Assignment 6

Brian Detweiler

May 4, 2017

1. Factors of diabetes: For this problem please use the `diabetes.csv` data file uploaded on the blackboard. Our goal is to understand what factors are important for diabetes. To serve this purpose, write a short report that has the following information

```
set.seed(48548493)
dat <- read.csv('diabetes.csv')
```

a) Summary or overview of the data

```
summary(dat)
```

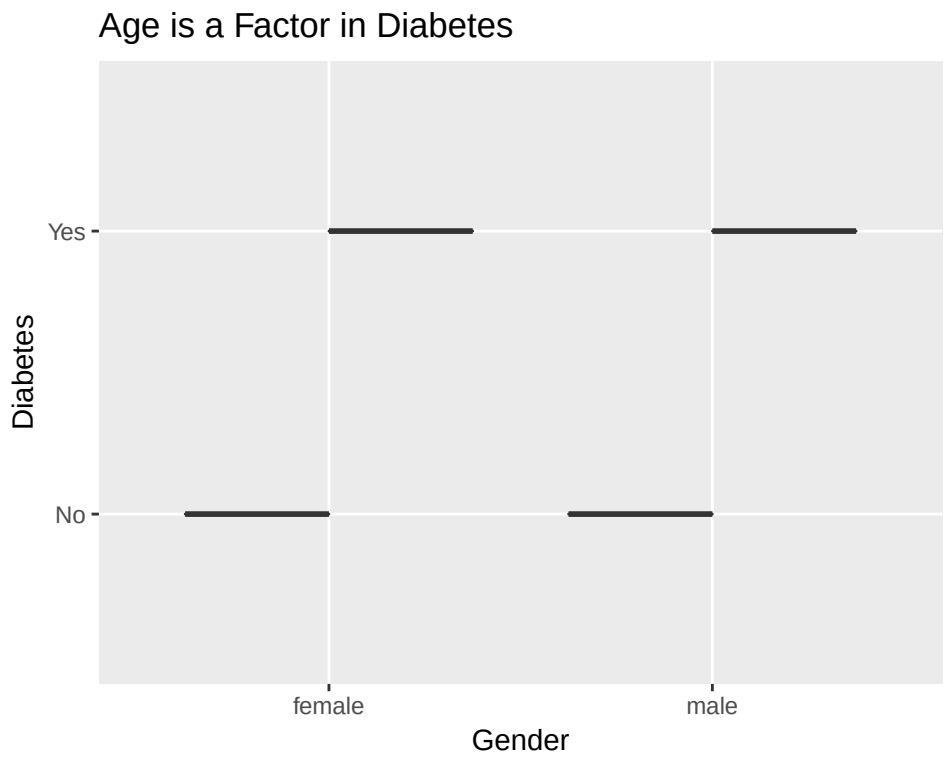
```
##        ID              Age           Gender     Diabetes         BMI
##  Min.   :51624   Min.   : 0.00   female:5020   No  :9098   Min.   :12.88
##  1st Qu.:56905   1st Qu.:17.00   male  :4980   Yes : 760   1st Qu.:21.58
##  Median :62160   Median :36.00                 NA's: 142   Median :25.98
##  Mean   :61945   Mean   :36.74                             Mean   :26.66
##  3rd Qu.:67039   3rd Qu.:54.00                             3rd Qu.:30.89
##  Max.   :71915   Max.   :80.00                             Max.   :81.25
##                                                            NA's   :366
##         HHIncome    PhysActive       Race1              Work
##  more 99999 :2220   No  :3677   Black   :1197   Looking   : 311
##  75000-99999:1084   Yes :4649   Hispanic: 610   NotWorking:2847
##  25000-34999: 958   NA's:1674   Mexican :1015   Working   :4613
##  35000-44999: 863               Other   : 806   NA's      :2229
##  45000-54999: 784               White   :6372
##  (Other)    :3280
##  NA's       : 811
##     BPSysAve        BPDiaAve          Pulse
##  Min.   : 76.0   Min.   :  0.00   Min.   : 40.00
##  1st Qu.:106.0   1st Qu.: 61.00   1st Qu.: 64.00
##  Median :116.0   Median : 69.00   Median : 72.00
##  Mean   :118.2   Mean   : 67.48   Mean   : 73.56
##  3rd Qu.:127.0   3rd Qu.: 76.00   3rd Qu.: 82.00
##  Max.   :226.0   Max.   :116.00   Max.   :136.00
##  NA's   :1449    NA's   :1449     NA's   :1437
```

```
# Exploratory Data Analysis Step
# ggpairs(dat)
```

