Regression 1 Final Project Code

Data Wrangling

```r
# loading data from the source
data_raw <- read.csv('./data/raw_data.csv')

# loading a data dictionary with more readable column names
dic <- openxlsx::read.xlsx("./data/data_dictionary.xlsx")

# cleaning data
data <-
  data_raw |>
  dplyr::mutate(
    dplyr::across(
      dplyr::where(is.character),
      ~factor(stringr::str_to_title(.x))
    ),
    # ordering factors for visualization & intuitive dummy creation
    dplyr::across(
      .cols = c(CAEC, CALC),
      .fns = ~factor(.x, level = c("No","Sometimes","Frequently","Always"))
    ),
    # converting numeric counts to integers (see first paragraph of the results section)
    dplyr::across(
      .cols = c(FCVC, TUE, NCP, CH2O, FAF, Age),
      .fns = as.integer
    ),
    # ordering transit types by their frequency
    MTRANS = forcats::fct_inorder(factor(MTRANS)),
    BMI = Weight/(Height^2)
    ) |>
  # removing unneeded variables
```

```
  dplyr::select(-c(Height, Weight, NObeyesdad))

# converting names to the human readable
names(data) <- dic$Name

# generating a "dirty" copy without integer conversions
data_dirty <-
  data_raw |>
  dplyr::mutate(
    dplyr::across(
      dplyr::where(is.character),
      ~factor(stringr::str_to_title(.x))
    ),
    dplyr::across(
      .cols = c(CAEC, CALC),
      .fns = ~factor(.x, level = c("No","Sometimes","Frequently","Always"))
    ),
    MTRANS = forcats::fct_inorder(factor(MTRANS)),
    BMI = Weight/(Height^2)
  ) |>
  dplyr::select(-c(Height, Weight, NObeyesdad))

names(data_dirty) <- dic$Name
```

Exploratory data analysis

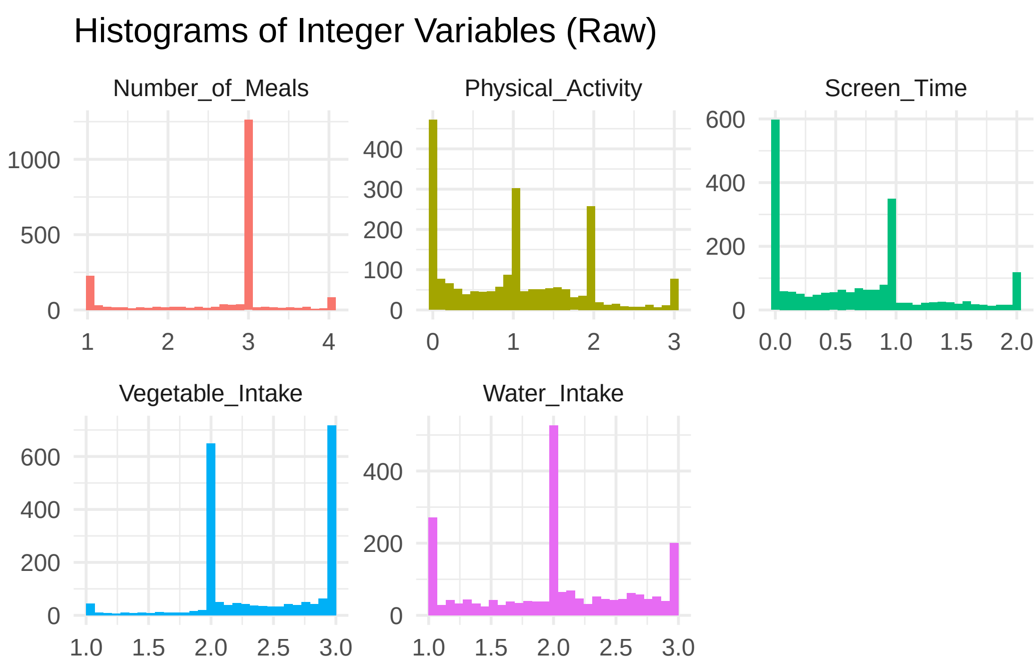
Univariate Analysis

```
psych::describe(data) |>
  dplyr::select(-c(median, trimmed, mad))
```

```
                     vars    n  mean   sd min   max range  skew kurtosis   se
Gender*                 1 2111  1.51 0.50   1  2.00  1.00 -0.02    -2.00 0.01
Age                     2 2111 23.97 6.31  14 61.00 47.00  1.56     2.97 0.14
Family_History*         3 2111  1.82 0.39   1  2.00  1.00 -1.64     0.70 0.01
High_Caloric_Food*      4 2111  1.88 0.32   1  2.00  1.00 -2.40     3.74 0.01
Vegetable_Intake        5 2111  2.21 0.60   1  3.00  2.00 -0.12    -0.47 0.01
Number_of_Meals         6 2111  2.52 0.83   1  4.00  3.00 -0.88    -0.46 0.02
Snacking*               7 2111  2.14 0.47   1  4.00  3.00  1.90     5.38 0.01
Smoking*                8 2111  1.02 0.14   1  2.00  1.00  6.70    42.95 0.00
```

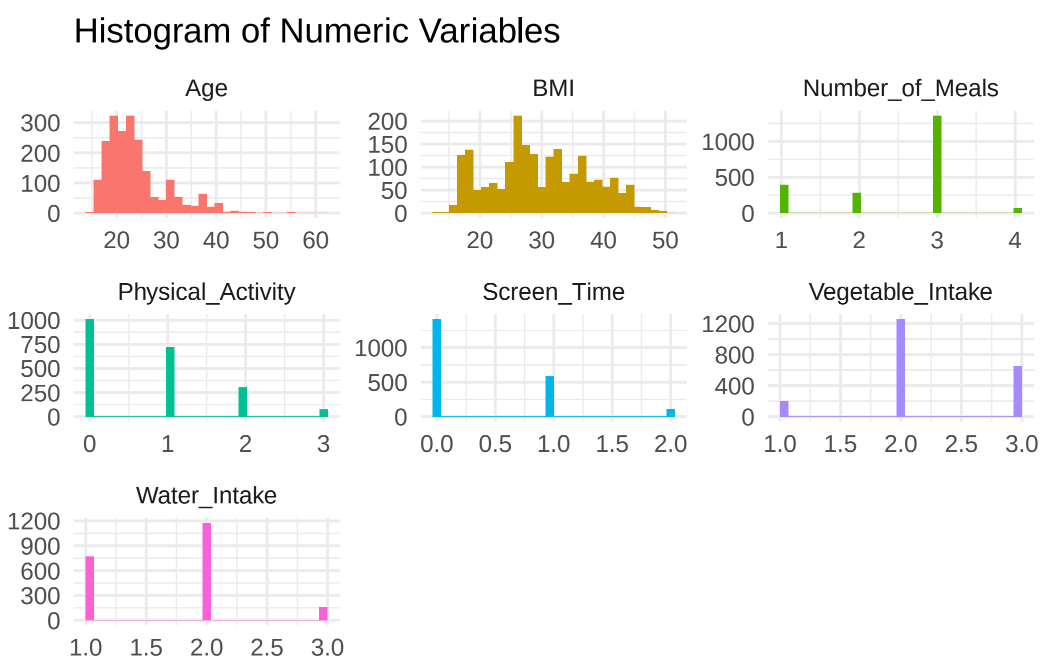
```
Water_Intake            9 2111  1.71 0.60   1  3.00  2.00  0.21    -0.60 0.01
Calorie_Monitoring*    10 2111  1.05 0.21   1  2.00  1.00  4.36    17.02 0.00
Physical_Activity      11 2111  0.73 0.83   0  3.00  3.00  0.90     0.00 0.02
Screen_Time            12 2111  0.38 0.58   0  2.00  2.00  1.25     0.55 0.01
Alcohol_Consumption*   13 2111  1.73 0.52   1  4.00  3.00 -0.24    -0.33 0.01
Transportation_Type*   14 2111  1.49 0.87   1  5.00  4.00  1.36     0.32 0.02
BMI                    15 2111 29.70 8.01  13 50.81 37.81  0.15    -0.81 0.17
```

```r
# figure B
data_dirty |>
  tidyr::pivot_longer(cols = dplyr::where(is.numeric)) |>
  dplyr::filter(!(name %in% c("Age","BMI"))) |>
  ggplot2::ggplot(ggplot2::aes(value, fill = name)) +
  ggplot2::geom_histogram() +
  ggplot2::facet_wrap(~name, scales = "free") +
  ggplot2::labs(title = "Histograms of Integer Variables (Raw)") +
  ggplot2::theme(legend.position = "none",
                 axis.title = element_blank())
```




```r
# CH2O, FAF, FCVC, NCP, TUE are discrete
# Age, BMI, Height, Weight are continuous, normal or log normal distributed
```

```r
# figure A
data |>
  tidyr::pivot_longer(cols = dplyr::where(is.numeric)) |>
  ggplot2::ggplot(ggplot2::aes(value, fill = name)) +
  ggplot2::geom_histogram() +
  ggplot2::facet_wrap(~name, scales = "free") +
  labs(title = "Histogram of Numeric Variables") +
  ggplot2::theme(legend.position = "none",
                 axis.title = element_blank())
```




```r
factors <- c("Alcohol_Consumption","Transportation_Type",
             "Calorie_Monitoring", "Snacking","Smoking",
             "Family_History","High_Caloric_Food", "Gender")

# bivariate frequency table (part 1 of Table A)
frequencies <-
  purrr::map_df(factors, \(i){
    f <- data |>
      dplyr::pull(var = i) |>
      table() |>
      t() |>
      data.frame() |>
```

```
      dplyr::mutate(
        Question = i,
        Total = sum(Freq),
        Proportion = round(Freq/sum(Freq), digits = 2)
      )
    mean <- data |>
      dplyr::summarise(
        Mean_BMI = mean(BMI),
        .by = i
      ) |>
      tidyr::pivot_longer(i, names_to = "Question", values_to = "Var2")
    dplyr::left_join(f, mean)
  }) |>
  dplyr::select(Question, Var2, Freq, Proportion, Mean_BMI)
```

