R Project Regression

4375 Machine Learning with Dr. Mazidi

Joshua Jan

3/27/2022

Data website link: https://www.kaggle.com/spittman1248/cdc-data-nutrition-physical-activity-obesity

Data cleaning comment: The data set consists of 53392 observations and 16 attributes. The columns are filled with different data types such as integers, decimals, and characters. For data cleaning, I experimented using tibble because it is supposedly easier to work with large data. The select() function moves variables to a new variable, filter() function returns rows with matching conditions, and arrange() function sorts a variable in descending order.

```
library(tibble)
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.1.3
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(class)
library(e1071)

CDC <- read.csv("cdc.csv")
# Basic R functions for data exploration
str(CDC)
```

```
## 'data.frame':    53392 obs. of  33 variables:
##  $ YearStart                 : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ YearEnd                   : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ LocationAbbr              : chr  "AL" "AL" "AL" "AL" ...
##  $ LocationDesc              : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
##  $ Datasource                : chr  "Behavioral Risk Factor Surveillance System" "Behavioral Risk Fa
##  $ Class                     : chr  "Obesity / Weight Status" "Obesity / Weight Status" "Obesity / W
##  $ Topic                     : chr  "Obesity / Weight Status" "Obesity / Weight Status" "Obesity / W
##  $ Question                  : chr  "Percent of adults aged 18 years and older who have obesity" "Pe
##  $ Data_Value_Unit           : logi  NA NA NA NA NA NA ...
##  $ Data_Value_Type           : chr  "Value" "Value" "Value" "Value" ...
##  $ Data_Value                : num  32 32.3 31.8 33.6 32.8 33.8 26.4 16.3 35.2 35.5 ...
##  $ Data_Value_Alt            : num  32 32.3 31.8 33.6 32.8 33.8 26.4 16.3 35.2 35.5 ...
##  $ Data_Value_Footnote_Symbol: chr  "" "" "" "" ...
##  $ Data_Value_Footnote       : chr  "" "" "" "" ...
##  $ Low_Confidence_Limit      : num  30.5 29.9 30 29.9 30.2 31 23.7 12.6 30.7 31.6 ...
##  $ High_Confidence_Limit     : num  33.5 34.7 33.6 37.6 35.6 36.8 29.3 20.9 40 39.6 ...
##  $ Sample_Size               : int  7304 2581 4723 1153 2402 1925 1812 356 598 865 ...
##  $ Total                     : chr  "Total" "" "" "" ...
##  $ Age.years.                : chr  "" "" "" "" ...
##  $ Education                 : chr  "" "" "" "Less than high school" ...
##  $ Gender                    : chr  "" "Male" "Female" "" ...
##  $ Income                    : chr  "" "" "" "" ...
##  $ Race.Ethnicity            : chr  "" "" "" "" ...
##  $ GeoLocation               : chr  "(32.84057112200048, -86.63186076199969)" "(32.84057112200048, -8
##  $ ClassID                   : chr  "OWS" "OWS" "OWS" "OWS" ...
##  $ TopicID                   : chr  "OWS1" "OWS1" "OWS1" "OWS1" ...
##  $ QuestionID                : chr  "Q036" "Q036" "Q036" "Q036" ...
##  $ DataValueTypeID           : chr  "VALUE" "VALUE" "VALUE" "VALUE" ...
##  $ LocationID                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ StratificationCategory1   : chr  "Total" "Gender" "Gender" "Education" ...