b) Five plots showing important factors of diabetes. Include brief descriptions of what each plot is revealing.

Age vs. Diabetes

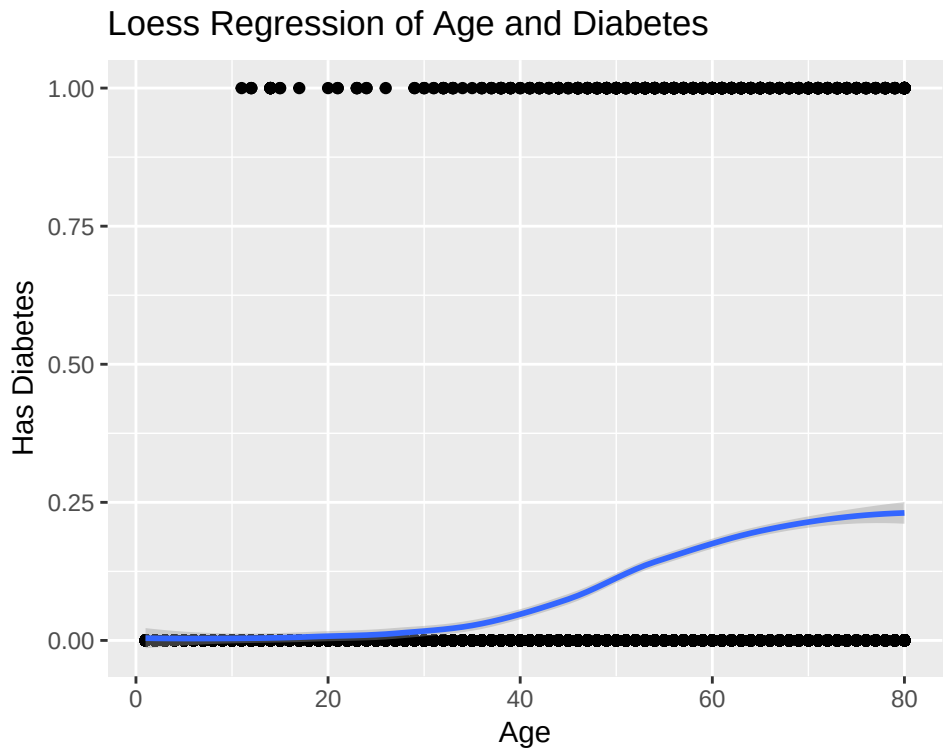
```
dat.mod <- dat %>% filter(!is.na(Diabetes))

ggplot(dat.mod, aes(x=Gender, y=Diabetes)) +
  geom_boxplot() +
  labs(title="Age is a Factor in Diabetes")
```




```
  labs(title="Age is a Factor in Diabetes")
```

```
## $title
## [1] "Age is a Factor in Diabetes"
##
## attr(,"class")
## [1] "labels"
```