Bivariate Analysis

```
# Part 2 of Table A
tests <-
  purrr::map_df(factors, \(i){
    q <- colnames(data[,i])
    bmi <- aov(
      formula = as.formula(paste("BMI ~ ", i)),
      data = data
    )
    tibble::tibble(
      Question = i,
      P_Value = c(summary(bmi)[[1]][["Pr(>F)"]][1])
    )
  })

analysis <-
  dplyr::left_join(
    x = frequencies,
    y = tests
  ) |>
  dplyr::mutate(dplyr::across(c(4:6), ~round(.x, digits = 2)))

analysis |> gt::gt()
```

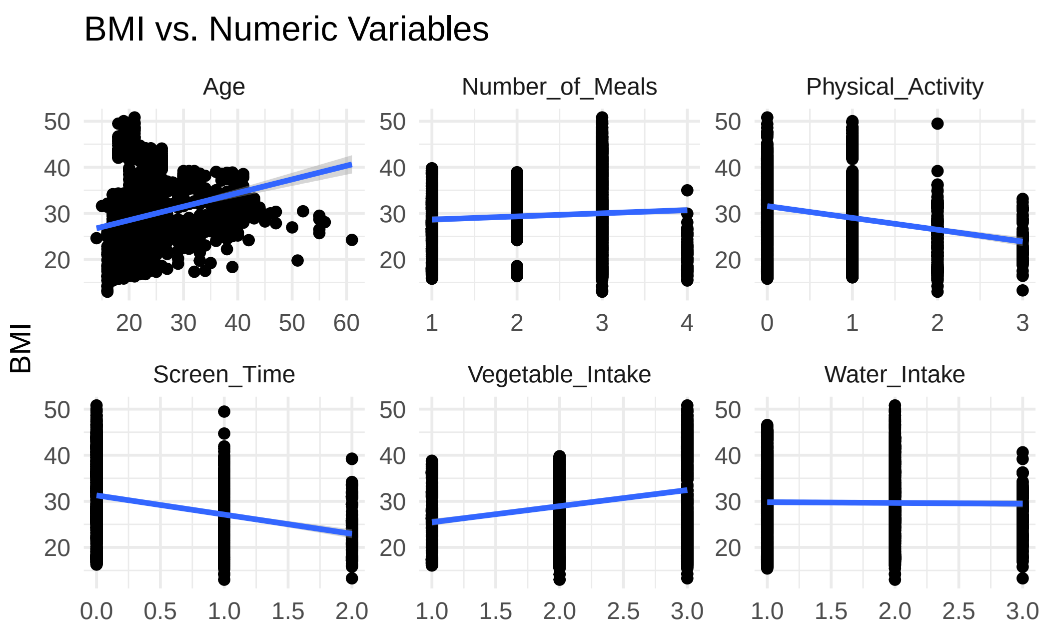
Question	Var2	Freq	Proportion	Mean_BMI	P_Value
Alcohol_Consumption	No	639	0.30	27.06	0.00
Alcohol_Consumption	Sometimes	1401	0.66	31.04	0.00
Alcohol_Consumption	Frequently	70	0.03	26.98	0.00
Alcohol_Consumption	Always	1	0.00	22.49	0.00
Transportation_Type	Public_transportation	1580	0.75	30.11	0.00
Transportation_Type	Walking	56	0.03	23.66	0.00
Transportation_Type	Automobile	457	0.22	29.19	0.00
Transportation_Type	Motorbike	11	0.01	25.76	0.00
Transportation_Type	Bike	7	0.00	25.17	0.00
Calorie_Monitoring	No	2015	0.95	30.02	0.00
Calorie_Monitoring	Yes	96	0.05	22.94	0.00
Snacking	No	51	0.02	25.43	0.00
Snacking	Sometimes	1765	0.84	31.19	0.00
Snacking	Frequently	242	0.11	20.90	0.00
Snacking	Always	53	0.03	24.32	0.00
Smoking	No	2067	0.98	29.70	0.97
Smoking	Yes	44	0.02	29.66	0.97
Family_History	No	385	0.18	21.50	0.00
Family_History	Yes	1726	0.82	31.53	0.00
High_Caloric_Food	No	245	0.12	24.26	0.00
High_Caloric_Food	Yes	1866	0.88	30.41	0.00
Gender	Female	1043	0.49	30.13	0.01
Gender	Male	1068	0.51	29.28	0.01

```r
# testing diff between bikes and motorbikes to finalize the merge
transit <- data |>
  dplyr::filter(Transportation_Type %in% c("Motorbike", "Bike"))

t.test(transit$BMI ~ transit$Transportation_Type) # 0.8402
```

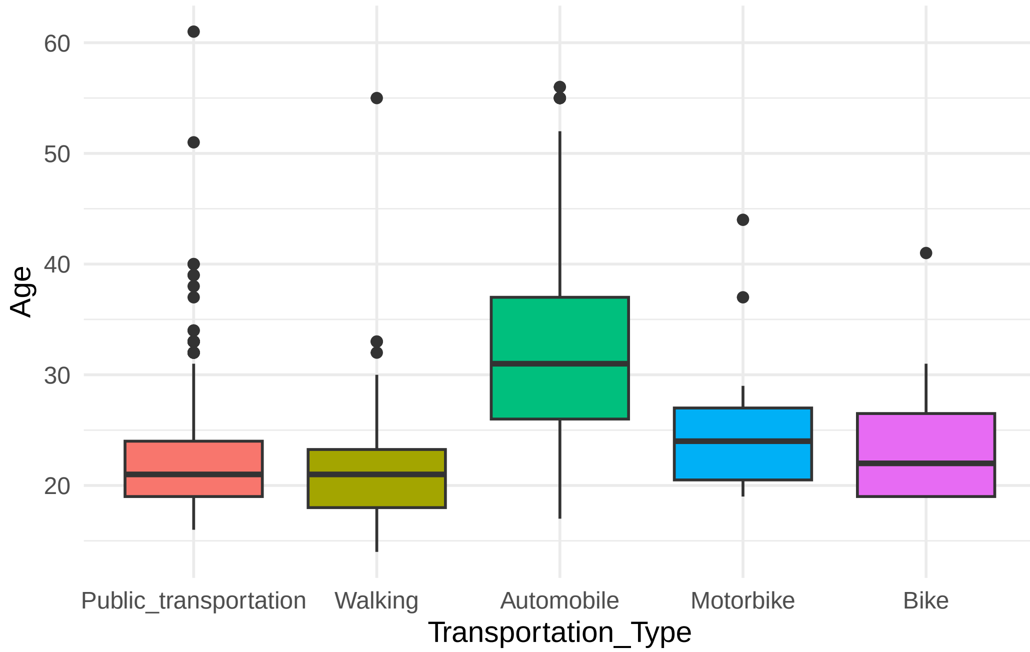
```
	Welch Two Sample t-test

data:  transit$BMI by transit$Transportation_Type
t = 0.20697, df = 10.064, p-value = 0.8402
alternative hypothesis: true difference in means between group Motorbike and group Bike is no
95 percent confidence interval:
 -5.790771  6.977841
sample estimates:
```

```
mean in group Motorbike      mean in group Bike
               25.76255                25.16902
```

