Obesity and how to prevent it*

Arsh Lakhanpal

25 April 2022

Abstract

This paper will study to correlation between eating habits and activities that an individual takes part in throughout their day. We observe that the the individuals meals per day, their physical activity frequency, their consumption of food between meals and consumption of alcohol to be the main factors which attributed to higher BMIs. This is significant because it can inform the readers about activities they can do or not do to live a healthy life. The consequences of obesity go past one's physical health, it includes their mental health, sleep and the environment.

Keywords: obesity, body mass index, weight, food, health

```
## # A tibble: 2,015 x 25
##    Gender   Age Height Weight family_history_with~ FAVC   FCVC   NCP CAEC  SMOKE
##    <chr>  <dbl>  <dbl>  <dbl> <chr>                <chr> <dbl> <dbl> <chr> <chr>
##  1 Female    21   1.62   64   yes                  no        2     3 Some~ no
##  2 Male      23   1.8    77   yes                  no        2     3 Some~ no
##  3 Male      27   1.8    87   no                   no        3     3 Some~ no
##  4 Male      22   1.78   89.8 no                   no        2     1 Some~ no
##  5 Male      29   1.62   53   no                   yes       2     3 Some~ no
##  6 Female    23   1.5    55   yes                  yes       3     3 Some~ no
##  7 Male      22   1.64   53   no                   no        2     3 Some~ no
##  8 Male      24   1.78   64   yes                  yes       3     3 Some~ no
##  9 Male      22   1.72   68   yes                  yes       2     3 Some~ no
## 10 Male      26   1.85  105   yes                  yes       3     3 Freq~ no
## # ... with 2,005 more rows, and 15 more variables: CH2O <dbl>, SCC <chr>,
## #   FAF <dbl>, TUE <dbl>, CALC <chr>, MTRANS <chr>, NObeyesdad <chr>,
## #   Indicator <dbl>, new_FAVC <dbl>, new_history <dbl>, new_CAEC <dbl>,
## #   new_CALC <dbl>, new_FAF <chr>, new_MTRANS <dbl>, BMI <dbl>
```

```
## # A tibble: 2 x 2
##   `clean$SCC`     n
##   <chr>       <int>
## 1 no           2015
## 2 yes            96
```

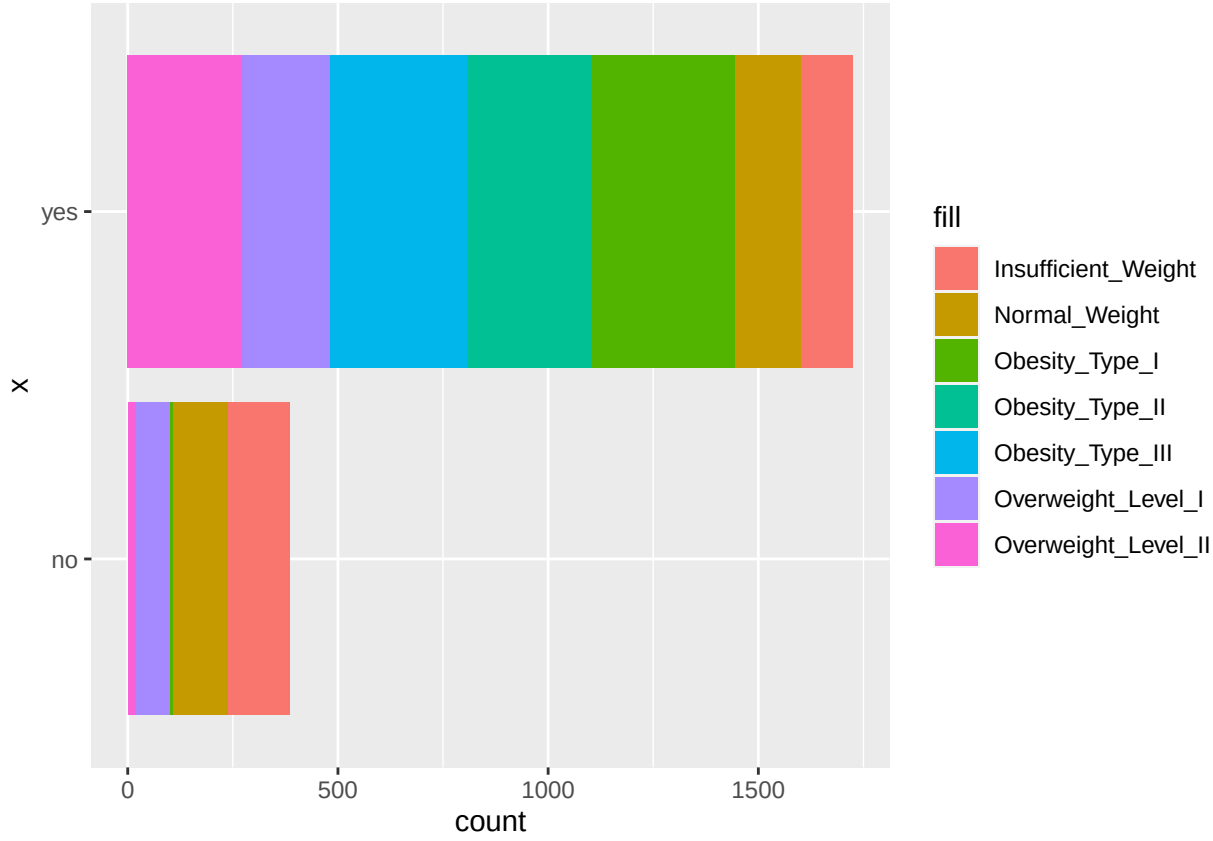
```
## # A tibble: 2 x 2
##   `filtered_obese$SCC == "no"`     n
##   <lgl>                        <int>
## 1 FALSE                            3
## 2 TRUE                           971
```

```
## # A tibble: 2 x 2
##   `overweight$SCC == "no"`     n
##   <lgl>                    <int>
## 1 FALSE                       34
```

