# Regression

Umar Ali-Salaam

9/25/2022

Source: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

This is a dataset based off of 70,000 records of patient data (Heart Related). Columns (13): ID, Age, Height(cm), Weight(kg), Gender, Systolic Blood Pressure (AP_HIGH), Diastolic Blood Pressure (AP LOW), Cholesterol, Glucose, Smoking, Alcohol Intake, Physical Activity, Presence or Absence of cardiovascular disease.

The .csv file needed to be edited a bit in Microsoft Excel before using it in R. I just performed a split column delimiter function around semicolons, to divide the singular column that existed into 13. Each row had 13 variables in 1 column separated by semicolons, the function I ran split it up into 13 columns, making a 70,000 x 13 table.

https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Visit the website above to better understand Systolic and Diastolic Blood Pressure

```
library(ggplot2)

# Read in .csv file
heart <- read.csv("cardio_train.csv")


# Clean out any rows that have an unrealistic blood pressure (AP_HIGH & AP_LOW)
# They looked to be input errors by the person who made the data set
s <- subset(heart, AP_HIGH > 50)

s1 <- subset(s, AP_HIGH < 200)

s2 <- subset(s1, AP_LOW > 25)

s3 <- subset(s2, AP_LOW < 200)


# Removing little people(4'10") and Giants(7'3"), values are in cm
s4 <- subset(s3, HEIGHT > 147)

s5 <- subset(s4, HEIGHT < 220)


# Removing anyone below 90 lbs and above 375 lbs, the values are in kg
s6 <- subset(s5, WEIGHT > 40)

h1 <- subset(s6, WEIGHT < 180)
```

```r
# AGE is in days so to get years i just divide by 365
h1$AGE <- (h1$AGE / 365)

# Removing people under 40
h <- subset(h1, AGE > 39)

# Checking for any NA values
# There is none
colSums(is.na(h))
```

```
##                ID               AGE            GENDER            HEIGHT
##                 0                 0                 0                 0
##            WEIGHT           AP_HIGH            AP_LOW       CHOLESTEROL
##                 0                 0                 0                 0
##           GLUCOSE             SMOKE           ALCOHOL PHYSICAL_ACTIVITY
##                 0                 0                 0                 0
##    CARDIO_DISEASE
##                 0
```

```r
# Everything that should be factored is factored
h$GENDER <- factor(h$GENDER)
h$CHOLESTEROL <- factor(h$CHOLESTEROL)
h$GLUCOSE <- factor(h$GLUCOSE)
h$SMOKE <- factor(h$SMOKE)
h$ALCOHOL <- factor(h$ALCOHOL)
h$PHYSICAL_ACTIVITY <- factor(h$PHYSICAL_ACTIVITY)
h$CARDIO_DISEASE <- factor(h$CARDIO_DISEASE)


# There is now 67,685 rows
```

Splitting the data into an 80/20 split

```r
#Split data in 80/20 train/test
splitt <- round(nrow(h) * 0.8)

train <- h[1:splitt,]

test <- h[(splitt + 1):nrow(h),]
```

Performing tests: Pearson Correlation, summary, str, Distribution of age and gender

```r
# Pearson correlation test between systolic and diastolic blood pressure
cor.test(train$AP_HIGH, train$AP_LOW)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$AP_HIGH and train$AP_LOW
## t = 218.81, df = 54146, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.6805382 0.6894789
## sample estimates:
##       cor
## 0.6850343
```

```
# Summary revealing distributions of the training data
summary(train)
```

```
##        ID              AGE        GENDER        HEIGHT          WEIGHT
##  Min.   :    0   Min.   :39.11   1:35215   Min.   :148.0   Min.   : 41.00
##  1st Qu.:19970   1st Qu.:48.34   2:18933   1st Qu.:159.0   1st Qu.: 65.00
##  Median :39994   Median :53.96             Median :165.0   Median : 72.00
##  Mean   :39968   Mean   :53.29             Mean   :164.7   Mean   : 74.25
##  3rd Qu.:59969   3rd Qu.:58.40             3rd Qu.:170.0   3rd Qu.: 82.00
##  Max.   :79861   Max.   :64.91             Max.   :207.0   Max.   :178.00
##     AP_HIGH          AP_LOW       CHOLESTEROL GLUCOSE   SMOKE     ALCOHOL
##  Min.   : 60.0   Min.   : 30.00   1:40700     1:46052   0:49358   0:51252
##  1st Qu.:120.0   1st Qu.: 80.00   2: 7332     2: 3977   1: 4790   1: 2896
##  Median :120.0   Median : 80.00   3: 6116     3: 4119
##  Mean   :126.4   Mean   : 81.33
##  3rd Qu.:140.0   3rd Qu.: 90.00
##  Max.   :197.0   Max.   :190.00
##  PHYSICAL_ACTIVITY CARDIO_DISEASE
##  0:10648           0:27414
##  1:43500           1:26734
##
##
##
##
```

```
# Revealing which variable types are in the data set
str(train)
```

```
## 'data.frame':    54148 obs. of  13 variables:
##  $ ID                : int  0 1 2 3 4 8 9 12 13 14 ...
##  $ AGE               : num  50.4 55.4 51.7 48.3 47.9 ...
##  $ GENDER            : Factor w/ 2 levels "1","2": 2 1 1 2 1 1 1 2 1 1 ...
##  $ HEIGHT            : int  168 156 165 169 156 151 157 178 158 164 ...
##  $ WEIGHT            : num  62 85 64 82 56 67 93 95 71 68 ...
##  $ AP_HIGH           : int  110 140 130 150 100 120 130 130 110 110 ...
##  $ AP_LOW            : int  80 90 70 100 60 80 80 90 70 60 ...
##  $ CHOLESTEROL       : Factor w/ 3 levels "1","2","3": 1 3 3 1 1 2 3 3 1 1 ...
##  $ GLUCOSE           : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 1 3 1 1 ...
##  $ SMOKE             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ALCOHOL           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PHYSICAL_ACTIVITY: Factor w/ 2 levels "0","1": 2 2 1 2 1 1 2 2 2 1 ...
##  $ CARDIO_DISEASE    : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 1 1 ...
```

```
# Finding the distribution of men and women
women <- which(train$GENDER == 1)
men <- which(train$GENDER == 2)
```

```
numW <- length(women) / (length(men) + length(women))

numM <- length(men) / (length(men) + length(women))

cat("\n The percentage of women in this data set is", round(numW * 100, 4),"%.\n",
"The percentage of men in this data set is", round(numM * 100, 4), "%.\n",
"There is a clear bias towards women in this data set, we have to keep that in\n",
 "mind going forward.")
```

```
##
##  The percentage of women in this data set is 65.0347 %.
##  The percentage of men in this data set is 34.9653 %.
##  There is a clear bias towards women in this data set, we have to keep that in
##  mind going forward.
```

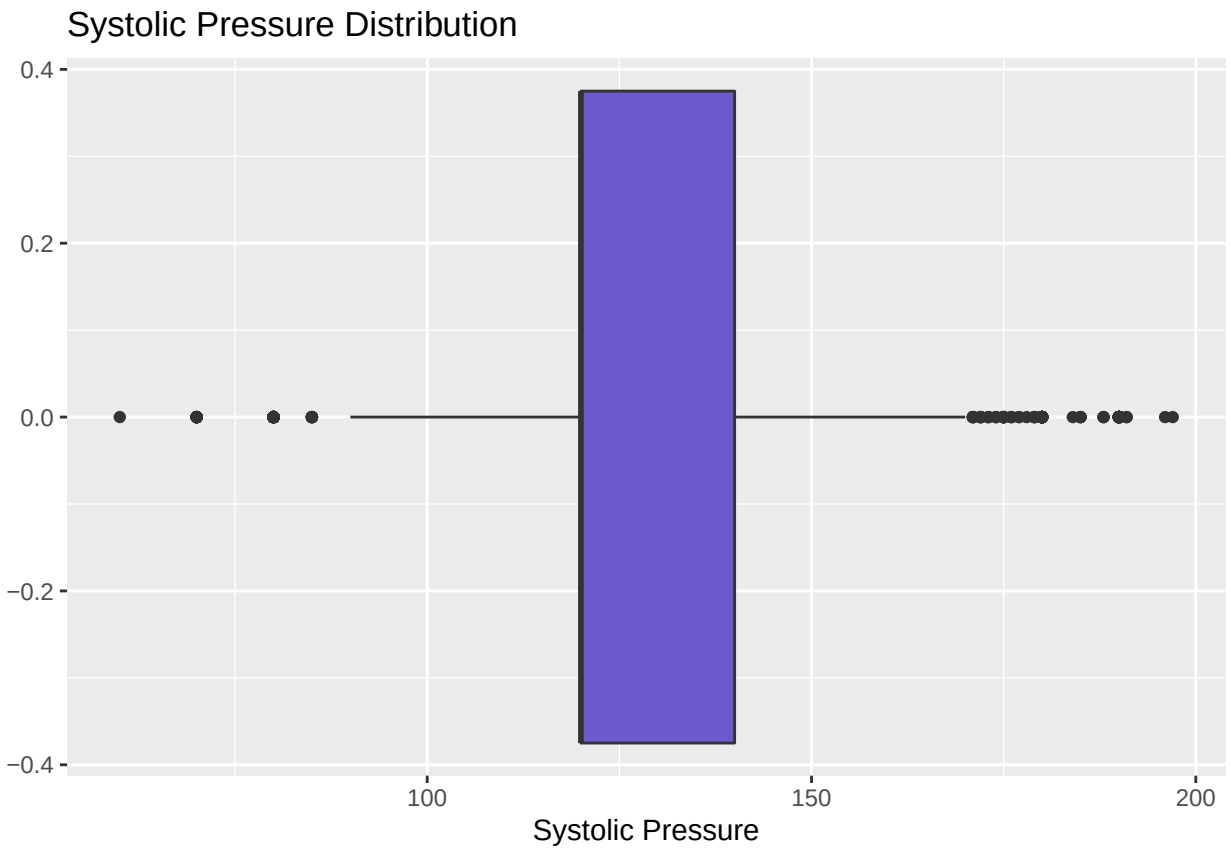
```
# Understanding the distribution of age
hist(train$AGE)
```


```
cat("\nFairly normal curve in terms of age, slightly skewed left (younger)")
```

