Predicting Anxiety Diagnosis from Survey Data

Wayne Huynh

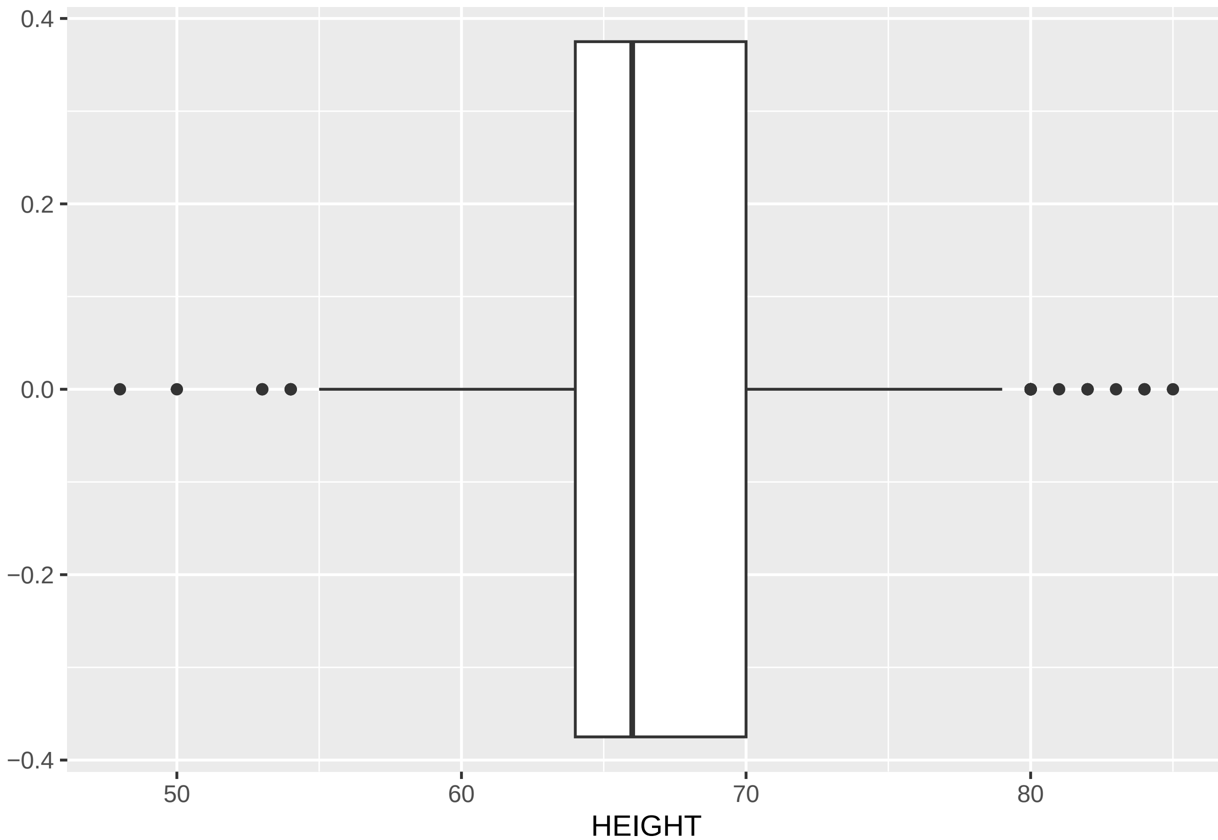
2023-08-09

```
#I am looking to find the relationship between people who have been diagnosed with anxiety
#vs their veteran status, height and weight.
#Anxiety diagnosis(ADANXEV) is my response variable
#Veteran status(VETERAN3), height(HEIGHT3), and weight(WEIGHT2) are my predictor variables

#Anxiety diagnosis(ADANXEV) and Veteran status(VETERAN3) won't have outliers since they
#are yes/no
#I will choose to not remove any outliers because part of my project will be seeing if
#extreme heights and weights contribute to an anxiety diagnosis
```

```
#Clean data
#Anxiety diagnosis(ADANXEV)
#Only keep people who answered Yes(1) or No(2)
data <- subset(data, ADANXEV == 1 | ADANXEV == 2)

#Veteran status(VETERAN3)