*Code and data are available at: https://github.com/lakhan99

```
## 2 TRUE                        532
```

```
## # A tibble: 2 x 2
##   `not_obese$SCC == "yes"`     n
##   <lgl>                    <int>
## 1 FALSE                      263
## 2 TRUE                        37
```




```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   515
```

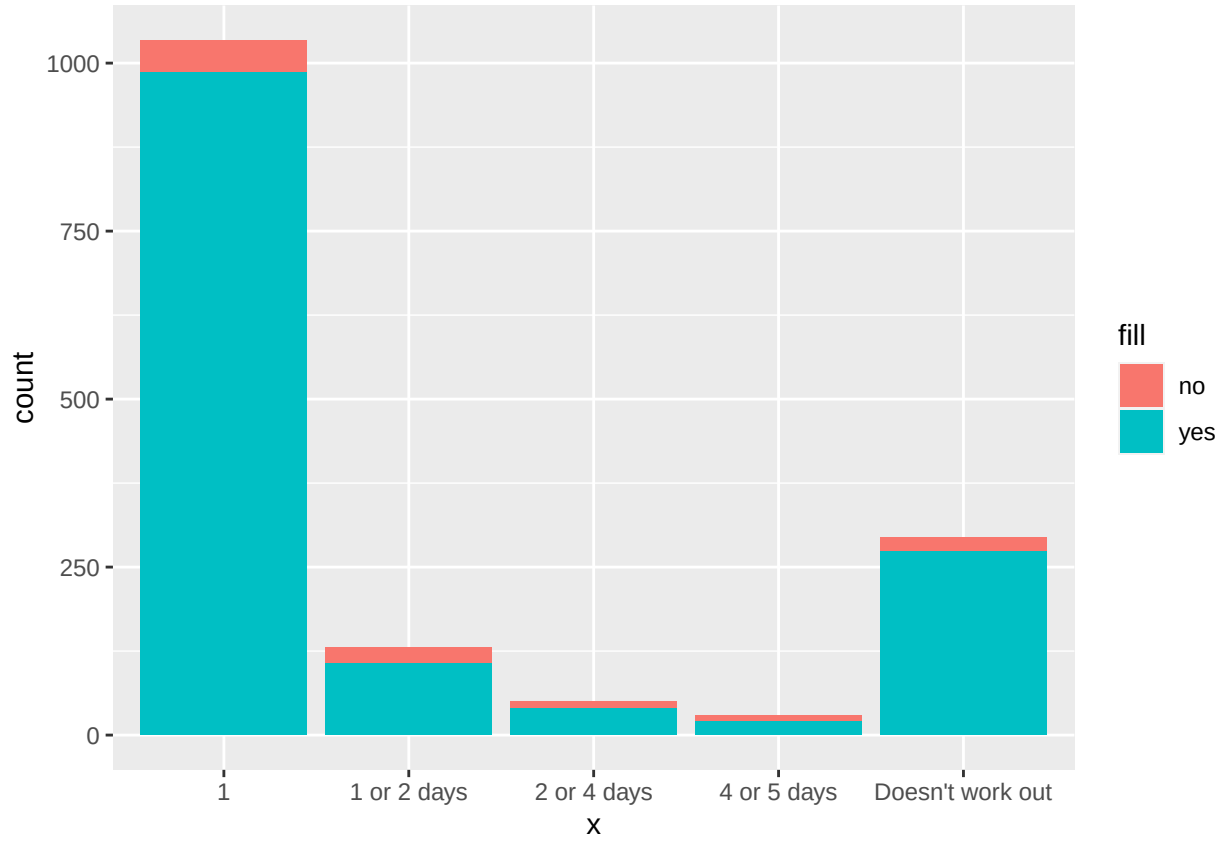
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    71
```

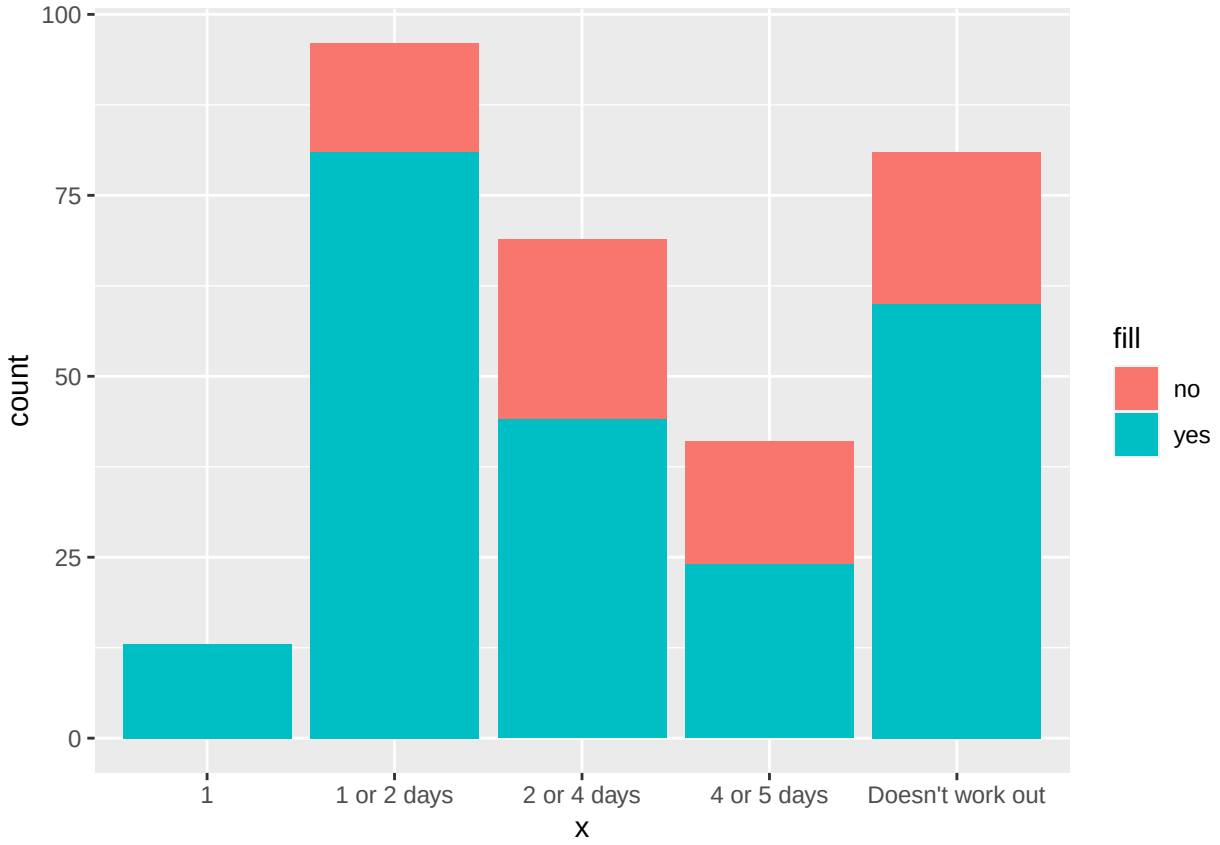
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    58
```

