Exploration

Joe Martin

10/14/2021

Begin with a summary of the processed dataset.

```
summary(df)
```

```
##  SwollenLymphNodes ChestCongestion ChillsSweats NasalCongestion CoughYN
##  No :418           No :323         No :130      No :167         No : 75
##  Yes:312           Yes:407         Yes:600      Yes:563         Yes:655
##
##
##
##
##  Sneeze    Fatigue   SubjectiveFever Headache      Weakness   WeaknessYN
##  No :339   No : 64   No :230         No :115   None    : 49   No : 49
##  Yes:391   Yes:666   Yes:500         Yes:615   Mild    :223   Yes:681
##                                                Moderate:338
##                                                Severe  :120
##
##
##   CoughIntensity CoughYN2      Myalgia    MyalgiaYN RunnyNose AbPain
##  None    : 47    No : 47   None    : 79   No : 79   No :211   No :639
##  Mild    :154    Yes:683   Mild    :213   Yes:651   Yes:519   Yes: 91
##  Moderate:357              Moderate:325
##  Severe  :172              Severe  :113
##
##
##  ChestPain Diarrhea  EyePn     Insomnia  ItchyEye  Nausea    EarPn
##  No :497   No :631   No :617   No :315   No :551   No :475   No :568
##  Yes:233   Yes: 99   Yes:113   Yes:415   Yes:179   Yes:255   Yes:162
##
##
##
##
##  Hearing   Pharyngitis Breathless ToothPn   Vision    Vomit     Wheeze
##  No :700   No :119     No :436    No :565   No :711   No :652   No :510
##  Yes: 30   Yes:611     Yes:294    Yes:165   Yes: 19   Yes: 78   Yes:220
##
##
##
##
##     BodyTemp
##  Min.   : 97.20
```

```
##  1st Qu.: 98.20
##  Median : 98.50
##  Mean   : 98.94
##  3rd Qu.: 99.30
##  Max.   :103.10
```

My first step in this analysis is to identify the most important variables so I can produce relevant numerical outputs. Most of the variables present in this dataset are binary, describing the presence of a symptom in a patient. There are three variables (Weakness,CoughIntensity,and Myalgia) with four factor levels and one variable (BodyTemp) with continuous numerical data. The goal of this analysis is to create statistical models with tidymodels.

For the purposes of selecting significant variables, I'll begin by examining the significance between BodyTemp and Weakness, CoughIntensity, and Myalgia using regression analysis. I'll then examine bit variables with high positivity rates. For example, more than 6 out of 7 patients responded Yes to having Weakness, Cough, Myalgia, Headache, and Pharyngitis.

I began by testing body temperature against Weakness, CoughIntensity and Myalgia. In this statistical summary, we can see that there is a p-value of about .02, meaning there are no strong correlations between any of these variables and BodyTemp. A Weakness rating of WeaknessSevere had the strongest correlation compared to the rest.

```
temp_lm1 <- lm(BodyTemp ~ Weakness + CoughIntensity + Myalgia, data = df)
summary(temp_lm1)
```

```
##
## Call:
## lm(formula = BodyTemp ~ Weakness + CoughIntensity + Myalgia,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8675 -0.7422 -0.3946  0.3478  4.2360
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              98.1916     0.2553 384.590  < 2e-16 ***
## WeaknessMild              0.2024     0.1927   1.050  0.29400
## WeaknessModerate          0.3043     0.1992   1.528  0.12700
## WeaknessSevere            0.6301     0.2304   2.735  0.00639 **
## CoughIntensityMild        0.3457     0.1983   1.743  0.08169 .
## CoughIntensityModerate    0.2784     0.1849   1.505  0.13265
## CoughIntensitySevere      0.3208     0.1961   1.636  0.10226
## MyalgiaMild               0.2798     0.1600   1.748  0.08087 .
## MyalgiaModerate           0.1243     0.1644   0.756  0.44987
## MyalgiaSevere             0.1001     0.2016   0.497  0.61968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.188 on 720 degrees of freedom
## Multiple R-squared:  0.02536,    Adjusted R-squared:  0.01318
## F-statistic: 2.082 on 9 and 720 DF,  p-value: 0.02895
```