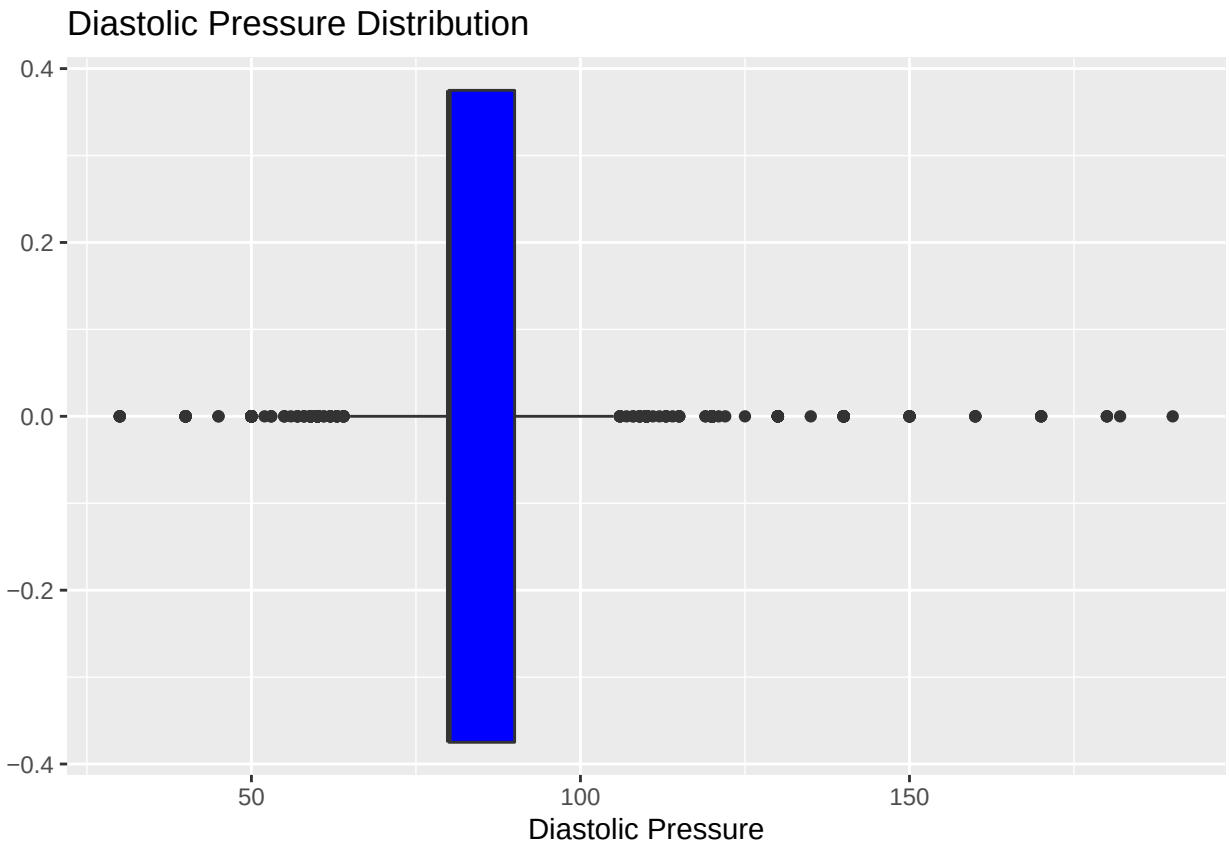
```
##
## Fairly normal curve in terms of age, slightly skewed left (younger)
```

Graphs for finding distributions and relationships

```
# Looking at distribution of Systolic and Diastolic Pressure
ggplot(train, aes(x = AP_HIGH)) +
  geom_boxplot(fill = "slateblue") +
  xlab("Systolic Pressure") +
  ggtitle("Systolic Pressure Distribution")
```

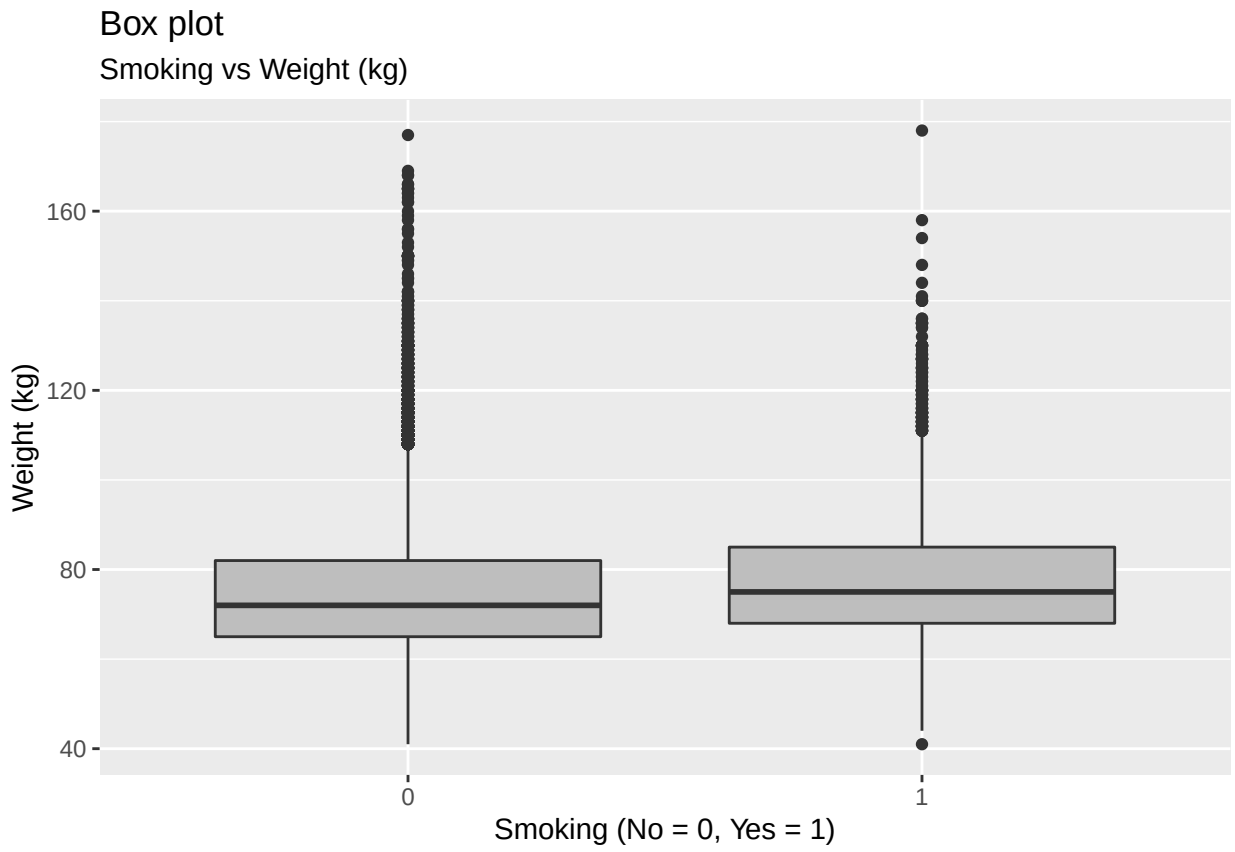



```
ggplot(train, aes(x = AP_LOW)) +
  geom_boxplot(fill = "blue") +
  xlab("Diastolic Pressure") +
  ggtitle("Diastolic Pressure Distribution")
```




```
# Finding distribution of age and gender
  ggplot(train, aes(GENDER, AGE)) +
   geom_violin(fill = "Pink") +
  labs(title = "Violin plot",
       subtitle = "Gender vs Age",
       x = "Gender (1 = Female, 2 = Male)",
       y = "Age")
```


```
# Finding correlation between smoking and weight
   ggplot(train, aes(SMOKE, WEIGHT)) +
   geom_boxplot(fill = "gray") +
  labs(title = "Box plot",
       subtitle = "Smoking vs Weight (kg)",
       x = "Smoking (No = 0, Yes = 1)",
       y = "Weight (kg)")
```