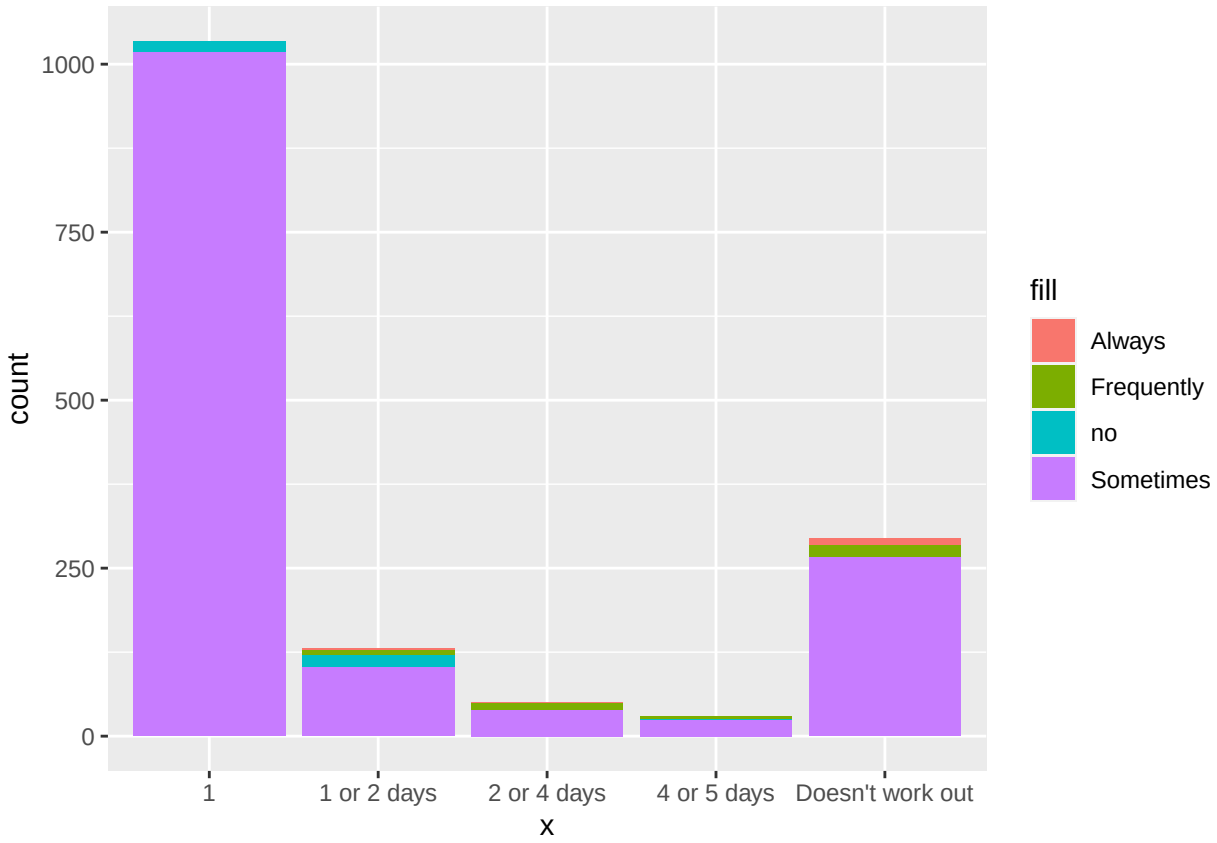
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    70
```

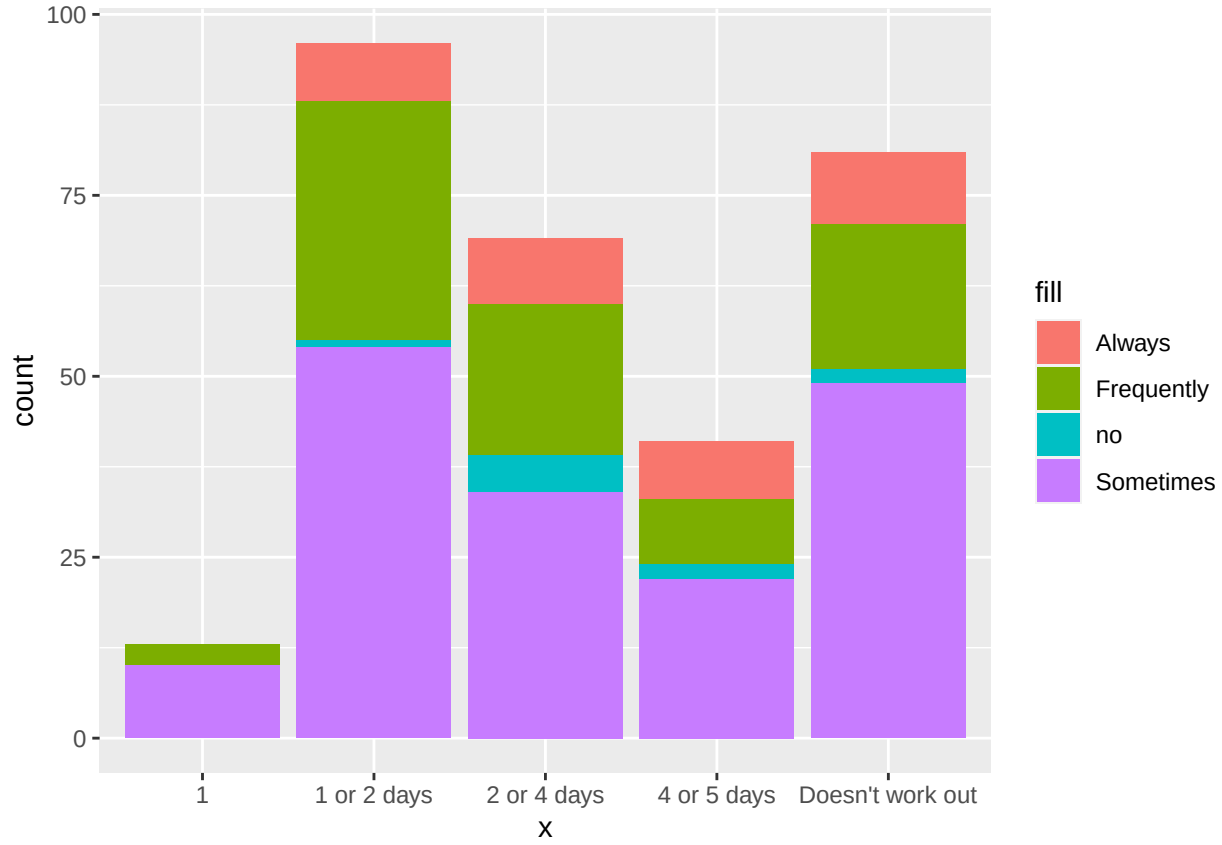
```
## # A tibble: 1 x 1
```

