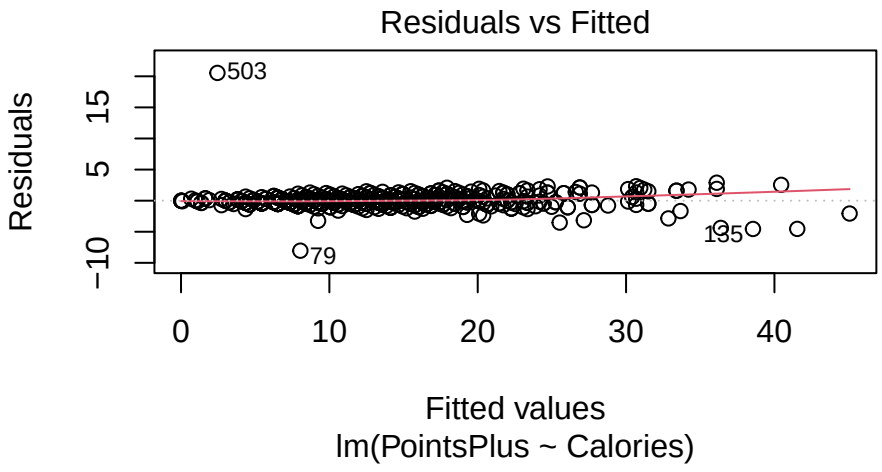
```
## (Intercept)    Calories
##   0.9333009   0.0000000
```

The p-value of the intercept is greater than 0, so the intercept model coefficient β_0 is insignificant. The p-value of the calories (independent variable) is 0, so the dependent variable model coefficient β_1 is significant.

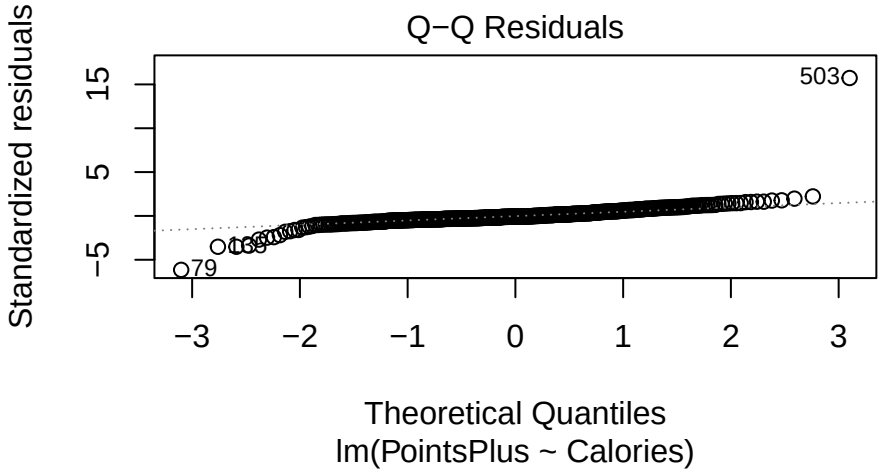
Essentially, for every 100 calories present in a menu item, you can expect to lose roughly 3 of your Weight Watchers Plus Points.

Finally, we visualize the residual plots.

```
plot(reg, which=1)
```




```
plot(reg, which=2)
```



The residuals pass the test of independence (Residuals vs Fitted). From the Q-Q plot, it seems fairly normal but with outliers at the extremes.

Research Question

```
print(plus_points_summary)
```

```
## # A tibble: 5 x 5
##   Company     minimum_pp median_pp mean_pp max_pp
##   <chr>            <dbl>     <dbl>   <dbl>  <dbl>
## 1 A&W                  0         9    9.57     24
## 2 Dairy Queen          2        14   16.1      43
```

```
## 3 Mcdonalds             1        12   12.9      31
## 4 Subway                1        13   13.6      28
## 5 Tim Hortons         Inf        NA   NaN      -Inf
```

Since we do not have the Weight Watcher Points of Tim Hortons, we can use the regression analysis between calories and Weight Watcher Plus Points to predict the points value for items in Tim Hortons' menu.

```r
# extract Tim Hortons Data
th = filter(fast_food_df, Company == "Tim Hortons")

# predict the plus points
model_coeff = summary(reg)$coefficients[, "Estimate"]
th$PointsPlus <- model_coeff[1] + model_coeff[2] * th$Calories

# visualize the predicted Plus Points distribution
hist(th$PointsPlus, main='Distribution of Predicted Plus Points',
     xlab = "Predicted Plus Points")
```