#Only keep people who answered Yes(1) or No(2)
data <- subset(data, VETERAN3 == 1 | VETERAN3 == 2)

#Height(HEIGHT3)
#Only keep people with a reported height (200-711) and (9000-9998)
data <- subset(data, (HEIGHT3 >=200 & HEIGHT3 <=711) | (HEIGHT3 >= 9000 & HEIGHT3 <= 9998))

#Convert reported heights to inches
to_in <- function(number) {
  result <-
    #convert ft/in to in
    ifelse(number >= 200 & number <= 711,
           (number %/% 100) * 12 + (number %% 100),
           #convert 9/m/cm to cm then to in
           ifelse(number >= 9000 & number <= 9998,
                  round(0.393701*(number %% 1000),0),
                  number))
  return(result)
}

#Apply the to_in function to the HEIGHT3 column
data <- data %>%
  mutate(HEIGHT = to_in(HEIGHT3))


#Weight(WEIGHT3)
#Only keep people with a reported weight (50-0999) and (9000-9998)
```

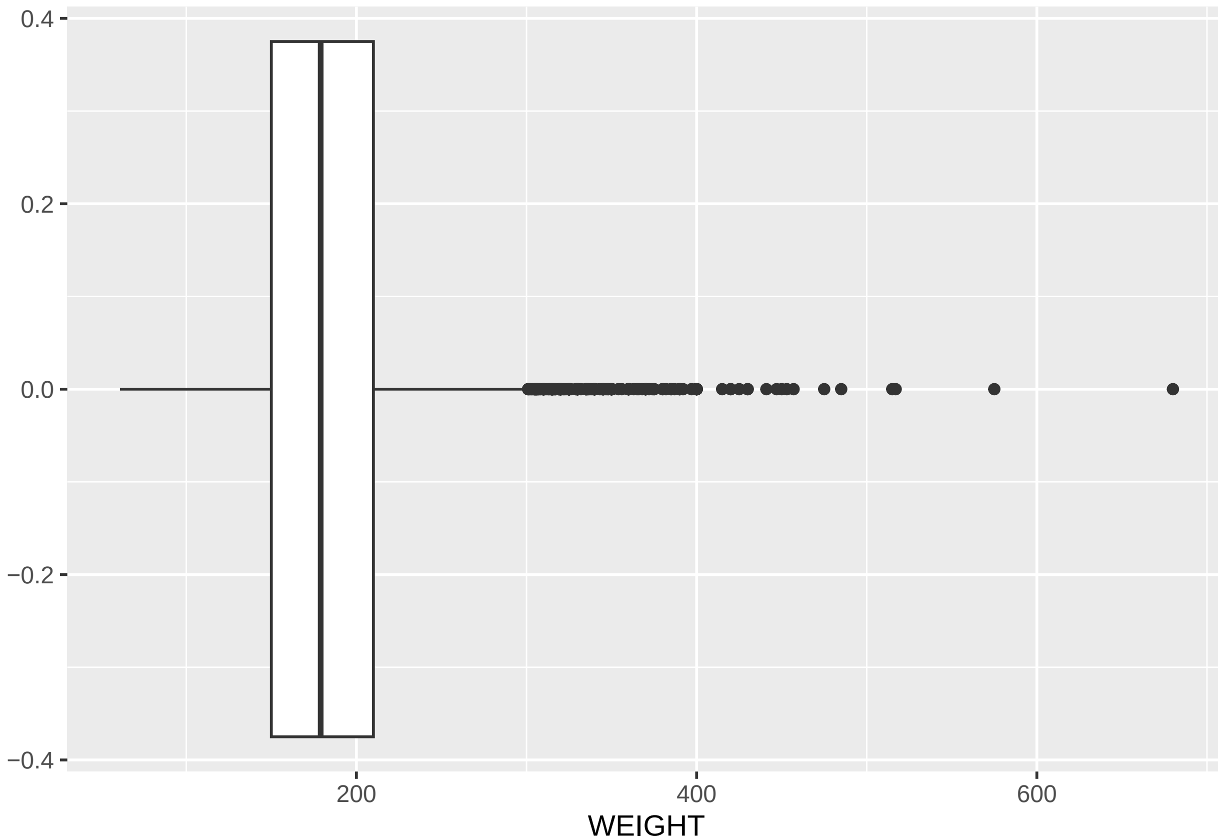
```r
data <- subset(data, (WEIGHT2 >=50 & WEIGHT2 <=999) | (WEIGHT2 >= 9000 & WEIGHT2 <= 9998))

#Convert reported weights from kilograms to pounds
to_lbs <- function(number) {
  result <-
    #convert kgs to lbs
    ifelse(number >= 9000 & number <= 9998,
           round(2.20462262185*(number %% 1000),0),
           number)
  return(result)
}
#Apply the to_lbs function to the WEIGHT2 column
data <- data %>%
  mutate(WEIGHT = to_lbs(WEIGHT2))
```

```r
ggplot(data) +
  geom_boxplot(mapping = aes(HEIGHT))
```



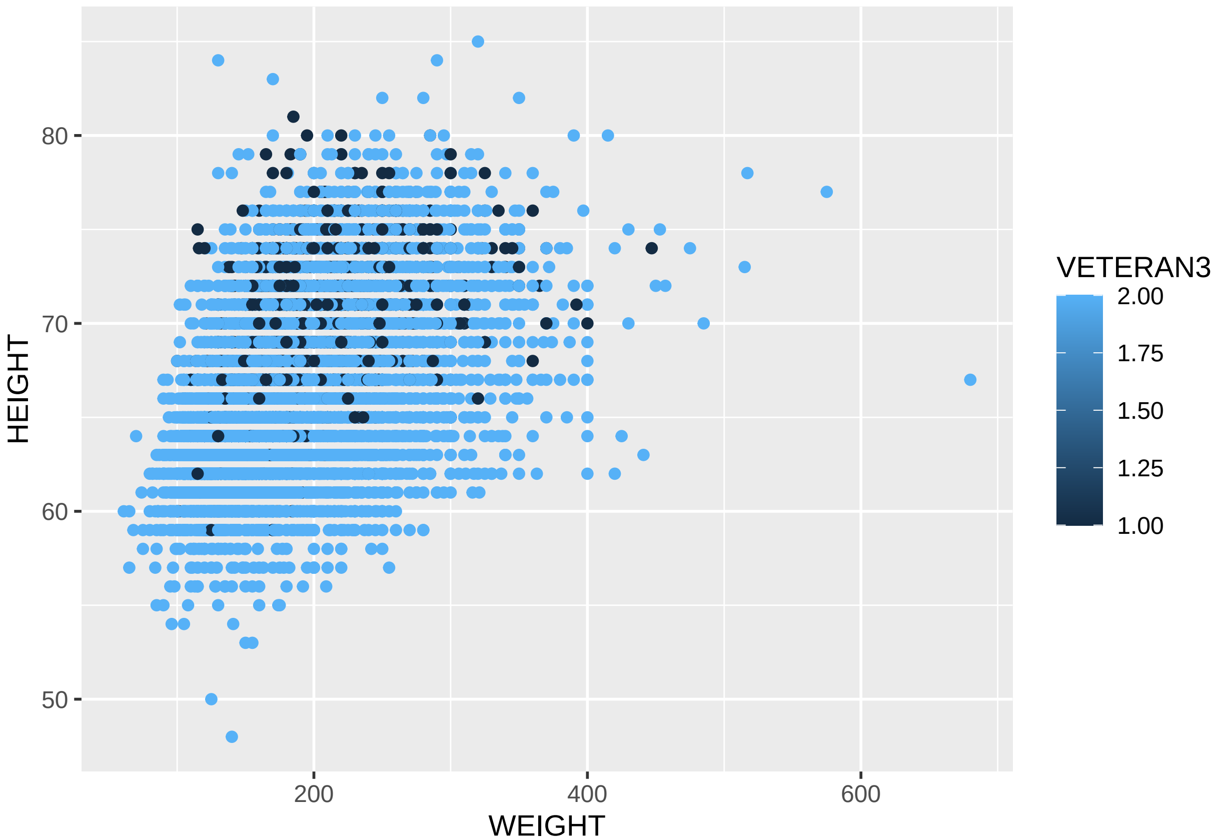

```r
#There are 4 low outliers and 6 higher outliers in terms of height, but overall it seems
#decently normally distributed
```

```r
ggplot(data) +
  geom_boxplot(mapping = aes(WEIGHT))