```
##       n
##   <int>
## 1   154
```


1000
750
500
250
0
count
1
1 or 2 days
2 or 4 days
4 or 5 days
Doesn't work out
x
fill
no
yes


100
75
50
25
0
count
1
1 or 2 days
2 or 4 days
4 or 5 days
Doesn't work out
x
fill
no
yes


1000
750
500
250
0
count
1
1 or 2 days
2 or 4 days
4 or 5 days
Doesn't work out
x
fill
Always
Frequently
no
Sometimes



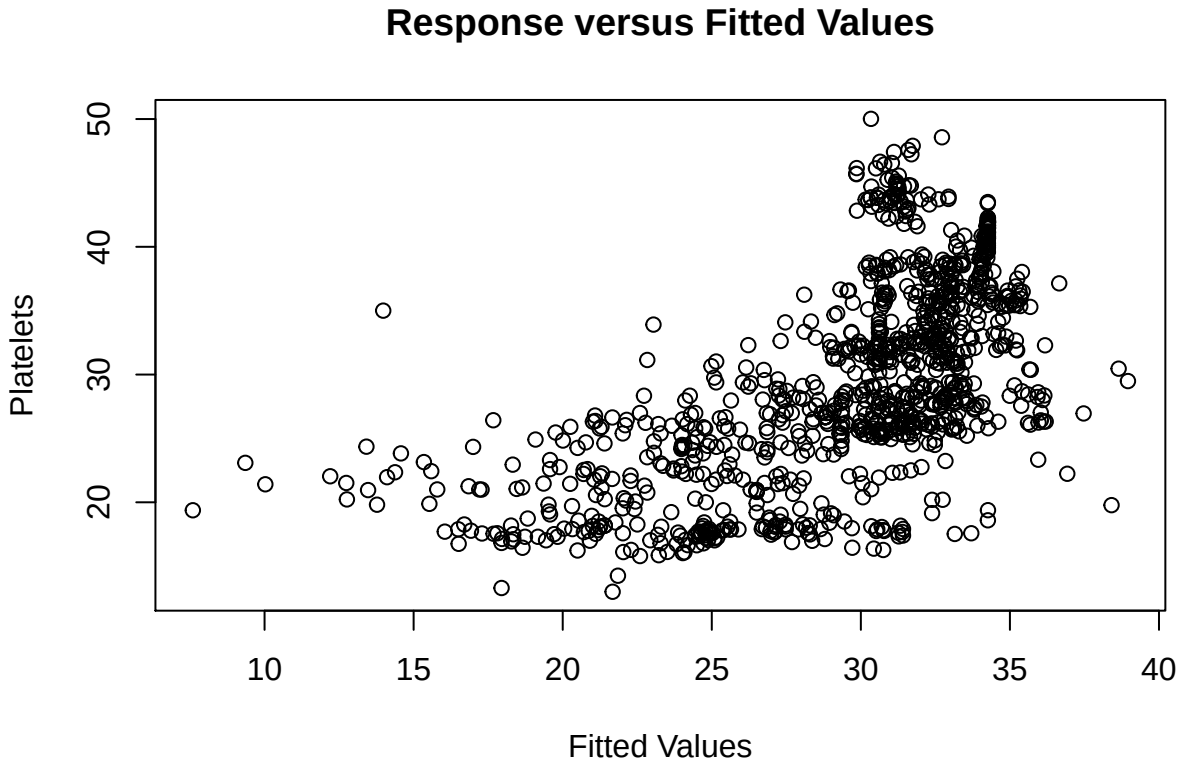

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   166
```