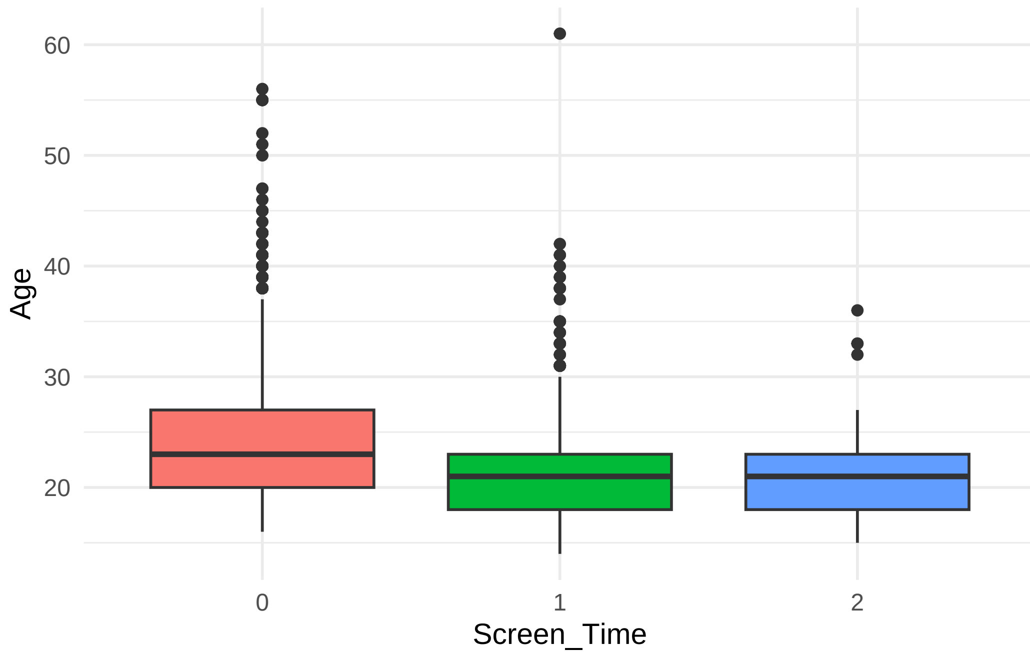
```
# figure C
data |>
  tidyr::pivot_longer(cols = c(2,5,6,9,11,12)) |>
  ggplot(aes(value, BMI)) +
  geom_point() +
  facet_wrap(~name, scales = "free") +
  geom_smooth(method = "lm") +
  labs(x = "",
       title = "BMI vs. Numeric Variables")
```




```
# looking for additional patterns
data |>
  ggplot(aes(Transportation_Type, Age, fill = Transportation_Type)) +
  geom_boxplot() +
  theme(legend.position = "none")
```




```
data |>
  ggplot(aes(factor(Screen_Time), Age, fill = factor(Screen_Time))) +
  geom_boxplot() +
  labs(x = "Screen_Time") +
  theme(legend.position = "none")
```

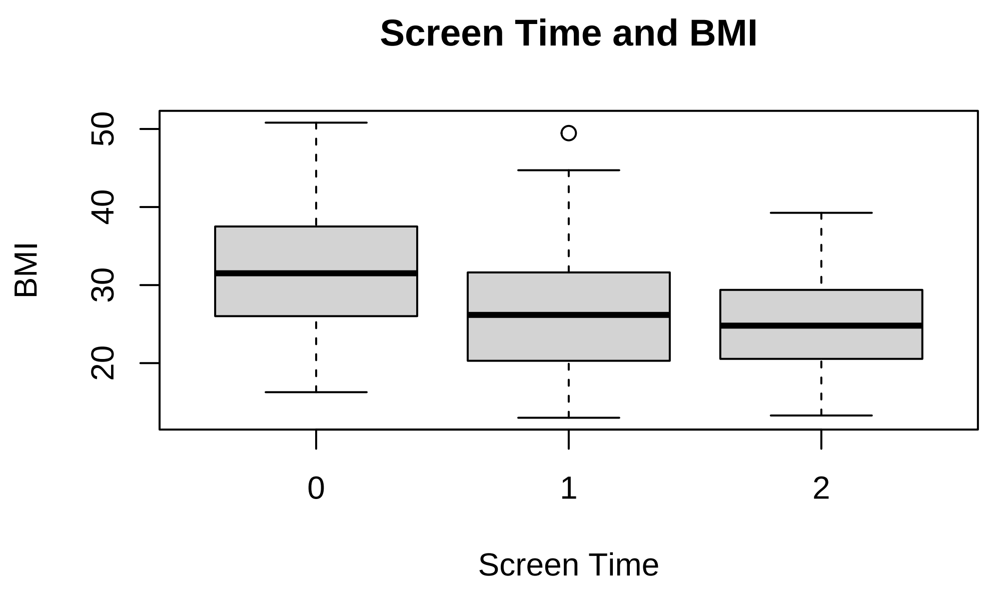



```r
cor(data |> dplyr::select(dplyr::where(is.numeric)))
```

```
                          Age Vegetable_Intake Number_of_Meals Water_Intake
Age                1.00000000      -0.01323971     -0.07063207  -0.09067179
Vegetable_Intake  -0.01323971       1.00000000      0.13851033   0.03749487
Number_of_Meals   -0.07063207       0.13851033      1.00000000   0.06743107
Water_Intake      -0.09067179       0.03749487      0.06743107   1.00000000
Physical_Activity -0.16330684       0.01934405      0.12688822   0.26609713
Screen_Time       -0.23495124      -0.15012044      0.02804751   0.09575291
BMI                0.23246073       0.26107584      0.07063317  -0.01273720
                  Physical_Activity Screen_Time         BMI
Age                     -0.16330684 -0.23495124  0.23246073
Vegetable_Intake         0.01934405 -0.15012044  0.26107584
Number_of_Meals          0.12688822  0.02804751  0.07063317
Water_Intake             0.26609713  0.09575291 -0.01273720
Physical_Activity        1.00000000  0.13437002 -0.26689086
Screen_Time              0.13437002  1.00000000 -0.30292843
BMI                     -0.26689086 -0.30292843  1.00000000
```

```r
#Screen time and BMI
raw_data <- read.csv("./data/raw_data.csv") |>
  dplyr::mutate(
```

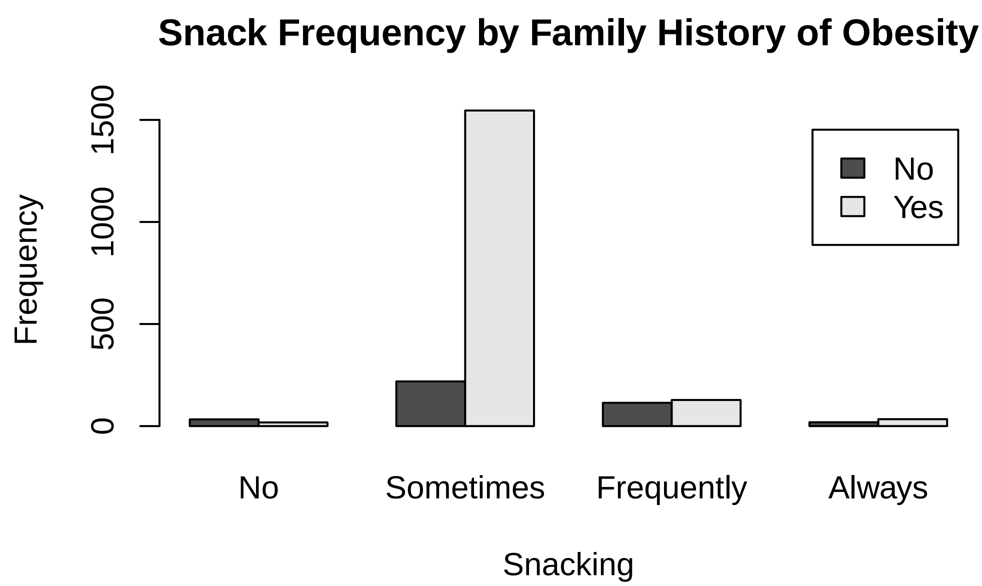
```r
    dplyr::across(
      dplyr::where(is.character),
      ~factor(stringr::str_to_title(.x))
    ),
    dplyr::across(
      .cols = c(FCVC, TUE, NCP, CH2O, FAF, Age),
      .fns = as.integer
    ),
    dplyr::across(
      .cols = c(CAEC, CALC),
      .fns = ~factor(.x, level = c("No","Sometimes","Frequently","Always"))
    ),
    MTRANS = forcats::fct_inorder(factor(MTRANS)),
    BMI = Weight/(Height^2)
  ) |>
  dplyr::select(-c(Height, Weight, NObeyesdad))

screen_bmi <- boxplot(BMI ~ TUE, data = raw_data, main = "Screen Time and BMI", xlab = "Scre
```




```r
screen_bmi_anova <- aov(BMI ~ factor(TUE), data = raw_data)
summary(screen_bmi_anova)
```

```
               Df Sum Sq Mean Sq F value Pr(>F)
factor(TUE)     2  13237    6619   114.2 <2e-16 ***
Residuals    2108 122186      58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

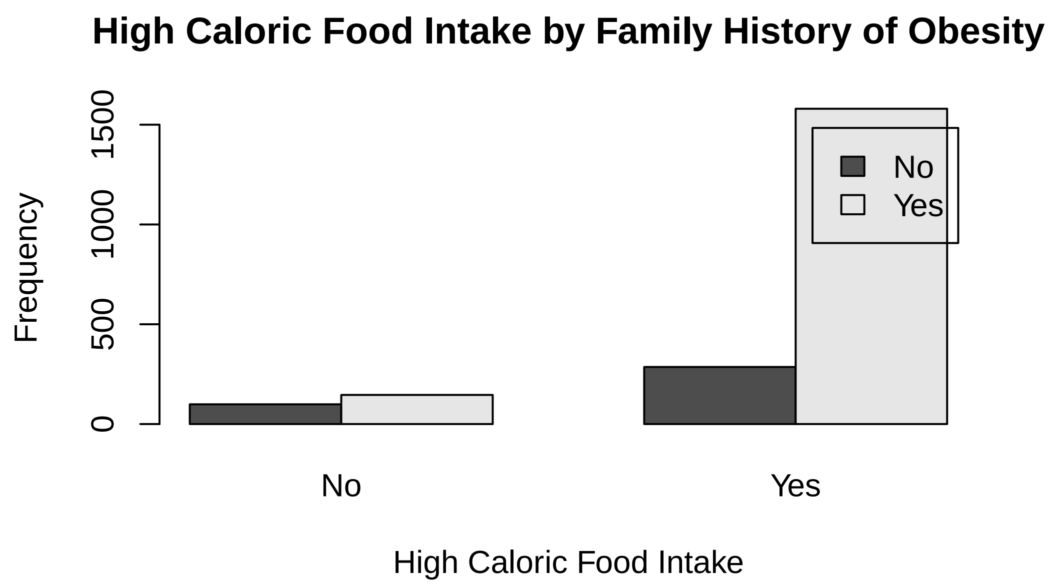
```
#Family history BMI and snacking
barplot(table(raw_data$family_history_with_overweight, raw_data$CAEC),
        beside = T,
        legend.text = T,
        xlab = "Snacking",
        ylab = "Frequency",
        main = "Snack Frequency by Family History of Obesity")
```