First Linear Regression Model

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Relationship between weight and Systolic Blood Pressure
# Systolic Blood Pressure (AP_HIGH) is the blood pressure while the heart has
# pumped blood, Diastolic Pressure is when it's between each heart beat.
fit = lm(WEIGHT ~ AP_HIGH, train)

summary(fit)
```

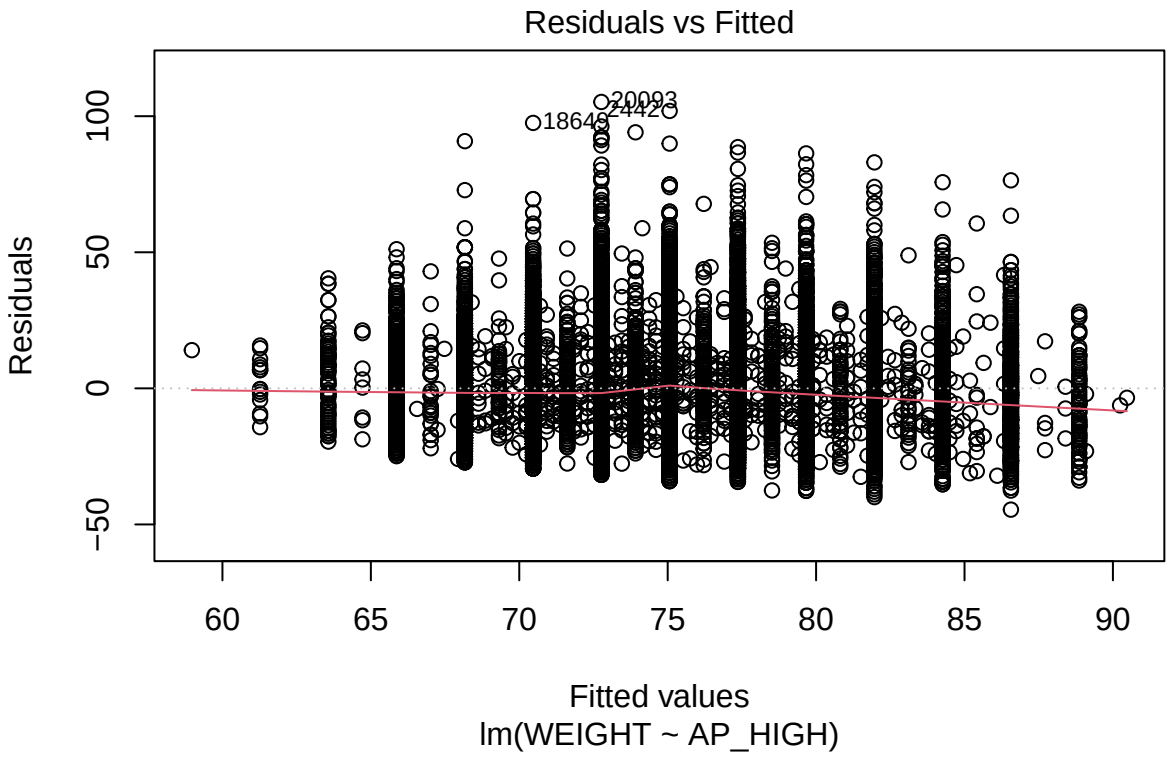
```
##
## Call:
```

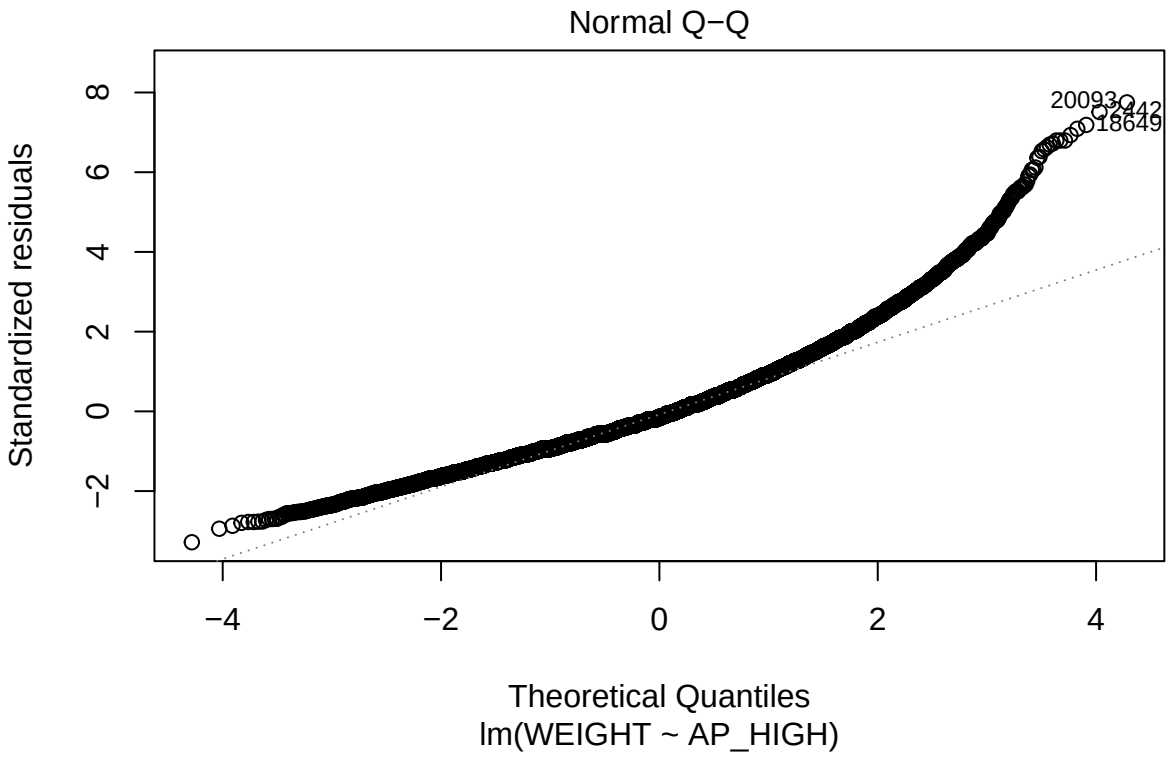
```
## lm(formula = WEIGHT ~ AP_HIGH, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.563  -9.365  -2.066   7.234 105.234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.17231    0.45519   99.24   <2e-16 ***
## AP_HIGH      0.22995    0.00357   64.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.57 on 54146 degrees of freedom
## Multiple R-squared:  0.07116,    Adjusted R-squared:  0.07114
## F-statistic:  4148 on 1 and 54146 DF,  p-value: < 2.2e-16
```

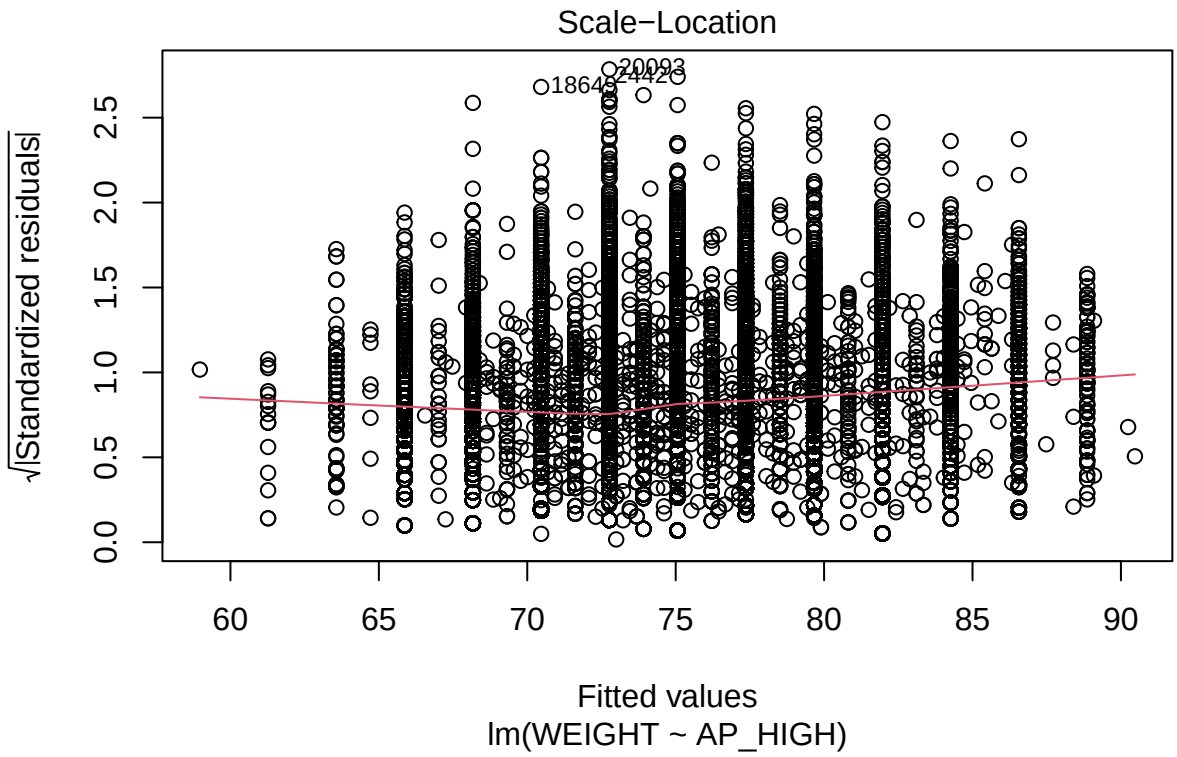
```
cat("The p-value of the F-statistic is < 2.2e^(-16) which is good, it means that\n
there is at least one data point that's significantly related to the outcome\n
variable. The median residual is somewhat close to 0 at -2.066 which is somewhat\n
symmetrical. The higher values seem to stray away from the line the most. The\n
residual standard error is 13.57, having all predictions be off by and average\n
of 13.57 units of Systolic Blood Pressure, won't really produce and accurate\n
model. Since we're only really looking at one predictor, we can use the multiple\n
R-squared value of 0.07116. This shows how much the independent variable can \n