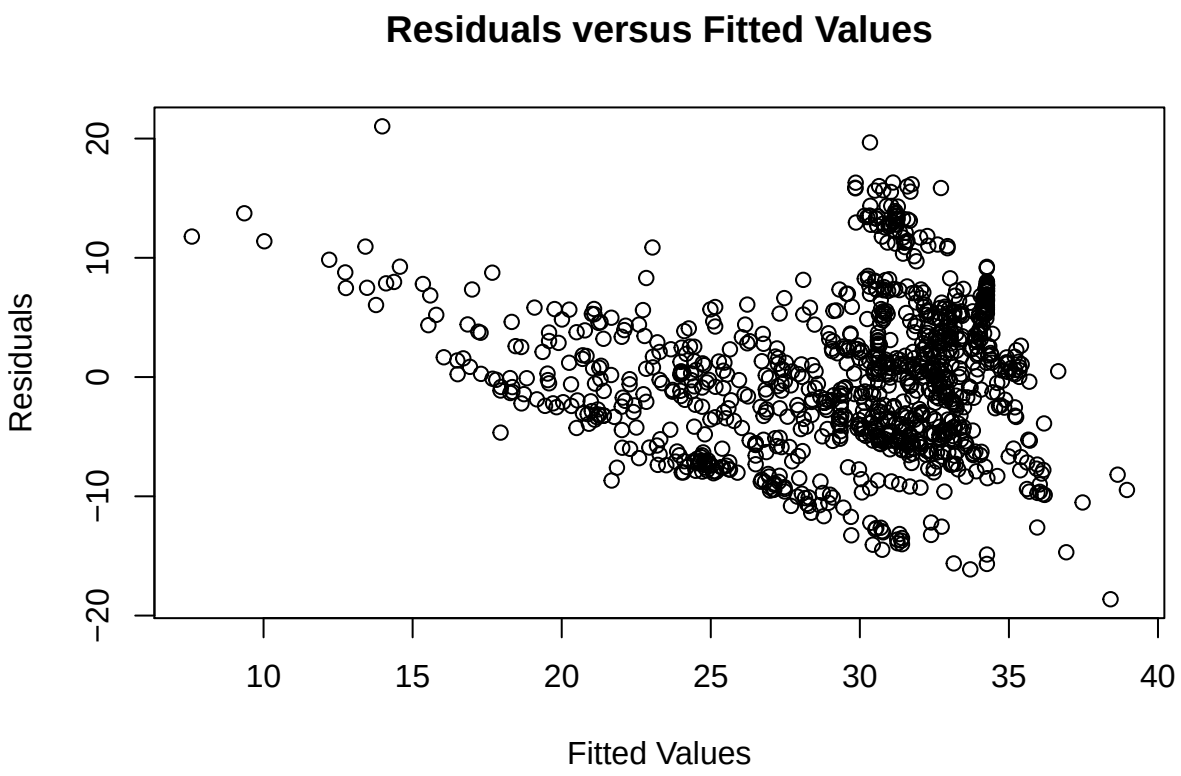
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   111
```

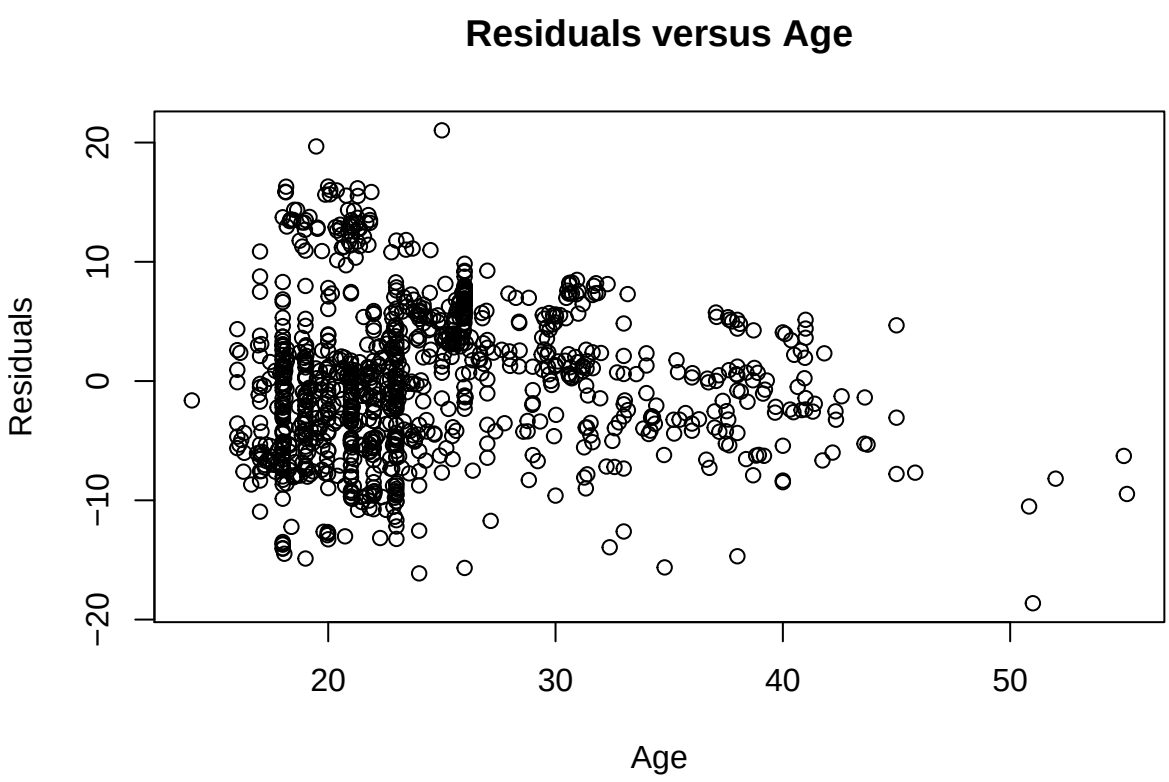
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   376
```

