Generational Obesity Analysis - Final Model

Ken Jones

April 4, 2024

Data Preparation

```
# Read in the dataset
data <- read_csv("ObesityDataSet.csv")
```

```
## Rows: 2111 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): Gender, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CAL...
## dbl (8): Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Preliminary Checks

```
# Perform preliminary checks
head(data)
```

```
## # A tibble: 6 x 17
##   Gender   Age Height Weight family_history_with_overw~1 FAVC   FCVC   NCP CAEC
##   <chr>  <dbl>  <dbl>  <dbl> <chr>                       <chr> <dbl> <dbl> <chr>
## 1 Female    21   1.62   64   yes                         no        2     3 Some~
## 2 Female    21   1.52   56   yes                         no        3     3 Some~
## 3 Male      23   1.8    77   yes                         no        2     3 Some~
## 4 Male      27   1.8    87   no                          no        3     3 Some~
## 5 Male      22   1.78   89.8 no                          no        2     1 Some~
## 6 Male      29   1.62   53   no                          yes       2     3 Some~
## # i abbreviated name: 1: family_history_with_overweight
## # i 8 more variables: SMOKE <chr>, CH2O <dbl>, SCC <chr>, FAF <dbl>, TUE <dbl>,
## #   CALC <chr>, MTRANS <chr>, NObeyesdad <chr>
```

```
summary(data)
```

```
##     Gender               Age            Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00
##  Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47
```

```
##  Mode  :character   Median :22.78   Median :1.700   Median : 83.00
##                     Mean   :24.31   Mean   :1.702   Mean   : 86.59
##                     3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
##                     Max.   :61.00   Max.   :1.980   Max.   :173.00
##  family_history_with_overweight     FAVC               FCVC
##  Length:2111                    Length:2111        Min.   :1.000
##  Class :character               Class :character   1st Qu.:2.000
##  Mode  :character               Mode  :character   Median :2.386
##                                                    Mean   :2.419
##                                                    3rd Qu.:3.000
##                                                    Max.   :3.000
##       NCP            CAEC              SMOKE               CH2O
##  Min.   :1.000   Length:2111        Length:2111        Min.   :1.000
##  1st Qu.:2.659   Class :character   Class :character   1st Qu.:1.585
##  Median :3.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   :2.686                                         Mean   :2.008
##  3rd Qu.:3.000                                         3rd Qu.:2.477
##  Max.   :4.000                                         Max.   :3.000
##      SCC                 FAF              TUE             CALC
##  Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111
##  Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character
##  Mode  :character   Median :1.0000   Median :0.6253   Mode  :character
##                     Mean   :1.0103   Mean   :0.6579
##                     3rd Qu.:1.6667   3rd Qu.:1.0000
##                     Max.   :3.0000   Max.   :2.0000
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
sum(is.na(data))
```

```
## [1] 0
```

```
# Convert character columns to factors and age to integer
# NOTE: We aren't using any ordinal variables from the dataset, so we don't need to worry about the ord
for (i in 1:ncol(data)) {
  print(paste(colnames(data)[i], typeof(data[[i]]), class(data[[i]])))
  if (typeof(data[[i]]) == "character") {
    data[[i]] <- as.factor(data[[i]])
  } else if (colnames(data)[i] == "Age") {
    data[[i]] <- as.integer(data[[i]])
  }
  print(paste(colnames(data)[i], typeof(data[[i]]), class(data[[i]])))
  print("----------------------------------------------------")
}
```

```
## [1] "Gender character character"
## [1] "Gender integer factor"
## [1] "----------------------------------------------------"
```

```
## [1] "Age double numeric"
## [1] "Age integer integer"
## [1] "----------------------------------------------------"
## [1] "Height double numeric"
## [1] "Height double numeric"
## [1] "----------------------------------------------------"
## [1] "Weight double numeric"
## [1] "Weight double numeric"
## [1] "----------------------------------------------------"
## [1] "family_history_with_overweight character character"
## [1] "family_history_with_overweight integer factor"
## [1] "----------------------------------------------------"
## [1] "FAVC character character"
## [1] "FAVC integer factor"
## [1] "----------------------------------------------------"
## [1] "FCVC double numeric"
## [1] "FCVC double numeric"
## [1] "----------------------------------------------------"
## [1] "NCP double numeric"
## [1] "NCP double numeric"
## [1] "----------------------------------------------------"
## [1] "CAEC character character"
## [1] "CAEC integer factor"
## [1] "----------------------------------------------------"
## [1] "SMOKE character character"
## [1] "SMOKE integer factor"
## [1] "----------------------------------------------------"
## [1] "CH2O double numeric"
## [1] "CH2O double numeric"
## [1] "----------------------------------------------------"
## [1] "SCC character character"
## [1] "SCC integer factor"
## [1] "----------------------------------------------------"
## [1] "FAF double numeric"
## [1] "FAF double numeric"
## [1] "----------------------------------------------------"
## [1] "TUE double numeric"
## [1] "TUE double numeric"
## [1] "----------------------------------------------------"
## [1] "CALC character character"
## [1] "CALC integer factor"
## [1] "----------------------------------------------------"
## [1] "MTRANS character character"
## [1] "MTRANS integer factor"
## [1] "----------------------------------------------------"
## [1] "NObeyesdad character character"
## [1] "NObeyesdad integer factor"
## [1] "----------------------------------------------------"
```

Exploratory Data Analysis

```r
# Function to plot exploratory data analysis plots for variables
eda_plot <- function(data, variable, label) {
  if (is.numeric(data[[variable]])) {
    # For numerical data, create a histogram
    ggplot(data, aes(x = .data[[variable]])) +
      geom_histogram(bins = 30, fill = "blue", color = "black") +
      ggtitle(paste("Distribution of", label)) +
      theme_minimal()
  } else {
    # For categorical data, create a bar plot
    ggplot(data, aes(x = .data[[variable]], fill = .data[[variable]])) +
      geom_bar() +
      ggtitle(paste("Distribution of", label)) +
      theme_minimal() +
      theme(legend.position = "none")
  }
}

# Apply the function to create plots for specified variables
eda_plot(data, "Age", "Age") # Required Predictor
```

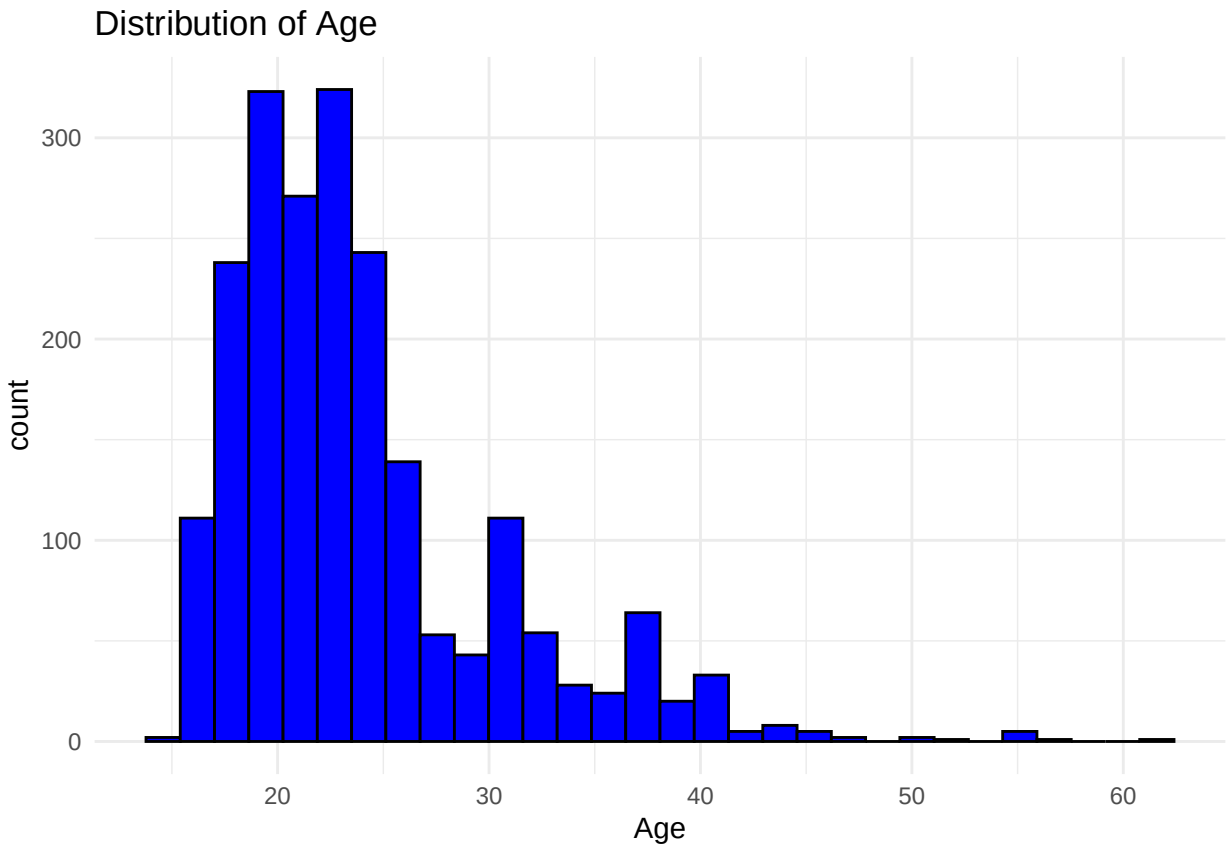


Figure 1: Exploratory Data Analysis Plots

