EDA and Visuals

Aleksander Rodriguez

2024-04-24

Exploratory Data Analysis

Within this exploratory analysis statistical tests and visuals will be created based on the data provided. Variables will be labeled to be strong or weak depending on their p-values, f-values, and correlation numbers. The visuals will be made based off some of the strongest variables to give information and the relation the that variable has to BMI.

```
# EDA Part

obesity <- read.csv("C:/Users/Alek4/OneDrive/Documents/CsvFiles/ObesityDataSet.csv")

# Wanted to see how the data is spread out. I didn't want the data to be lopsided
# toward one demographic which could make models less accurate. The table
# shows that there is a good spread within the data.
table(obesity$Obese_level)
```

```
##
## Insufficient_Weight       Normal_Weight      Obesity_Type_I     Obesity_Type_II
##                 272                 287                 351                 297
##    Obesity_Type_III  Overweight_Level_I Overweight_Level_II
##                 324                 290                 290
```

```
# Adding the BMI column by calculating it with weight and height.
# BMI stands for Body Mass Index. It measures an individual's body
#weight relative to their height.
obesity$BMI <- obesity$Weight/(obesity$Height ^ 2)
obesity$Weight_Status <- ifelse(obesity$BMI  >= 30, 1,0)
head(obesity)
```

```
##   Age Gender Height Weight    Alcohol High_Cal Veggies_consumed Meals_Daily
## 1  21 Female   1.62   64.0         no       no                2           3
## 2  21 Female   1.52   56.0  Sometimes       no                3           3
## 3  23   Male   1.80   77.0 Frequently       no                2           3
## 4  27   Male   1.80   87.0 Frequently       no                3           3
## 5  22   Male   1.78   89.8  Sometimes       no                2           1
## 6  29   Male   1.62   53.0  Sometimes      yes                2           3
##   Monitor_Cals SMOKE Water_Consumed family_history_with_overweight
## 1           no    no              2                            yes
## 2          yes   yes              3                            yes
## 3           no    no              2                            yes
```

```
## 4           no    no              2                            no
## 5           no    no              2                            no
## 6           no    no              2                            no
##   Physical_Activity TUE Food_Between_Meals     Tranportaion_Used
## 1                 0   1          Sometimes Public_Transportation
## 2                 3   0          Sometimes Public_Transportation
## 3                 2   1          Sometimes Public_Transportation
## 4                 2   0          Sometimes               Walking
## 5                 0   0          Sometimes Public_Transportation
## 6                 0   0          Sometimes            Automobile
##           Obese_level      BMI Weight_Status
## 1       Normal_Weight 24.38653             0
## 2       Normal_Weight 24.23823             0
## 3       Normal_Weight 23.76543             0
## 4  Overweight_Level_I 26.85185             0
## 5 Overweight_Level_II 28.34238             0
## 6       Normal_Weight 20.19509             0
```

Testing Variables

Checking if the variable's represent a good p-value and f-value so we can make good analysts over the data given. Correlation test's will also be used for continuous data to determine if the data is good to use. A strong relation or correlation is what we are looking for.

Age has strong a correlation

```
cor_results <- cor.test(obesity$BMI, obesity$Age)
cor_results
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Age
## t = 11.563, df = 2109, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2036214 0.2838689
## sample estimates:
##       cor
## 0.2441631
```

Not a strong relation between BMI and gender

```
t.test(BMI ~ Gender, data = obesity)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by Gender
## t = 2.4282, df = 1823.5, p-value = 0.01527
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
```

```
##  0.1633912 1.5358581
## sample estimates:
## mean in group Female   mean in group Male
##             30.13000             29.28038
```

Weight has a strong correlation to BMI

```
cor_results <- cor.test(obesity$BMI, obesity$Weight)
cor_results
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Weight
## t = 120.87, df = 2109, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9292008 0.9399808
## sample estimates:
##       cor
## 0.9348057
```

Height has a strong correlation but is not as strong as others to BMI

```
cor_results <- cor.test(obesity$BMI, obesity$Height)
cor_results
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Height
## t = 6.1053, df = 2109, p-value = 1.218e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.08962567 0.17347201
## sample estimates:
##       cor
## 0.1317845
```