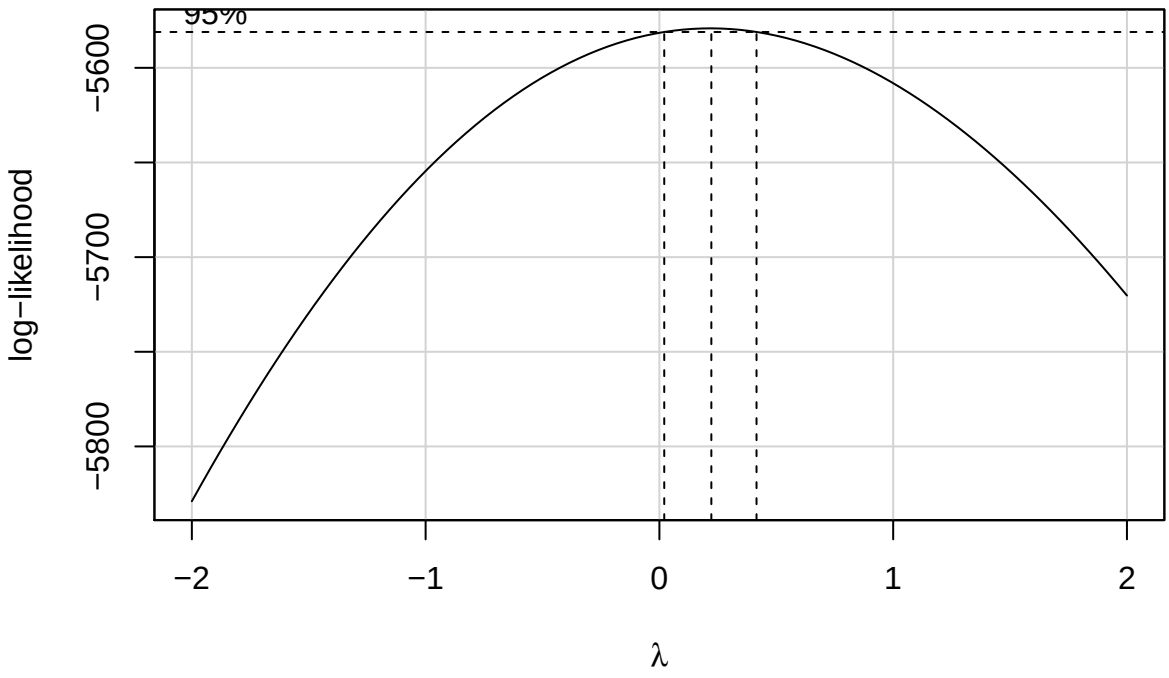
```
##
## Call:
## lm(formula = BMI ~ new_FAVC + new_CAEC + new_history + FAF +
##     Age + MTRANS, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6308  -4.5007  -0.0965   3.9553  21.0231
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                14.12750    1.60348   8.811  < 2e-16 ***
## new_FAVC                    3.14326    0.61889   5.079 4.49e-07 ***
```

```
## new_CAEC                    -4.10268    0.43647  -9.400  < 2e-16 ***
## new_history                  6.96811    0.53899  12.928  < 2e-16 ***
## FAF                         -0.80408    0.23353  -3.443 0.000598 ***
## Age                          0.35596    0.03963   8.983  < 2e-16 ***
## MTRANSBike                   2.65106    3.67157   0.722 0.470425
## MTRANSMotorbike              7.86293    2.86879   2.741 0.006232 **
## MTRANSPublic_Transportation  4.86976    0.59159   8.232 5.46e-16 ***
## MTRANSWalking                2.73566    1.22463   2.234 0.025702 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.261 on 1046 degrees of freedom
## Multiple R-squared:  0.3743, Adjusted R-squared:  0.3689
## F-statistic: 69.53 on 9 and 1046 DF,  p-value: < 2.2e-16
```