##  $ Stratification1           : chr  "Total" "Male" "Female" "Less than high school" ...
##  $ StratificationCategoryId1 : chr  "OVR" "GEN" "GEN" "EDU" ...
##  $ StratificationID1         : chr  "OVERALL" "MALE" "FEMALE" "EDUHS" ...
```

```
names(CDC)
```

```
##  [1] "YearStart"                 "YearEnd"
##  [3] "LocationAbbr"              "LocationDesc"
##  [5] "Datasource"                "Class"
##  [7] "Topic"                     "Question"
```

```
##  [9] "Data_Value_Unit"            "Data_Value_Type"
## [11] "Data_Value"                 "Data_Value_Alt"
## [13] "Data_Value_Footnote_Symbol" "Data_Value_Footnote"
## [15] "Low_Confidence_Limit"       "High_Confidence_Limit"
## [17] "Sample_Size"                "Total"
## [19] "Age.years."                 "Education"
## [21] "Gender"                     "Income"
## [23] "Race.Ethnicity"             "GeoLocation"
## [25] "ClassID"                    "TopicID"
## [27] "QuestionID"                 "DataValueTypeID"
## [29] "LocationID"                 "StratificationCategory1"
## [31] "Stratification1"            "StratificationCategoryId1"
## [33] "StratificationID1"
```

```
summary(CDC)
```

```
##    YearStart       YearEnd     LocationAbbr       LocationDesc
##  Min.   :2011   Min.   :2011   Length:53392       Length:53392
##  1st Qu.:2012   1st Qu.:2012   Class :character   Class :character
##  Median :2013   Median :2013   Mode  :character   Mode  :character
##  Mean   :2013   Mean   :2013
##  3rd Qu.:2015   3rd Qu.:2015
##  Max.   :2016   Max.   :2016
##
##   Datasource           Class              Topic             Question
##  Length:53392       Length:53392       Length:53392       Length:53392
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Data_Value_Unit Data_Value_Type      Data_Value    Data_Value_Alt
##  Mode:logical    Length:53392       Min.   : 0.90   Min.   : 0.90
##  NA's:53392      Class :character   1st Qu.:24.10   1st Qu.:24.10
##                  Mode  :character   Median :30.70   Median :30.70
##                                     Mean   :31.16   Mean   :31.16
##                                     3rd Qu.:37.00   3rd Qu.:37.00
##                                     Max.   :77.60   Max.   :77.60
##                                     NA's   :5046    NA's   :5046
##  Data_Value_Footnote_Symbol Data_Value_Footnote Low_Confidence_Limit
##  Length:53392               Length:53392        Min.   : 0.30
##  Class :character           Class :character    1st Qu.:20.00
##  Mode  :character           Mode  :character    Median :26.45
##                                                 Mean   :26.89
##                                                 3rd Qu.:32.90
##                                                 Max.   :69.50
##                                                 NA's   :5046
##  High_Confidence_Limit  Sample_Size        Total            Age.years.
##  Min.   : 3.00         Min.   :    50   Length:53392       Length:53392
##  1st Qu.:28.20         1st Qu.:   566   Class :character   Class :character
##  Median :35.60         Median :  1209   Mode  :character   Mode  :character
##  Mean   :35.99         Mean   :  3889
##  3rd Qu.:42.20         3rd Qu.:  2519
```

```
##  Max.   :87.70          Max.   :476876
##  NA's   :5046           NA's   :5046
##   Education             Gender              Income          Race.Ethnicity
##  Length:53392       Length:53392       Length:53392       Length:53392
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  GeoLocation           ClassID            TopicID           QuestionID
##  Length:53392       Length:53392       Length:53392       Length:53392
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  DataValueTypeID      LocationID    StratificationCategory1 Stratification1
##  Length:53392       Min.   : 1.00   Length:53392            Length:53392
##  Class :character   1st Qu.:17.00   Class :character        Class :character
##  Mode  :character   Median :30.00   Mode  :character        Mode  :character
##                     Mean   :30.28
##                     3rd Qu.:44.00
##                     Max.   :78.00
##
##  StratificationCategoryId1 StratificationID1
##  Length:53392              Length:53392
##  Class :character          Class :character
##  Mode  :character          Mode  :character
##
##
##
##
```

```
head(CDC)
```

```