```
chisq.test(raw_data$family_history_with_overweight, raw_data$CAEC)
```

```

    Pearson's Chi-squared test

data:  raw_data$family_history_with_overweight and raw_data$CAEC
X-squared = 260.36, df = 3, p-value < 2.2e-16
```

```r
#Family History and High Caloric Food Intake
barplot(table(raw_data$family_history_with_overweight, raw_data$FAVC),
        beside = T,
        legend.text = T,
        xlab = "High Caloric Food Intake",
        ylab = "Frequency",
        main = "High Caloric Food Intake by Family History of Obesity")
```

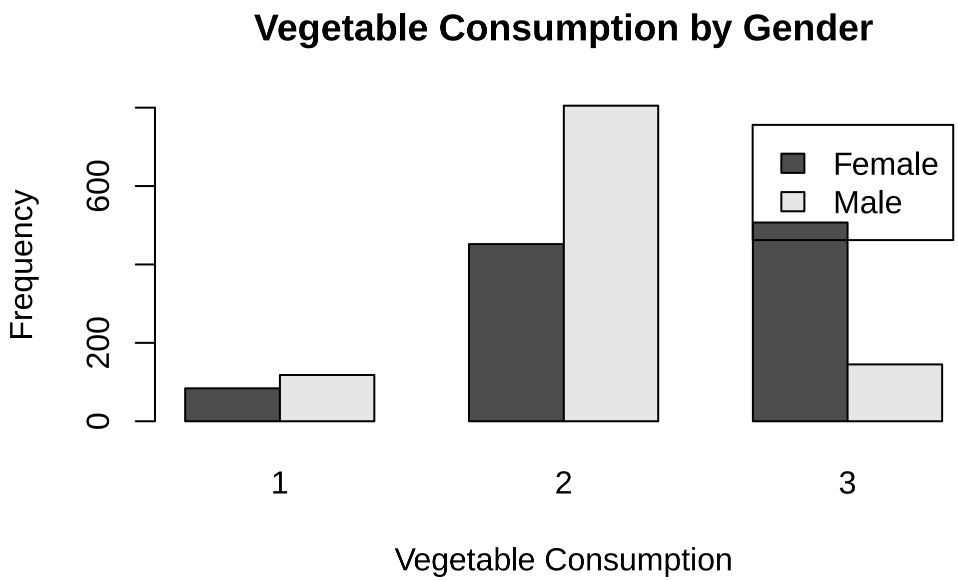



```r
chisq.test(raw_data$family_history_with_overweight, raw_data$FAVC)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  raw_data$family_history_with_overweight and raw_data$FAVC
X-squared = 89.687, df = 1, p-value < 2.2e-16
```

```r
#Vegetable consumption and gender
barplot(table(raw_data$Gender, raw_data$FCVC),
        beside = T,
        legend.text = T,
        xlab = "Vegetable Consumption",
```

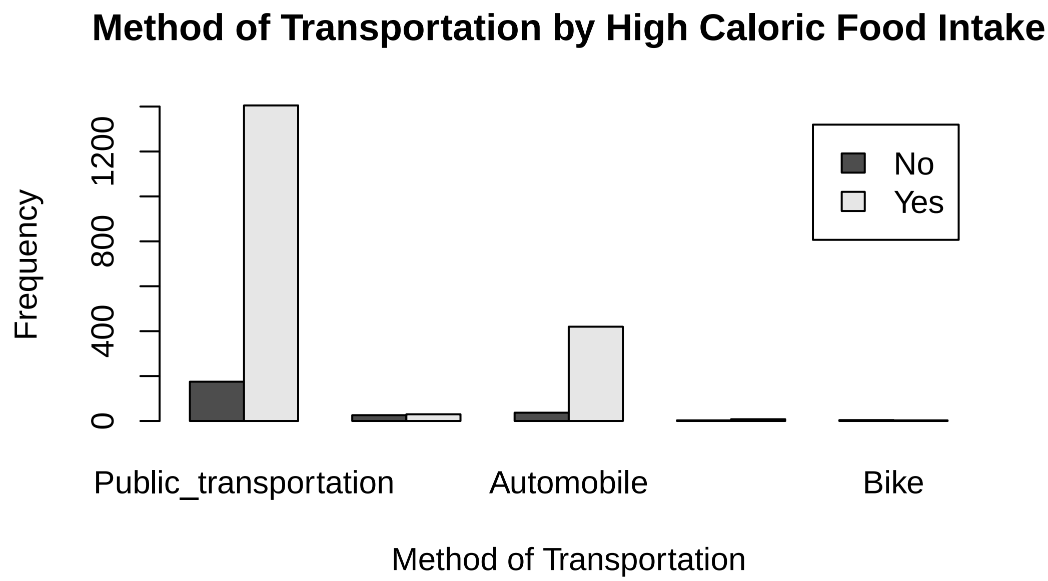
```
        ylab = "Frequency",
        main = "Vegetable Consumption by Gender")
```




```
chisq.test(raw_data$Gender, raw_data$FCVC)
```

```

    Pearson's Chi-squared test

data:  raw_data$Gender and raw_data$FCVC
X-squared = 305.59, df = 2, p-value < 2.2e-16
```

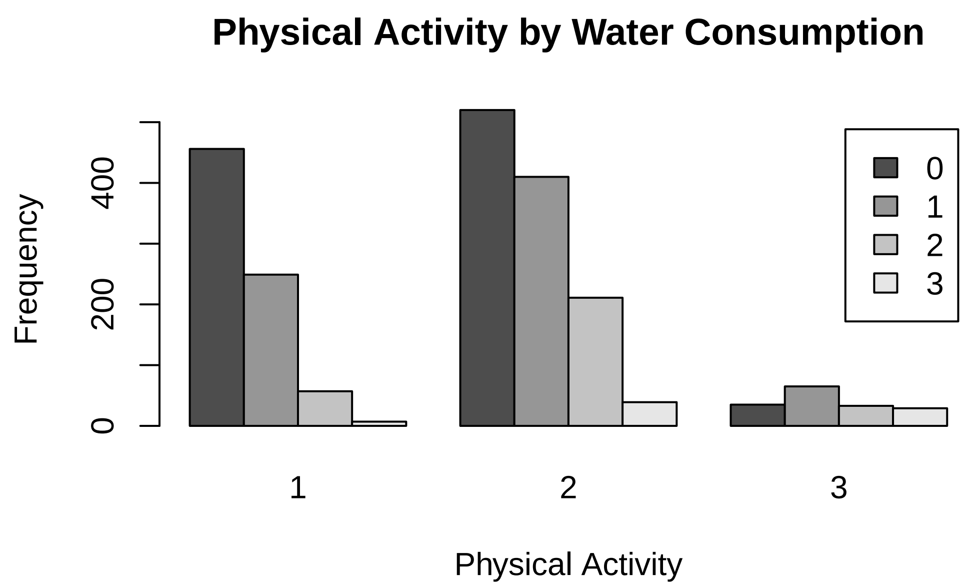
```
#Method of transportation and High Caloric Food Intake
barplot(table(raw_data$FAVC, raw_data$MTRANS),
        beside = T,
        legend.text = T,
        xlab = "Method of Transportation",
        ylab = "Frequency",
        main = "Method of Transportation by High Caloric Food Intake")
```




```
fisher.test(table(raw_data$MTRANS, raw_data$FAVC))
```

```
    Fisher's Exact Test for Count Data

data:  table(raw_data$MTRANS, raw_data$FAVC)
p-value = 6.617e-13
alternative hypothesis: two.sided
```

```
#Water intake and physical activity
barplot(table(raw_data$FAF, raw_data$CH2O),
        beside = T,
        legend.text = T,
        xlab = "Physical Activity",
        ylab = "Frequency",
        main = "Physical Activity by Water Consumption")
```




```
chisq.test(raw_data$CH2O, raw_data$FAF)
```

```

	Pearson's Chi-squared test

data:  raw_data$CH2O and raw_data$FAF
X-squared = 199.94, df = 6, p-value < 2.2e-16
```

Model Selection

Full Model

```
# doing some additional pre-processing
# using the recipes package to make additional transformations easier later
full_rec <-
  recipes::recipe(BMI ~ ., data = data) |>
  recipes::step_other(Transportation_Type, Alcohol_Consumption, threshold = 0.01) |>
  recipes::step_dummy(recipes::all_nominal_predictors())

prepped_data <- recipes::prep(full_rec) |> recipes::bake(data)
```

```r
# starting with a full model - this should be the first of the two models we'll need
full_mod <- lm(BMI ~ ., data = prepped_data)
summary(full_mod)
```