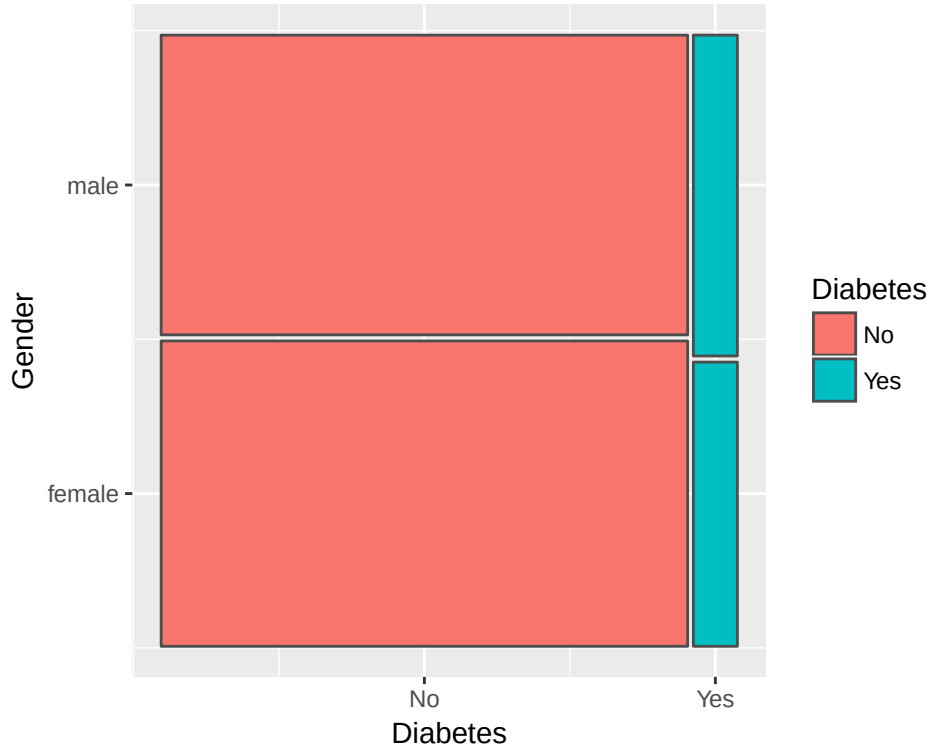
```
ggplot(data = dat.mod, aes(x=Age, y=as.numeric(Diabetes)-1)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title="Loess Regression of Age and Diabetes", y="Has Diabetes")
```



We can see that by the upper and lower quantiles of the "Yes" and "No" categorical variables, diabetes tends to have most of its effect on people over the age of 50.

Gender vs. Diabetes

```
# Female = 0
# Male = 1

dat.moz <- dat.mod %>% group_by(Gender, Diabetes) %>% summarise(counts = n())
prodplot(data=dat.moz, counts~Gender + Diabetes, c("vspine","hspine"), na.rm=T,subset=(level==2)) +
  aes(fill=Diabetes)
```




```
# Male has diabetes
O11 <- dat.moz[[4,3]]
# Female has diabetes
O12 <- dat.moz[[2,3]]
# Male no diabetes
O21 <- dat.moz[[3,3]]
# Female no diabetes
O22 <- dat.moz[[1,3]]


C1 <- O11 + O21
C2 <- O12 + O22

n1 <- O11 + O12
n2 <- O21 + O22
N <- n1 + n2

T1 <- (sqrt(N) * (O11 * O22 - O12 * O21)) / sqrt(as.numeric(n1) * as.numeric(n2) * as.numeric(C1) * as.
```

There appears to be a larger proportion of males who have diabetes.

	Male	Female	Total
Has Diabetes	403	357	760
No Diabetes	4506	4592	9098
—	—	—	—
Total	4909	4949	9858

We can perform a Chi-squared Test for Differences in Probabilities using the 2x2 contingency table, such that

H_0 : Equal probability that a randomly selected element will be in class females or males (0, or 1)