```r
eda_plot(data, "family_history_with_overweight", "Family History with Overweight")
```

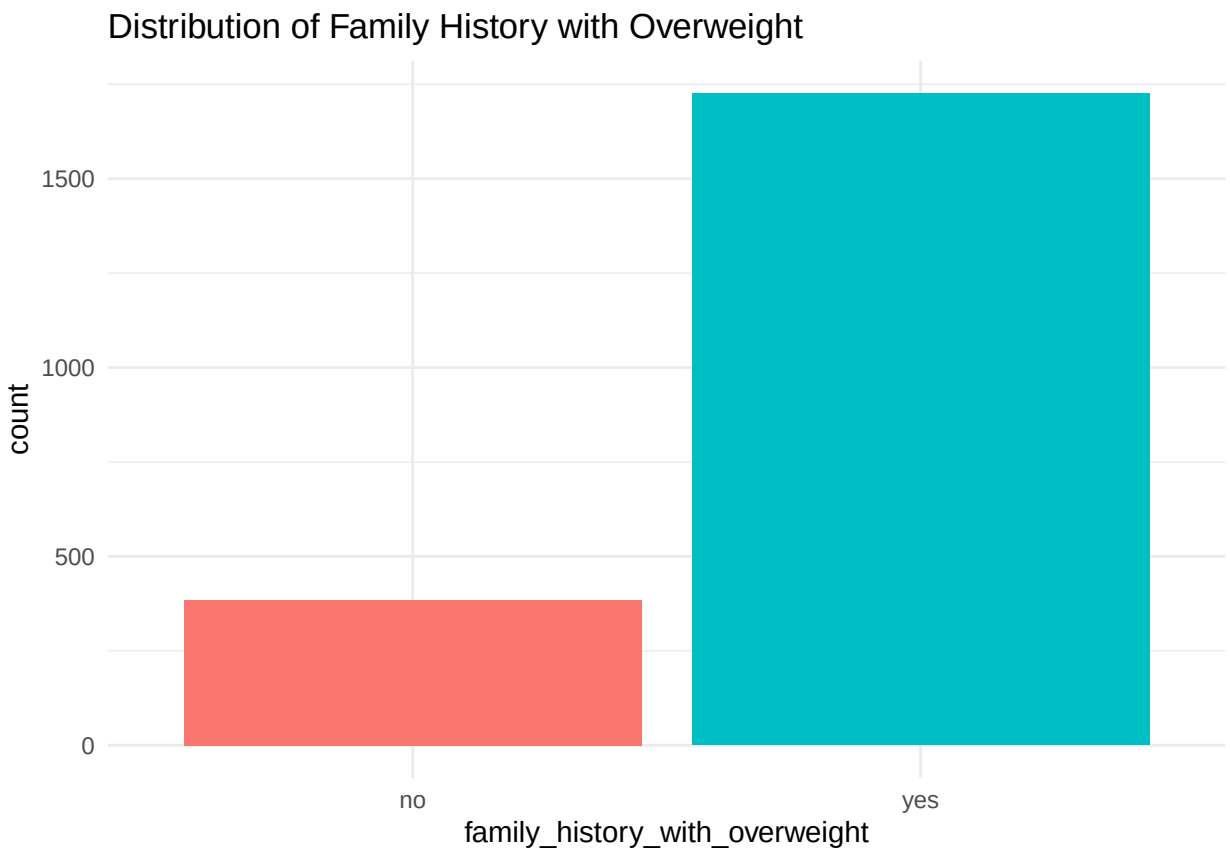


Figure 2: Exploratory Data Analysis Plots