explain the dependent variable, and 7.116% is fairly awful in my opinion. So,\n
I don't think I need to go into more detail to know the linear relationship \n
between Systolic Blood Pressure and Weight isn't that strong.")
```

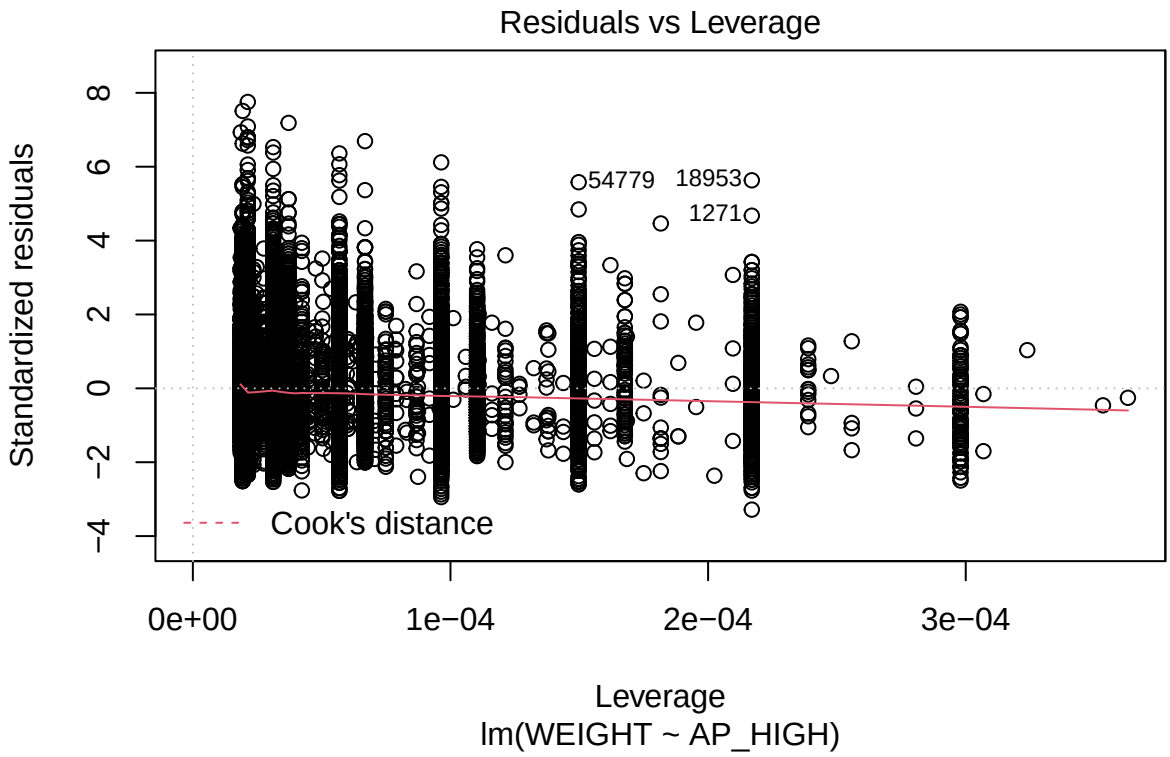
```
## The p-value of the F-statistic is < 2.2e^(-16) which is good, it means that
##
## there is at least one data point that's significantly related to the outcome
##
## variable. The median residual is somewhat close to 0 at -2.066 which is somewhat
##
## symmetrical. The higher values seem to stray away from the line the most. The
##
## residual standard error is 13.57, having all predictions be off by and average
##
## of 13.57 units of Systolic Blood Pressure, won't really produce and accurate
##
## model. Since we're only really looking at one predictor, we can use the multiple
##
## R-squared value of 0.07116. This shows how much the independent variable can
##
## explain the dependent variable, and 7.116% is fairly awful in my opinion. So,
##
## I don't think I need to go into more detail to know the linear relationship
##
## between Systolic Blood Pressure and Weight isn't that strong.
```

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(WEIGHT ~ AP_HIGH)
20093
24426
18646
100
50
0
−50
60
65
70
75
80
85
90


Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
20093
18649
−4
−2
0
2
4
Theoretical Quantiles
lm(WEIGHT ~ AP_HIGH)


Scale−Location
20093
24423
18648
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
60
65
70
75
80
85
90
Fitted values
lm(WEIGHT ~ AP_HIGH)