Testing BodyTemp against Weakness, it becomes more clear that there is significance between body temperature and severe weakness (p-value of .0086).

```
temp_weak <- lm(BodyTemp ~ Weakness, data = df)
summary(temp_weak)
```

```
##
## Call:
## lm(formula = BodyTemp ~ Weakness, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8275 -0.7275 -0.4082  0.3749  4.2354
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       98.6082     0.1699 580.428  < 2e-16 ***
## WeaknessMild       0.2564     0.1876   1.367  0.17218
## WeaknessModerate   0.3170     0.1818   1.744  0.08163 .
## WeaknessSevere     0.6193     0.2016   3.072  0.00221 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.189 on 726 degrees of freedom
## Multiple R-squared:  0.01595,    Adjusted R-squared:  0.01188
## F-statistic: 3.922 on 3 and 726 DF,  p-value: 0.008565
```

I'll repeat this process for the bit variables. In this model, the p-value is .0074, meaning at least one of the relationships is significant. This read-out suggests the greatest significance is between BodyTemp and Pharyngitis.

```
temp_lm2 <- lm(BodyTemp ~ WeaknessYN + CoughYN2 + MyalgiaYN + Headache + Pharyngitis, data = df)
summary(temp_lm2)
```

```
##
## Call:
## lm(formula = BodyTemp ~ WeaknessYN + CoughYN2 + MyalgiaYN + Headache +
##     Pharyngitis, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7559 -0.7559 -0.3559  0.3441  4.0994
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    97.88774    0.28688 341.213   <2e-16 ***
## WeaknessYNYes   0.25654    0.18408   1.394   0.1638
## CoughYN2Yes     0.31115    0.17949   1.733   0.0834 .
## MyalgiaYNYes    0.25533    0.15015   1.700   0.0895 .
## HeadacheYes     0.05046    0.12331   0.409   0.6825
## PharyngitisYes  0.29472    0.11907   2.475   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.187 on 724 degrees of freedom
```

```
## Multiple R-squared:  0.02157,    Adjusted R-squared:  0.01481
## F-statistic: 3.192 on 5 and 724 DF,  p-value: 0.007386
```

Upon further examination, we see that the p-value for a regression with BodyTemp and Pharyngitis is .0182.

```
temp_phar <- lm(BodyTemp ~ Pharyngitis, data = df)
summary(temp_phar)
```

```
##
## Call:
## lm(formula = BodyTemp ~ Pharyngitis, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7812 -0.7812 -0.3812  0.3188  4.1188
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     98.6983     0.1093 902.801   <2e-16 ***
## PharyngitisYes   0.2829     0.1195   2.367   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.193 on 728 degrees of freedom
## Multiple R-squared:  0.007638,   Adjusted R-squared:  0.006275
## F-statistic: 5.603 on 1 and 728 DF,  p-value: 0.01819
```

I'll complete my analysis by testing remaining bit variables against BodyTemp and seeing if there is any significance. In this regression, the highest significance is between BodyTemp and SubjectiveFever and BodyTemp and Sneeze. BodyTemp and Fatigue also seems significant.

```
remainder <- lm(BodyTemp ~ SwollenLymphNodes + ChestCongestion + ChillsSweats + NasalCongestion + Sneeze
summary(remainder)
```

```
##
## Call:
## lm(formula = BodyTemp ~ SwollenLymphNodes + ChestCongestion +
##     ChillsSweats + NasalCongestion + Sneeze + Fatigue + SubjectiveFever +
##     RunnyNose + AbPain + ChestPain + Diarrhea + EyePn + Insomnia +
##     ItchyEye + Nausea + EarPn + Hearing + Breathless + ToothPn +
##     Vision + Vomit + Wheeze, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9559 -0.7332 -0.3080  0.3856  4.3917
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          98.48295    0.19326 509.587  < 2e-16 ***
## SwollenLymphNodesYes -0.12848    0.08959  -1.434 0.151988
## ChestCongestionYes    0.10787    0.09376   1.151 0.250309
## ChillsSweatsYes       0.15715    0.12381   1.269 0.204771
```

```
## NasalCongestionYes   -0.19113    0.11346  -1.685 0.092513 .
## SneezeYes            -0.34871    0.09792  -3.561 0.000394 ***
## FatigueYes            0.33125    0.15432   2.146 0.032177 *
## SubjectiveFeverYes    0.46710    0.10115   4.618 4.61e-06 ***
## RunnyNoseYes         -0.10390    0.10787  -0.963 0.335776
## AbPainYes             0.01505    0.13867   0.109 0.913618
## ChestPainYes          0.09249    0.10431   0.887 0.375515
## DiarrheaYes          -0.15715    0.12861  -1.222 0.222177
## EyePnYes              0.13545    0.12916   1.049 0.294681
## InsomniaYes           0.00550    0.08819   0.062 0.950292
## ItchyEyeYes           0.01492    0.11026   0.135 0.892382
## NauseaYes            -0.01594    0.10079  -0.158 0.874378
## EarPnYes              0.12721    0.11222   1.134 0.257336
## HearingYes            0.24364    0.22269   1.094 0.274306
## BreathlessYes         0.09954    0.09949   1.001 0.317394
## ToothPnYes           -0.04906    0.11312  -0.434 0.664654
## VisionYes            -0.26993    0.27546  -0.980 0.327450
## VomitYes              0.15172    0.15039   1.009 0.313391
## WheezeYes            -0.01681    0.10364  -0.162 0.871234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.15 on 707 degrees of freedom
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.07586
## F-statistic:  3.72 on 22 and 707 DF,  p-value: 2.832e-08
```