H_a : Probability of being in males or females is not equal

With the test statistic

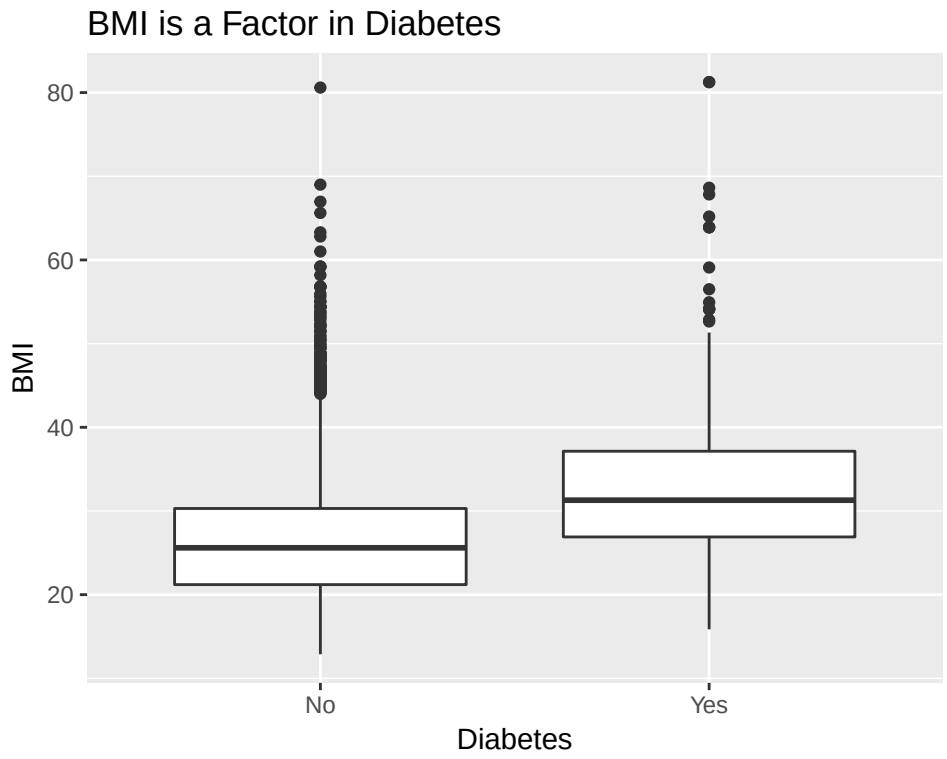
$$\begin{aligned} T_1 &= \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}} \\ &= \frac{\sqrt{9858}(403 \cdot 357 - 357 \cdot 4506)}{\sqrt{760 \cdot 9098 \cdot 4909 \cdot 4949}} \\ &= 1.8533452 \end{aligned}$$

Our test statistic produces a Z-value of 1.8533452, which is less than 1.96, so our p-value is $0.0646 > 0.05$. We cannot say that these populations are different from each other.

BMI vs. Diabetes

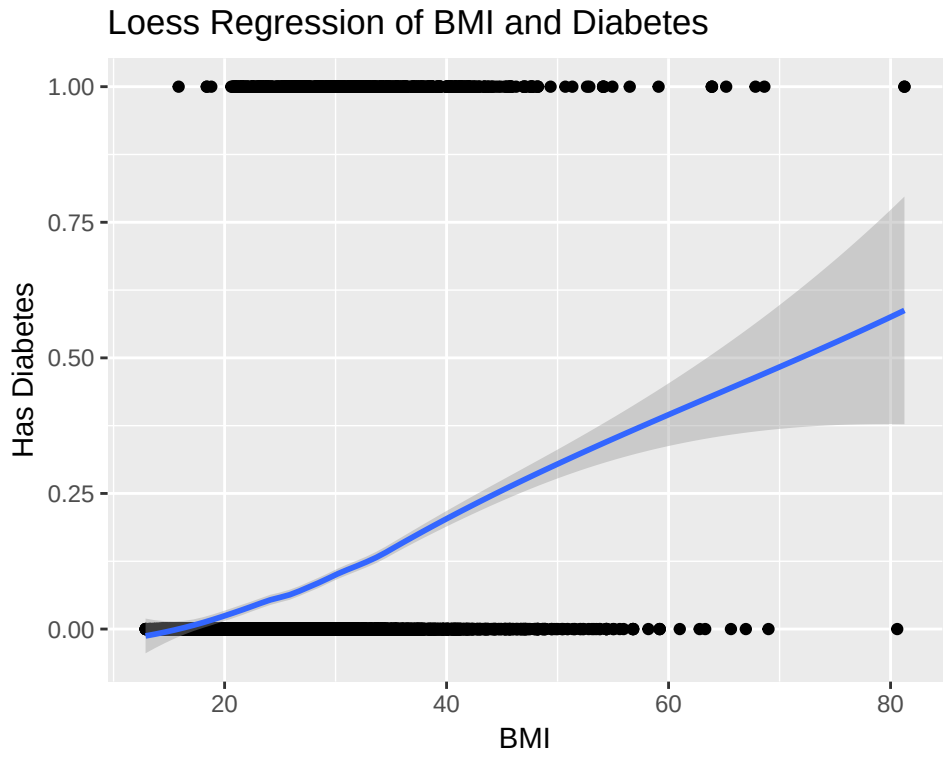
```
ggplot(dat.mod, aes(x=Diabetes, y=BMI)) +
  geom_boxplot() +
  labs(title="BMI is a Factor in Diabetes")
```

```
## Warning: Removed 229 rows containing non-finite values (stat_boxplot).
```




```
dat.mod <- dat.mod %>% filter(!is.na(BMI))

ggplot(data = dat.mod, aes(x=BMI, y=as.numeric(Diabetes)-1)) +
```

```
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title="Loess Regression of BMI and Diabetes", y="Has Diabetes")
```