```r
eda_plot(data, "FAVC", "Frequent Consumption of High Caloric Food")
```

```r
eda_plot(data, "TUE", "Time Using Technology Devices (Hrs/Day)")
```

```r
eda_plot(data, "MTRANS", "Transportation Method Used Daily")
```

```r
# eda_plot(data, "Gender", "Gender")
# eda_plot(data, "CAEC", "Consumption of Food Between Meals")
# eda_plot(data, "SMOKE", "Smoking")
# eda_plot(data, "Weight", "Weight")
# eda_plot(data, "Height", "Height")
# eda_plot(data, "FCVC", "Frequency of Vegetables Consumption")
# eda_plot(data, "NCP", "Number of Main Meals")
# eda_plot(data, "CH2O", "Consumption of Water")
# eda_plot(data, "FAF", "Physical Activity Frequency")
# eda_plot(data, "CALC", "Consumption of Alcohol")
# eda_plot(data, "NObeyesdad", "Obesity Level") # Will be dropped as it is the categorical interpretati
```

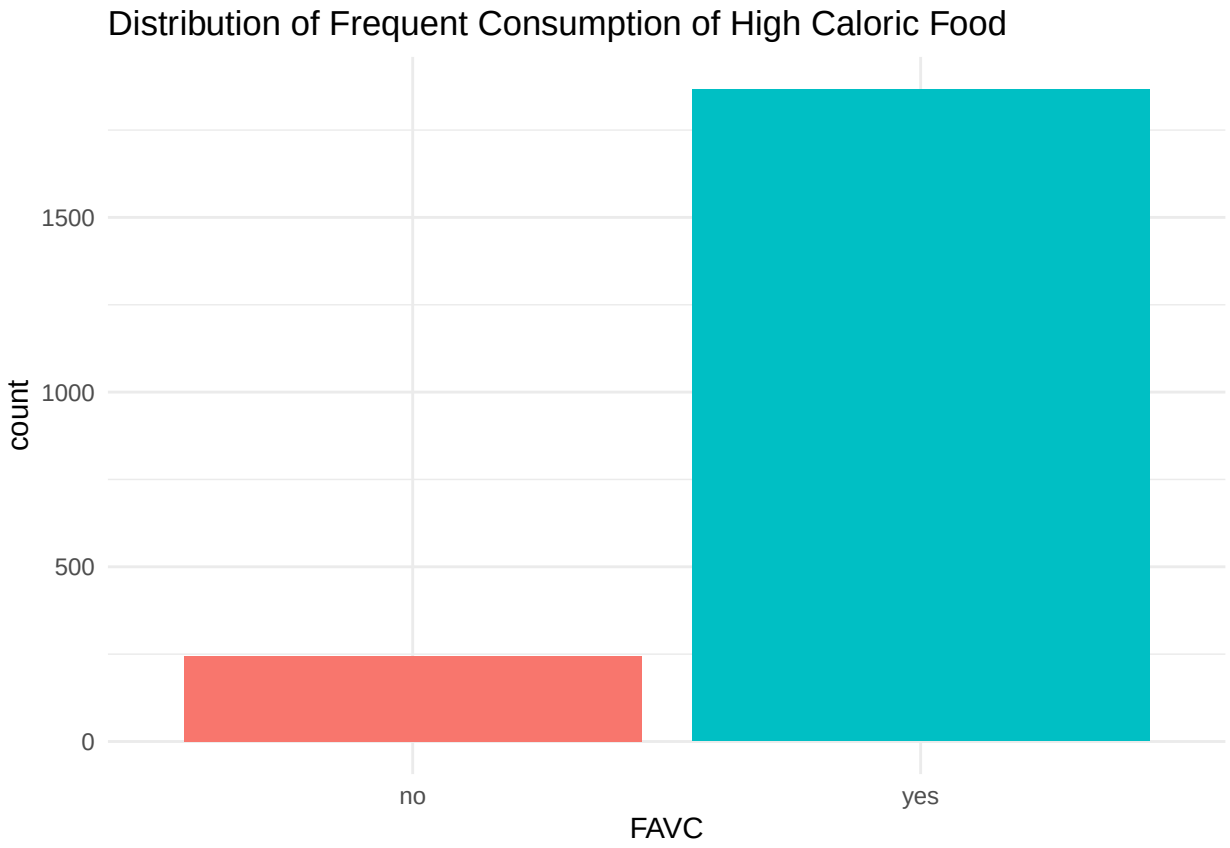


Figure 3: Exploratory Data Analysis Plots

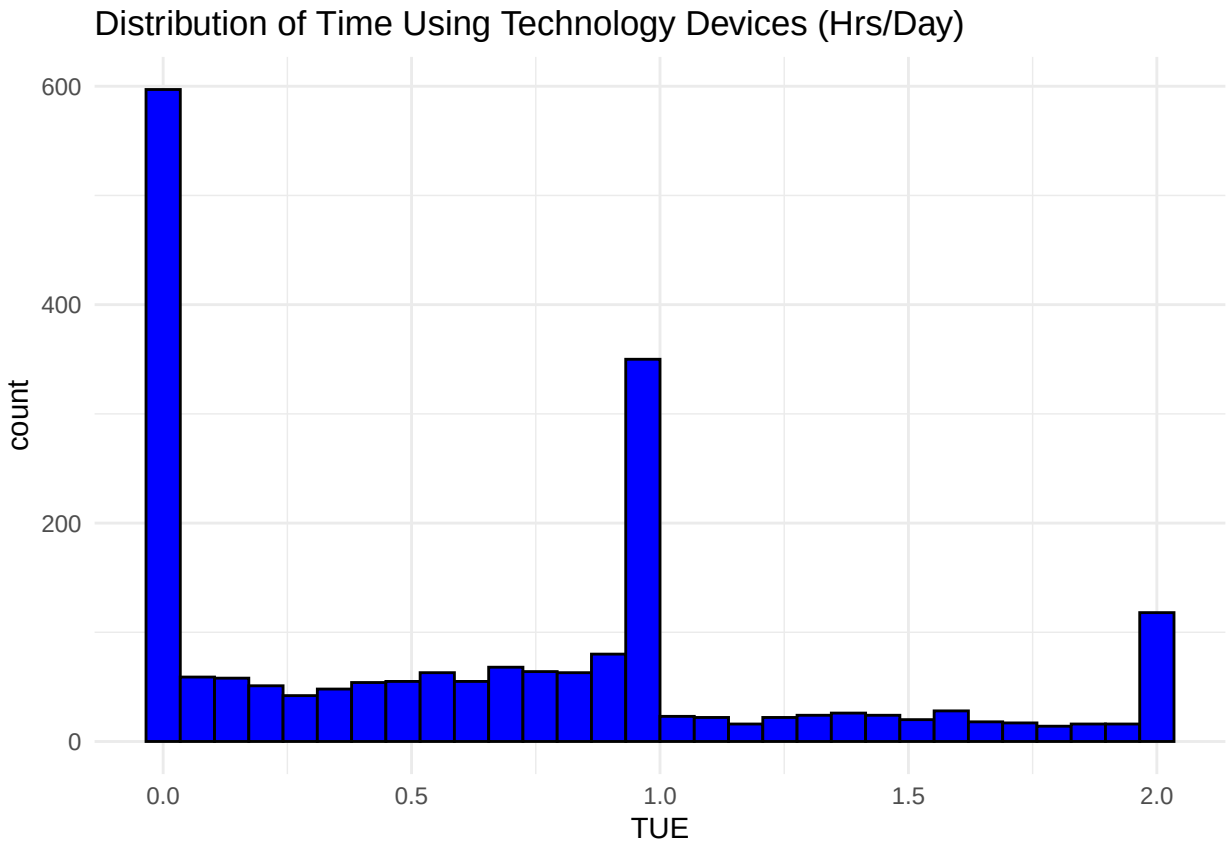


Figure 4: Exploratory Data Analysis Plots

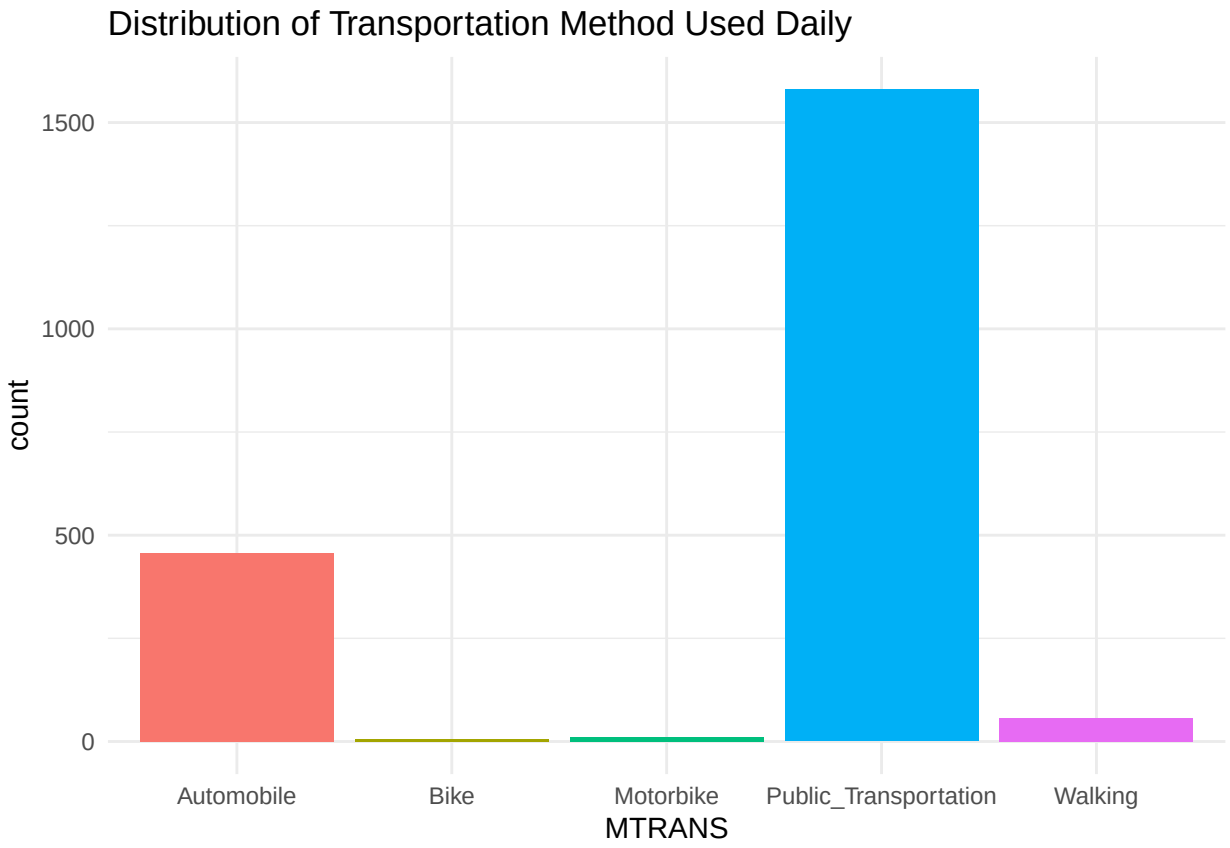
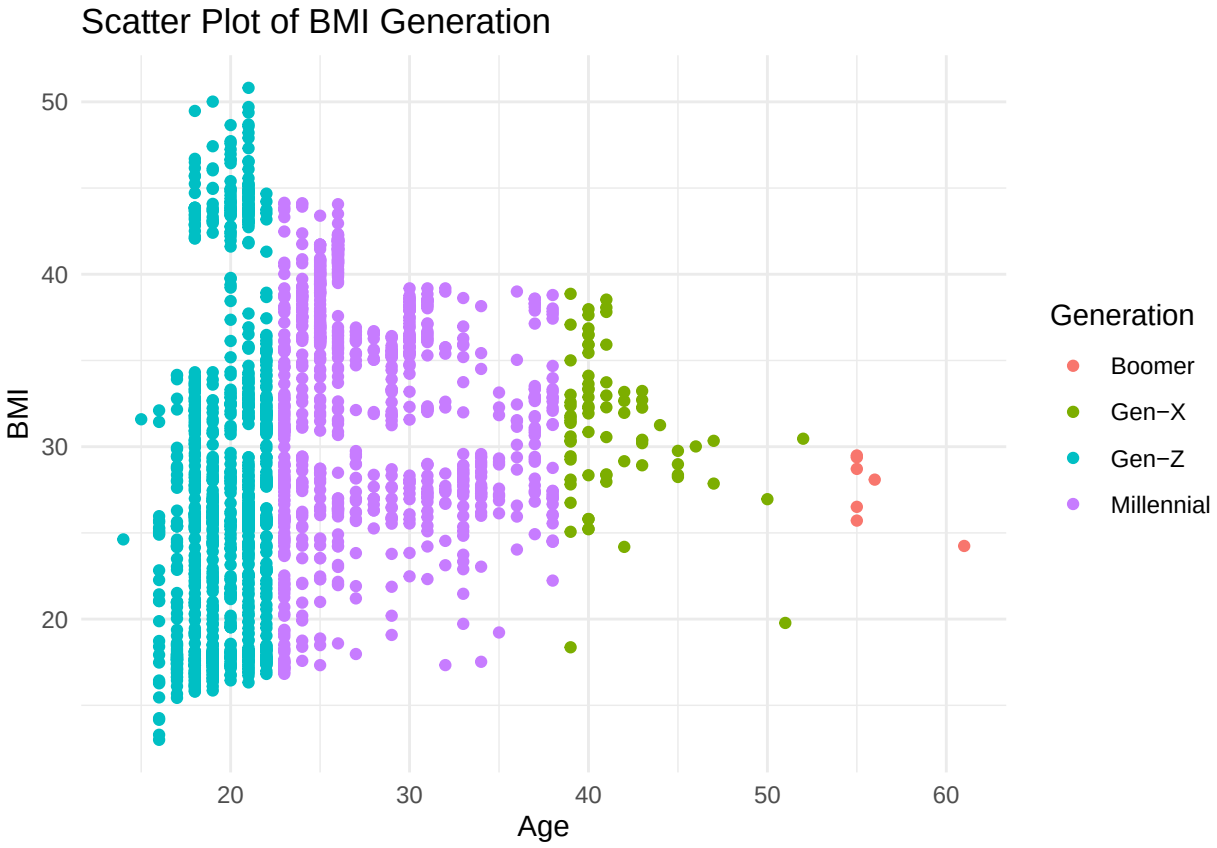


Figure 5: Exploratory Data Analysis Plots

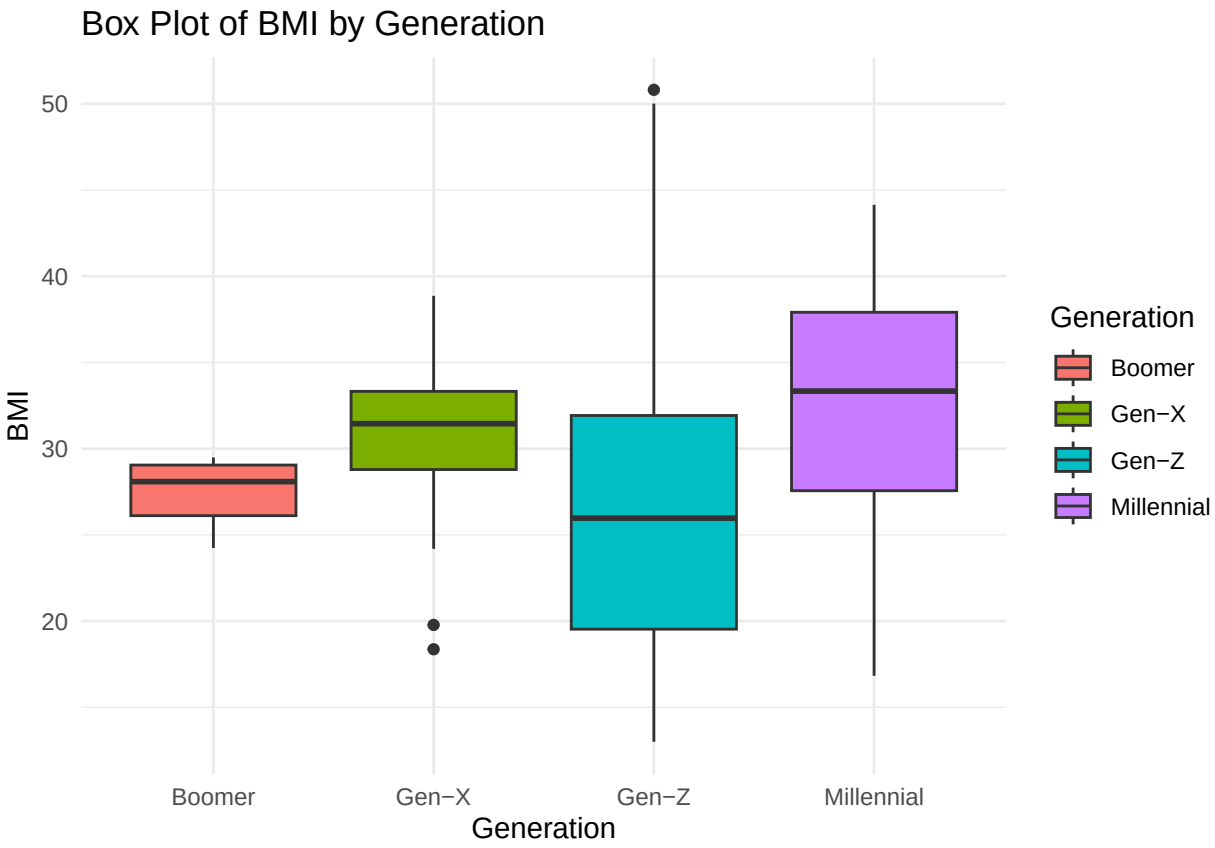
Feature Engineering