```
cat("The way the Systolic Blood Pressure (AP_HIGH) was recorded means there will\n
be stacked values vertically at different intervals. The red trend line is \n
fairly straight in the 'Residuals vs Fitted' graph, which means it most likely\n
is a linear relationship vs a non-linear relationship. At higher values, the \n
'Normal Q-Q' graph strays away from the trend line, which worries me, this shows\n
that the relationship might not be linear. The line in the 'Scale-Location' plot\n
is fairly horizontal and the distribution is somewhat random, which is good. I \n
don't see the Cook's line on my last graph, which is good, but it may have \n
something to do with size of the data. I'm not sure, but I lean towards it being\n
a good thing.")
```

```
## The way the Systolic Blood Pressure (AP_HIGH) was recorded means there will
##
## be stacked values vertically at different intervals. The red trend line is
##
## fairly straight in the 'Residuals vs Fitted' graph, which means it most likely
##
## is a linear relationship vs a non-linear relationship. At higher values, the
##
## 'Normal Q-Q' graph strays away from the trend line, which worries me, this shows
##
## that the relationship might not be linear. The line in the 'Scale-Location' plot
##
## is fairly horizontal and the distribution is somewhat random, which is good. I
##
```

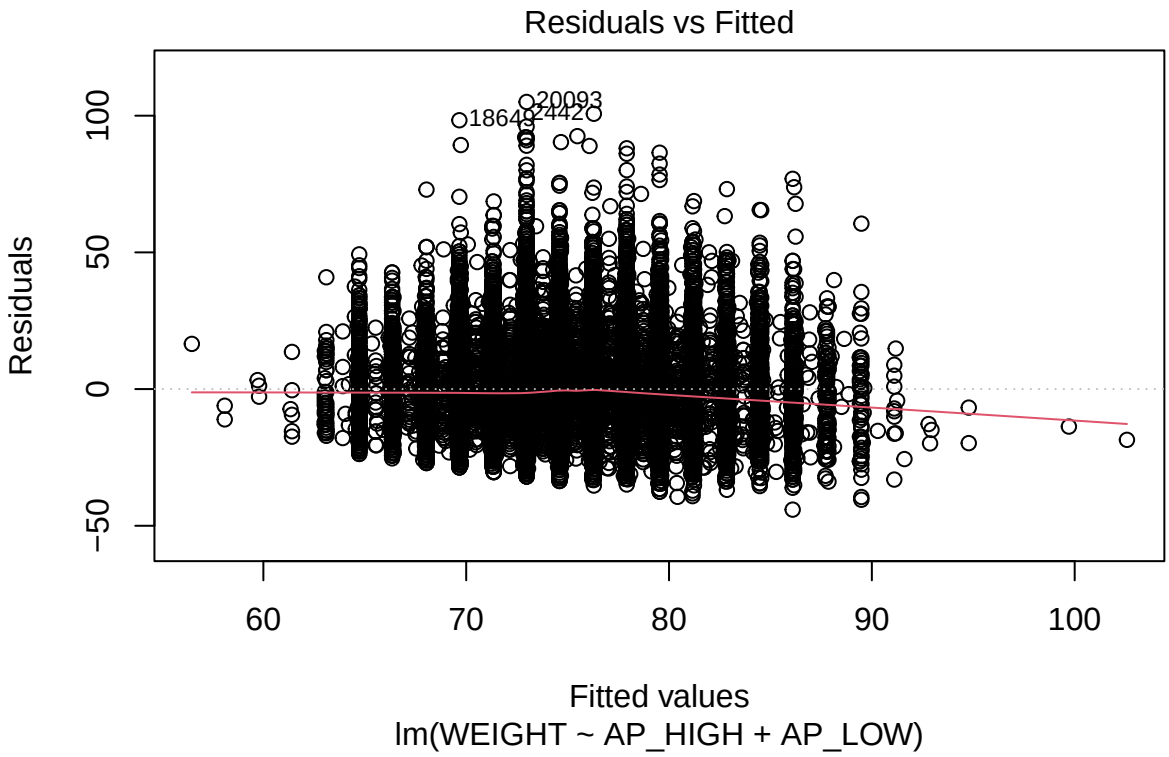
```
## don't see the Cook's line on my last graph, which is good, but it may have
##
## something to do with size of the data. I'm not sure, but I lean towards it being
##
## a good thing.
```

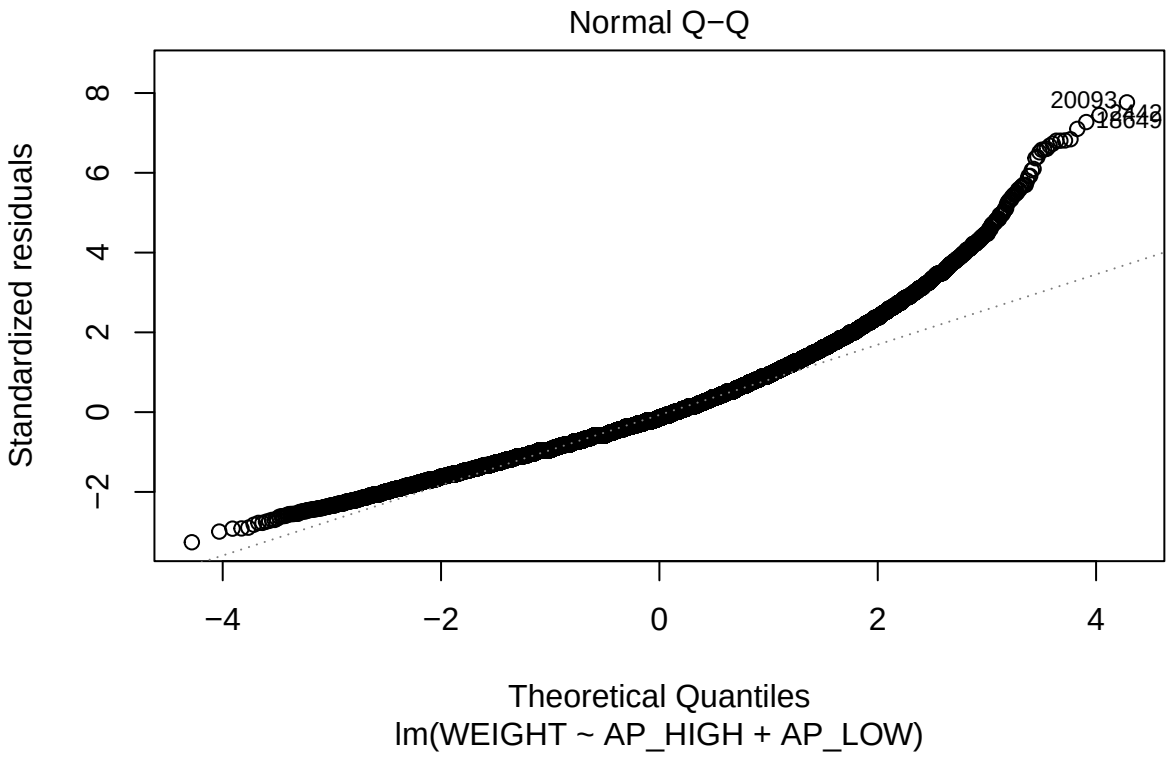
Adding Diastolic Blood Pressure to my model

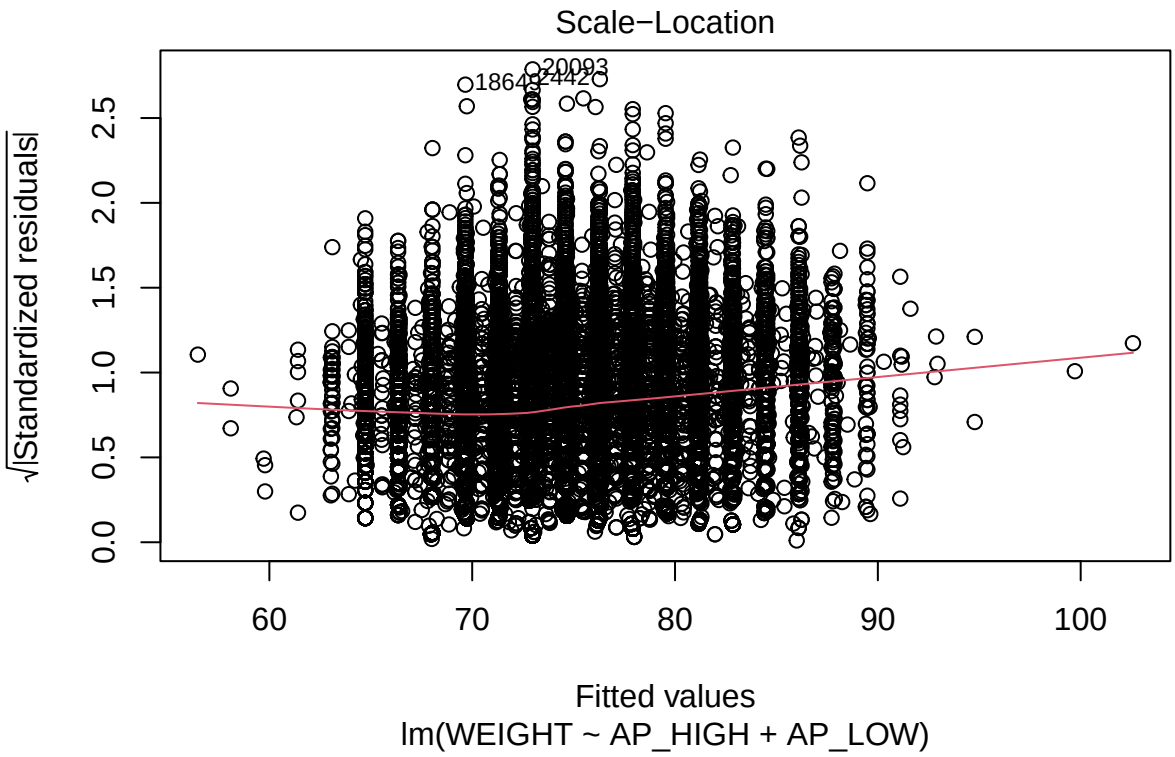
```
# Adding Diastolic Blood Pressure seemed to help a little bit
fit1 = lm(WEIGHT ~ AP_HIGH + AP_LOW, train)

summary(fit1)
```

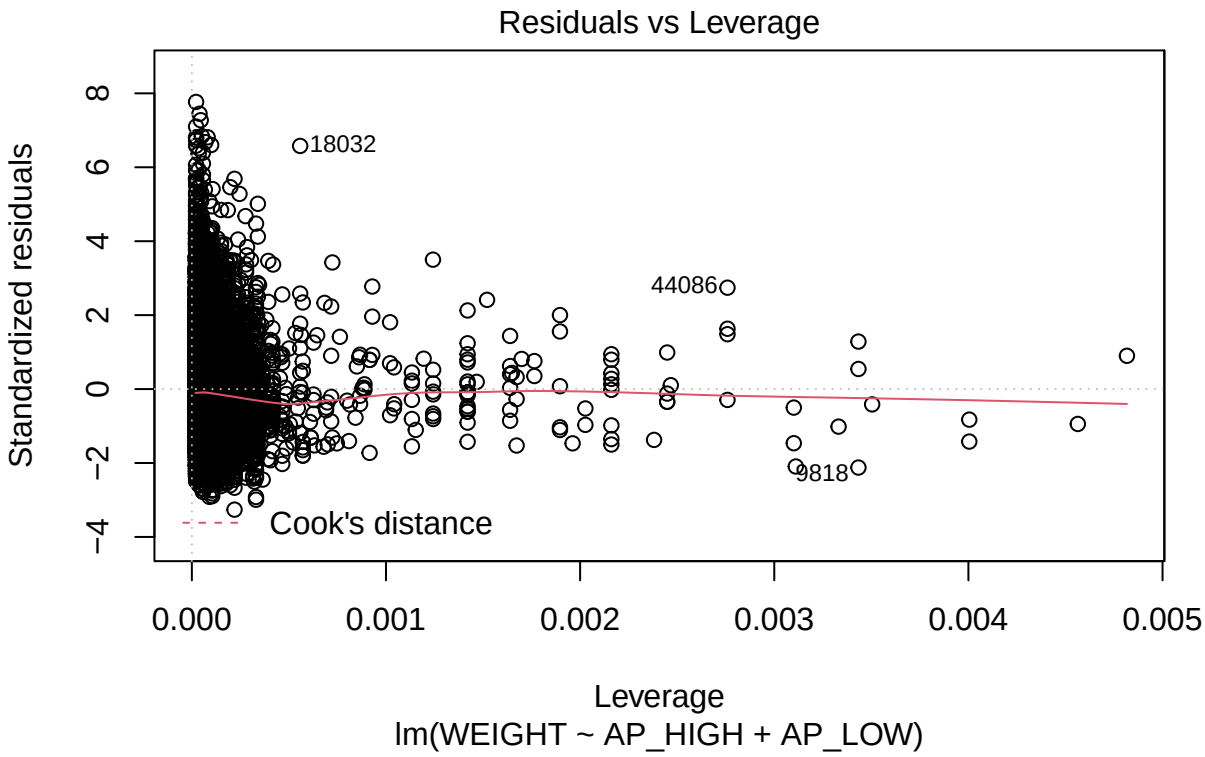
```
##
## Call:
## lm(formula = WEIGHT ~ AP_HIGH + AP_LOW, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.097  -8.978  -1.978   7.093 105.022
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.968109   0.521198   76.69   <2e-16 ***
## AP_HIGH      0.162197   0.004883   33.22   <2e-16 ***
## AP_LOW       0.169330   0.008359   20.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.52 on 54145 degrees of freedom
## Multiple R-squared:  0.07814,    Adjusted R-squared:  0.07811
## F-statistic:  2295 on 2 and 54145 DF,  p-value: < 2.2e-16
```