p-value of .00000002329

```
temp_fever <- lm(BodyTemp ~ SubjectiveFever, data = df)
summary(temp_fever)
```

```
##
## Call:
## lm(formula = BodyTemp ~ SubjectiveFever, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9012 -0.7739 -0.3739  0.3988  4.4261
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        98.57391    0.07726 1275.799  < 2e-16 ***
## SubjectiveFeverYes  0.52729    0.09336    5.648 2.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.172 on 728 degrees of freedom
## Multiple R-squared:  0.04198,    Adjusted R-squared:  0.04066
## F-statistic:  31.9 on 1 and 728 DF,  p-value: 2.329e-08
```

p-value of .000006037

```
temp_sneeze <- lm(BodyTemp ~ Sneeze, data = df)
summary(temp_sneeze)
```

```
##
## Call:
## lm(formula = BodyTemp ~ Sneeze, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9490 -0.7496 -0.3490  0.3504  4.2504
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 99.14897    0.06411 1546.478  < 2e-16 ***
## SneezeYes   -0.39935    0.08760   -4.559 6.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.18 on 728 degrees of freedom
## Multiple R-squared:  0.02775,    Adjusted R-squared:  0.02642
## F-statistic: 20.78 on 1 and 728 DF,  p-value: 6.037e-06
```

p-value of .0144

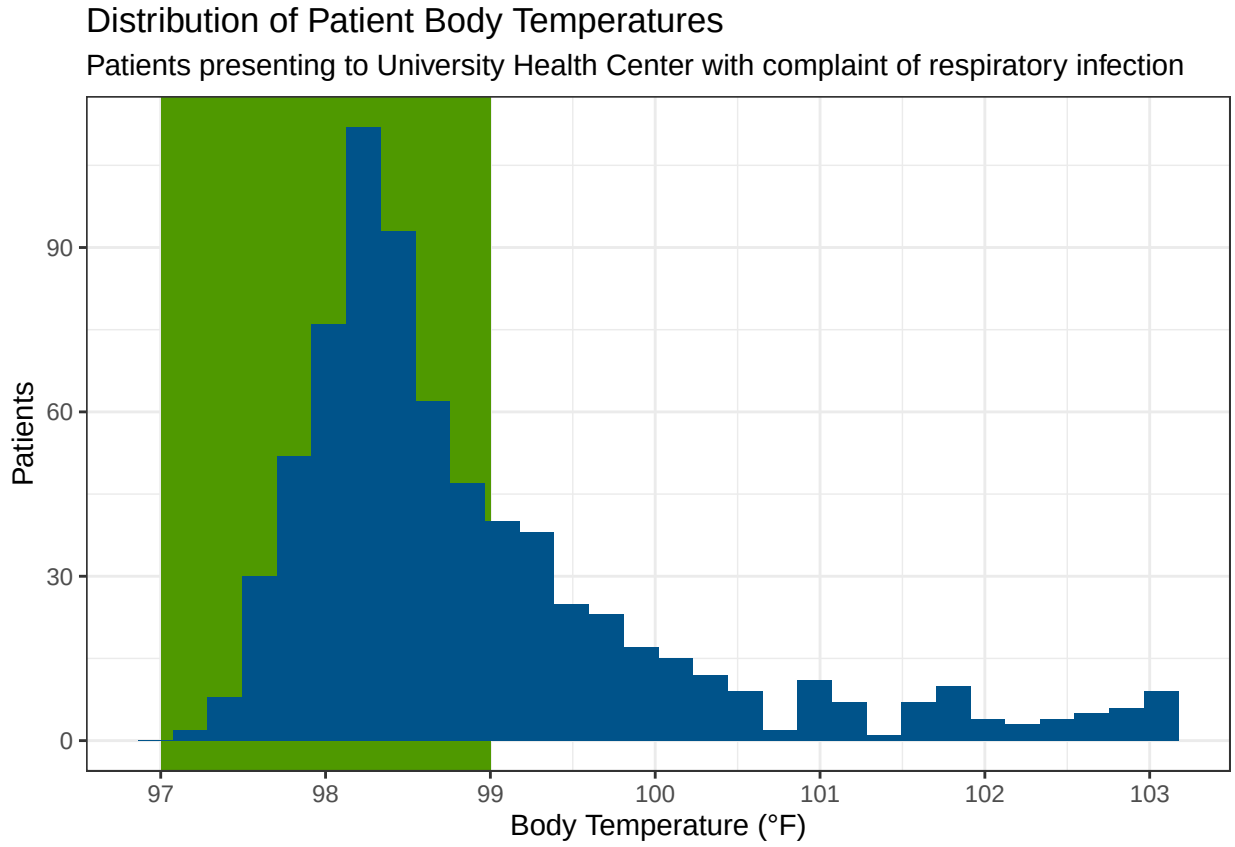
```
Fatigue <- lm(BodyTemp ~ Fatigue, data = df)
summary(Fatigue)
```

```
##
## Call:
## lm(formula = BodyTemp ~ Fatigue, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7686 -0.7686 -0.3686  0.3314  4.1314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.5859     0.1490 661.509   <2e-16 ***
## FatigueYes    0.3827     0.1560   2.453   0.0144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 728 degrees of freedom
## Multiple R-squared:  0.008195,   Adjusted R-squared:  0.006833
## F-statistic: 6.015 on 1 and 728 DF,  p-value: 0.01441
```