Alcohol has a strong relation with BMI

```
anova_results <- aov(BMI ~ factor(Alcohol), data = obesity)
summary(anova_results)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(Alcohol)    3   7538  2512.6    41.4 <2e-16 ***
## Residuals       2107 127885    60.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

High_Cal has a strong relation with BMI

```r
t.test(BMI ~ High_Cal, data = obesity)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by High_Cal
## t = -16.435, df = 425.22, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -6.889981 -5.418010
## sample estimates:
##  mean in group no mean in group yes
##          24.26039          30.41438
```

Veggies_consumed has a strong correlation to BMI

```r
cor.test(obesity$BMI, obesity$Veggies_consumed)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Veggies_consumed
## t = 12.552, df = 2109, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2235020 0.3029065
## sample estimates:
##       cor
## 0.2636508
```

Meals_Daily surprisingly does not have a good correlation with BMI

```r
cor.test(obesity$BMI, obesity$Meals_Daily)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Meals_Daily
## t = 1.837, df = 2109, p-value = 0.06635
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.002698094  0.082491390
## sample estimates:
##        cor
## 0.03996928
```

Monitor_Cals has a strong relation with BMI

```r
t.test(BMI ~ Monitor_Cals, data = obesity)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by Monitor_Cals
## t = 15.765, df = 133.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  6.195710 7.973315
## sample estimates:
##  mean in group no mean in group yes
##          30.02233          22.93782
```

Smoke has a weak relation with BMI

```
t.test(BMI ~ SMOKE, data = obesity)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by SMOKE
## t = 0.045462, df = 45.762, p-value = 0.9639
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -1.987413  2.079248
## sample estimates:
##  mean in group no mean in group yes
##          29.70112          29.65520
```

Water_consumed has a strong correlation but is not as strong as others

```
cor.test(obesity$BMI, obesity$Water_Consumed)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Water_Consumed
## t = 6.6922, df = 2109, p-value = 2.809e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1021660 0.1857205
## sample estimates:
##       cor
## 0.1442003
```

family_history_with_overweight has a strong relation with BMI

```
t.test(BMI ~ family_history_with_overweight, data = obesity)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  BMI by family_history_with_overweight
## t = -35.768, df = 1007, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -10.578880  -9.478471
## sample estimates:
##  mean in group no mean in group yes
##          21.50049          31.52917
```

Physical_Activity has a strong correlation to BMI

```
cor.test(obesity$BMI, obesity$Physical_Activity)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$Physical_Activity
## t = -8.2848, df = 2109, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2185448 -0.1359039
## sample estimates:
##        cor
## -0.1775373
```

TUE has a strong correlation but is not as strong as others

```
cor.test(obesity$BMI, obesity$TUE)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity$BMI and obesity$TUE
## t = -4.6025, df = 2109, p-value = 4.423e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1417800 -0.0573014
## sample estimates:
##         cor
## -0.09972039
```

Transportation_Used has a strong relation to BMI

```
anova_results <- aov(BMI ~ factor(Tranportaion_Used), data = obesity)
summary(anova_results)
```

```
##                             Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Tranportaion_Used)    4   2741   685.2   10.88 9.97e-09 ***
## Residuals                 2106 132682    63.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary of Testing

Throughout the EDA of this data, the discovery of many variables with a strong correlation and relation to BMI were found. The EDA process found that the best variables are the amount of consumption of alcohol, veggies, and high-calorie meals. Others were the amount of physical activity, family history of being overweight, and the mode of transportation used. These variables make sense due to the fact these correlate with healthy habits. Eating veggies and low calorie meals, with a significant amount of physical activity can help with weight loss. Mode of transportation can also help lower BMI levels due having to walking or biking places. Having family history of being overweight is another variable that makes sense. If one grows up in a house with others that do not maintain a low BMI then most likely the eating habits of their family members would ware off on them. Drinking water was a strong correlation but was not as important as others.

A surprising variable, that was one of the worst in finding a correlation with BMI, was the amount of meals one eats daily. This could provide insight into the fact that it might not be about how often one eats, but rather what one eats. For example, if you eat five times a day but consume veggies and low-calorie food, you might not increase your BMI, compared to eating three times a day with high-calorie food. Smoking and usage of technology were other weak variables.