##   YearStart YearEnd LocationAbbr LocationDesc
## 1      2011    2011           AL      Alabama
## 2      2011    2011           AL      Alabama
## 3      2011    2011           AL      Alabama
## 4      2011    2011           AL      Alabama
## 5      2011    2011           AL      Alabama
## 6      2011    2011           AL      Alabama
##                                   Datasource                    Class
## 1 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 2 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 3 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 4 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 5 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 6 Behavioral Risk Factor Surveillance System Obesity / Weight Status
##                     Topic
## 1 Obesity / Weight Status
## 2 Obesity / Weight Status
```

```
## 3 Obesity / Weight Status
## 4 Obesity / Weight Status
## 5 Obesity / Weight Status
## 6 Obesity / Weight Status
##                                                 Question Data_Value_Unit
## 1 Percent of adults aged 18 years and older who have obesity              NA
## 2 Percent of adults aged 18 years and older who have obesity              NA
## 3 Percent of adults aged 18 years and older who have obesity              NA
## 4 Percent of adults aged 18 years and older who have obesity              NA
## 5 Percent of adults aged 18 years and older who have obesity              NA
## 6 Percent of adults aged 18 years and older who have obesity              NA
##   Data_Value_Type Data_Value Data_Value_Alt Data_Value_Footnote_Symbol
## 1           Value       32.0           32.0
## 2           Value       32.3           32.3
## 3           Value       31.8           31.8
## 4           Value       33.6           33.6
## 5           Value       32.8           32.8
## 6           Value       33.8           33.8
##   Data_Value_Footnote Low_Confidence_Limit High_Confidence_Limit Sample_Size
## 1                                    30.5                  33.5        7304
## 2                                    29.9                  34.7        2581
## 3                                    30.0                  33.6        4723
## 4                                    29.9                  37.6        1153
## 5                                    30.2                  35.6        2402
## 6                                    31.0                  36.8        1925
##   Total Age.years.                      Education Gender Income
## 1 Total
## 2                                            Male
## 3                                          Female
## 4                          Less than high school
## 5                           High school graduate
## 6                Some college or technical school
##   Race.Ethnicity                            GeoLocation ClassID TopicID
## 1                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 2                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 3                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 4                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 5                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 6                (32.84057112200048, -86.63186076199969)     OWS    OWS1
##   QuestionID DataValueTypeID LocationID StratificationCategory1
## 1       Q036           VALUE          1                   Total
## 2       Q036           VALUE          1                  Gender
## 3       Q036           VALUE          1                  Gender
## 4       Q036           VALUE          1               Education
## 5       Q036           VALUE          1               Education
## 6       Q036           VALUE          1               Education
##                    Stratification1 StratificationCategoryId1 StratificationID1
## 1                            Total                       OVR           OVERALL
## 2                             Male                       GEN              MALE
## 3                           Female                       GEN            FEMALE
## 4            Less than high school                       EDU             EDUHS
## 5             High school graduate                       EDU         EDUHSGRAD
## 6 Some college or technical school                       EDU          EDUCOTEC
```

```
tail(CDC)
```