```



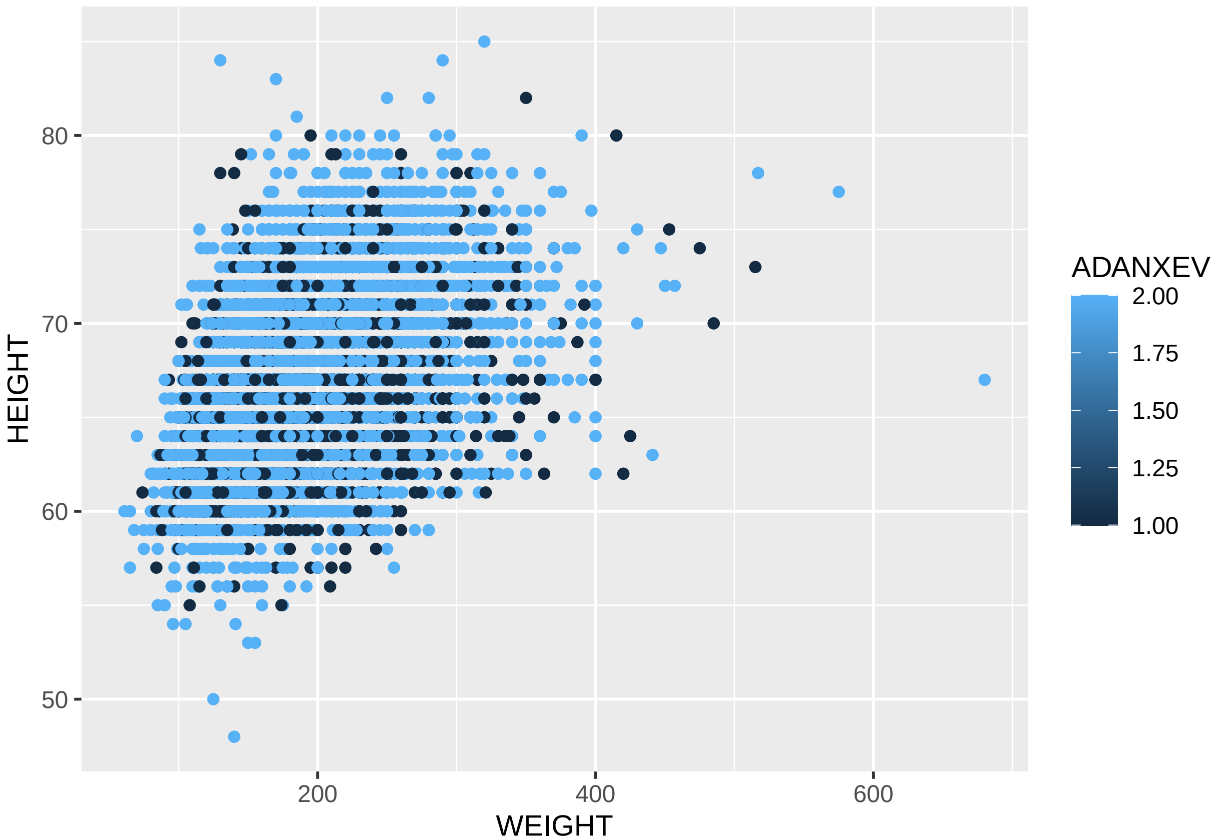

```
#The weight distribution is right-skewed
```

```
ggplot(data) +
  geom_point(aes(x = WEIGHT, y = HEIGHT, color = VETERAN3))
```



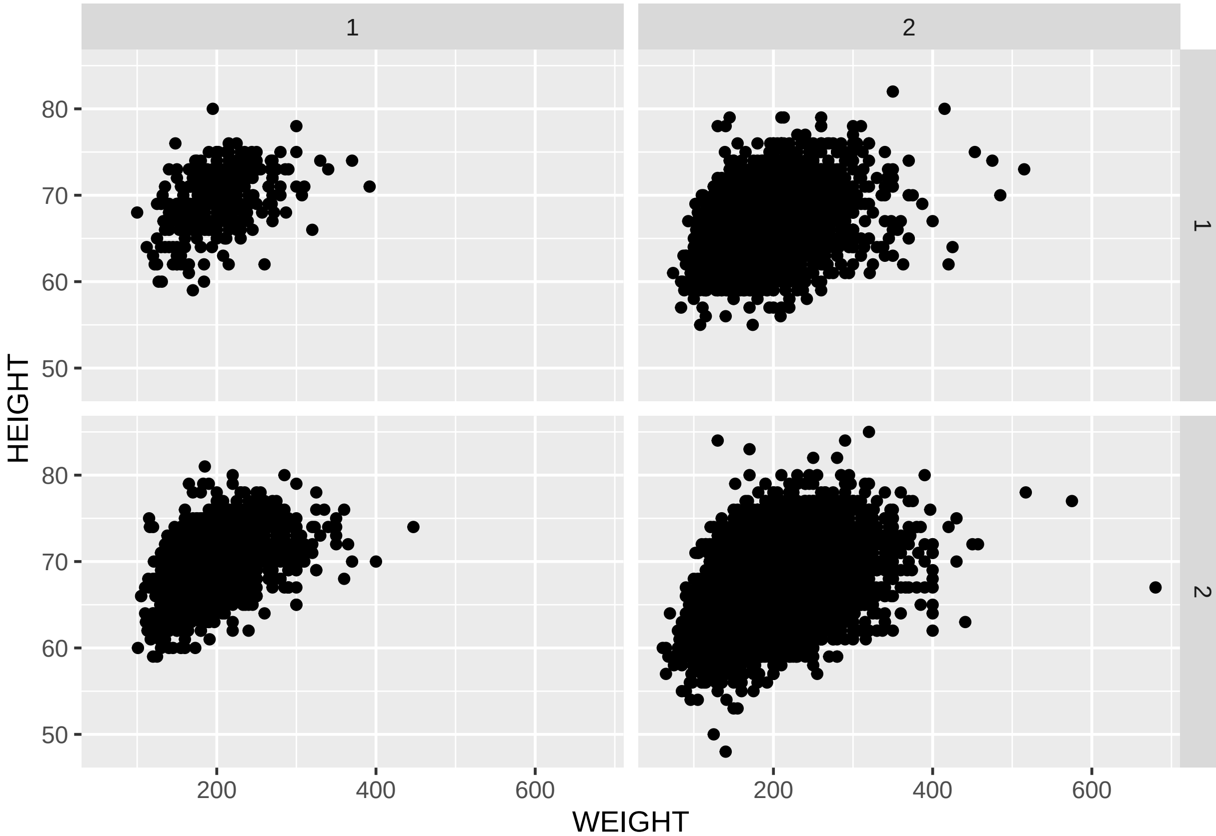

```
#This plot shows people's height/weight measurement color coded by their Veteran status
#Black is a veteran, blue is not
#Veterans' height/weight measurements are mostly in the middle of the plot
```

```
ggplot(data) +
  geom_point(aes(x = WEIGHT, y = HEIGHT, color = ADANXEV))
```




```
#This plot shows people's height/weight measurement color coded by their anxiety diagnosis
#Black has anxiety, blue does not
#Initially, it does seem like there are more people with anxiety in the bottom right area
#This shows roughly that shorter people and heavier people tend to have anxiety more often