```
plot(fit1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(WEIGHT ~ AP_HIGH + AP_LOW)
20093
24422
60
70
80
90
100
−50
0
50
100


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(WEIGHT ~ AP_HIGH + AP_LOW)
20093
18649


Scale−Location
20093
18642
24420
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
60
70
80
90
100
Fitted values
lm(WEIGHT ~ AP_HIGH + AP_LOW)



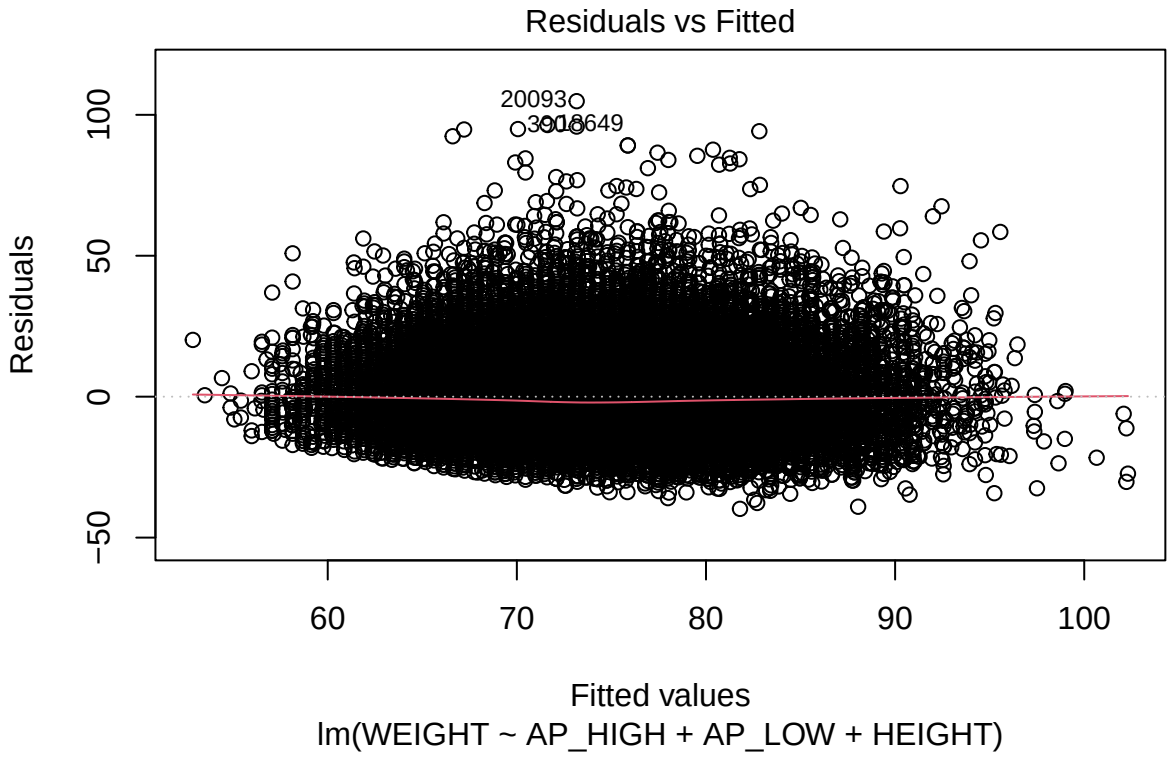

Adding Diastolic Blood Pressure and Height to my model

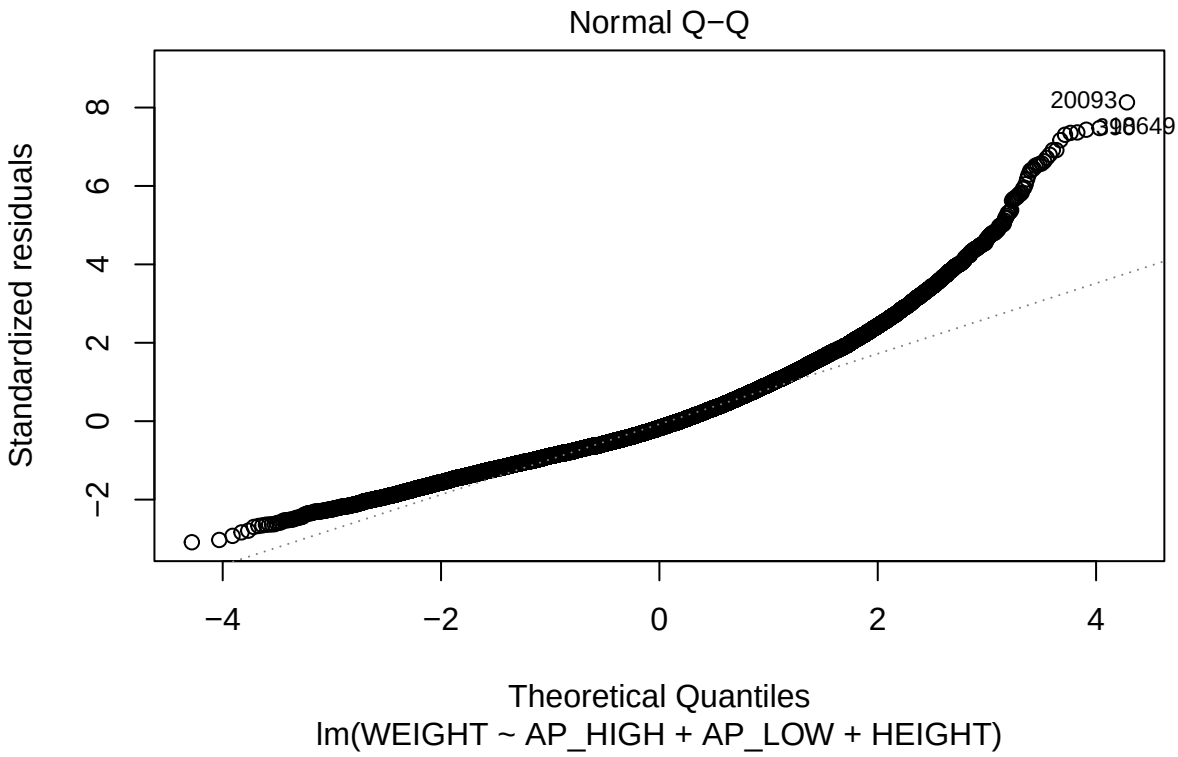
```
# Adding Diastolic Blood Pressure and Height seemed to help a bit more
fit2 = lm(WEIGHT ~ AP_HIGH + AP_LOW + HEIGHT, train)

summary(fit2)
```

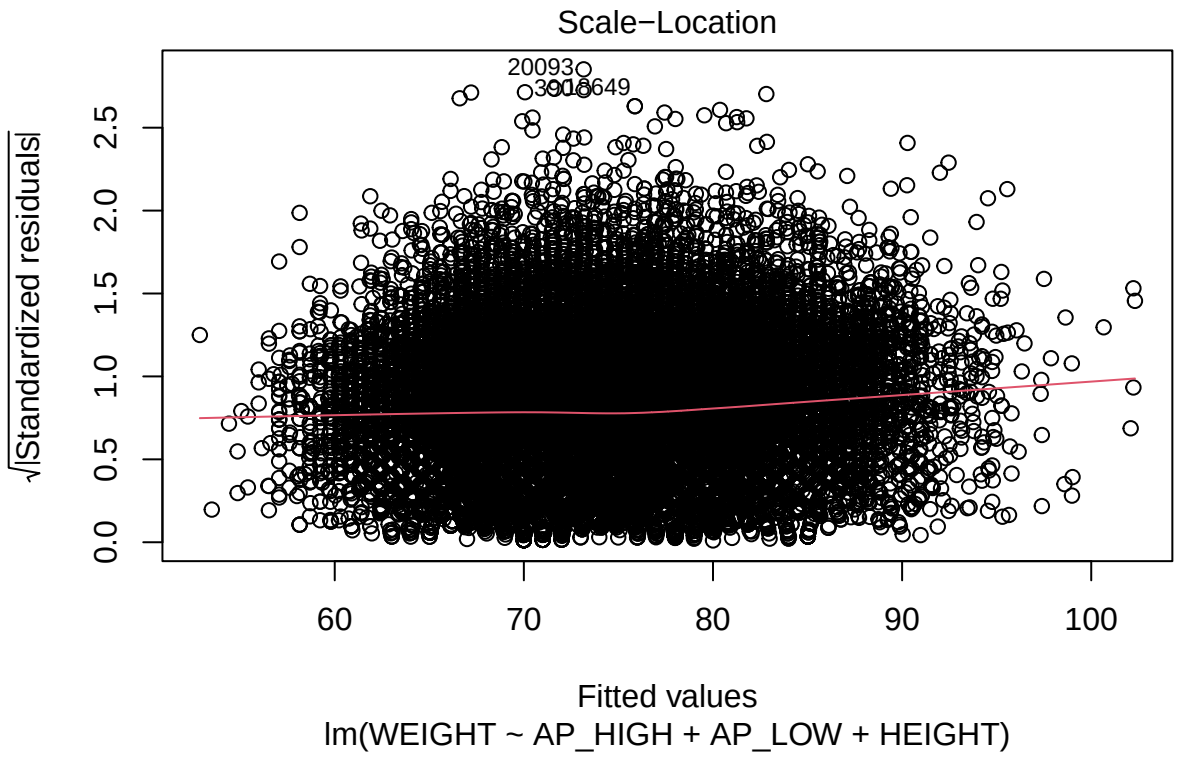
```
##
## Call:
## lm(formula = WEIGHT ~ AP_HIGH + AP_LOW + HEIGHT, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.799  -8.786  -2.210   6.838 104.838
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48.061509   1.293172  -37.17   <2e-16 ***
## AP_HIGH       0.165375   0.004655   35.53   <2e-16 ***
## AP_LOW        0.151049   0.007973   18.95   <2e-16 ***
## HEIGHT        0.541180   0.007340   73.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.89 on 54144 degrees of freedom
## Multiple R-squared:  0.1623, Adjusted R-squared:  0.1622
```