```
##       YearStart YearEnd LocationAbbr   LocationDesc
## 53387      2016    2016           VI Virgin Islands
## 53388      2016    2016           VI Virgin Islands
## 53389      2016    2016           VI Virgin Islands
## 53390      2016    2016           VI Virgin Islands
## 53391      2016    2016           VI Virgin Islands
## 53392      2016    2016           VI Virgin Islands
##                                     Datasource             Class
## 53387 Behavioral Risk Factor Surveillance System Physical Activity
## 53388 Behavioral Risk Factor Surveillance System Physical Activity
## 53389 Behavioral Risk Factor Surveillance System Physical Activity
## 53390 Behavioral Risk Factor Surveillance System Physical Activity
## 53391 Behavioral Risk Factor Surveillance System Physical Activity
## 53392 Behavioral Risk Factor Surveillance System Physical Activity
##                              Topic
## 53387 Physical Activity - Behavior
## 53388 Physical Activity - Behavior
## 53389 Physical Activity - Behavior
## 53390 Physical Activity - Behavior
## 53391 Physical Activity - Behavior
## 53392 Physical Activity - Behavior
##                                                                 Question
## 53387 Percent of adults who engage in no leisure-time physical activity
## 53388 Percent of adults who engage in no leisure-time physical activity
## 53389 Percent of adults who engage in no leisure-time physical activity
## 53390 Percent of adults who engage in no leisure-time physical activity
## 53391 Percent of adults who engage in no leisure-time physical activity
## 53392 Percent of adults who engage in no leisure-time physical activity
##       Data_Value_Unit Data_Value_Type Data_Value Data_Value_Alt
## 53387              NA           Value       30.3           30.3
## 53388              NA           Value         NA             NA
## 53389              NA           Value         NA             NA
## 53390              NA           Value         NA             NA
## 53391              NA           Value         NA             NA
## 53392              NA           Value         NA             NA
##       Data_Value_Footnote_Symbol
## 53387
## 53388                          ~
## 53389                          ~
## 53390                          ~
## 53391                          ~
## 53392                          ~
##                                        Data_Value_Footnote
## 53387
## 53388 Data not available because sample size is insufficient.
## 53389 Data not available because sample size is insufficient.
## 53390 Data not available because sample size is insufficient.
## 53391 Data not available because sample size is insufficient.
## 53392 Data not available because sample size is insufficient.
##       Low_Confidence_Limit High_Confidence_Limit Sample_Size Total Age.years.
## 53387                 20.2                  42.9         178
```

```
## 53388                   NA                     NA          NA
## 53389                   NA                     NA          NA
## 53390                   NA                     NA          NA
## 53391                   NA                     NA          NA
## 53392                   NA                     NA          NA
##       Education Gender Income                Race.Ethnicity
## 53387                                              Hispanic
## 53388                                                 Asian
## 53389                            Hawaiian/Pacific Islander
## 53390                          American Indian/Alaska Native
## 53391                                       2 or more races
## 