```
Call:
lm(formula = BMI ~ ., data = prepped_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4051  -3.9656   0.4073   3.5768  23.9357 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     7.51355    1.26730   5.929 3.56e-09 ***
Age                             0.27599    0.02580  10.696  < 2e-16 ***
Vegetable_Intake                3.17290    0.22137  14.333  < 2e-16 ***
Number_of_Meals                 0.61828    0.15292   4.043 5.47e-05 ***
Water_Intake                    0.34172    0.22138   1.544 0.122843    
Physical_Activity              -1.12920    0.16111  -7.009 3.23e-12 ***
Screen_Time                    -1.85734    0.22637  -8.205 3.99e-16 ***
Gender_Male                    -0.02599    0.27021  -0.096 0.923398    
Family_History_Yes              6.54154    0.35235  18.566  < 2e-16 ***
High_Caloric_Food_Yes           2.17354    0.40453   5.373 8.60e-08 ***
Snacking_Sometimes              1.43007    0.83856   1.705 0.088270 .  
Snacking_Frequently            -5.44934    0.90378  -6.030 1.94e-09 ***
Snacking_Always                -2.17257    1.12344  -1.934 0.053266 .  
Smoking_Yes                    -0.22930    0.86035  -0.267 0.789864    
Calorie_Monitoring_Yes         -1.97486    0.60526  -3.263 0.001121 ** 
Alcohol_Consumption_Sometimes   1.87430    0.28061   6.679 3.06e-11 ***
Alcohol_Consumption_Frequently  1.19977    0.71125   1.687 0.091781 .  
Alcohol_Consumption_other       5.75064    5.60115   1.027 0.304686    
Transportation_Type_Walking    -2.61205    0.79000  -3.306 0.000961 ***
Transportation_Type_Automobile -4.33177    0.38106 -11.368  < 2e-16 ***
Transportation_Type_other      -2.00068    1.33339  -1.500 0.133650    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.521 on 2090 degrees of freedom
Multiple R-squared:  0.5296,    Adjusted R-squared:  0.5251 
F-statistic: 117.7 on 20 and 2090 DF,  p-value: < 2.2e-16
```

```r
BIC(full_mod)
```

```
[1] 13351.5
```

BIC Selected Model w/ No Transformations

```r
# perform best subset selection
best_subset <- leaps::regsubsets(BMI ~ ., data = prepped_data, nvmax = 20, method = "exhaust
results <- summary(best_subset)

# extract results
n <- nrow(prepped_data)
p <- 20
results_df <-
  tibble::tibble(
    predictors = 1:p,
    adj_R2 = results$adjr2,
    bic = results$bic,
    aic = n*log(results$rss/n) + (1:p)*2
  )

# training the bic selected model
form <- paste("BMI~", paste(names(which(results$which[which.min(results_df$bic),-1])), colla
  as.formula()

bic_mod <- lm(form, data = prepped_data)
summary(bic_mod)
```

```
Call:
lm(formula = form, data = prepped_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.3613  -3.9867   0.3698   3.6075  23.8470 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     9.18157    0.94368   9.730  < 2e-16 ***
```

```
Age                              0.27438    0.02538  10.809  < 2e-16 ***
Vegetable_Intake                 3.23026    0.20950  15.419  < 2e-16 ***
Number_of_Meals                  0.63847    0.15154   4.213 2.62e-05 ***
Physical_Activity               -1.08922    0.15455  -7.048 2.46e-12 ***
Screen_Time                     -1.76232    0.22342  -7.888 4.90e-15 ***
Family_History_Yes               6.71901    0.34119  19.693  < 2e-16 ***
High_Caloric_Food_Yes            2.19047    0.40274   5.439 5.98e-08 ***
Snacking_Frequently             -6.82686    0.40615 -16.809  < 2e-16 ***
Snacking_Always                 -3.46848    0.78521  -4.417 1.05e-05 ***
Calorie_Monitoring_Yes          -1.96777    0.60065  -3.276  0.00107 **
Alcohol_Consumption_Sometimes    1.75692    0.27269   6.443 1.45e-10 ***
Transportation_Type_Walking     -2.46590    0.78160  -3.155  0.00163 **
Transportation_Type_Automobile -4.24159    0.37681 -11.257  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.526 on 2097 degrees of freedom
Multiple R-squared:  0.5272,    Adjusted R-squared:  0.5243
F-statistic: 179.9 on 13 and 2097 DF,  p-value: < 2.2e-16
```

BIC Selected Model w/ Log Transformation

```r
full_mod2 <- lm(log(BMI) ~ ., data = prepped_data)
summary(full_mod2)
```

```
Call:
lm(formula = log(BMI) ~ ., data = prepped_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72004 -0.12574  0.02526  0.12871  0.73143

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.6013988  0.0443136  58.704  < 2e-16 ***
Age                            0.0114191  0.0009022  12.656  < 2e-16 ***
Vegetable_Intake               0.0940204  0.0077406  12.146  < 2e-16 ***
Number_of_Meals                0.0113759  0.0053471   2.127  0.03350 *
Water_Intake                   0.0160370  0.0077411   2.072  0.03842 *
```

```
Physical_Activity             -0.0411310  0.0056336  -7.301 4.04e-13 ***
Screen_Time                   -0.0627265  0.0079155  -7.925 3.69e-15 ***
Gender_Male                    0.0152891  0.0094485   1.618  0.10578    
Family_History_Yes             0.2442277  0.0123205  19.823  < 2e-16 ***
High_Caloric_Food_Yes          0.0636985  0.0141450   4.503 7.06e-06 ***
Snacking_Sometimes             0.0352580  0.0293219   1.202  0.22933    
Snacking_Frequently           -0.2218182  0.0316022  -7.019 3.01e-12 ***
Snacking_Always               -0.0843369  0.0392831  -2.147  0.03192 *  
Smoking_Yes                   -0.0054726  0.0300836  -0.182  0.85567    
Calorie_Monitoring_Yes        -0.0543980  0.0211641  -2.570  0.01023 *  
Alcohol_Consumption_Sometimes  0.0583724  0.0098119   5.949 3.15e-09 ***
Alcohol_Consumption_Frequently  0.0579349  0.0248701   2.330  0.01993 *  
Alcohol_Consumption_other      0.1866724  0.1958548   0.953  0.34064    
Transportation_Type_Walking   -0.0810451  0.0276240  -2.934  0.00338 ** 
Transportation_Type_Automobile -0.1552961  0.0133245 -11.655  < 2e-16 ***
Transportation_Type_other     -0.0622513  0.0466244  -1.335  0.18197    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.193 on 2090 degrees of freedom
Multiple R-squared:  0.5361,    Adjusted R-squared:  0.5317 
F-statistic: 120.8 on 20 and 2090 DF,  p-value: < 2.2e-16
```

```r
# perform best subset selection
best_subset2 <- leaps::regsubsets(log(BMI) ~ ., data = prepped_data, nvmax = 20, method = "e
results2 <- summary(best_subset2)

# extract and plot results
n <- nrow(prepped_data)
p <- 20
results_df2 <-
  tibble::tibble(
    predictors = 1:p,
    adj_R2 = results2$adjr2,
    bic = results2$bic,
    aic = n*log(results2$rss/n) + (1:p)*2
  )

# training the bic selected model
form <- paste("log(BMI)~", paste(names(which(results2$which[which.min(results_df2$bic),-1]))
  as.formula()
```

```r
bic_mod2 <- lm(form, data = prepped_data)
summary(bic_mod2)
```

```

Call:
lm(formula = form, data = prepped_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7281 -0.1285  0.0265  0.1259  0.7457 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     2.6788647  0.0312415  85.747  < 2e-16 ***
Age                             0.0112776  0.0008849  12.744  < 2e-16 ***
Vegetable_Intake                0.0916109  0.0072785  12.586  < 2e-16 ***
Physical_Activity              -0.0374721  0.0053432  -7.013 3.13e-12 ***
Screen_Time                    -0.0605584  0.0077747  -7.789 1.05e-14 ***
Family_History_Yes              0.2587828  0.0118366  21.863  < 2e-16 ***
High_Caloric_Food_Yes           0.0741247  0.0138873   5.338 1.04e-07 ***
Snacking_Frequently            -0.2561759  0.0142068 -18.032  < 2e-16 ***
Snacking_Always                -0.1170597  0.0274689  -4.262 2.12e-05 ***
Alcohol_Consumption_Sometimes   0.0577526  0.0094327   6.123 1.10e-09 ***
Transportation_Type_Automobile -0.1464172  0.0131432 -11.140  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1943 on 2100 degrees of freedom
Multiple R-squared:  0.5279,	Adjusted R-squared:  0.5257 
F-statistic: 234.9 on 10 and 2100 DF,  p-value: < 2.2e-16
```

BIC Selected Model w/ BoxCox Transformation