```r
# Feature engineering to create necessary transformations and new variables
data_fe <- data %>%
  mutate(
    BMI = Weight / (Height ^ 2), # Calculate BMI
    Year_Born = 2019 - Age, # Derive year of birth use date of research study as starting point
    Generation = factor(case_when( # Classify into generations
      Year_Born >= 1946 & Year_Born <= 1964 ~ "Boomer",
      Year_Born >= 1965 & Year_Born <= 1980 ~ "Gen-X",
      Year_Born >= 1981 & Year_Born <= 1996 ~ "Millennial",
      Year_Born >= 1997 & Year_Born <= 2010 ~ "Gen-Z",
      Year_Born >= 2011 ~ "Gen-Alpha"
    ),)# levels = c("Boomer", "Gen-X", "Millennial", "Gen-Z", "Gen-Alpha"), ordered = TRUE) # if we ass
  )

ggplot(data_fe, aes(x = Age, y = BMI, color = Generation)) +
  geom_point() +
  ggtitle("Scatter Plot of BMI Generation") +
  xlab("Age") +
  ylab("BMI") +
  theme_minimal()
```



```r
ggplot(data_fe, aes(x = Generation, y = BMI, fill = Generation)) +
  geom_boxplot() +
  ggtitle("Box Plot of BMI by Generation") +
  xlab("Generation") +
  ylab("BMI") +
  theme_minimal()
```



Custom Regression Plot Function

```r
# Function to create a residuals vs fitted plot
reg_plot_with_residuals <- function(data, formula, model_name){
  model <- lm(formula, data)
  aug_data <- augment(model, data)

  # Create the residuals vs fitted plot
  p <- ggplot(aug_data, aes(x = .fitted, y = .resid)) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "gray") + # Dash the zero line
    geom_point(color = "red") +
    stat_smooth() + # Add a smooth line
    labs(title = model_name, x = "Fitted Values", y = "Residuals") +
    theme_minimal()

  # Print the plot
```

```
  print(p)

  # Print the summary of the model
  print("Model Summary")
  print(summary(model))

  robust_errors <- coeftest(model, vcov = vcovHC(model, type = "HC1"))
  print("Robust Standard Errors")
  print(robust_errors)

  # Studentized Breusch-Pagan test for heteroscedasticity
  print("H0: Homoscedasticity vs. H1: Heteroscedasticity")
  print(bptest(model))