53392                                                 Other
##                   GeoLocation ClassID TopicID QuestionID DataValueTypeID
## 53387 (18.335765, -64.896335)      PA     PA1       Q047           VALUE
## 53388 (18.335765, -64.896335)      PA     PA1       Q047           VALUE
## 53389 (18.335765, -64.896335)      PA     PA1       Q047           VALUE
## 53390 (18.335765, -64.896335)      PA     PA1       Q047           VALUE
## 53391 (18.335765, -64.896335)      PA     PA1       Q047           VALUE
## 53392 (18.335765, -64.896335)      PA     PA1       Q047           VALUE
##       LocationID StratificationCategory1               Stratification1
## 53387         78          Race/Ethnicity                      Hispanic
## 53388         78          Race/Ethnicity                         Asian
## 53389         78          Race/Ethnicity     Hawaiian/Pacific Islander
## 53390         78          Race/Ethnicity American Indian/Alaska Native
## 53391         78          Race/Ethnicity               2 or more races
## 53392         78          Race/Ethnicity                         Other
##       StratificationCategoryId1 StratificationID1
## 53387                      RACE           RACEHIS
## 53388                      RACE           RACEASN
## 53389                      RACE           RACEHPI
## 53390                      RACE           RACENAA
## 53391                      RACE         RACE2PLUS
## 53392                      RACE           RACEOTH
```

```
CDC2 <- select(CDC, YearEnd, LocationDesc, LocationAbbr, Question, Sample_Size, Data_Value, Low_Confide
as_tibble(CDC2)
```

```
## # A tibble: 53,392 x 16
##    YearEnd LocationDesc LocationAbbr Question             Sample_Size Data_Value
##      <int> <chr>        <chr>        <chr>                      <int>      <dbl>
##  1    2011 Alabama      AL           Percent of adults a~        7304       32
##  2    2011 Alabama      AL           Percent of adults a~        2581       32.3
##  3    2011 Alabama      AL           Percent of adults a~        4723       31.8
##  4    2011 Alabama      AL           Percent of adults a~        1153       33.6
##  5    2011 Alabama      AL           Percent of adults a~        2402       32.8
##  6    2011 Alabama      AL           Percent of adults a~        1925       33.8
##  7    2011 Alabama      AL           Percent of adults a~        1812       26.4
##  8    2011 Alabama      AL           Percent of adults a~         356       16.3
##  9    2011 Alabama      AL           Percent of adults a~         598       35.2
## 10    2011 Alabama      AL           Percent of adults a~         865       35.5
## # ... with 53,382 more rows, and 10 more variables: Low_Confidence_Limit <dbl>,
## #   Education <chr>, Gender <chr>, Income <chr>, Age.years. <chr>,
## #   Race.Ethnicity <chr>, Stratification1 <chr>,
## #   StratificationCategoryId1 <chr>, StratificationID1 <chr>, GeoLocation <chr>
```

```
CDC_overweight <- select(CDC2, YearEnd, LocationDesc, LocationAbbr, Question, Sample_Size, Data_Value, S

CDC_overweight <- filter(CDC_overweight, Question == "Percent of adults aged 18 years and older who have

CDC_obese <- select(CDC2, YearEnd, LocationDesc, LocationAbbr, Question, Sample_Size, Data_Value, Strati

CDC_obese <- filter(CDC_obese, Question == "Percent of adults aged 18 years and older who have obesity"

arrange(CDC_overweight, YearEnd)
```

```