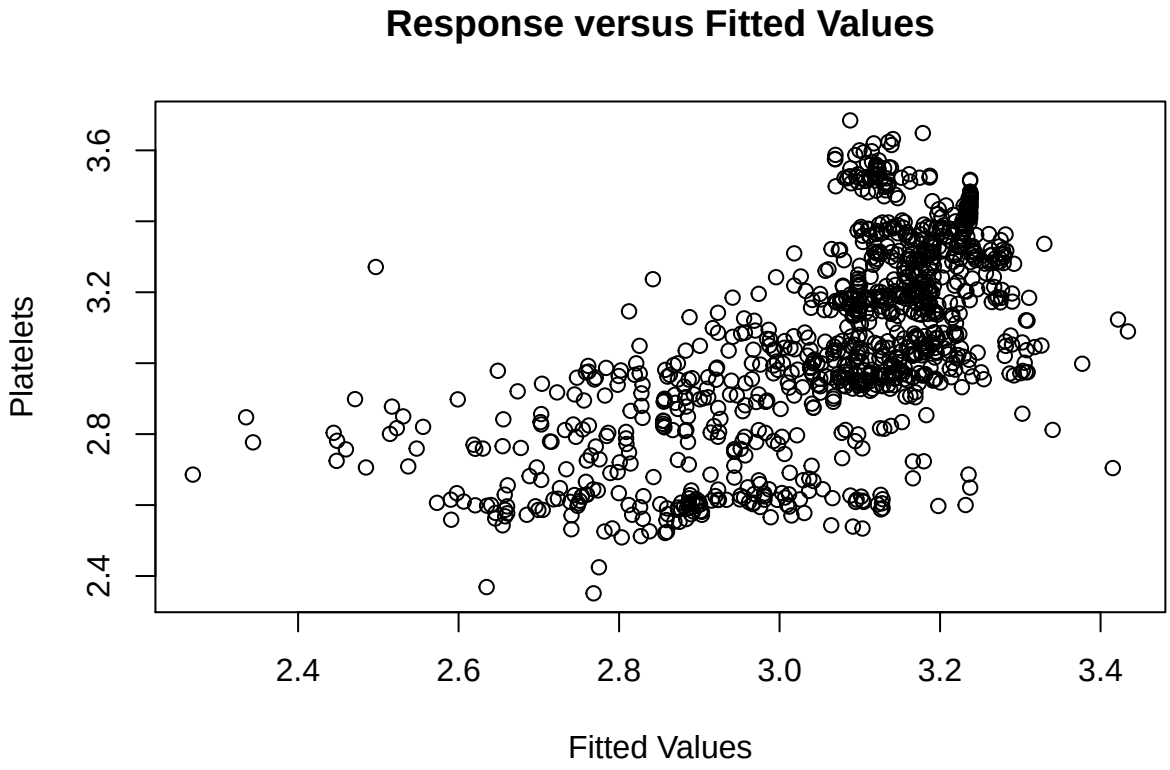

Residuals versus Fitted Values
Residuals
Fitted Values
20
10
0
−10
−20
10
15
20
25
30
35
40


Residuals versus Age
Residuals
20
10
0
−10
−20
20
30
40
50
Age




```
##
## Call:
## lm(formula = BMI_box ~ new_FAVC + new_CAEC + new_history + FAF +
##     Age + MTRANS, data = train_box)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71094 -0.14344  0.00743  0.15011  0.77433
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.509567   0.055881  44.909  < 2e-16 ***
## new_FAVC                     0.102305   0.021568   4.743 2.39e-06 ***
## new_CAEC                    -0.154368   0.015211 -10.148  < 2e-16 ***
## new_history                  0.255116   0.018784  13.582  < 2e-16 ***
## FAF                         -0.030431   0.008138  -3.739 0.000195 ***
## Age                          0.013638   0.001381   9.875  < 2e-16 ***
## MTRANSBike                   0.102940   0.127954   0.805 0.421287
## MTRANSMotorbike              0.299284   0.099977   2.994 0.002823 **
## MTRANSPublic_Transportation  0.170293   0.020617   8.260 4.37e-16 ***
## MTRANSWalking                0.104147   0.042678   2.440 0.014841 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2182 on 1046 degrees of freedom
## Multiple R-squared:  0.3994, Adjusted R-squared:  0.3942
```

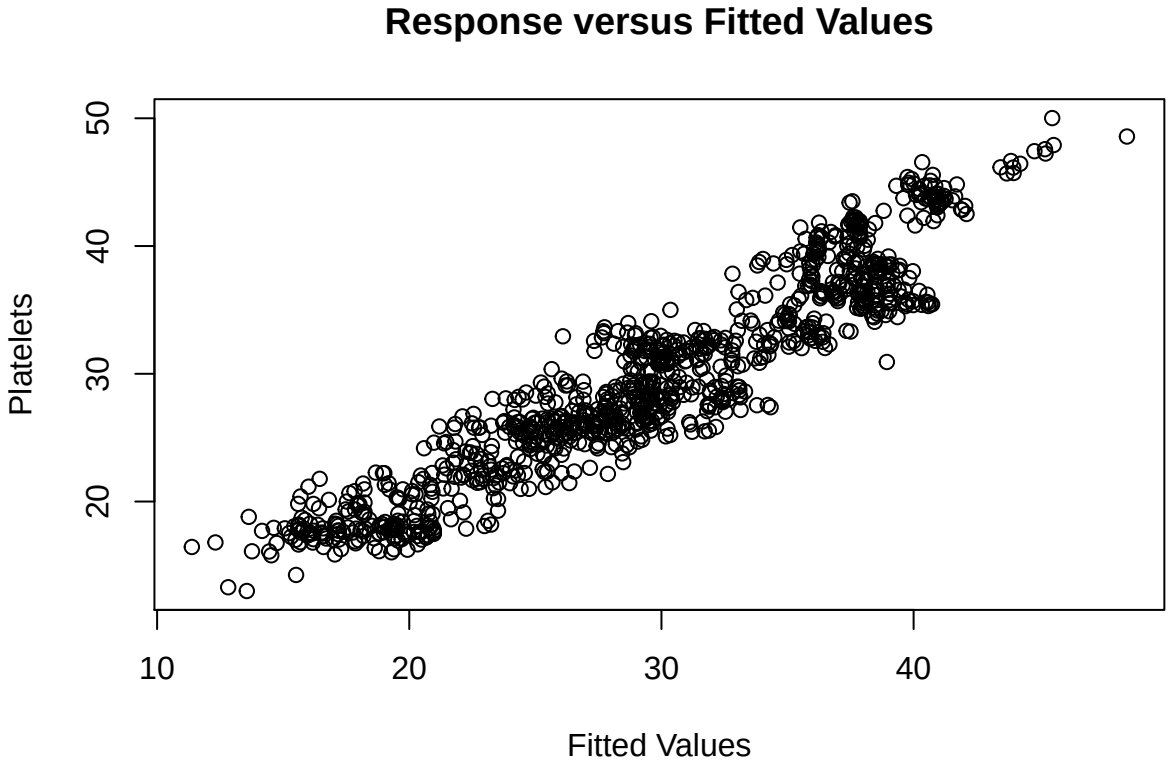
```
## F-statistic: 77.27 on 9 and 1046 DF,  p-value: < 2.2e-16
```




```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.4643        0.50       0.3164       0.6122
## Y2    0.3466        0.33       0.2153       0.4779
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                               LRT df      pval
## LR test, lambda = (0 0) 43.07354  2 4.433e-10
##
## Likelihood ratio test that no transformations are needed
##                               LRT df       pval
## LR test, lambda = (1 1) 94.82952  2 < 2.22e-16
```

```
##
## Call:
## lm(formula = BMI ~ new_FAVC + new_CAEC + new_history + FAF +
##     Weight_power, data = train_power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0153 -1.9292 -0.1135  1.9624  6.8405
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -38.28121    0.96645 -39.610  < 2e-16 ***
## new_FAVC      -0.76042    0.26013  -2.923  0.00354 **
## new_CAEC      -0.61061    0.18949  -3.222  0.00131 **
## new_history    0.05474    0.24478   0.224  0.82308
## FAF           -1.17895    0.09500 -12.411  < 2e-16 ***
## Weight_power  16.07452    0.21918  73.338  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.639 on 1050 degrees of freedom
## Multiple R-squared:  0.8884, Adjusted R-squared:  0.8879
## F-statistic:  1672 on 5 and 1050 DF,  p-value: < 2.2e-16
```