  # Checking for Multicollinearity via VIF
  if (length(model$coefficients) > 1) {
    print("Multicollinearity: VIF > 4")
    print(vif(model)>4)
  }

  return(list(model = model, residuals_plot = p, coeftest_robust = robust_errors))
}
```

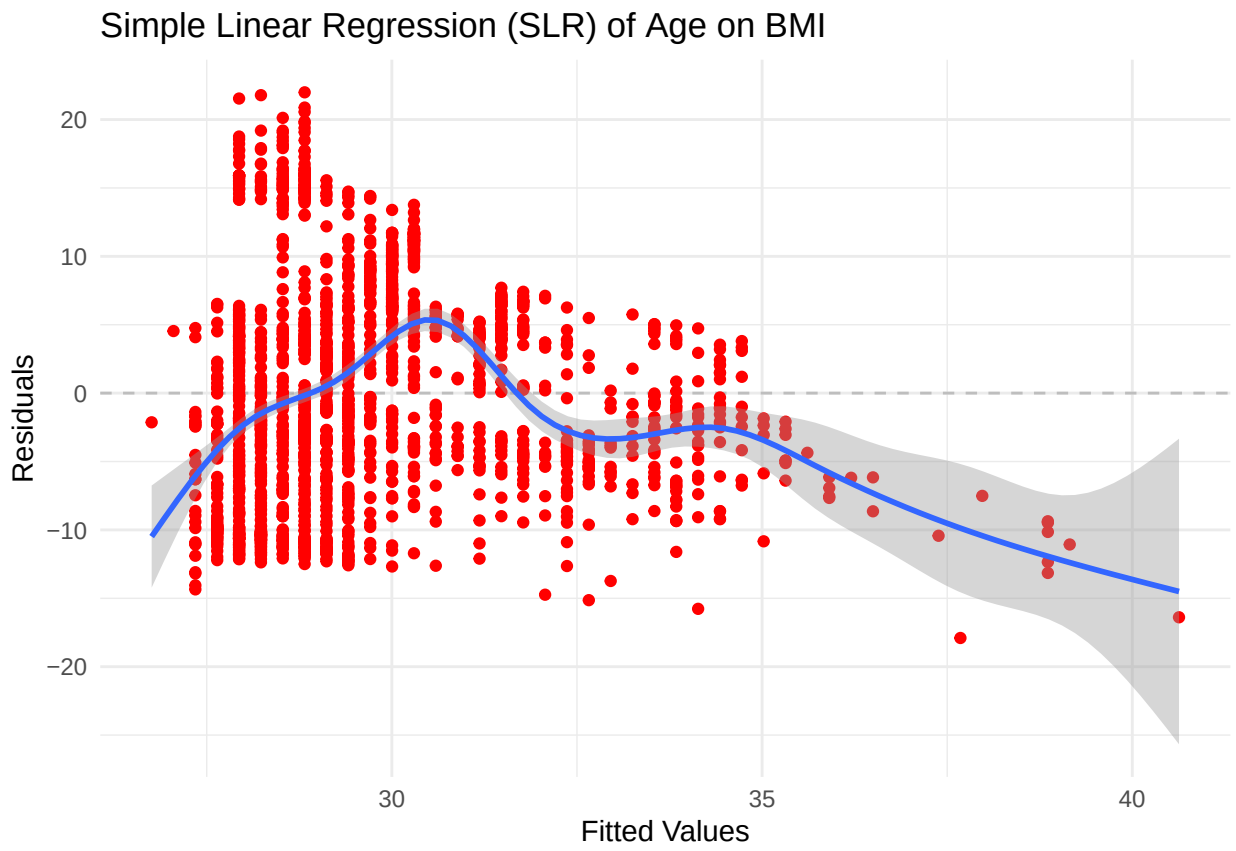
Regression Model Building

Model with Age Only

- Investigates the direct linear relationship between age and BMI to understand how BMI changes with age.
- The chart shows a pattern in the residuals, indicating that the linear model may not fully capture the relationship between age and BMI, particularly for middle ages.
- The model summary indicates a significant positive relationship between age and BMI (p-value $<$ 2e-16). However, the low R-squared value (0.0617) suggests that age alone explains only a small portion of the variance in BMI.
- The pattern in the residuals suggests that a simple linear model may not fully capture the relationship between age and BMI.

```
model_1 <- reg_plot_with_residuals(data_fe, BMI ~ Age, "Simple Linear Regression (SLR) of Age on BMI")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## [1] "Model Summary"
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.901  -5.832  -1.284   5.389  21.989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.62346    0.66667   33.94   <2e-16 ***
## Age          0.29520    0.02689   10.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.794 on 2109 degrees of freedom
## Multiple R-squared:  0.05404,    Adjusted R-squared:  0.05359
## F-statistic: 120.5 on 1 and 2109 DF,  p-value: < 2.2e-16
##
## [1] "Robust Standard Errors"
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 22.623456   0.671893  33.671 < 2.2e-16 ***
```

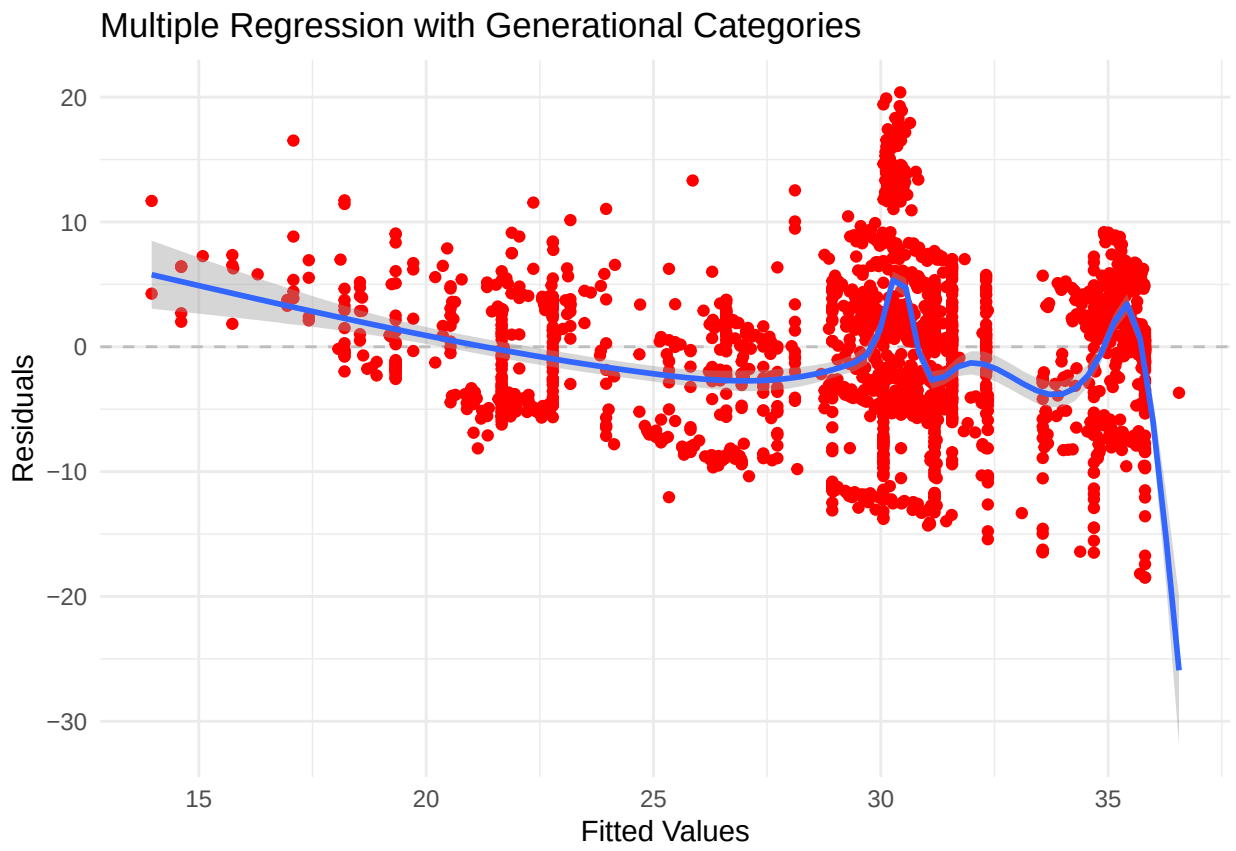
```
## Age          0.295201   0.025643  11.512 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## [1] "H0: Homoscedasticity vs. H1: Heteroscedasticity"
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 63.093, df = 1, p-value = 1.972e-15
##
## [1] "Multicollinearity: VIF > 4"
##   Age
## FALSE
```

Multiple Regression with Generational Categories

- Explores the combined effect of generational categories and lifestyle factors on BMI.
- The chart shows a distribution of residuals similar to the previous multiple regression model, indicating that the addition of generational categories does not significantly improve the model fit.
- This model demonstrates that technology use (TUE), frequent consumption of high-caloric food (FAVC), family history with overweight, transportation method (MTRANS), and generational categories are significant predictors of BMI (p-values < 0.05). The R-squared value (0.3545) is the highest among the old model plots, indicating that this model explains the largest portion of the variance in BMI and highlights the combined impact of these factors.
- The distribution of residuals is similar to the previous multiple regression model, indicating that the addition of generational categories does not significantly improve the model fit.