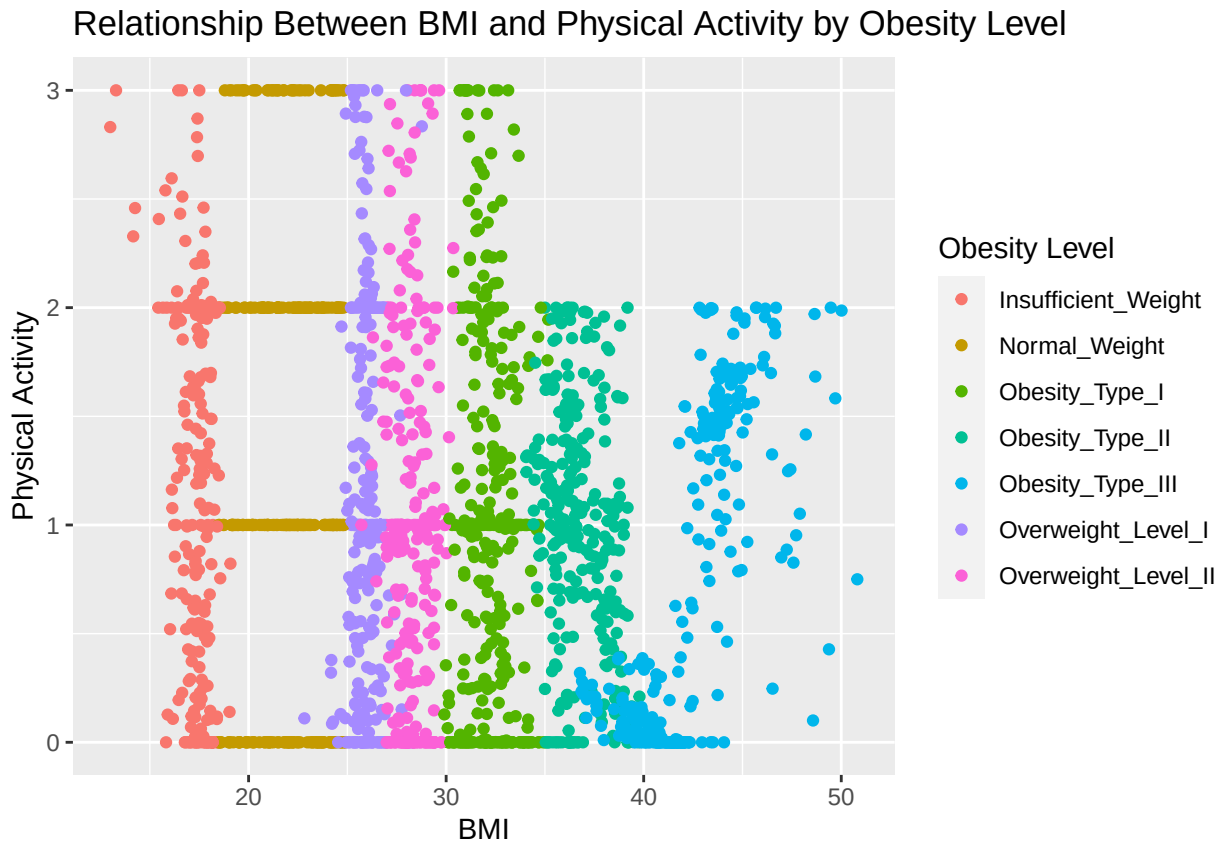
Most of the variables made sense and had a reasonable relation to BMI. With the conclusion of the EDA process, the process of building visuals and models based on the strong variables can commence to help communicate findings within the data.Then eventually give ideas, actions, and reasons of how to lower their BMI according to the data.

Visuals to Illustrate The EDA

The following graph represents the importance of physical activity. The y-axis contains ones physical activity, x-axis contains ones BMI, and the dots calories according to their obesity level. It shows show that to lower ones BMI and stray away form obesity one should be more physically active. The graph gives evidence to why someone should keep active physically.