```r
# trying a log transformation of BMI ----------
# trying to lessen the heteroskedacicity
recipe3 <- full_rec |>
  recipes::step_BoxCox(BMI)

prepped_data3 <- recipes::prep(recipe3) |> recipes::bake(data)
```

```r
full_mod3 <- lm(BMI ~ ., data = prepped_data3)
summary(full_mod3)
```

```
Call:
lm(formula = BMI ~ ., data = prepped_data3)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6627 -0.9331  0.1296  0.8675  5.4592 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     5.162573   0.305600  16.893  < 2e-16 ***
Age                             0.072072   0.006222  11.583  < 2e-16 ***
Vegetable_Intake                0.717671   0.053381  13.444  < 2e-16 ***
Number_of_Meals                 0.120020   0.036876   3.255  0.00115 ** 
Water_Intake                    0.093977   0.053385   1.760  0.07849 .  
Physical_Activity              -0.278710   0.038851  -7.174 1.01e-12 ***
Screen_Time                    -0.442835   0.054587  -8.112 8.37e-16 ***
Gender_Male                     0.042556   0.065160   0.653  0.51376    
Family_History_Yes              1.627848   0.084966  19.159  < 2e-16 ***
High_Caloric_Food_Yes           0.491640   0.097548   5.040 5.06e-07 ***
Snacking_Sometimes              0.307650   0.202212   1.521  0.12831    
Snacking_Frequently            -1.405860   0.217938  -6.451 1.38e-10 ***
Snacking_Always                -0.548227   0.270908  -2.024  0.04313 *  
Smoking_Yes                    -0.048072   0.207466  -0.232  0.81679    
Calorie_Monitoring_Yes         -0.436824   0.145954  -2.993  0.00280 ** 
Alcohol_Consumption_Sometimes   0.432199   0.067666   6.387 2.08e-10 ***
Alcohol_Consumption_Frequently  0.333487   0.171512   1.944  0.05198 .  
Alcohol_Consumption_other       1.341951   1.350673   0.994  0.32056    
Transportation_Type_Walking    -0.604704   0.190503  -3.174  0.00152 ** 
Transportation_Type_Automobile -1.060642   0.091889 -11.543  < 2e-16 ***
Transportation_Type_other      -0.465696   0.321536  -1.448  0.14767    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.331 on 2090 degrees of freedom
Multiple R-squared:  0.5335,	Adjusted R-squared:  0.529 
F-statistic: 119.5 on 20 and 2090 DF,  p-value: < 2.2e-16
```

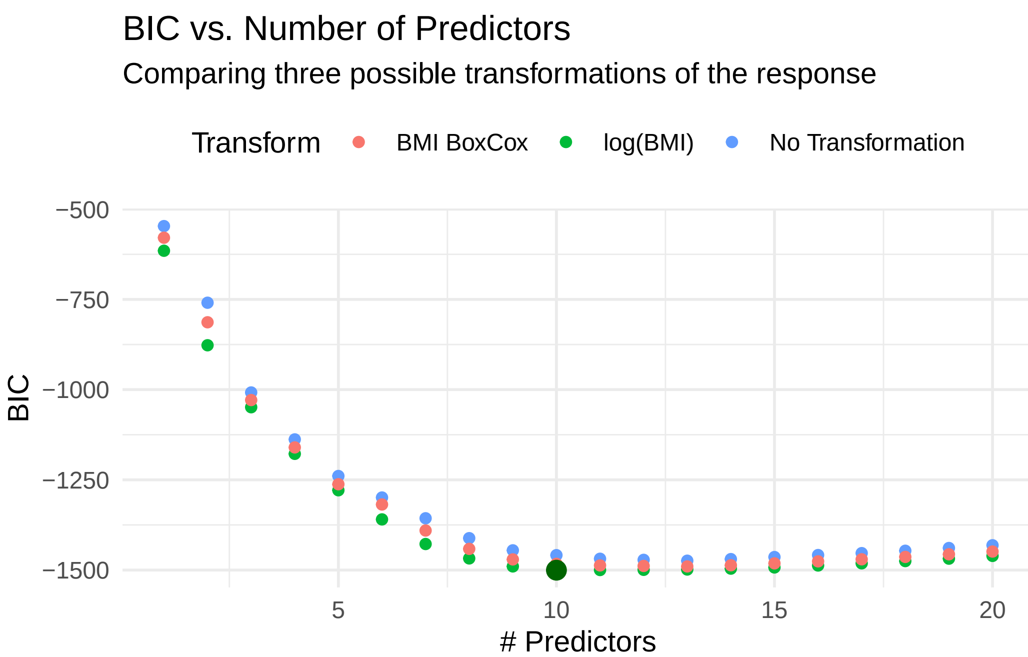
```r
# perform best subset selection
best_subset3 <- leaps::regsubsets(BMI ~ ., data = prepped_data3, nvmax = 20, method = "exhaus
results3 <- summary(best_subset3)

# extract and plot results
n <- nrow(prepped_data3)
p <- 20
results_df3 <-
  tibble::tibble(
    predictors = 1:p,
    adj_R2 = results3$adjr3,
    bic = results3$bic,
    aic = n*log(results3$rss/n) + (1:p)*2
  )

# training the bic selected model
form <- paste("BMI~", paste(names(which(results3$which[which.min(results_df3$bic),-1])), col
  as.formula()

bic_mod3 <- lm(form, data = prepped_data3)
summary(bic_mod3)
```

```
Call:
lm(formula = form, data = prepped_data3)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6119 -0.9423  0.1307  0.8767  5.4691 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     5.587409   0.227647  24.544  < 2e-16 ***
Age                             0.071812   0.006124  11.727  < 2e-16 ***
Vegetable_Intake                0.719174   0.050539  14.230  < 2e-16 ***
Number_of_Meals                 0.125631   0.036556   3.437  0.00060 ***
Physical_Activity              -0.260430   0.037283  -6.985 3.80e-12 ***
Screen_Time                    -0.418009   0.053896  -7.756 1.36e-14 ***
Family_History_Yes              1.674634   0.082307  20.346  < 2e-16 ***
High_Caloric_Food_Yes           0.496403   0.097154   5.109 3.52e-07 ***
Snacking_Frequently            -1.709179   0.097977 -17.445  < 2e-16 ***
Snacking_Always                -0.815858   0.189420  -4.307 1.73e-05 ***
```

```
Calorie_Monitoring_Yes         -0.437928   0.144899  -3.022  0.00254 **
Alcohol_Consumption_Sometimes   0.401656   0.065781   6.106 1.21e-09 ***
Transportation_Type_Walking    -0.562542   0.188548  -2.984  0.00288 **
Transportation_Type_Automobile -1.034504   0.090899 -11.381  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.333 on 2097 degrees of freedom
Multiple R-squared:  0.5307,    Adjusted R-squared:  0.5278
F-statistic: 182.4 on 13 and 2097 DF,  p-value: < 2.2e-16
```

Model Comparison

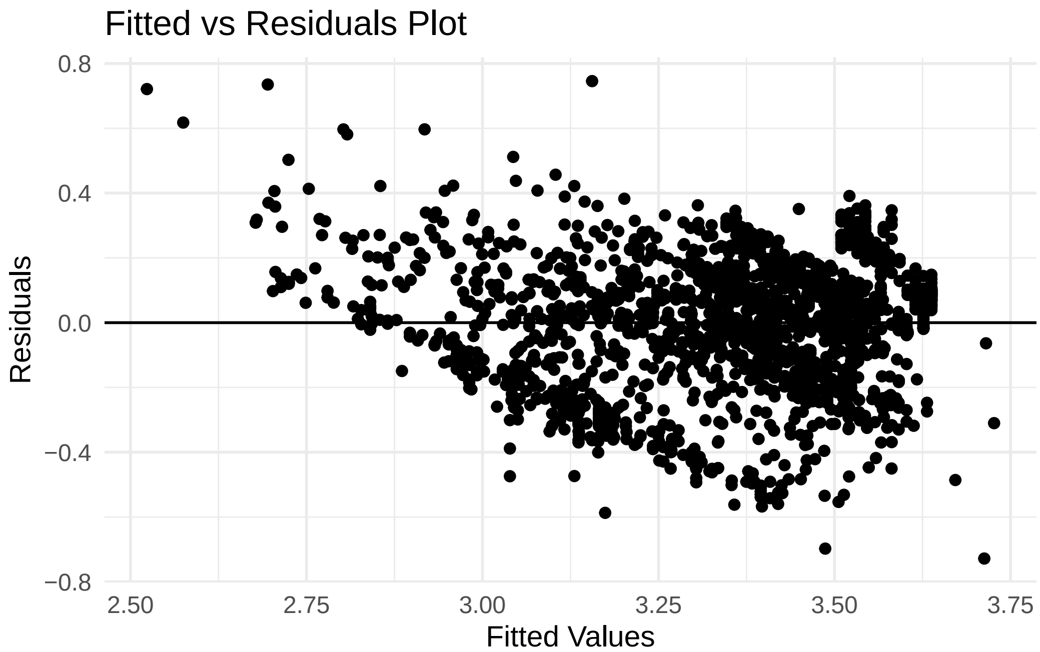
```r
# Figure 1
dplyr::bind_rows(results_df, results_df2, results_df3, .id = "Transform") |>
  dplyr::mutate(Transform =
                  dplyr::case_when(
                    Transform == 1 ~ "No Transformation",
                    Transform == 2 ~ "log(BMI)",
                    Transform == 3 ~ "BMI BoxCox",
                    FALSE ~ NA
                  )
                  ) |>
  ggplot(aes(predictors, bic, color = Transform)) +
  geom_point() +
  geom_point(data = results_df2[which.min(results_df2$bic), ], color="darkgreen", size = 3) +
  labs(title = "BIC vs. Number of Predictors",
       subtitle = "Comparing three possible transformations of the response",
       x = "# Predictors", y = "BIC") +
  theme(legend.position = "top")
```



Model Diagnostics

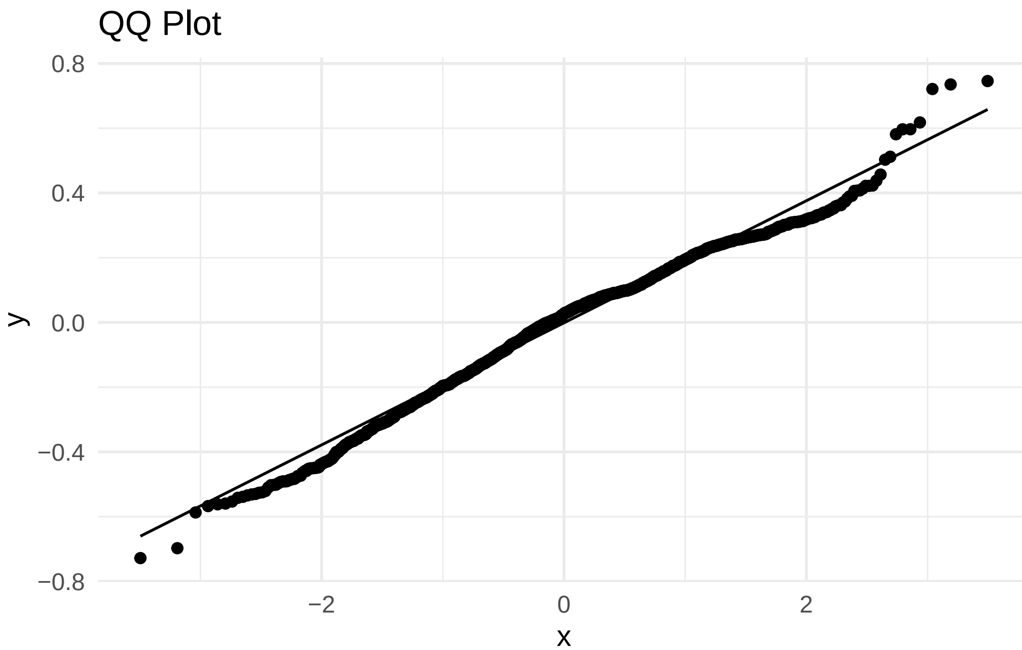
Equal Variance of Errors