```
model_2 <- reg_plot_with_residuals(data_fe, BMI ~  Generation + TUE + FAVC + family_history_with_overwe
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## [1] "Model Summary"
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4831  -4.4232   0.1776   3.8891  20.3839
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       19.64377    2.47397   7.940 3.26e-15 ***
## GenerationGen-X                    0.81996    2.55619   0.321   0.7484
## GenerationGen-Z                   -4.55652    2.46685  -1.847   0.0649 .
## GenerationMillennial               0.07258    2.46067   0.029   0.9765
## TUE                               -1.12468    0.24149  -4.657 3.40e-06 ***
## FAVCyes                            3.45712    0.45896   7.533 7.34e-14 ***
## family_history_with_overweightyes  8.39588    0.38257  21.946  < 2e-16 ***
## MTRANSBike                         1.65572    2.47225   0.670   0.5031
## MTRANSMotorbike                    0.66996    1.97717   0.339   0.7348
## MTRANSPublic_Transportation        4.24274    0.38612  10.988  < 2e-16 ***
## MTRANSWalking                      0.65065    0.95105   0.684   0.4940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.453 on 2100 degrees of freedom
```

```
## Multiple R-squared:  0.3543, Adjusted R-squared:  0.3512
## F-statistic: 115.2 on 10 and 2100 DF,  p-value: < 2.2e-16
##
## [1] "Robust Standard Errors"
##
## t test of coefficients:
##
##                                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                      19.643773   1.225474 16.0295 < 2.2e-16 ***
## GenerationGen-X                   0.819961   1.276645  0.6423 0.5207626
## GenerationGen-Z                  -4.556517   1.213983 -3.7534 0.0001792 ***
## GenerationMillennial              0.072584   1.198647  0.0606 0.9517191
## TUE                              -1.124676   0.228821 -4.9151 9.561e-07 ***
## FAVCyes                           3.457122   0.378358  9.1372 < 2.2e-16 ***
## family_history_with_overweightyes  8.395878   0.304764 27.5488 < 2.2e-16 ***
## MTRANSBike                        1.655716   1.553230  1.0660 0.2865539
## MTRANSMotorbike                   0.669963   1.687480  0.3970 0.6913932
## MTRANSPublic_Transportation       4.242738   0.348051 12.1900 < 2.2e-16 ***
## MTRANSWalking                     0.650653   0.756322  0.8603 0.3897299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## [1] "H0: Homoscedasticity vs. H1: Heteroscedasticity"
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 205.02, df = 10, p-value < 2.2e-16
##
## [1] "Multicollinearity: VIF > 4"
##                   GenerationGen-X                   GenerationGen-Z
##                              TRUE                              TRUE
##              GenerationMillennial                               TUE
##                              TRUE                             FALSE
##                           FAVCyes family_history_with_overweightyes
##                             FALSE                             FALSE
##                         MTRANSBike                   MTRANSMotorbike
##                             FALSE                             FALSE
##       MTRANSPublic_Transportation                      MTRANSWalking
##                             FALSE                             FALSE
```

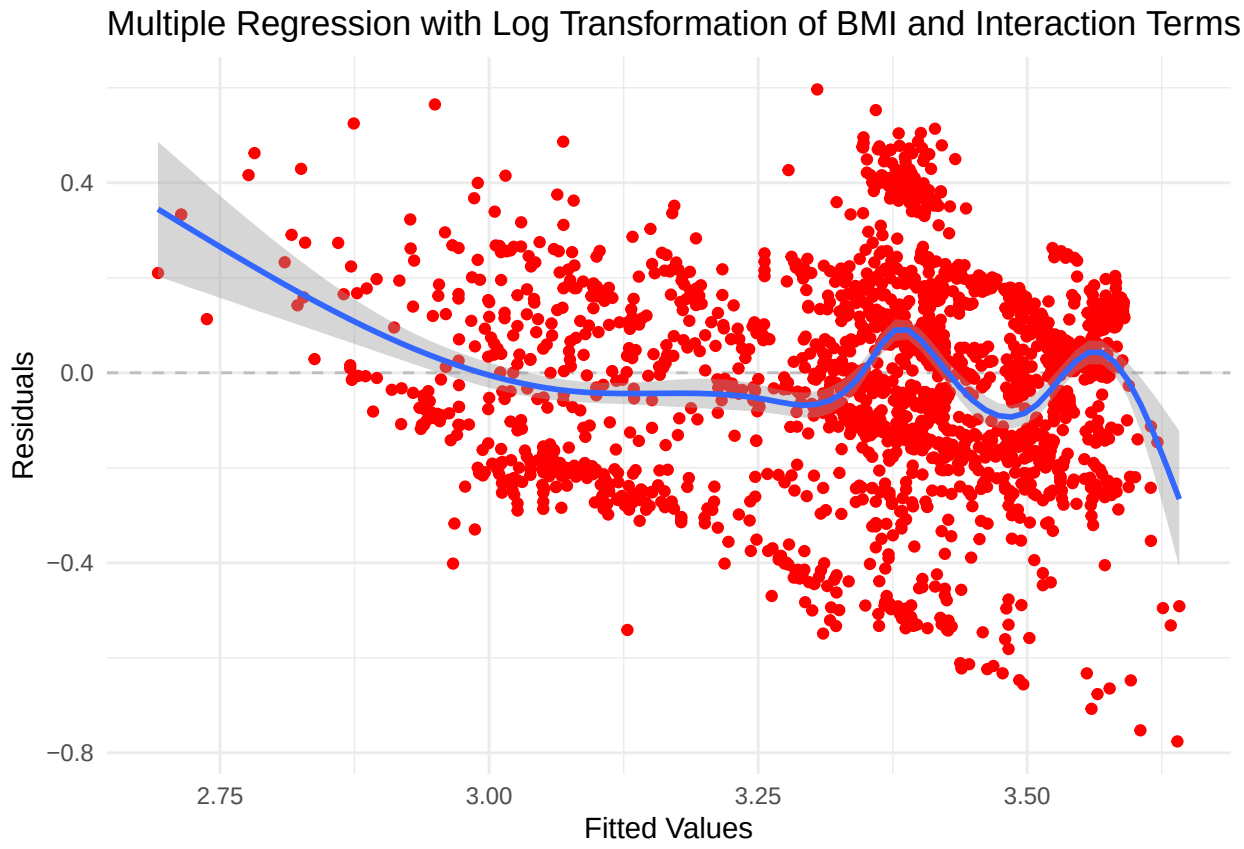
Multiple Regression with Log Transformation of BMI and Interaction Terms

- Explores the impact of generational categories, lifestyle factors, and their interactions on the logarithmic transformation of BMI.
- The use of the logarithmic transformation of BMI aims to linearize its relationship with the predictors and stabilize the variance of the residuals.
- This model introduces interaction terms between TUE (Time Using Technology Devices) and Age, as well as between FAVC (Frequent Consumption of High Caloric Food) and Gender, to capture the nuanced effects of these factors on BMI.
- The interaction between MTRANS (Method of Transportation) and FAF (Frequency of Physical Activity) is included to examine how the impact of transportation method on BMI might vary based on physical activity levels.

- The chart shows a distribution of residuals that suggests a better fit than the previous models, indicating that the log transformation and interaction terms might be capturing more of the complexity in the data.
- This model demonstrates that while generational categories are not significant predictors of BMI, the interactions between TUE and Age, FAVC and Gender, as well as MTRANS and FAF, are significant (p-values < 0.05). The R-squared value (0.4097) indicates that this model explains a substantial portion of the variance in the logarithmic transformation of BMI.