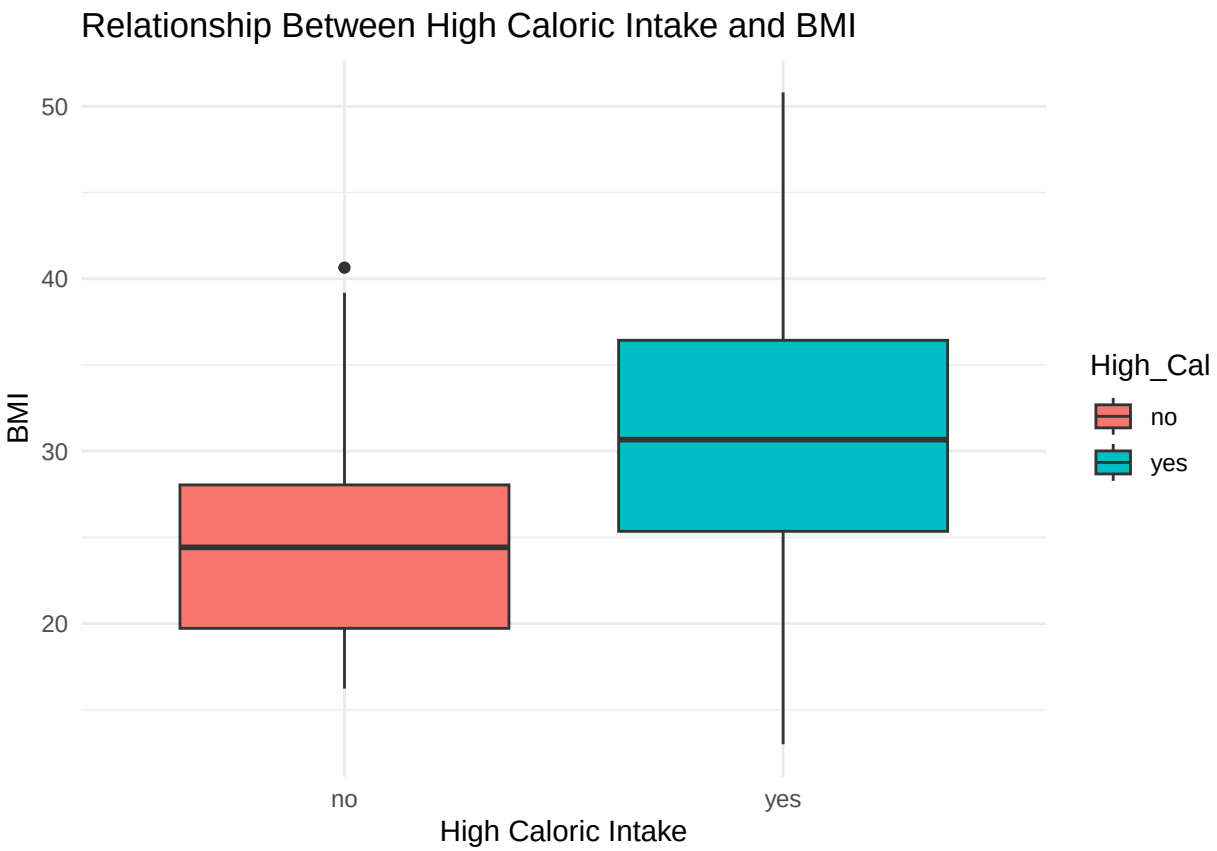
```
library(ggplot2)

ggplot(obesity, aes(x = BMI, y = Physical_Activity, color = factor(Obese_level))) +
  geom_point() +
  labs(x = "BMI", y = "Physical Activity",
       title = "Relationship Between BMI and Physical Activity by Obesity Level") +
  scale_color_discrete(name = "Obesity Level")
```