#than taller and/or light people
```

```
ggplot(data = data) +
  geom_point(mapping = aes(x = WEIGHT, y = HEIGHT)) +
  facet_grid(ADANXEV ~ VETERAN3)
```

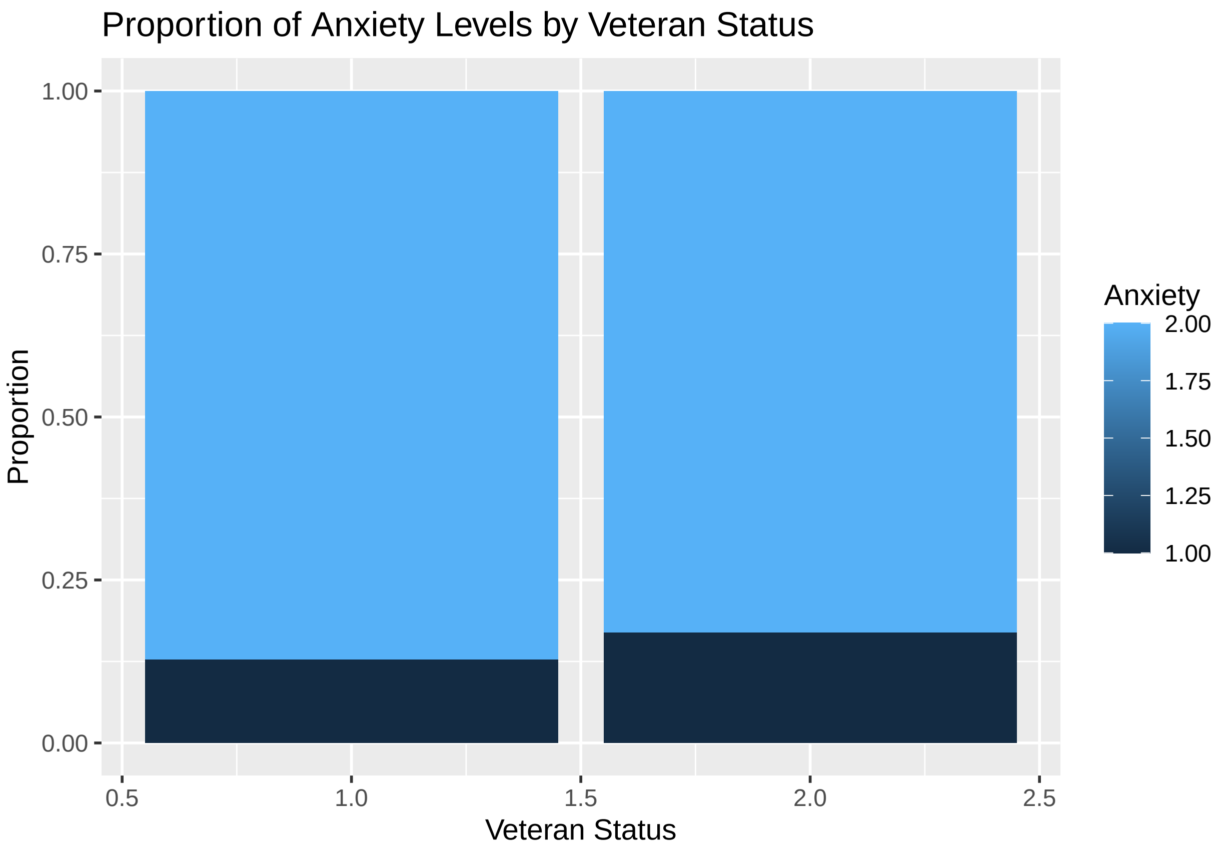



```
#1 on the right = anxiety, 2 on the right = no anxiety diagnosis
#1 on top = veteran, 2 on top = not a veteran
#This confirms that veterans look much more alike each other in terms of height/weight
#than non-veterans
#It also shows that anxiety diagnoses are slightly more spread out from the big cluster
#than non-anxiety diagnoses
```

```
#Calculate the proportion of anxiety diagnoses within each veteran status
summary_data <- data %>%
  group_by(VETERAN3, ADANXEV) %>%
  summarise(proportion = n()) %>%
  ungroup() %>%
  mutate(proportion = proportion / sum(proportion))
```