```
data_diagnostic <- broom::augment(bic_mod2)

ggplot(aes(x = .fitted, y = .resid), data = data_diagnostic) +
  geom_point() + geom_hline(yintercept = 0) +
  ggtitle("Fitted vs Residuals Plot") +
  labs(x = "Fitted Values", y = "Residuals")
```



Normal Residuals

```
#normality
ggplot(aes(sample = .resid), data = data_diagnostic) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("QQ Plot")
```




```
shapiro.test(data_diagnostic$.resid)
```

```
	Shapiro-Wilk normality test

data:  data_diagnostic$.resid
W = 0.98647, p-value = 3.018e-13
```

Independent Residuals

```
library(gghalfnorm)
library(faraway)
x <- model.matrix(bic_mod2)[,-1]
# looking at vif
faraway::vif(x) |>
  round(digits=2) |>
  sort(decreasing=TRUE) |>
  data.frame() |>
  dplyr::rename(VIF=1) |>
  tibble::rownames_to_column(var="Variable")
```

```
                         Variable  VIF
1                             Age 1.74
2  Transportation_Type_Automobile 1.64
3              Family_History_Yes 1.17
4                     Screen_Time 1.15
5             Snacking_Frequently 1.15
6               Physical_Activity 1.11
7             High_Caloric_Food_Yes 1.11
8   Alcohol_Consumption_Sometimes 1.11
9                Vegetable_Intake 1.06
10                Snacking_Always 1.03
```

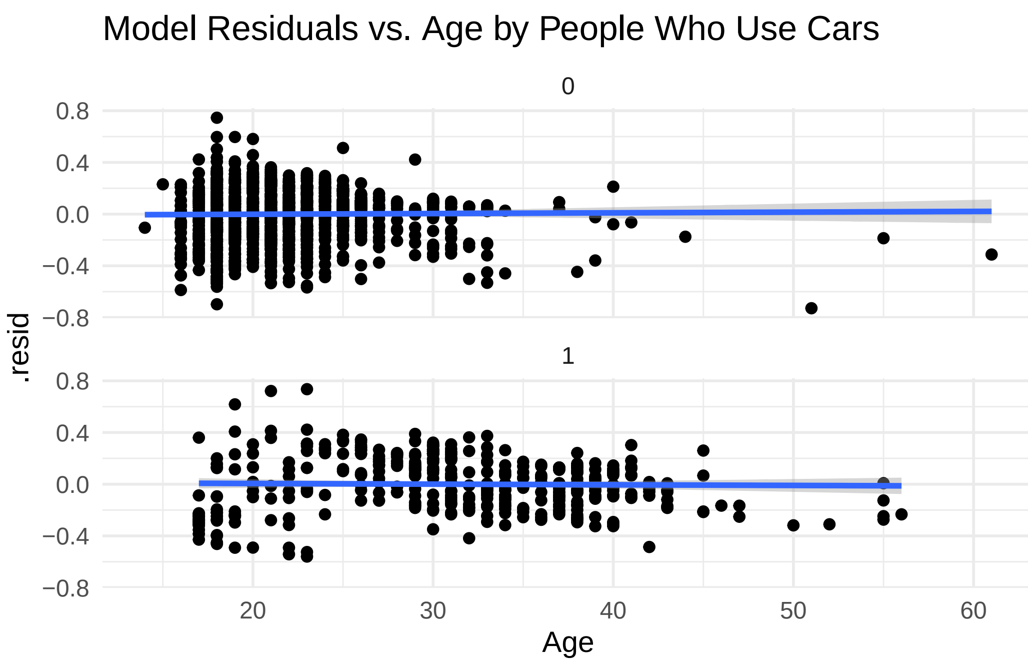
```
# nothing especially concerning

# looking at pairwise correlations of predictors
cor(x)
```

```
                                      Age Vegetable_Intake Physical_Activity
Age                            1.00000000     -0.013239705      -0.163306843
Vegetable_Intake              -0.01323971      1.000000000       0.019344048
Physical_Activity             -0.16330684      0.019344048       1.000000000
Screen_Time                   -0.23495124     -0.150120443       0.134370020
Family_History_Yes             0.19555239      0.008331892      -0.128375257
High_Caloric_Food_Yes          0.05587190     -0.073481896      -0.156302053
Snacking_Frequently           -0.11442188      0.077944676       0.086072453
Snacking_Always               -0.02282438      0.038916459       0.076585622
Alcohol_Consumption_Sometimes -0.01772641      0.087590575      -0.158171068
Transportation_Type_Automobile 0.60427406     -0.098691575       0.004464302
                              Screen_Time Family_History_Yes
Age                           -0.23495124        0.195552391
Vegetable_Intake              -0.15012044        0.008331892
Physical_Activity              0.13437002       -0.128375257
Screen_Time                    1.00000000       -0.097282976
Family_History_Yes            -0.09728298        1.000000000
High_Caloric_Food_Yes         -0.05478303        0.208035507
Snacking_Frequently            0.10650696       -0.269018294
Snacking_Always                0.09768498       -0.073188529
Alcohol_Consumption_Sometimes -0.18464410       -0.024636667
Transportation_Type_Automobile -0.11903070       0.099326516
                              High_Caloric_Food_Yes Snacking_Frequently
Age                                      0.05587190         -0.11442188
Vegetable_Intake                        -0.07348190          0.07794468
```

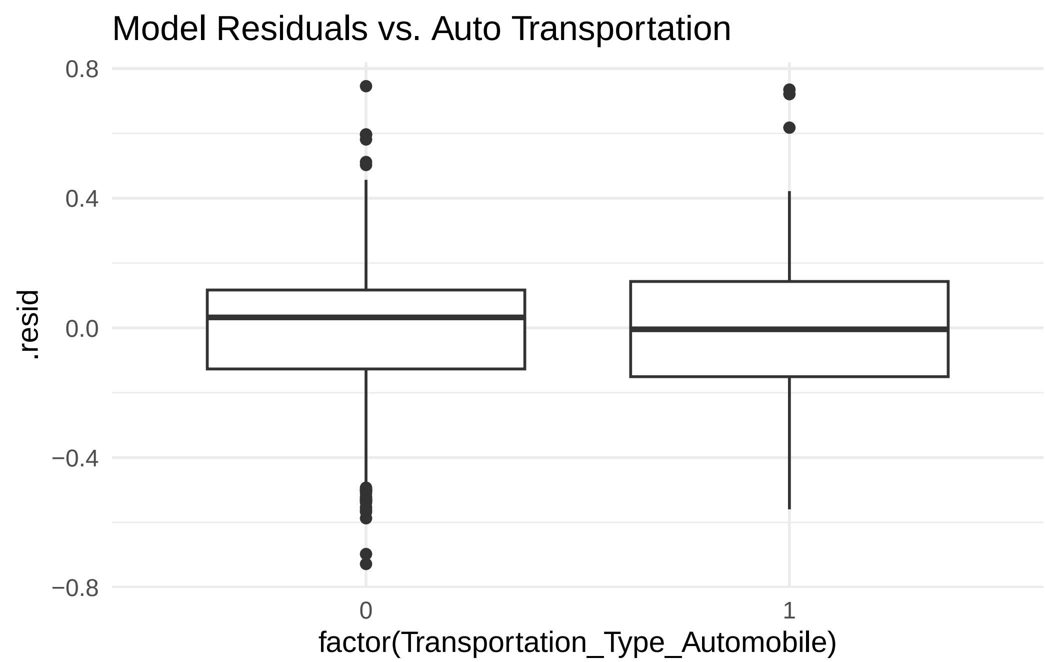
```
Physical_Activity                          -0.15630205          0.08607245
Screen_Time                                -0.05478303          0.10650696
Family_History_Yes                          0.20803551         -0.26901829
High_Caloric_Food_Yes                       1.00000000         -0.18065106
Snacking_Frequently                        -0.18065106          1.00000000
Snacking_Always                            -0.05529196         -0.05774552
Alcohol_Consumption_Sometimes               0.13961092         -0.12780025
Transportation_Type_Automobile              0.05759657         -0.09888545
                               Snacking_Always Alcohol_Consumption_Sometimes
Age                               -0.022824376                   -0.01772641
Vegetable_Intake                   0.038916459                    0.08759058
Physical_Activity                  0.076585622                   -0.15817107
Screen_Time                        0.097684980                   -0.18464410
Family_History_Yes                -0.073188529                   -0.02463667
High_Caloric_Food_Yes             -0.055291958                    0.13961092
Snacking_Frequently               -0.057745518                   -0.12780025
Snacking_Always                    1.000000000                   -0.04597909
Alcohol_Consumption_Sometimes     -0.045979095                    1.00000000
Transportation_Type_Automobile     0.003869257                   -0.07862441
                               Transportation_Type_Automobile
Age                                               0.604274062
Vegetable_Intake                                 -0.098691575
Physical_Activity                                 0.004464302
Screen_Time                                      -0.119030701
Family_History_Yes                                0.099326516
High_Caloric_Food_Yes                             0.057596565
Snacking_Frequently                              -0.098885449
Snacking_Always                                   0.003869257
Alcohol_Consumption_Sometimes                    -0.078624414
Transportation_Type_Automobile                    1.000000000
```