##   YearEnd LocationDesc LocationAbbr
## 1    2011     National           US
## 2    2012     National           US
## 3    2013     National           US
## 4    2014     National           US
## 5    2015     National           US
## 6    2016     National           US
##                                                                                 Question
## 1 Percent of adults aged 18 years and older who have an overweight classification
## 2 Percent of adults aged 18 years and older who have an overweight classification
## 3 Percent of adults aged 18 years and older who have an overweight classification
## 4 Percent of adults aged 18 years and older who have an overweight classification
## 5 Percent of adults aged 18 years and older who have an overweight classification
## 6 Percent of adults aged 18 years and older who have an overweight classification
##   Sample_Size Data_Value Stratification1
## 1      470531       35.8           Total
## 2      442230       35.7           Total
## 3      457487       35.5           Total
## 4      425875       35.2           Total
## 5      398316       35.7           Total
## 6      438479       35.2           Total
```

```
arrange(CDC_obese, YearEnd)
```

```
##   YearEnd LocationDesc LocationAbbr
## 1    2011     National           US
## 2    2012     National           US
## 3    2013     National           US
## 4    2014     National           US
## 5    2015     National           US
## 6    2016     National           US
##                                                 Question Sample_Size
## 1 Percent of adults aged 18 years and older who have obesity      470700
## 2 Percent of adults aged 18 years and older who have obesity      442230
## 3 Percent of adults aged 18 years and older who have obesity      457487
## 4 Percent of adults aged 18 years and older who have obesity      425875
## 5 Percent of adults aged 18 years and older who have obesity      398316
## 6 Percent of adults aged 18 years and older who have obesity      438479
##   Data_Value Stratification1
## 1       27.4           Total
## 2       27.7           Total
## 3       28.3           Total
```

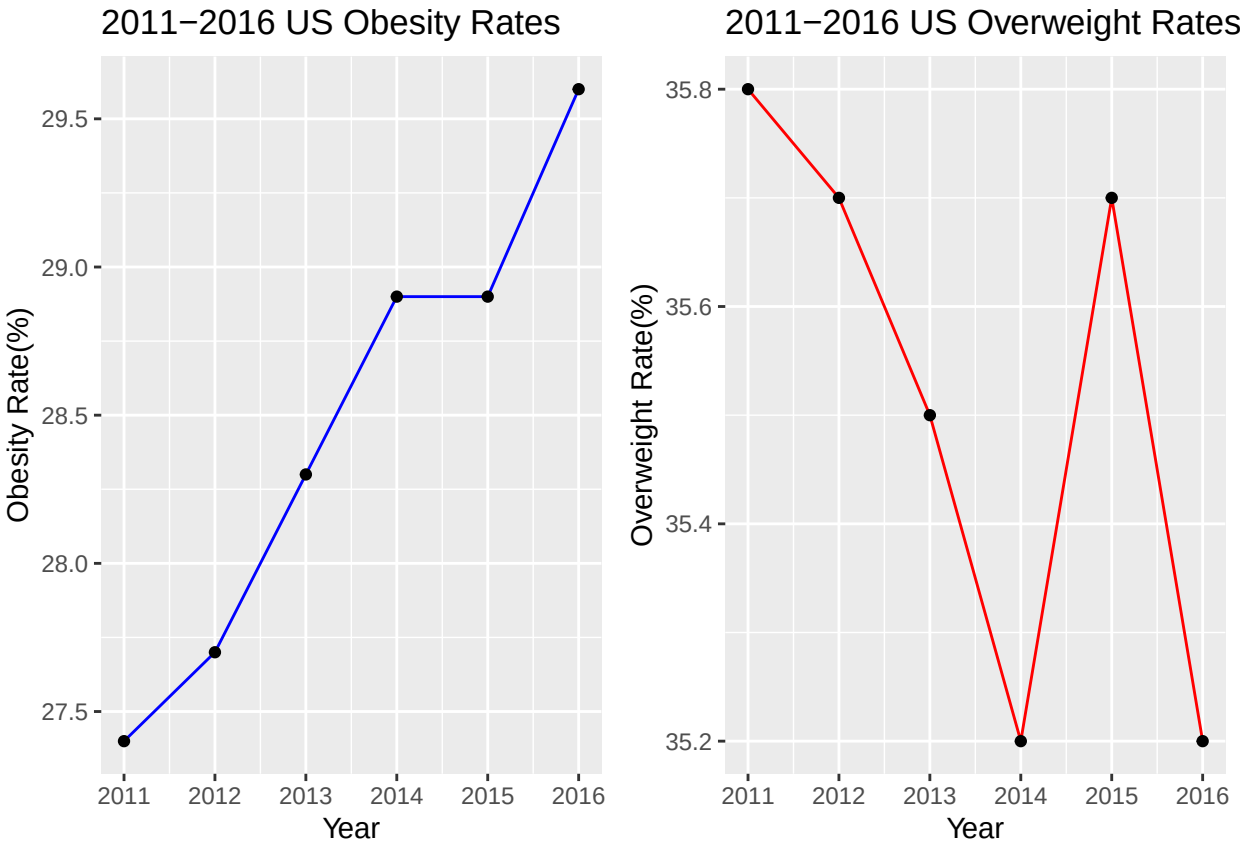
```
## 4       28.9           Total
## 5       28.9           Total
## 6       29.6           Total
```

Graphs comment: According to the graph, overweight rates have decreased until 2014, which infers that Americans are losing weight annually. However, the obesity rates have been increasing throughout the years, so Americans are gaining weight. In other words, people classified as overweight are transitioning to obese.

```
overweight_plot <- ggplot(data=CDC_overweight, aes(x = YearEnd, y = Data_Value, group=1)) +
geom_line(color="red")+
geom_point()+labs(title = "2011-2016 US Overweight Rates", x = "Year", y = "Overweight Rate(%)")

obese_plot  <- ggplot(data=CDC_obese, aes(x = YearEnd, y = Data_Value, group=1)) +
geom_line(color="blue")+
geom_point()+labs(title = "2011-2016 US Obesity Rates", x = "Year", y = "Obesity Rate(%)")

grid.arrange(arrangeGrob(obese_plot, overweight_plot, ncol = 2))
```



Linear regression comment: I decided to choose these features because it seemed logical to assume that there would be a positive relationship between adults not consuming vegetables and the obesity rate. As a result, I wanted to affirm that my hypothesis was correct. The linear regression verifies that not eating vegetables could lead to an increase in obesity. Obs_rate = 0.04512 * Veg + 28.34208. As the percentage of not eating vegetables increases, so will the obesity rate. The RSE is 7.143, making the percentage error (7.143/50.81648) * 100 = 14.056. The RMSE states that our test data is off by a 7.128568 obesity rate on average. Everything looks decent except the fact that the p-value is not very small.