```
## `summarise()` has grouped output by 'VETERAN3'. You can override using the
## `.groups` argument.
```

```
# Create a percent stacked bar plot
ggplot(summary_data, aes(x = VETERAN3, y = proportion, fill = ADANXEV)) +
  geom_bar(position = "fill", stat = "identity") +
  labs(x = "Veteran Status", y = "Proportion", fill = "Anxiety") +
  ggtitle("Proportion of Anxiety Levels by Veteran Status")
```

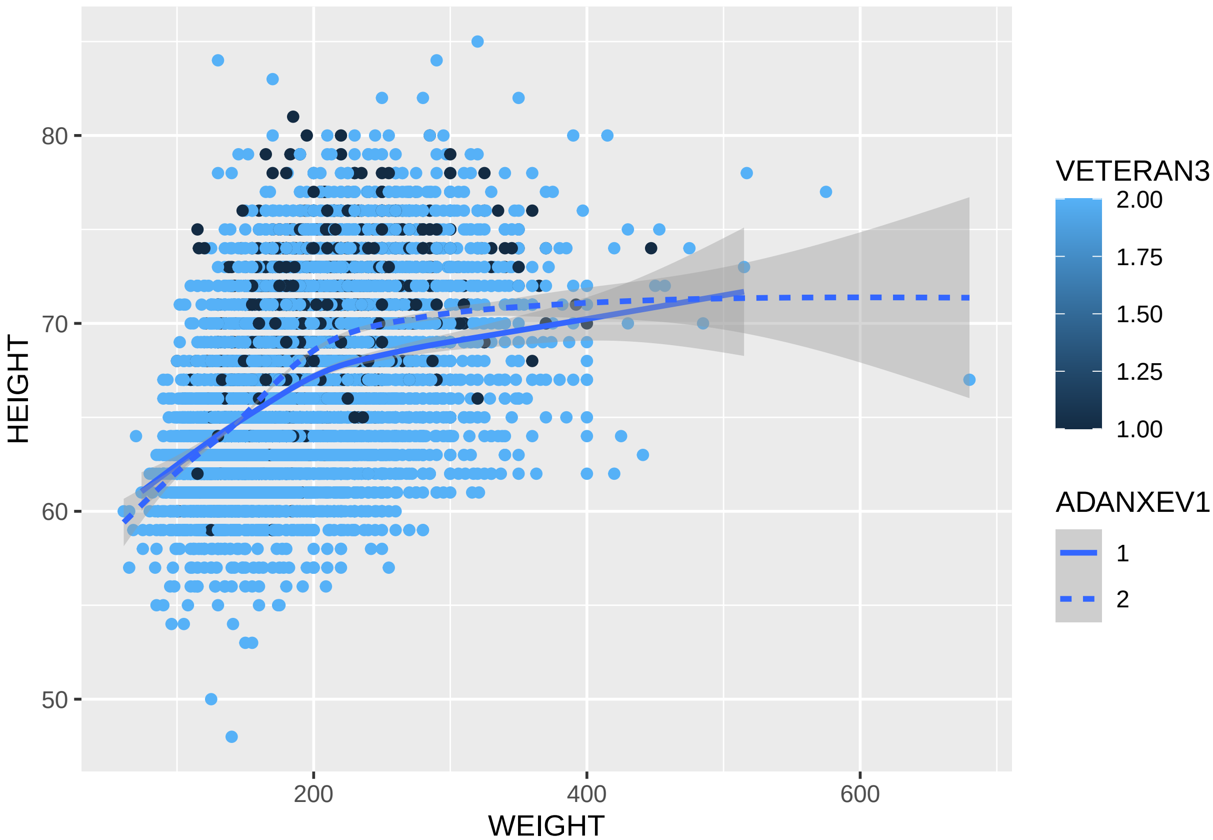



```r
#Veteran = 1 means veteran; Veteran = 2 means not a veteran
#Anxiety = 1 or black means diagnosed anxiety
#Anxiety = 2 or blue means no anxiety diagnosis
#This plot shows that veterans (left side) have a slightly lower anxiety proportion than
#the non-veteran group (right side)
#We do not know if this difference is significant though
```

```r
data$ADANXEV1 <- factor(data$ADANXEV)
ggplot(data = data, mapping = aes(x = WEIGHT, y = HEIGHT, color = VETERAN3)) +
  geom_point() +
  geom_smooth(mapping = aes(linetype = ADANXEV1))
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
#ADANXEV1 = 1 means anxiety; ADANXEV1 = 2 means no anxiety diagnosis
#The anxiety trendline is lower which might signal that shorter people get diagnosed with
#anxiety more
#Because veterans look to be taller on average than a non-veteran, this also might signal
#that veterans get diagnosed with anxiety less than a non-veteran
```

```
#Anxiety diagnosis(ADANXEV)
#Calculate descriptive statistics of 'ADANXEV'
summary(data$ADANXEV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.836   2.000   2.000
```

```
#1st quartile is 2 indicating a majority of people are not diagnosed with anxiety

#Count of people with (1) and without (2) anxiety diagnosis
table(data$ADANXEV)
```

```
##
##     1     2
##  3145 15982
```

```
#With: 3,145 people
#Without: 15,982 people
```

```
#Veteran status(VETERAN3)
#Calculate descriptive statistics of 'VETERAN3'
summary(data$VETERAN3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   1.000   2.000   2.000   1.865   2.000   2.000
```

```r
#1st quartile is 2 indicating an overwhelming majority of people are not veterans

#Count of veterans (1) and non-veterans (2)
table(data$VETERAN3)
```

```
##
##     1     2
##  2585 16542
```

```r
#Veterans: 2,585 people
#Non-veterans: 16,542 people
```

```r
#Height(HEIGHT)
#Calculate descriptive statistics of 'HEIGHT'
summary(data$HEIGHT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   64.00   66.00   66.87   70.00   85.00
```

```r
#Shortest person is 48in (4')
#Median person is 66in (5'6")
#Tallest person is 85in (7'1")

table(data$HEIGHT)
```

```
##
##   48   50   53   54   55   56   57   58   59   60   61   62   63   64   65   66
##    1    1    2    3    7   18   30   40  174  596  639 1459 1474 1767 1587 1776
##   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82
## 1515 1351 1290 1296 1108 1233  654  500  277  180   74   33   22   12    1    3
##   83   84   85
##    1    2    1
```

```r
#There are relatively few people shorter than 60in (5') or taller than 74in (6'2") which
#is true of American society so this height dataset feels valid
```

```r
#Weight (WEIGHT)
#Calculate descriptive statistics of 'WEIGHT'
summary(data$WEIGHT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    61.0   150.0   179.0   183.3   210.0   680.0
```

```r
#Lightest person is 61lbs
#Median person is 179lbs
#Heaviest person is 680lbs

#Count of veterans (1) and non-veterans (2)
table(data$WEIGHT)
```

```
##
##   61   65   68   70   74   75   80   82   84   85   86   87   88   89   90   91
##    1    2    1    1    1    2    3    3    2    5    4    2    1    1   14    3
##   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107
##    4    2    4   14    9    5   11    3   50    5   18   11   20   70   23   19
##  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123
##   31   16  127   10   52   17   27  151   26   34   54   20  264   15   42   35
```

```