```
model_3 <- reg_plot_with_residuals(data_fe, log(BMI) ~  Generation + TUE*Age + FAVC*Gender + family_his
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## [1] "Model Summary"
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77596 -0.14384  0.02138  0.14211  0.59650
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.636718   0.137255  19.210  < 2e-16 ***
```

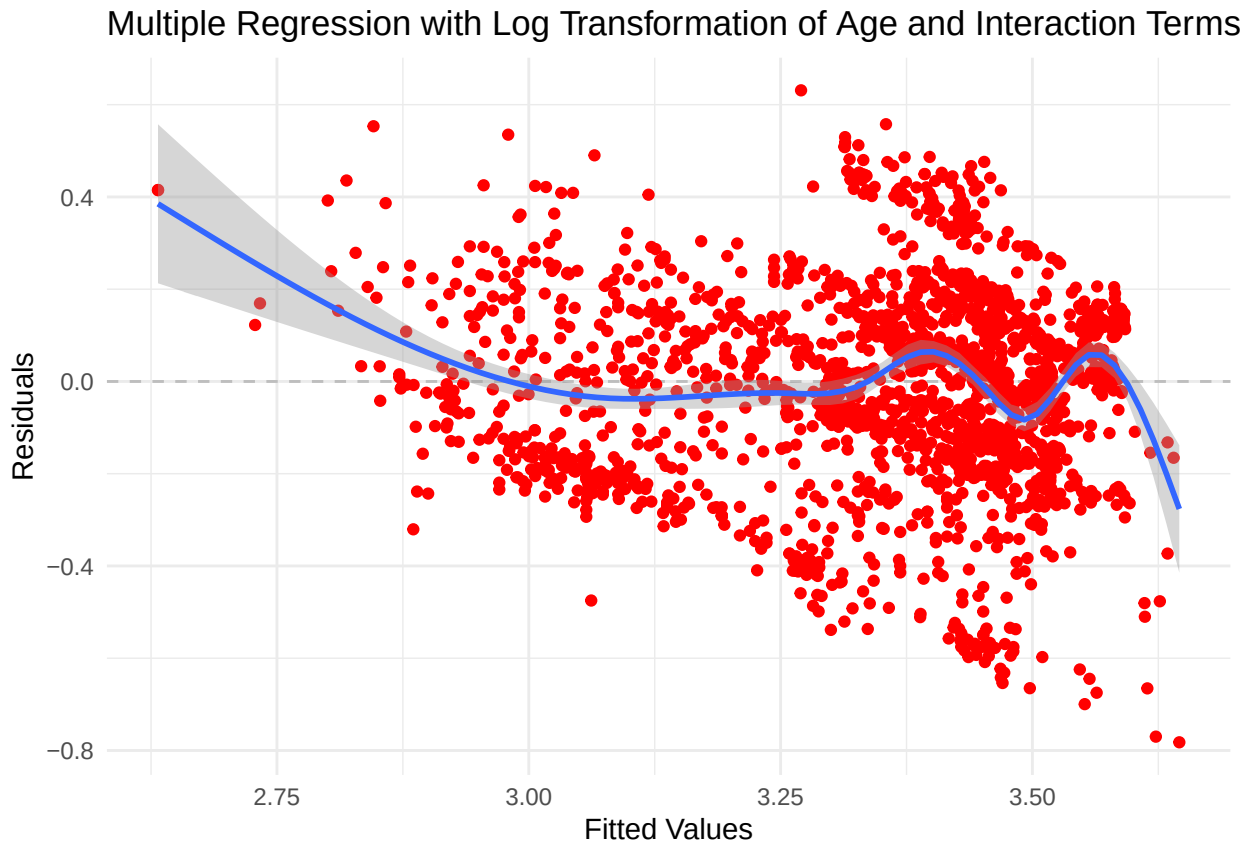
```
## GenerationGen-X                     0.123930   0.091531   1.354 0.175895
## GenerationGen-Z                     0.073976   0.103078   0.718 0.473044
## GenerationMillennial                0.176814   0.096963   1.824 0.068366 .
## TUE                                -0.169128   0.036091  -4.686 2.96e-06 ***
## Age                                 0.006114   0.001895   3.226 0.001273 **
## FAVCyes                             0.179348   0.020109   8.919  < 2e-16 ***
## GenderMale                          0.121762   0.028806   4.227 2.47e-05 ***
## family_history_with_overweightyes   0.305699   0.013049  23.427  < 2e-16 ***
## MTRANSBike                          0.081991   0.176058   0.466 0.641478
## MTRANSMotorbike                    -0.065346   0.081665  -0.800 0.423702
## MTRANSPublic_Transportation         0.132915   0.021427   6.203 6.65e-10 ***
## MTRANSWalking                      -0.018827   0.055716  -0.338 0.735464
## FAF                                -0.044050   0.011920  -3.695 0.000225 ***
## TUE:Age                             0.005830   0.001532   3.805 0.000146 ***
## FAVCyes:GenderMale                 -0.160950   0.030414  -5.292 1.34e-07 ***
## MTRANSBike:FAF                     -0.002812   0.078133  -0.036 0.971289
## MTRANSMotorbike:FAF                 0.115485   0.056271   2.052 0.040264 *
## MTRANSPublic_Transportation:FAF     0.028834   0.013571   2.125 0.033726 *
## MTRANSWalking:FAF                   0.058933   0.029874   1.973 0.048660 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2177 on 2091 degrees of freedom
## Multiple R-squared:  0.4097, Adjusted R-squared:  0.4044
## F-statistic: 76.39 on 19 and 2091 DF,  p-value: < 2.2e-16
##
## [1] "Robust Standard Errors"
##
## t test of coefficients:
##
##                                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                        2.6367178  0.1276176 20.6611 < 2.2e-16 ***
## GenerationGen-X                    0.1239298  0.0796170  1.5566 0.1197231
## GenerationGen-Z                    0.0739756  0.0915128  0.8084 0.4189732
## GenerationMillennial               0.1768142  0.0856500  2.0644 0.0391046 *
## TUE                               -0.1691277  0.0361176 -4.6827 3.013e-06 ***
## Age                                0.0061144  0.0018442  3.3155 0.0009303 ***
## FAVCyes                            0.1793478  0.0190798  9.3999 < 2.2e-16 ***
## GenderMale                         0.1217620  0.0262900  4.6315 3.854e-06 ***
## family_history_with_overweightyes  0.3056986  0.0122961 24.8613 < 2.2e-16 ***
## MTRANSBike                         0.0819905  0.0928796  0.8828 0.3774666
## MTRANSMotorbike                   -0.0653464  0.0666925 -0.9798 0.3272900
## MTRANSPublic_Transportation        0.1329146  0.0190536  6.9758 4.064e-12 ***
## MTRANSWalking                     -0.0188269  0.0516731 -0.3643 0.7156355
## FAF                               -0.0440500  0.0099320 -4.4352 9.678e-06 ***
## TUE:Age                            0.0058304  0.0014856  3.9246 8.969e-05 ***
## FAVCyes:GenderMale                -0.1609501  0.0283465 -5.6780 1.553e-08 ***
## MTRANSBike:FAF                    -0.0028125  0.0565023 -0.0498 0.9603052
## MTRANSMotorbike:FAF                0.1154854  0.0302890  3.8128 0.0001414 ***
## MTRANSPublic_Transportation:FAF    0.0288344  0.0118166  2.4402 0.0147631 *
## MTRANSWalking:FAF                  0.0589325  0.0267832  2.2004 0.0278907 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## [1] "H0: Homoscedasticity vs. H1: Heteroscedasticity"
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 253.69, df = 19, p-value < 2.2e-16
##
## [1] "Multicollinearity: VIF > 4"
##                    GenerationGen-X                    GenerationGen-Z
##                               TRUE                               TRUE
##              GenerationMillennial                                TUE
##                               TRUE                               TRUE
##                                Age                             FAVCyes
##                               TRUE                              FALSE
##                         GenderMale family_history_with_overweightyes
##                               TRUE                              FALSE
##                         MTRANSBike                     MTRANSMotorbike
##                               TRUE                              FALSE
##       MTRANSPublic_Transportation                        MTRANSWalking
##                              FALSE                              FALSE
##                                FAF                            TUE:Age
##                               TRUE                               TRUE
##                FAVCyes:GenderMale                      MTRANSBike:FAF
##                               TRUE                               TRUE
##               MTRANSMotorbike:FAF   MTRANSPublic_Transportation:FAF
##                              FALSE                               TRUE
##                 MTRANSWalking:FAF
##                              FALSE
```

Final Model Selection: Log Transformation of BMI & Age and Interaction Terms

```
model_4 <- reg_plot_with_residuals(data_fe, log(BMI) ~  log(Age) + TUE*Age + FAVC*Gender + family_histo
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## [1] "Model Summary"
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78225 -0.14473  0.02156  0.13950  0.63111
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.838268   0.341795  -2.453   0.0143 *
## log(Age)                         1.526082   0.147716  10.331  < 2e-16 ***
## TUE                             -0.085700   0.037317  -2.297   0.0217 *
## Age                             -0.044109   0.005303  -8.318  < 2e-16 ***
## FAVCyes                          0.185625   0.019910   9.323  < 2e-16 ***
## GenderMale                       0.132483   0.028501   4.648 3.56e-06 ***
## family_history_with_overweightyes  0.295247   0.012981  22.745  < 2e-16 ***
## MTRANSBike                       0.057523   0.173857   0.331   0.7408
## MTRANSMotorbike                 -0.065293   0.080831  -0.808   0.4193
## MTRANSPublic_Transportation      0.124386   0.020786   5.984 2.55e-09 ***
## MTRANSWalking                   -0.008395   0.055016  -0.153   0.8787
## FAF                             -0.040486   0.011779  -3.437   0.0006 ***
## TUE:Age                          0.002427   0.001579   1.537   0.1245
## FAVCyes:GenderMale              -0.173694   0.030148  -5.761 9.57e-09 ***
## MTRANSBike:FAF                  -0.006828   0.077229  -0.088   0.9296
```