The following histogram shows the range of body temperatures patients have when they present to the University Health Center with a complaint related to a respiratory infection. Values highlighted within the green area represent an approximate normal temperature, adjusting for natural variation.

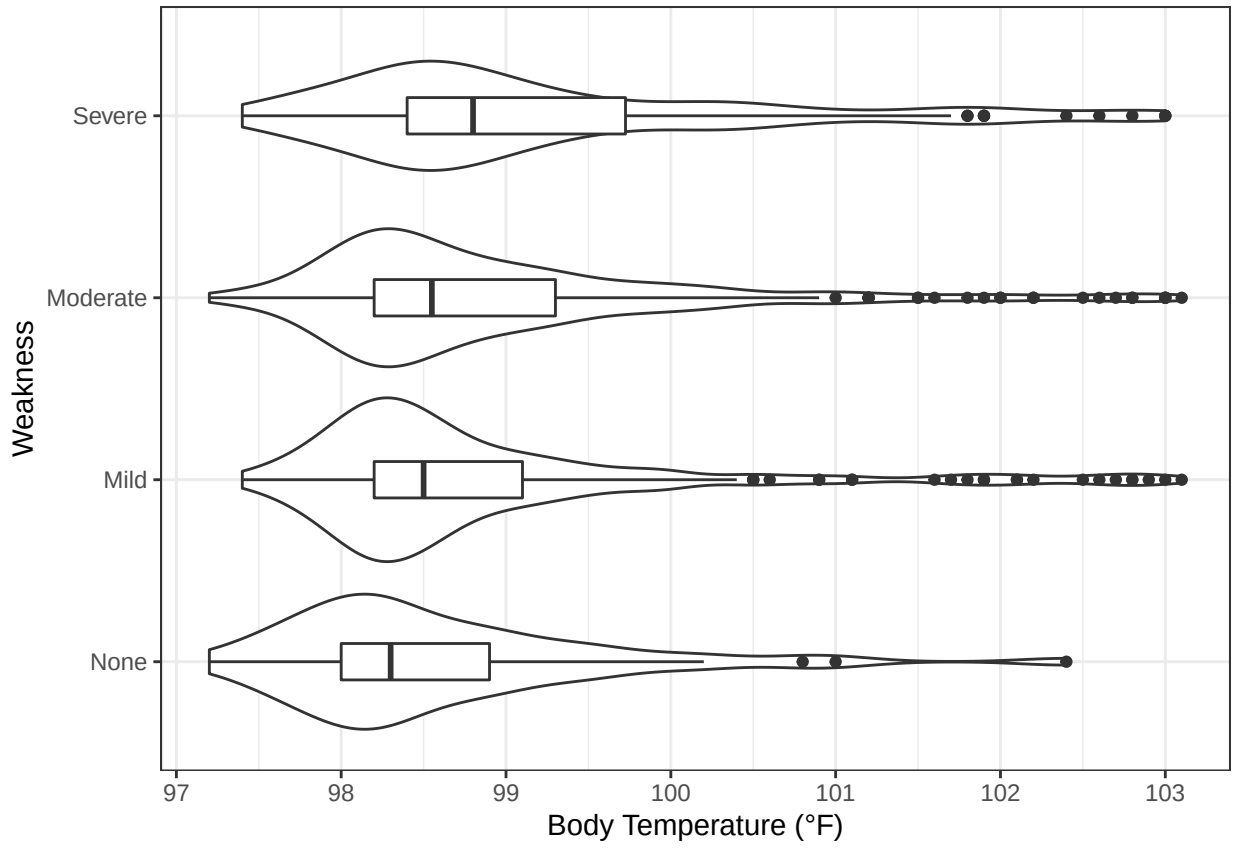
```r
#For each (important) continuous variable, create a histogram or density plot.