##   124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139
##    46  263   38   39   78   31  463   18   71   32   41  401   51   40   96   43
##   140  141  142  143  144  145  146  147  148  149  150  151  152  153  154  155
##   570   23   88   69   46  439   46   51  103   47  823   34   82   48   58  382
##   156  157  158  159  160  161  162  163  164  165  166  167  168  169  170  171
##    56   41   91   34  769   24   87   44   48  488   28   55  110   42  808   32
##   172  173  174  175  176  177  178  179  180  181  182  183  184  185  186  187
##   116   47   73  500   52   42   72   31  919   31   74   63   44  505   51   64
##   188  189  190  191  192  193  194  195  196  197  198  199  200  201  202  203
##    56   50  641   29   63   35   38  306   39   47   82   29 1050   26   46   36
##   204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219
##    36  198   30   29   39   23  416   16   59   25   45  251   32   23   49   19
##   220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235
##   485    9   27   17   24  252   17   16   23   13  391    5   25   14   11  136
##   236  237  238  239  240  241  242  243  244  245  246  247  248  249  250  251
##    19    8   25    8  374   10   19   11   12  121    8   14   20    9  367    7
##   252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267
##    10    9    8   50   11    8   16    6  175    8    8    8    5   67    4    5
##   268  269  270  271  272  273  274  275  276  277  278  279  280  282  283  284
##    13    7  139    3    8    5    6   58    8    3    9    3  126    3    3    2
##   285  286  287  288  289  290  291  292  293  294  295  296  297  298  299  300
##    38    4    7    2    9   80    4    3    1    2   21    2    3    2    2  142
##   301  302  303  304  305  306  307  308  309  310  311  312  313  314  315  316
##     3    2    2    2    8    7    3    2    1   27    1    1    2    2   15    4
##   317  319  320  321  322  323  324  325  326  329  330  332  335  336  337  338
##     3    3   30    1    2    1    1   21    1    2   16    1    6    3    1    1
##   339  340  343  344  345  346  347  348  350  354  356  360  363  365  366  368
##     1   22    1    1    9    1    1    2   27    1    1   13    1    1    1    1
##   370  372  374  375  380  382  385  387  390  392  397  400  415  420  425  430
##    11    1    1    2    2    1    2    1    4    1    1   12    1    2    1    2
##   441  447  450  453  457  475  485  515  517  575  680
##     1    1    1    1    1    1    1    1    1    1    1
```

```r
#Most people reported weights that are divisible by 5
#There are relatively few people that weigh less than 100lbs or more than 300lbs which
#is generally true of America
```

```r
#ADANXEV has to be 0 and 1 for the logistic regression to run
data$ANXIETY <- ifelse(data$ADANXEV == 2, 1, 0)
```

```r
#Logistic regression of how height and weight affect anxiety diagnosis frequency
model1 <- glm(ANXIETY ~ HEIGHT + WEIGHT, data = data, family = binomial)
# Display the summary of the logistic regression model
summary(model1)
```

```
##
## Call:
## glm(formula = ANXIETY ~ HEIGHT + WEIGHT, family = binomial, data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.1936732  0.3335161  -6.577 4.79e-11 ***
## HEIGHT       0.0677929  0.0054842  12.361  < 2e-16 ***
## WEIGHT      -0.0037819  0.0004527  -8.354  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17097  on 19126  degrees of freedom
## Residual deviance: 16935  on 19124  degrees of freedom
## AIC: 16941
##
## Number of Fisher Scoring iterations: 4
```

```r
#Logistic regression of how height and weight and veteran status affect anxiety diagnosis
#frequency
model2 <- glm(ANXIETY ~ HEIGHT + WEIGHT + VETERAN3, data = data, family = binomial)
# Display the summary of the logistic regression model
summary(model2)
```

```
##
## Call:
## glm(formula = ANXIETY ~ HEIGHT + WEIGHT + VETERAN3, family = binomial,
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.6030687  0.3983857  -4.024 5.72e-05 ***
## HEIGHT       0.0637003  0.0056765  11.222  < 2e-16 ***
## WEIGHT      -0.0037462  0.0004523  -8.283  < 2e-16 ***
## VETERAN3    -0.1728791  0.0650980  -2.656  0.00792 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17097  on 19126  degrees of freedom
## Residual deviance: 16928  on 19123  degrees of freedom
## AIC: 16936
##
## Number of Fisher Scoring iterations: 4
```