```r
# the corrplot indicates some potential collinearity with age and transportaiton
data_diagnostic |>
  ggplot(aes(Age, .resid)) +
  geom_point() +
  facet_wrap(~factor(Transportation_Type_Automobile), ncol = 1) +
  geom_smooth(method = "lm") +
  labs(title = "Model Residuals vs. Age by People Who Use Cars")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

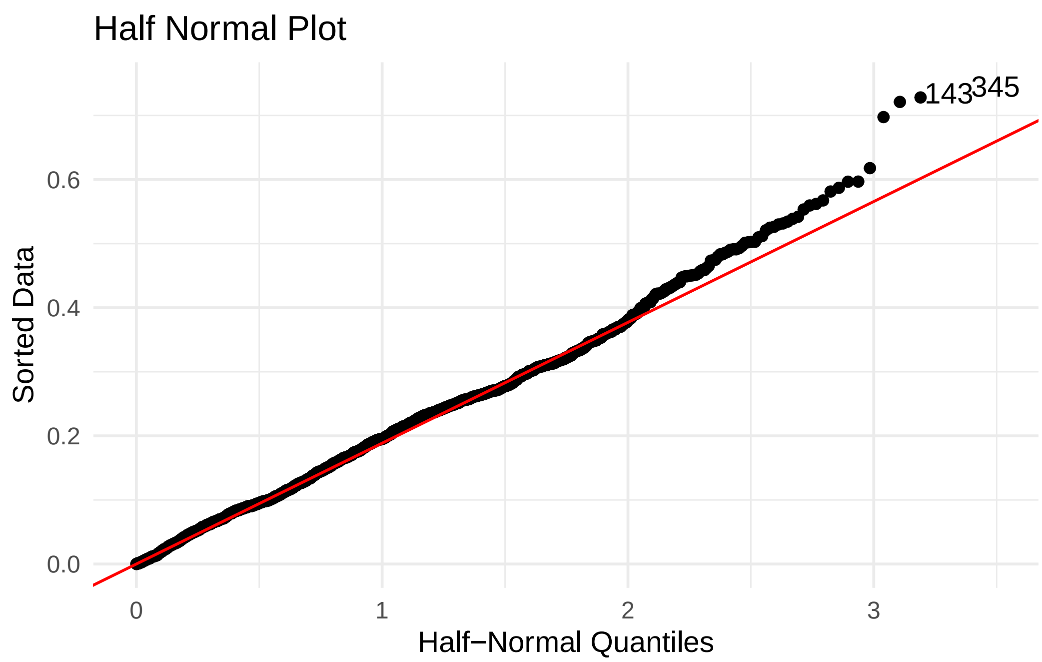



```r
data_diagnostic |>
  ggplot(aes(factor(Transportation_Type_Automobile), .resid)) +
  geom_boxplot() +
  labs(title = "Model Residuals vs. Auto Transportation")
```



Outliers

```
x <- data_diagnostic$.resid
gghalfnorm(x, nlab = 2, labs = as.character(seq_along(x)), repel = FALSE) +
  ggtitle("Half Normal Plot")
```




```r
# excluding top 3 points
exclude <- prepped_data[-c(345,143),]

exc_mod <- lm(log(BMI) ~ Age + Vegetable_Intake + Physical_Activity + Screen_Time +
                Family_History_Yes + High_Caloric_Food_Yes + Snacking_Frequently +
                Snacking_Always + Alcohol_Consumption_Sometimes + Transportation_Type_Automol
              data = exclude)

summary(exc_mod)
```

```
Call:
lm(formula = log(BMI) ~ Age + Vegetable_Intake + Physical_Activity +
    Screen_Time + Family_History_Yes + High_Caloric_Food_Yes +
    Snacking_Frequently + Snacking_Always + Alcohol_Consumption_Sometimes +
    Transportation_Type_Automobile, data = exclude)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73313 -0.12844  0.02468  0.12737  0.72962

Coefficients:
```

```
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    2.6811269  0.0310602  86.320  < 2e-16 ***
Age                            0.0113883  0.0008794  12.950  < 2e-16 ***
Vegetable_Intake               0.0909079  0.0072329  12.569  < 2e-16 ***
Physical_Activity             -0.0379583  0.0053111  -7.147 1.22e-12 ***
Screen_Time                   -0.0615743  0.0077262  -7.970 2.59e-15 ***
Family_History_Yes             0.2589521  0.0117722  21.997  < 2e-16 ***
High_Caloric_Food_Yes          0.0719168  0.0138024   5.210 2.07e-07 ***
Snacking_Frequently           -0.2624597  0.0141611 -18.534  < 2e-16 ***
Snacking_Always               -0.1165236  0.0272892  -4.270 2.04e-05 ***
Alcohol_Consumption_Sometimes  0.0574818  0.0093769   6.130 1.05e-09 ***
Transportation_Type_Automobile -0.1493053  0.0130756 -11.419  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.193 on 2098 degrees of freedom
Multiple R-squared:  0.5337,	Adjusted R-squared:  0.5315 
F-statistic: 240.1 on 10 and 2098 DF,  p-value: < 2.2e-16
```

```r
summary(bic_mod2)
```

```

Call:
lm(formula = form, data = prepped_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7281 -0.1285  0.0265  0.1259  0.7457 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    2.6788647  0.0312415  85.747  < 2e-16 ***
Age                            0.0112776  0.0008849  12.744  < 2e-16 ***
Vegetable_Intake               0.0916109  0.0072785  12.586  < 2e-16 ***
Physical_Activity             -0.0374721  0.0053432  -7.013 3.13e-12 ***
Screen_Time                   -0.0605584  0.0077747  -7.789 1.05e-14 ***
Family_History_Yes             0.2587828  0.0118366  21.863  < 2e-16 ***
High_Caloric_Food_Yes          0.0741247  0.0138873   5.338 1.04e-07 ***
Snacking_Frequently           -0.2561759  0.0142068 -18.032  < 2e-16 ***
Snacking_Always               -0.1170597  0.0274689  -4.262 2.12e-05 ***
Alcohol_Consumption_Sometimes  0.0577526  0.0094327   6.123 1.10e-09 ***
Transportation_Type_Automobile -0.1464172  0.0131432 -11.140  < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1943 on 2100 degrees of freedom
Multiple R-squared:  0.5279,	Adjusted R-squared:  0.5257 
F-statistic: 234.9 on 10 and 2100 DF,  p-value: < 2.2e-16
```

```
# no substantial changes between the models
```

Predictions

```
# getting median values for all predictors
x <- model.matrix(bic_mod2) |>
  as.data.frame() |>
  dplyr::summarise(dplyr::across(dplyr::everything(), median)) |>
  dplyr::select(-1)

predict(bic_mod2, x) |> exp()
```

```
       1 
31.92477 
```