```
## MTRANSMotorbike:FAF                0.107457   0.055716   1.929   0.0539 .
## MTRANSPublic_Transportation:FAF    0.025268   0.013356   1.892   0.0586 .
## MTRANSWalking:FAF                  0.052914   0.029557   1.790   0.0736 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2155 on 2093 degrees of freedom
## Multiple R-squared:  0.4209, Adjusted R-squared:  0.4162
## F-statistic: 89.48 on 17 and 2093 DF,  p-value: < 2.2e-16
##
## [1] "Robust Standard Errors"
##
## t test of coefficients:
##
##                                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                     -0.8382685  0.3042063 -2.7556 0.0059094 **
## log(Age)                         1.5260817  0.1306139 11.6839 < 2.2e-16 ***
## TUE                             -0.0857000  0.0336625 -2.5459 0.0109720 *
## Age                             -0.0441088  0.0046626 -9.4602 < 2.2e-16 ***
## FAVCyes                          0.1856247  0.0190377  9.7504 < 2.2e-16 ***
## GenderMale                       0.1324832  0.0261112  5.0738 4.243e-07 ***
## family_history_with_overweightyes  0.2952475  0.0120851 24.4307 < 2.2e-16 ***
## MTRANSBike                       0.0575227  0.1076837  0.5342 0.5932722
## MTRANSMotorbike                 -0.0652927  0.0617860 -1.0568 0.2907449
## MTRANSPublic_Transportation      0.1243861  0.0179701  6.9219 5.906e-12 ***
## MTRANSWalking                   -0.0083951  0.0566359 -0.1482 0.8821765
## FAF                             -0.0404862  0.0093713 -4.3203 1.631e-05 ***
## TUE:Age                          0.0024266  0.0013528  1.7937 0.0730049 .
## FAVCyes:GenderMale              -0.1736936  0.0283160 -6.1341 1.021e-09 ***
## MTRANSBike:FAF                  -0.0068278  0.0576595 -0.1184 0.9057499
## MTRANSMotorbike:FAF              0.1074571  0.0290094  3.7042 0.0002175 ***
## MTRANSPublic_Transportation:FAF  0.0252683  0.0113022  2.2357 0.0254763 *
## MTRANSWalking:FAF                0.0529142  0.0280207  1.8884 0.0591110 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## [1] "H0: Homoscedasticity vs. H1: Heteroscedasticity"
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 245.55, df = 17, p-value < 2.2e-16
##
## [1] "Multicollinearity: VIF > 4"
##                          log(Age)                                TUE
##                              TRUE                               TRUE
##                               Age                            FAVCyes
##                              TRUE                              FALSE
##                        GenderMale family_history_with_overweightyes
##                              TRUE                              FALSE
##                        MTRANSBike                    MTRANSMotorbike
##                              TRUE                              FALSE
##       MTRANSPublic_Transportation                      MTRANSWalking
##                             FALSE                              FALSE
```

```
##                              FAF                              TUE:Age
##                             TRUE                                 TRUE
##                FAVCyes:GenderMale                      MTRANSBike:FAF
##                             TRUE                                 TRUE
##               MTRANSMotorbike:FAF   MTRANSPublic_Transportation:FAF
##                            FALSE                                TRUE
##                 MTRANSWalking:FAF
##                            FALSE
```

Variable Selection and Model Construction Reasoning

- We used the log transformation of BMI and Age to linearize their relationship with other variables and stabilize the variance of the residuals, making them more homoscedastic (a required assumption of OLS). This transformation implies that we're modeling percentage changes rather than absolute changes.
- The interaction between TUE (Time Using Technology Devices) and Age is included to capture the varying impact of screen time on BMI across different ages. Younger individuals might be more susceptible to the negative effects of excessive screen time on BMI compared to older individuals.
- The interaction between FAVC (Frequent Consumption of High Caloric Food) and Gender is included to account for any differing effects of diet on BMI across genders, as males and females might respond differently to high-calorie diets.
- Family history with overweight is considered a predictor based on the assumption that genetic predispositions could influence an individual's BMI.
- The interaction between MTRANS (Method of Transportation) and FAF (Frequency of Physical Activity) is included to consider that the impact of transportation method on BMI could be moderated by physical activity levels.
- We did not include variables such as smoking, consumption of vegetables, etc., in the final model as their inclusion were either not of primary interest in our hypothesis testing or we could not verify the context of the variable. For example, smoking what? Cigarettes, marijuana, etc.?

Interpretation of Coefficients

- log(Age): For a 1% increase in age, BMI is expected to increase by about 1.526%, holding other variables constant. This relationship is multiplicative due to the logarithmic - transformation.
- TUE: Each additional hour spent using technology devices is associated with an 8.57% decrease in BMI, conditional on Age being zero. However, the interaction with Age modifies this effect.
- Age: Each additional year of age is associated with a 4.41% decrease in BMI, conditional on TUE being zero. This effect is part of a more complex relationship due to the inclusion of both Age and log(Age).
- FAVCyes: Individuals who frequently consume high-caloric food have a BMI roughly 18.56% higher than those who do not, all else being equal.
- GenderMale: Being male is associated with a 13.25% higher BMI compared to being female, controlling for other factors.
- family_history_with_overweightyes: Having a family history of overweight is associated with a 29.52% higher BMI.
- MTRANS (Public Transportation): Using public transportation is associated with an increase in BMI compared to using an automobile.
- FAF: Each unit increase in the frequency of physical activity is associated with a 4.05% decrease in BMI.
- TUE*Age: The effect of technology use on BMI changes with age, but this interaction is not statistically significant in our model.
- FAVCyes*GenderMale: The effect of frequently consuming high-caloric food on BMI is different for males compared to females, reducing the BMI for males.

- MTRANS*FAF: The impact of the transportation method on BMI varies with the frequency of physical activity, with some interactions being significant.

Model Diagnostics

- The Breusch-Pagan test indicates the presence of heteroscedasticity, which is why we used robust standard errors to obtain more reliable significance tests for our coefficients.
- The Variance Inflation Factor (VIF) values indicate some multicollinearity in our model, especially with the interaction terms. While this does not bias our coefficient estimates, it can inflate their standard errors, leading to less precise estimates.

Conclusion

- We decided on the last model as our final model because it explains about 42% of the variability in the percent change of BMI, which is a moderate amount. The adjusted R-squared value suggests that the model fits the data well without being overly complex.
- The presence of significant interaction terms indicates that the effect of some predictors on BMI is not simply additive but depends on the level of other variables.
- Future research could explore the inclusion of other potential predictors or different modeling techniques to further improve the explanatory power and predictive accuracy of the model.

Table 1: Tentative Models with Robust Standard Errors

	Dependent variable:			
	BMI		log(BMI)	
	(1)	(2)	(3)	(4)
log(Age)				1.526*** (0.131)
Age	0.295*** (0.026)		0.006*** (0.002)	−0.044*** (0.005)
GenerationGen-X		0.820 (1.277)	0.124 (0.080)	
GenerationGen-Z		−4.557*** (1.214)	0.074 (0.092)	
GenerationMillennial		0.073 (1.199)	0.177* (0.086)	
TUE		−1.125*** (0.229)	−0.169*** (0.036)	−0.086* (0.034)
FAVCyes		3.457*** (0.378)	0.179*** (0.019)	0.186*** (0.019)
GenderMale			0.122*** (0.026)	0.132*** (0.026)
family_history_with_overweightyes		8.396*** (0.305)	0.306*** (0.012)	0.295*** (0.012)
MTRANSBike		1.656 (1.553)	0.082 (0.093)	0.058 (0.108)
MTRANSMotorbike		0.670 (1.687)	−0.065 (0.067)	−0.065 (0.062)
MTRANSPublic_Transportation		4.243*** (0.348)	0.133*** (0.019)	0.124*** (0.018)
MTRANSWalking		0.651 (0.756)	−0.019 (0.052)	−0.008 (0.057)
FAF			−0.044*** (0.010)	−0.040*** (0.009)
TUE:Age			0.006*** (0.001)	0.002 (0.001)
FAVCyes:GenderMale			−0.161*** (0.028)	−0.174*** (0.028)
MTRANSBike:FAF			−0.003 (0.057)	−0.007 (0.058)
MTRANSMotorbike:FAF			0.115*** (0.030)	0.107*** (0.029)
MTRANSPublic_Transportation:FAF			0.029* (0.012)	0.025* (0.011)
MTRANSWalking:FAF			0.059* (0.027)	0.053 (0.028)
Constant	22.623*** (0.672)	19.644*** (1.225)	2.637*** (0.128)	−0.838** (0.304)
Observations	2,111	2,111	2,111	2,111
R^2	0.054	0.354	0.410	0.421
Adjusted R^2	0.054	0.351	0.404	0.416
Residual Std. Error	7.794 (df = 2109)	6.453 (df = 2100)	0.218 (df = 2091)	0.216 (df = 2093)
Note:				*p<0.05; **p<0.01; ***p<0.001