```r
#Logistic regression of how height and veteran status affect anxiety diagnosis frequency
model3 <- glm(ANXIETY ~ HEIGHT + VETERAN3, data = data, family = binomial)
# Display the summary of the logistic regression model
summary(model3)
```

```
##
## Call:
## glm(formula = ANXIETY ~ HEIGHT + VETERAN3, family = binomial,
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.85614    0.38804  -2.206  0.02736 *
## HEIGHT       0.04248    0.00505   8.411  < 2e-16 ***
## VETERAN3    -0.18505    0.06501  -2.846  0.00442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17097  on 19126  degrees of freedom
## Residual deviance: 16995  on 19124  degrees of freedom
## AIC: 17001
##
## Number of Fisher Scoring iterations: 4
```

```r
#Logistic regression of how weight and veteran status affect anxiety diagnosis frequency
model4 <- glm(ANXIETY ~ WEIGHT + VETERAN3, data = data, family = binomial)
# Display the summary of the logistic regression model
summary(model4)
```

```
##
## Call:
## glm(formula = ANXIETY ~ WEIGHT + VETERAN3, family = binomial,
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.5364809  0.1482581  17.109  < 2e-16 ***
## WEIGHT      -0.0013259  0.0004074  -3.255  0.00114 **
## VETERAN3    -0.3550936  0.0627469  -5.659 1.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17097  on 19126  degrees of freedom
## Residual deviance: 17056  on 19124  degrees of freedom
## AIC: 17062
##
## Number of Fisher Scoring iterations: 4
```

```r
model1$aic
```

```
## [1] 16941.43
```

```r
#16941.43

model2$aic
```

```
## [1] 16936.19
```

```r
#16936.19

model3$aic
```

```
## [1] 17000.64
```

```r
#17000.64

model4$aic
```

```
## [1] 17062.32
```

```r
#17062.32
```

```r
#Model 2 has the lowest AIC value so it is likely to be the best fit model

#The HEIGHT coefficient signals that for every 1" increase in HEIGHT, the log-odds of
#having an anxiety diagnosis increases by 0.0637003.
#HEIGHT is statistically significant with a P-value below 0.05 and the relationship is
#unlikely to be from random chance alone
#This does not line up with my prediction from the previous visualizations where I assumed
#that shorter people would be diagnosed with anxiety more
#This could be because there are more shorter people, so it seems like there are more
#anxiety diagnoses initially at the bottom, but the proportion of taller people having
#anxiety could be higher. But that is somewhat hard to discern off a visualization with
#thousands of data points

#The WEIGHT coefficient signals that for every 1lb increase in WEIGHT, the log-odds of
#having an anxiety diagnosis decreases by 0.0037462.
#WEIGHT is statistically significant with a P-value below 0.05 and the relationship is
#unlikely to be from random chance alone
#This does line up with my prediction that heavier people are more likely to have an
#anxiety diagnosis.

#The VETERAN3 coefficient signals that for every 1 unit increase in VETERAN3 (from yes
#to no), the log-odds of having an anxiety diagnosis decreases by 0.1728791.
#VETERAN3 is statistically significant with a P-value below 0.05 and the relationship is
#unlikely to be from random chance alone
#This does not line up with one of the visualizations showing that anxiety diagnosis rate
#increases from a veteran to a non-veteran
```

```r
model2.chi <- model2$null.deviance - model2$deviance

model2.df <- model2$df.null - model2$df.residual
cat("p-value = ", 1-pchisq(model2.chi, model2.df))
```

```
## p-value =  0
```

```r
#p-value =  0
#A p-value < 0.05 means the observed Chi-square difference is unlikely to have occurred by
#random chance alone
```

```r
cat("Chi-square difference = ", model2.chi)
```

```
## Chi-square difference =  169.0151
```

```r
#Chi-square difference =  169.0151
#169.0151 is the difference in deviances between the null model and Model 2, which
#includes all of the predictor variables (HEIGHT, WEIGHT, and VETERAN3)
#This is a relatively high Chi-square difference meaning that adding more predictor
#variables will improve my model fit in a logistic regression model
```