1 Introduction

Obesity is defined as the "abnormal or excessive fat accumulation that presents a risk to health [citation]" Although not as large a problem in previous years, as of 2017, over 4 million people have died as a result of obesity. According to the World Health Organization, an individual with a body mass index (referred to as BMI) that is greater than 30 is considered obese and an individual with a BMI greater than 25 is overweight [citation]. Similarly a BMI of less than 18.5 suggests that the individual is underweight. This value is calculated by dividing the individuals weight in kilograms by the square of their height which is measured in meters[citation]. The reason that BMI will be used in this paper is because of the fact that it accounts for the height of an individual which provides a standard for us to compare people with different heights[citation]. This allows results of this study to be relevant as height also is a determining factor in an individuals weight.

In this report, from data collected from individuals living in Peru, Mexico and Columbia, I plan to examine various factors and determine their significance in a person being overweight or obese. I start out my data section by creating a number of graphs and studying a lot of the variables from the data set. From this, I am able to identify the most significant variables that play a role in an individual having a higher weight. To be precise, this is because from the formula of BMI, a higher weight would indicate a higher BMI and thus resulting in someone being overweight or obese. With the variables that are most significant, I construct a linear model. With the benchmarks set by the BMI, this model could be used in determining what factors and the severity of those factors play a role in someone being overweight or obese.

"add more when more graphs"

Section 2 of this paper talks more about where and how the data was collected and if any modifications were made to this data set for the purpose of this study. This includes the construction of new variables such as "BMI" and removal of other variables. In section 3 of this paper, we will make a linear model of our data and discuss the results of our study. Section 4 talks about the limitations of this study using a measure like BMI and possible ethical concerns the reader consider before making any sort of decision.

2 Data

This dataset was obtained from the UCI Machine Learning Repository and was donated on August 27, 2019 [citation]. It consists of 17 variables and 2111 observations, all based of individuals aged 14-61 who lived in Mexico, Peru and Colombia. This data set was used for a paper by Fabio Palechor and Alexis Manotas in which they simply presented the data that was collected []. The responses were conducted using surveys which can be found in the appendix of this paper. I worked on this data on R [citation] and used readr[] to help load the data. Other packages such as tidyverse[] and dplyr[] were used to clean the data whereas ggplot2 [] was used to construct the graphs in this paper. Finally, the linear model was created using the stats package, something included in base R and the car package [].

2.1 Data Cleaning and Modifications

The raw data for this dataset did not contain any empty or ambiguous responses however, I did run the code to omit any "N/A's" that may have been in the dataset. Once this was done, I removed the variables that were insignificant which consisted only of the gender variable. Although women are more likely to become obese in comparison to men in general, this statistic does vary for many countries [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3649717/]. Along with that, this paper aims to focus on decisions people make in their day-to-day lives which involve their diet and physical activity, something which is accessible to both genders included in the study. Variables with an individuals height, weight, age and their family history with obesity were included in the data along with responses about their eating habits. These included their,

- Consumption of high caloric food
- Frequent consumption of vegetables
- Consumption of food between meals
- Consumption of water daily
- Consumption of Alcohol

There were also questions on their physical habits. The habitual questions were based on their:

- Calorie consumption monitoring
- Transportation Method Used
- Physical activity frequency
- Time using technological devices

There were various modifications that needed to be done for the data. The first modification was creating the "BMI" variable as the dataset itself only consisted of the variable in which it stated the weight class for every individual. To do this, as per the formula to calculate BMI, every individuals weight (kilograms) was divided

by the square of their height (metres). This variable was important for the modeling part of this section as the response variable for this study had to be a numerical value. Other variables that were created were those which pertained to whether or not an individual consumed high caloric food frequently, their consumption of food between meals, their family history with obesity, their consumption of alcohol and their method of transportation. These were all variables that were in the dataset however to use these variables in our linear model, we had to give numerical values to the responses that were recorded. The response for an individuals family history with obesity was either "yes" or "no" and the new variable simply records these as "1" and "0" respectively. Similarly, for the response about method of transportation, the responses "Automobile," "Motorbike," "Public Transportation," "Walking," and "Biking" were all given respective values of "5," "4," "3," "2," and "1" with the assumption that biking instead of driving to work is more beneficial to one's health. The other variables created were also created with this assumption. A variable for frequency of physical activity was also created which took responses in the numeric form to responses such as "4 to 5 days" which helped with Figure #insert figure.# Finally, an indicator variable was created to aid with splitting our data into training and test data sets for model validation.

3 Model

4 Results

5 Discussion

BMI not the perfect measure Doesn't take into account that muscle weighs more than fat Doesn't account for mood?

definition of "always"

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional details

B References