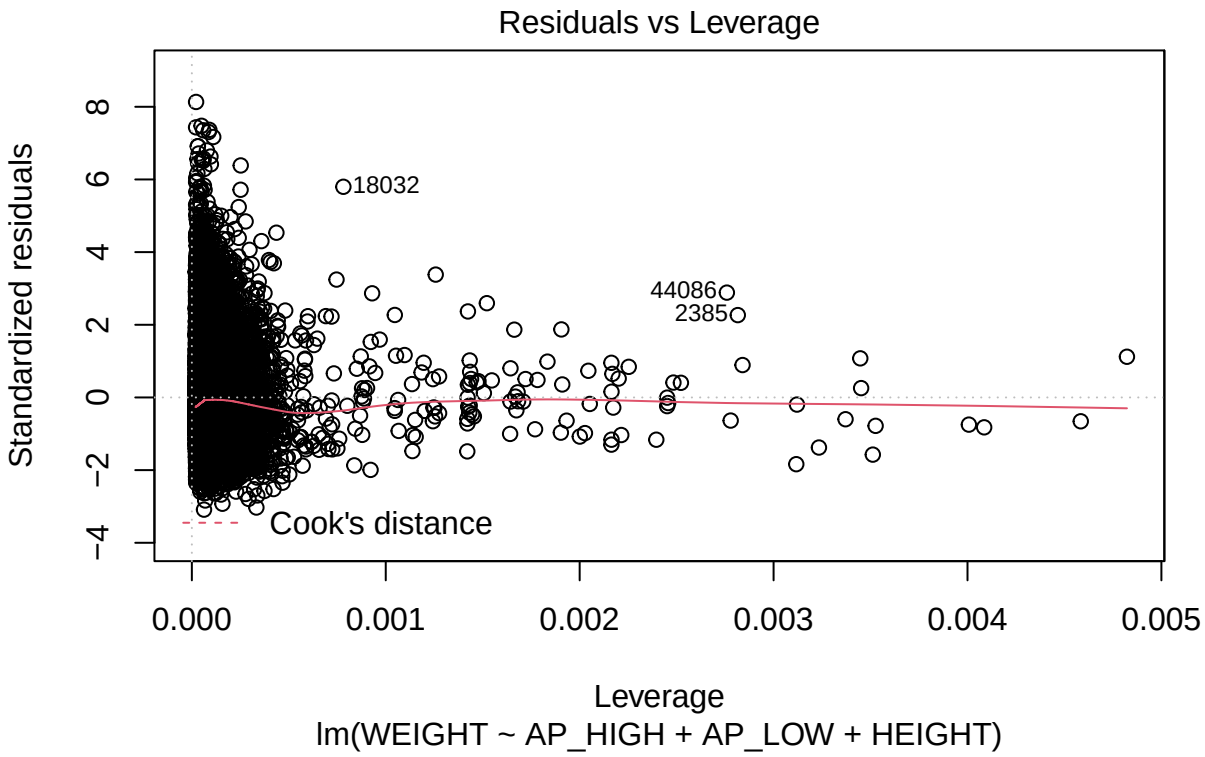
```
## F-statistic:  3496 on 3 and 54144 DF,  p-value: < 2.2e-16
```

```
plot(fit2)
```




Normal Q−Q
Standardized residuals
−2
0
2
4
6
8
20093
−4
−2
0
2
4
Theoretical Quantiles
lm(WEIGHT ~ AP_HIGH + AP_LOW + HEIGHT)


Scale−Location
20093
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
60
70
80
90
100
Fitted values
lm(WEIGHT ~ AP_HIGH + AP_LOW + HEIGHT)




Conclusion about which model is the best

```
cat("The more predictors I added to the models the more it improved, but I still\n
think that a linear model isn't the best fit for this relation ship in predicting\n
weight. To me, looking at the data I think something more exponential would fit \n
the relationship better. With that being said, I definitely think that the last\n
model is the best model. The more predictors helped it increase the significant\n
coefficients and the R-Squared value. It isn't really surprising to me, because\n
height would absolutely have a correlation to weight, since you need more mass \n
to be taller.")
```

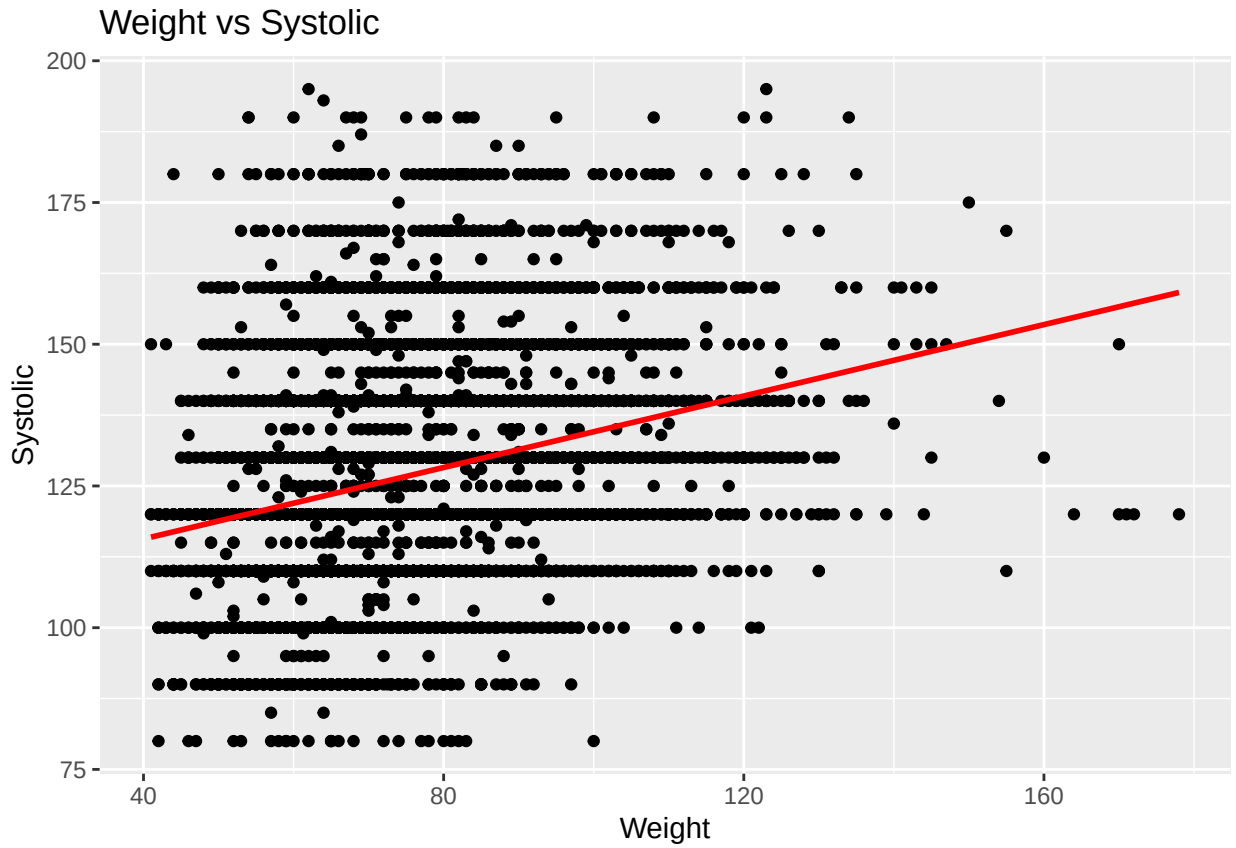
```
## The more predictors I added to the models the more it improved, but I still
##
## think that a linear model isn't the best fit for this relation ship in predicting
##
## weight. To me, looking at the data I think something more exponential would fit
##
## the relationship better. With that being said, I definitely think that the last
##
## model is the best model. The more predictors helped it increase the significant
##
## coefficients and the R-Squared value. It isn't really surprising to me, because
##
## height would absolutely have a correlation to weight, since you need more mass
##
## to be taller.
```

```r
# I think these functions and graphs help see the trend of adding more predictors
# Pearson Correlation test and a scatter plot and trend line graph
cor.test(test$WEIGHT, test$AP_HIGH)
```

```
##
##  Pearson's product-moment correlation
##
## data:  test$WEIGHT and test$AP_HIGH
## t = 33.683, df = 13535, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2624829 0.2935697
## sample estimates:
##       cor
## 0.2780991
```

```r
ggplot(data = test, mapping = aes(x = WEIGHT, y = AP_HIGH)) +
  geom_point() +
  geom_smooth(color = "red", method = "lm", stat = "smooth", position = "identity", se = FALSE) +
  labs(title = "Weight vs Systolic", x = "Weight", y = "Systolic")
```