body_temp_hist <- df %>% ggplot(aes(x=BodyTemp))+
  geom_rect(mapping=aes(xmin = 97, xmax = 99, ymin = -Inf, ymax = Inf), fill="#4f9900", alpha=.01, inhe
  geom_histogram(fill = "#00538a")+
  theme_bw()+
  scale_x_continuous(breaks = c(97:103))+
  labs(title = "Distribution of Patient Body Temperatures",subtitle= "Patients presenting to University

body_temp_hist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

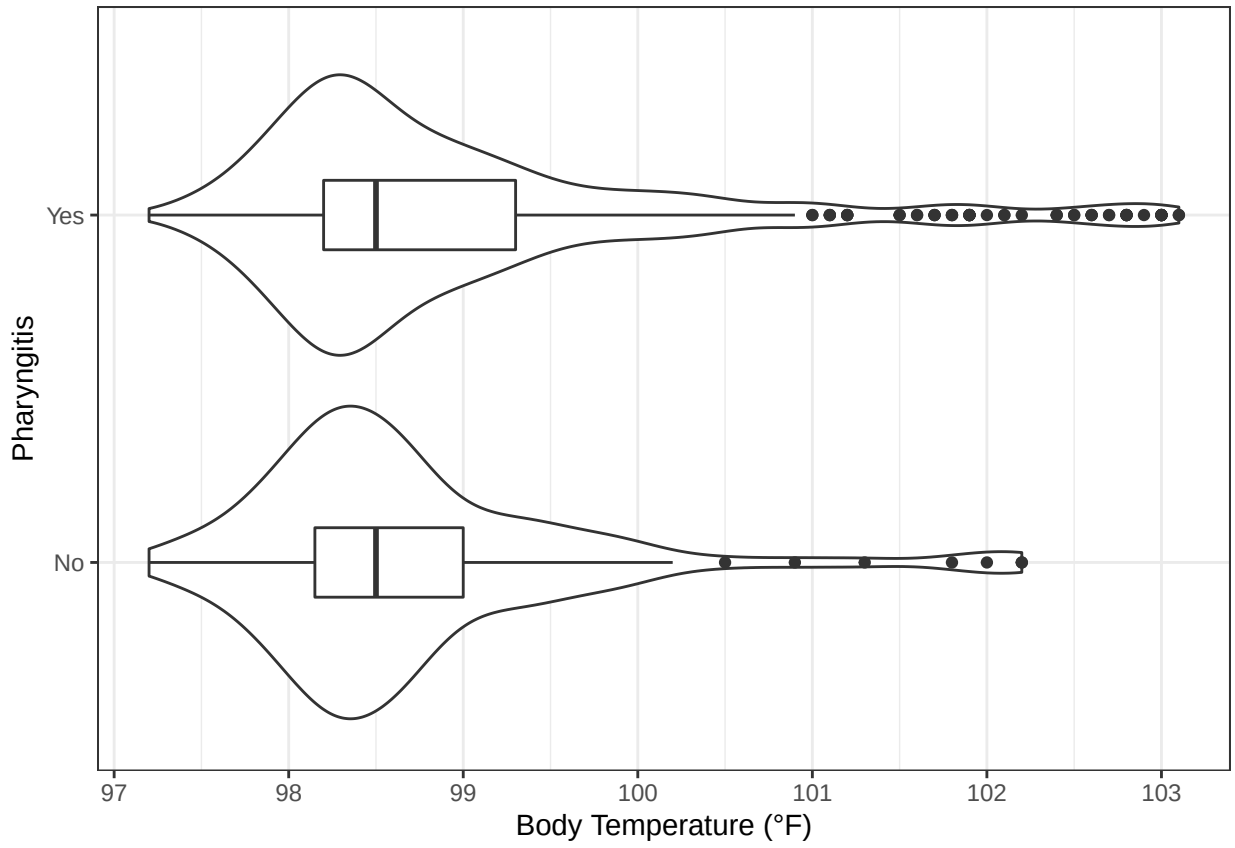


The following series of boxplots are meant to demonstrate the relationship between the body temperature of University Health Center Patients and their responses to symptoms they reported.

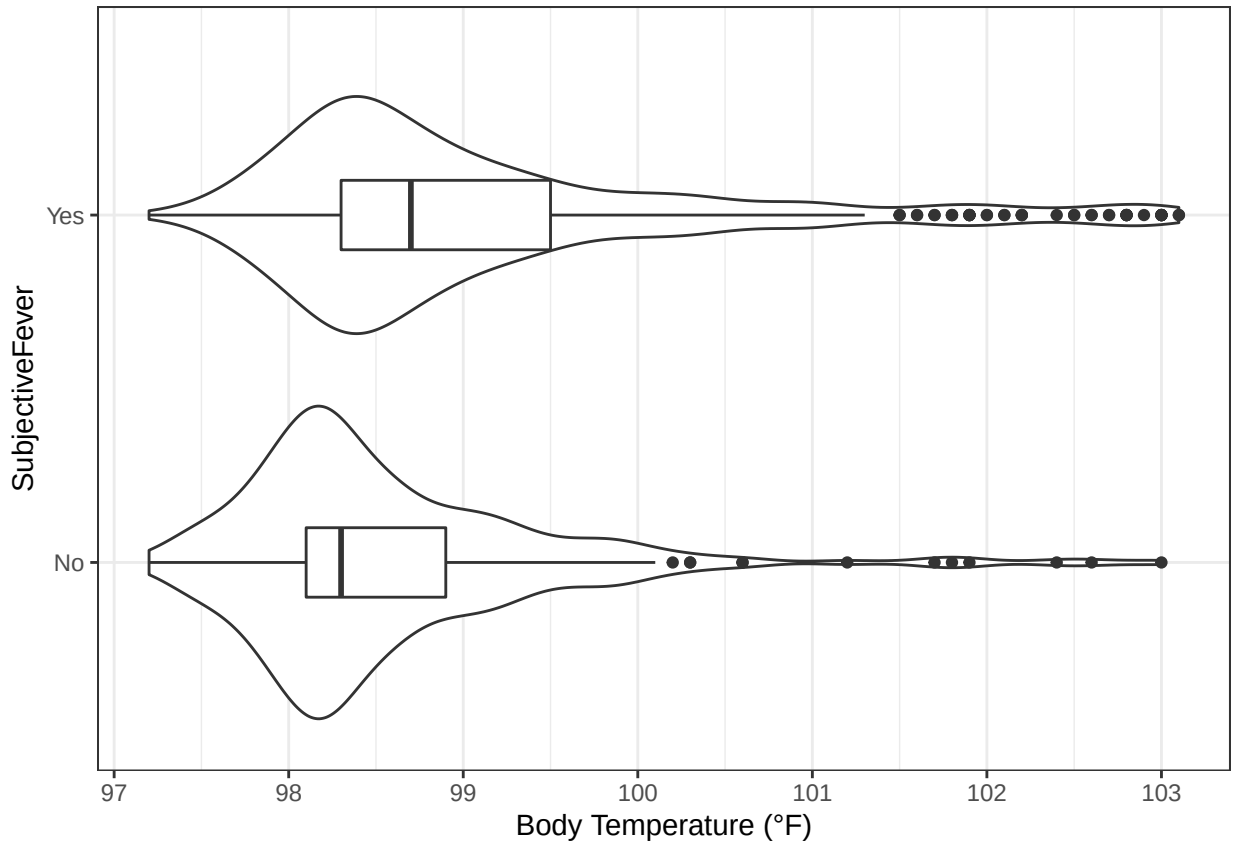
```r
#Create scatterplots or boxplots or similar such plots for the variable you decided is your main outcom