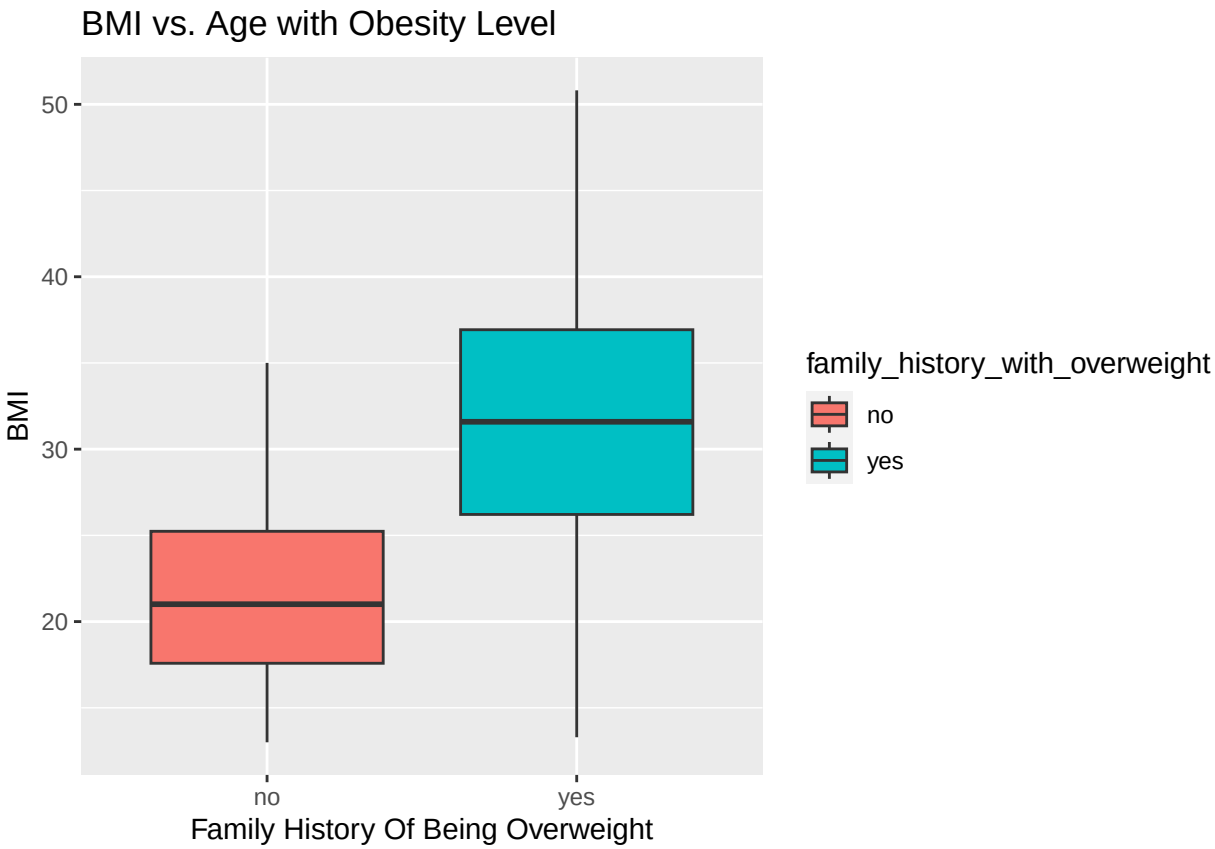
The following graph shows that higher calories within one's meals can effect their BMI's by increasing it. Higher calories results in a higher BMI. This can also give evidence as to why the amount of meals one eat does not matter but what they eat does. Less calories in one's meals is better for ones BMI. Food companies should be encouraged to push lower calorie meals and the public should be informed on how much calorie intake is good for their body.

```
ggplot(obesity, aes(x = High_Cal, y = BMI, fill = High_Cal)) +
  geom_boxplot() +
  labs(x = "High Caloric Intake", y = "BMI",
       title = "Relationship Between High Caloric Intake and BMI") +
  theme_minimal()
```



The following graph shows the relation between family history of being overweight and ones BMI. The findings illustrate the strong correlation between these two variables. It gives the idea that if you were born into a family that is overweight you will have a higher chance to be at a high BMI. If your family is overweight and maintains an unhealthy lifestyle you might partake or pick up those unhealthy habits as well.

```
ggplot(obesity, aes(x = family_history_with_overweight, y = BMI, fill =
                    family_history_with_overweight)) +
  geom_boxplot() +
  labs(x = "Family History Of Being Overweight", y = "BMI", title = "BMI vs. Age with Obesity Level") +
  scale_color_discrete(name = "Obesity Level")
```


BMI vs. Age with Obesity Level
50
40
30
20
BMI
no
yes
Family History Of Being Overweight
family_history_with_overweight
no
yes