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

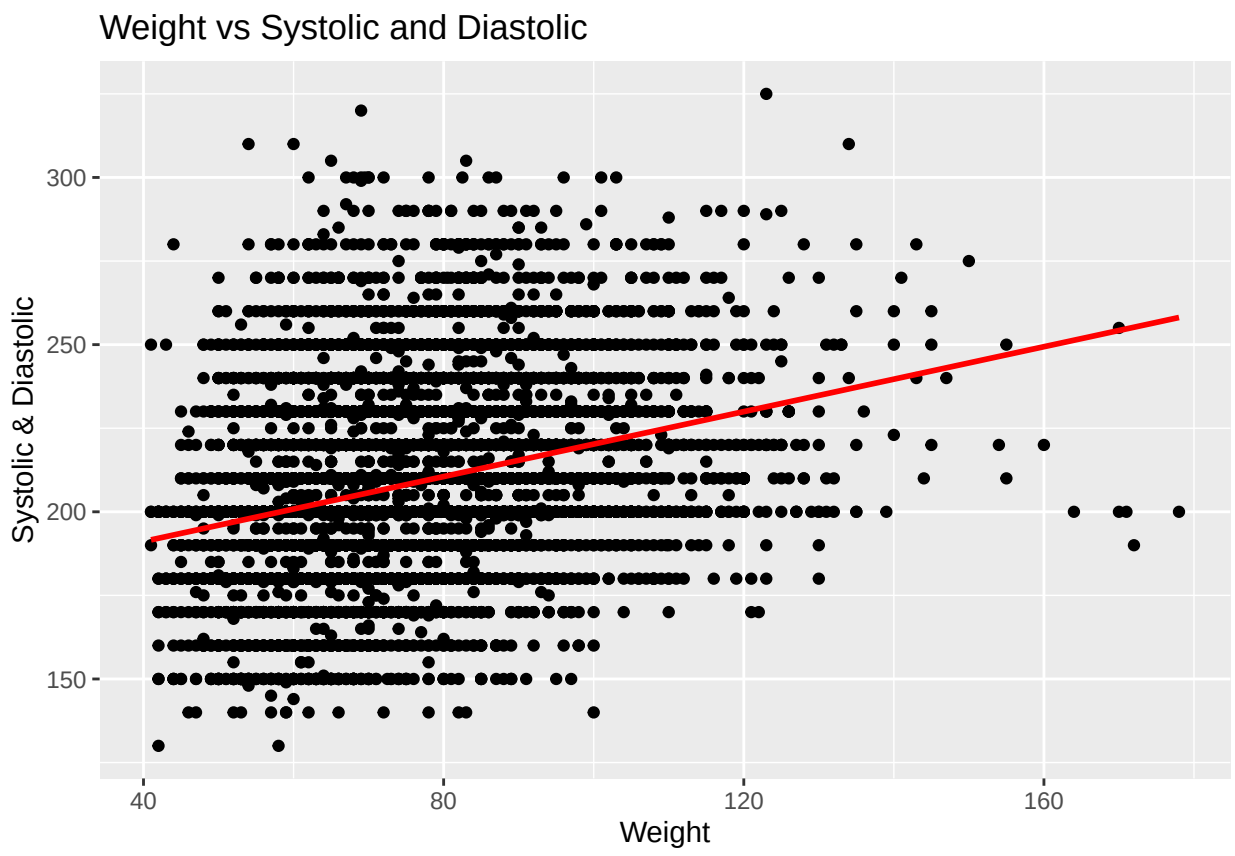


```
cor.test(test$WEIGHT, test$AP_HIGH + test$AP_LOW)
```

```
##
##  Pearson's product-moment correlation
##
## data:  test$WEIGHT and test$AP_HIGH + test$AP_LOW
## t = 35.372, df = 13535, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2753968 0.3062383
## sample estimates:
##       cor
## 0.2908931
```

```
ggplot(data = test, mapping = aes(x = WEIGHT, y = AP_LOW + AP_HIGH)) +
  geom_point() +
  geom_smooth(color = "red", method = "lm", stat = "smooth", position = "identity", se = FALSE) +
  labs(title = "Weight vs Systolic and Diastolic", x = "Weight", y = "Systolic & Diastolic")
```

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

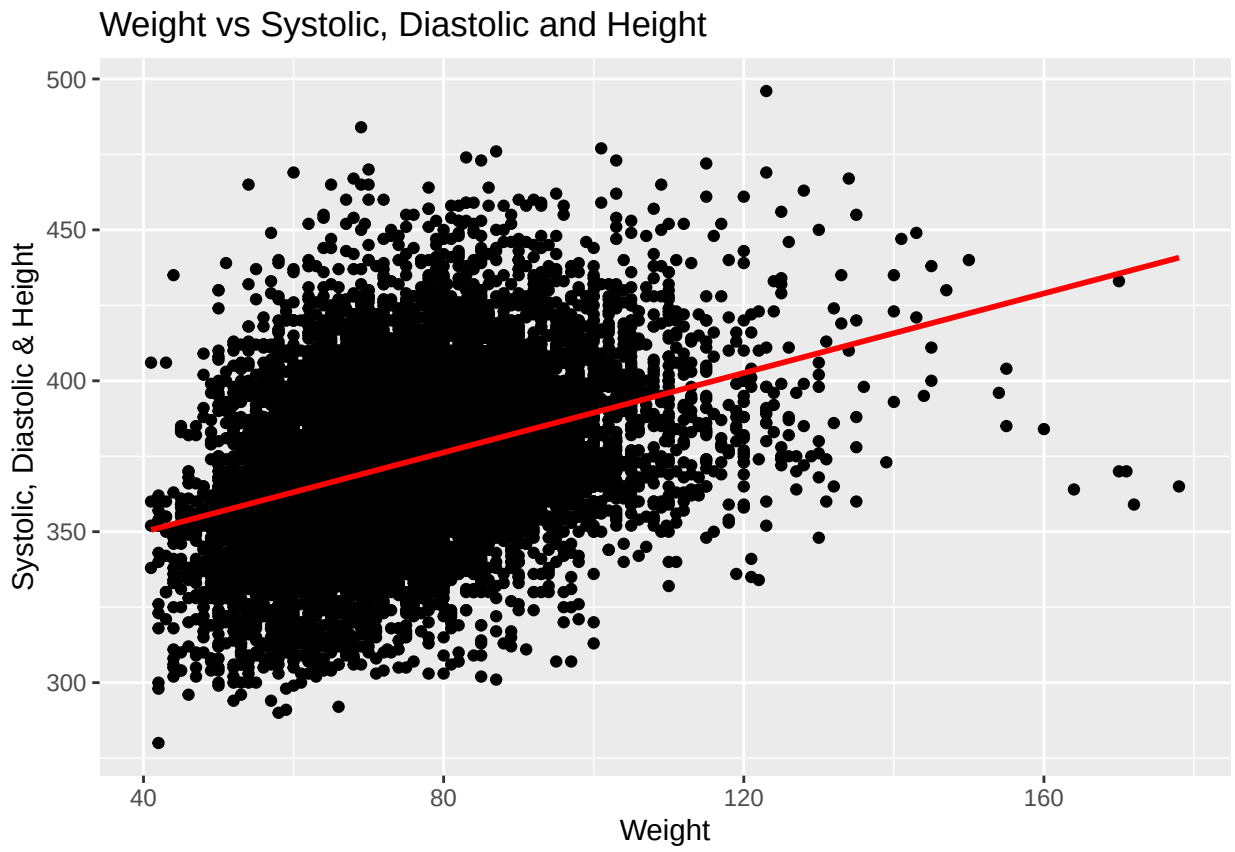



```
cor.test(test$WEIGHT, test$AP_HIGH + test$AP_LOW + test$HEIGHT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  test$WEIGHT and test$AP_HIGH + test$AP_LOW + test$HEIGHT
## t = 46.496, df = 13535, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3564968 0.3855495
## sample estimates:
##      cor
## 0.371114
```

```
ggplot(data = test, mapping = aes(x = WEIGHT, y = AP_LOW + AP_HIGH + HEIGHT)) +
  geom_point() +
  geom_smooth(color = "red", method = "lm", stat = "smooth", position = "identity", se = FALSE) +
  labs(title = "Weight vs Systolic, Diastolic and Height", x = "Weight", y = "Systolic, Diastolic & Hei
```

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




```
cat("They're extremely similar to the train data, I think that's because of my\n
sample size (70,000 rows). The size of the data set is large and random enough \n
to generate similar results in both the test and train data.")
```

```
## They're extremely similar to the train data, I think that's because of my
```

```
##
## sample size (70,000 rows). The size of the data set is large and random enough
##
## to generate similar results in both the test and train data.
```