weakness_boxplot <- df %>% ggplot(aes(x=BodyTemp, y = Weakness))+
  geom_violin()+
  geom_boxplot(width = .2)+
  theme_bw()+
  labs(x= "Body Temperature (°F)")
weakness_boxplot
```



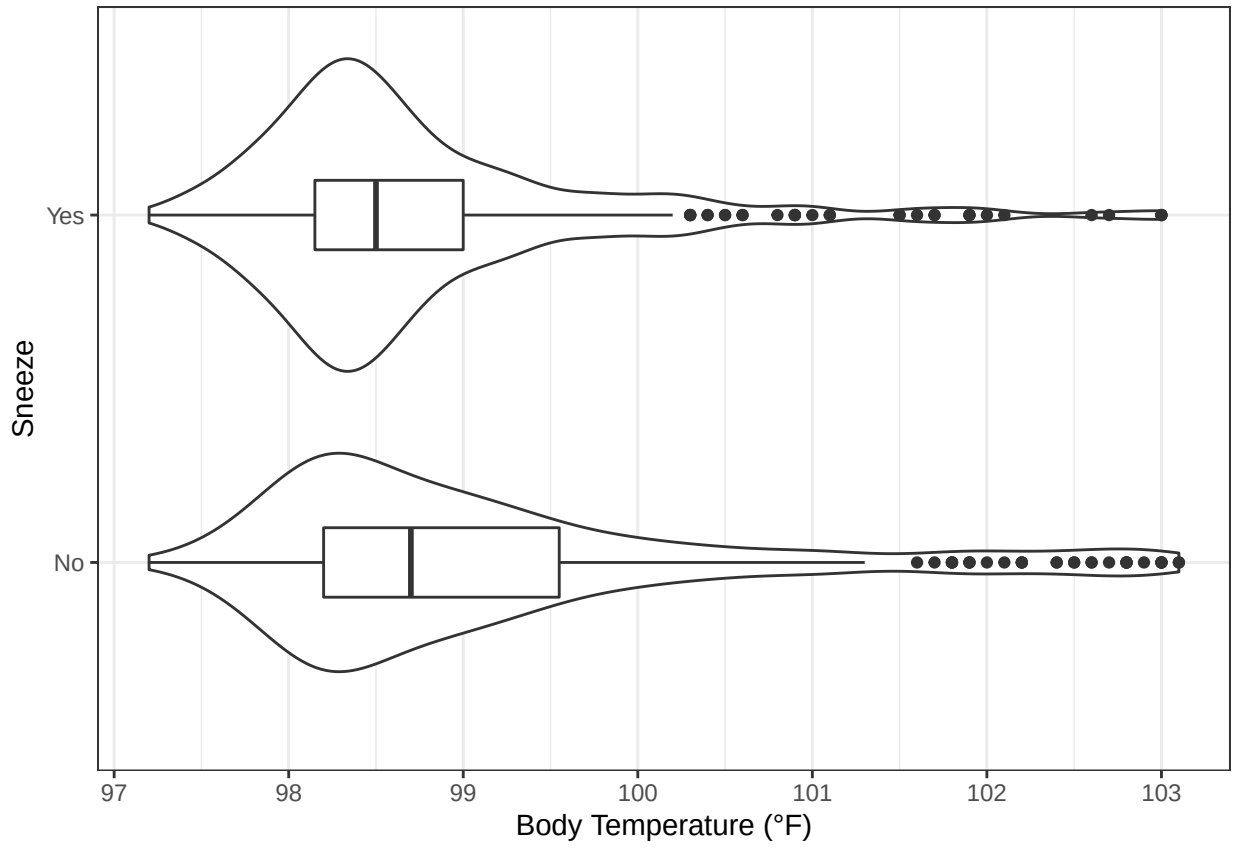