```r
Veg <- select(CDC2, Question, Data_Value)
Veg <- filter(Veg, Question=="Percent of adults who report consuming vegetables less than one time daily
Veg <- select(Veg, Data_Value)

Obs_rate <- select(CDC2, Question, Data_Value)
Obs_rate <- filter(Obs_rate, Question == "Percent of adults aged 18 years and older who have obesity")
Obs_rate <- select(Obs_rate, Data_Value)

Veg <- na.omit(Veg)
Obs_rate <- na.omit(Obs_rate)

Veg <- unlist(Veg)
Obs_rate <- unlist(Obs_rate)

Veg <- Veg[-c(500:3996)]
Obs_rate <- Obs_rate[-c(500:8127)]

beg <- proc.time()
lm1 <- lm(Obs_rate~Veg)
end <- proc.time()
summary(lm1)
```

```
##
## Call:
## lm(formula = Obs_rate ~ Veg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7611  -3.8328   0.1856   4.7247  24.7887
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.34208    1.07104  26.462   <2e-16 ***
## Veg          0.04512    0.04409   1.023    0.307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.143 on 497 degrees of freedom
## Multiple R-squared:  0.002103,   Adjusted R-squared:  9.5e-05
## F-statistic: 1.047 on 1 and 497 DF,  p-value: 0.3066
```

```r
mse1 <- mean(lm1$residuals^2)
mse1
```

```
## [1] 50.81648
```

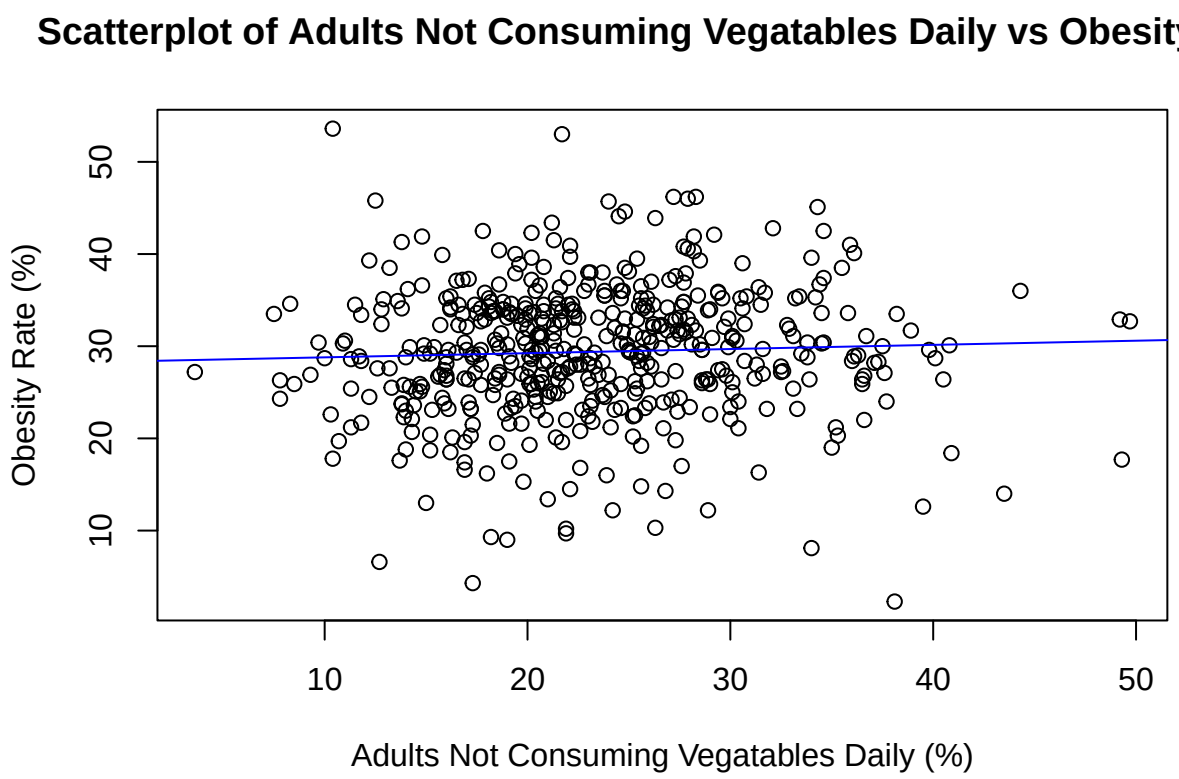
```r
rmse1 <- sqrt(mse1)
rmse1
```

```
## [1] 7.128568
```

```
time <- end - beg
time
```

```
##    user  system elapsed
##       0       0       0
```

```
plot(Veg, Obs_rate, main = "Scatterplot of Adults Not Consuming Vegatables Daily vs Obesity Rate", xlab
abline(lm1, col = "blue")
```



kNN regression comment: I would think the percentage of people not eating vegetables and not exercising would affect the obesity rate. However, the accuracy is not high even with the most optimal k-value, so we can assume that kNN might not be the right approach here. The time performance is a bit slower than linear regression as well. Another reason is that the data wasn't correctly scaled. A solution to find the best k is to use a for loop and traverse the kNN function within it. I manually changed the k value because the range for my k is not substantial.

```
Phy <- select(CDC2, Question, Data_Value)
Phy <- filter(Phy, Question == "Percent of adults who engage in no leisure-time physical activity")
Phy <- select(Phy, Data_Value)
Phy <- na.omit(Phy)

beg <- proc.time()
knn_pred <- knn(Veg, Phy, Obs_rate, k = 3)
end <- proc.time()
knn_pred <- knn_pred[-c(500:8127)]
```

```
knn_pred <- as.numeric(knn_pred)
mean(knn_pred == Obs_rate)
```

```
## [1] 0.002004008
```

```
time <- end - beg
time
```

```
##    user  system elapsed
##    0.01    0.00    0.02
```

SVM regression comment: I made gender a factor related to the year, sample size, data value, and confidence limits. I chose gender to be a factor because it is a binary value where (0 equals female and 1 equals male) so it can provide some valuable data combined with other data elements. The runtime is two milliseconds slower than kNN. Nonetheless, the accuracy is 88.4%, so the results suggest that our model is a good classifier for the data.

```
CDC3 <- select(CDC, YearEnd, Sample_Size, Data_Value, Low_Confidence_Limit, High_Confidence_Limit, Gende
CDC3$Gender <- as.factor(CDC3$Gender)

CDC3 <- CDC3[-c(500:nrow(CDC3)), ]
CDC3 <- CDC3[!apply(CDC3 == "", 1, all), ]
CDC3 <- na.omit(CDC3)

set.seed(1234)
i <- sample(1:nrow(CDC3), 0.75*nrow(CDC3), replace=FALSE)
train <- CDC3[i,]
test <- CDC3[-i,]

beg <- proc.time()
svm1 <- svm(Gender~ ., data = train, kernel = "linear", cost = 10, scale = TRUE)
end <- proc.time()
summary(svm1)
```

```
##
## Call:
## svm(formula = Gender ~ ., data = train, kernel = "linear", cost = 10,
##     scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  10
##
## Number of Support Vectors:  49
##
##  ( 16 19 14 )
##
##
## Number of Classes:  3
##
```

```
## Levels:
##   Female Male
```

```
pred <- predict(svm1, newdata = test)
table(pred == test$Gender)
```

```
##
## FALSE  TRUE
##    13    99
```

```
mean(pred == test$Gender)
```

```
## [1] 0.8839286
```

```
time <- end - beg
time
```

```
##    user  system elapsed
##    0.03    0.00    0.03
```

Results analysis: The processing order goes from linear regression, kNN, and SVM, where linear regression is the fastest. Linear regression is interpretable and straightforward but sensitive to outliers. kNN does not assume the shape of the data compared to linear regression, but extra steps like scaling the data are necessary to get good results. SVM deals with complex calculations and decision boundaries, so computing results prove to be stressful. Data doesn't show a correlation between not eating vegetables and not exercising with obesity prevalence. However, there could be some computation error or data faults. This would make sense because the energy obtained from eating foods is not expended. Thus, logically speaking, not eating vegetables and not exercising would correlate to the obesity rate in some way or another.