```
phar_boxplot <- df %>% ggplot(aes(x=BodyTemp, y = Pharyngitis))+
  geom_violin()+
  geom_boxplot(width = .2)+
  theme_bw()+
  labs(x= "Body Temperature (°F)")
phar_boxplot
```



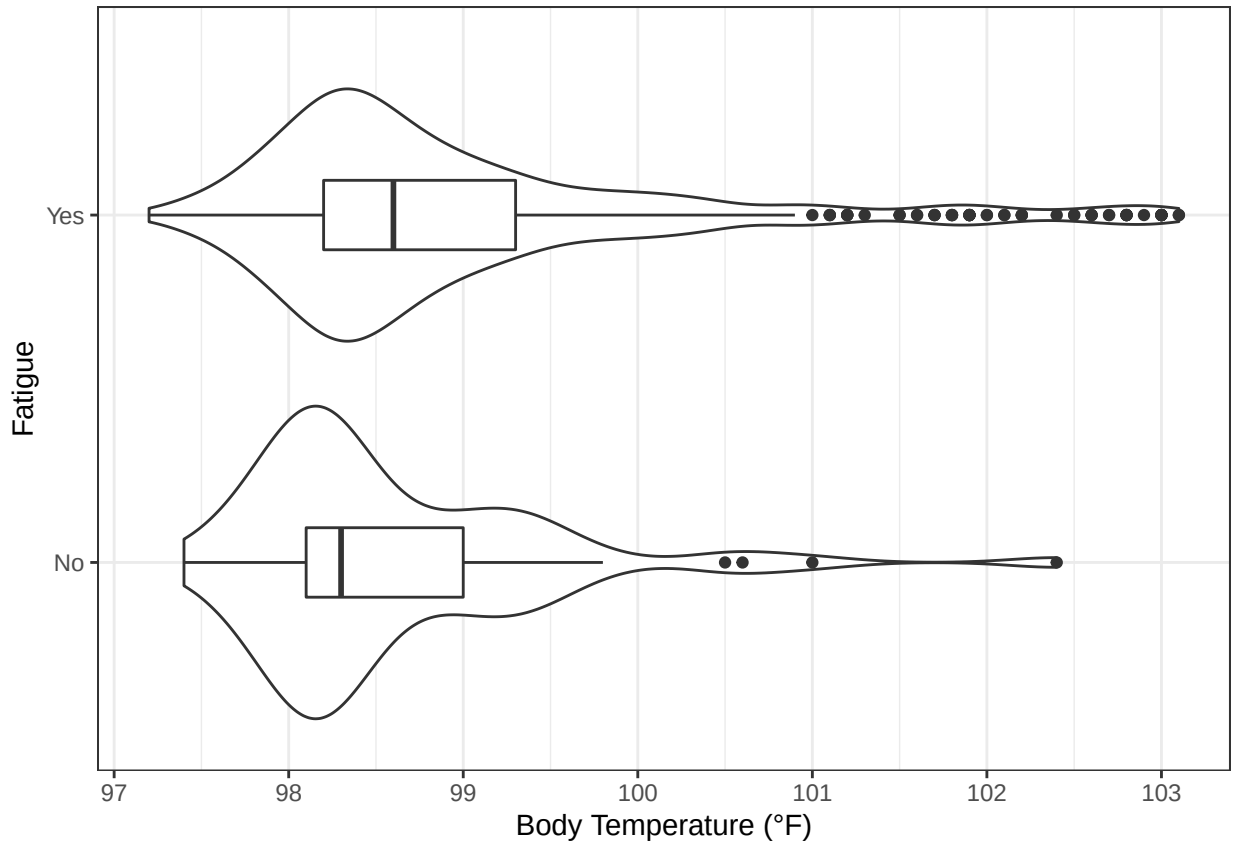

```
fever_boxplot <- df %>% ggplot(aes(x=BodyTemp, y = SubjectiveFever))+
  geom_violin()+
  geom_boxplot(width = .2)+
  theme_bw()+
  labs(x= "Body Temperature (°F)")
fever_boxplot
```




```
sneeze_boxplot <- df %>% ggplot(aes(x=BodyTemp, y = Sneeze))+
  geom_violin()+
  geom_boxplot(width = .2)+
  theme_bw()+
  labs(x= "Body Temperature (°F)")
sneeze_boxplot
```




```
fatigue_boxplot <- df %>% ggplot(aes(x=BodyTemp, y = Fatigue))+
  geom_violin()+
  geom_boxplot(width = .2)+
  theme_bw()+
  labs(x= "Body Temperature (°F)")
fatigue_boxplot
```



Through this exploration, there is a clear visual guide showing the strong relationship between body temperature and weakness. While the graphical representation for the relationship between body temperature and sneezing is less obvious in showing a relationship, the p-value is still one of the highest of all the variables tested. Similarly, the relationship between body temperature and perceived fever is strong. Given the obvious nature of this relationship, this variable will come secondary in modeling. The primary variable I will test is Sneezing.