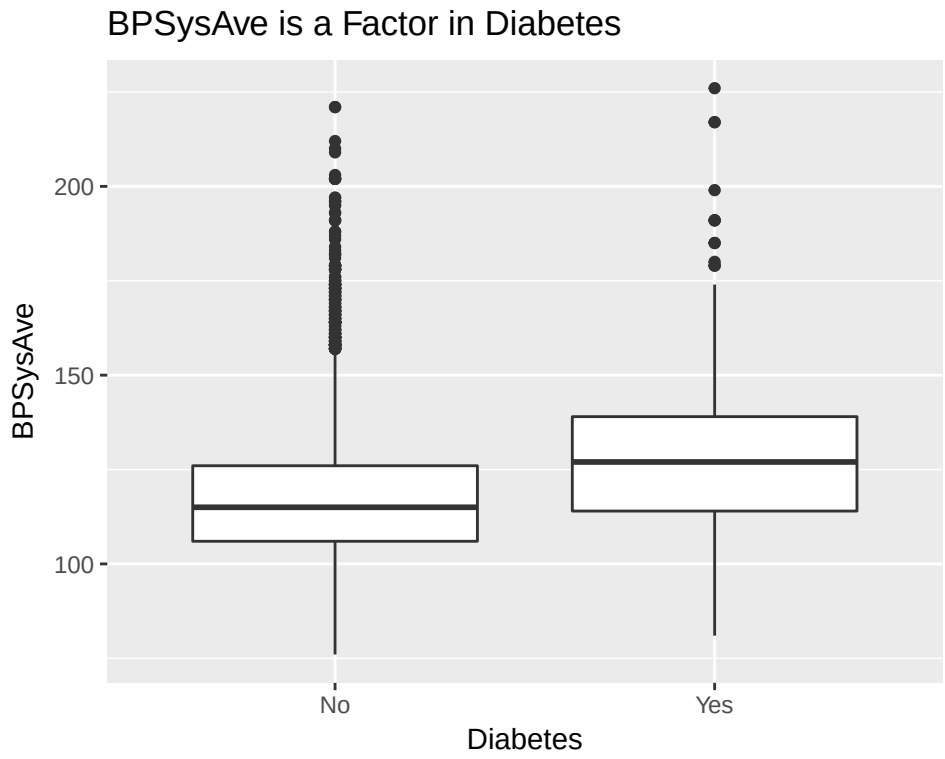


BMI does appear to have a positive effect on Diabetes.

BPSysAve vs. Diabetes

```
ggplot(dat.mod, aes(x=Diabetes, y=BPSysAve)) +
  geom_boxplot() +
  labs(title="BPSysAve is a Factor in Diabetes")
```

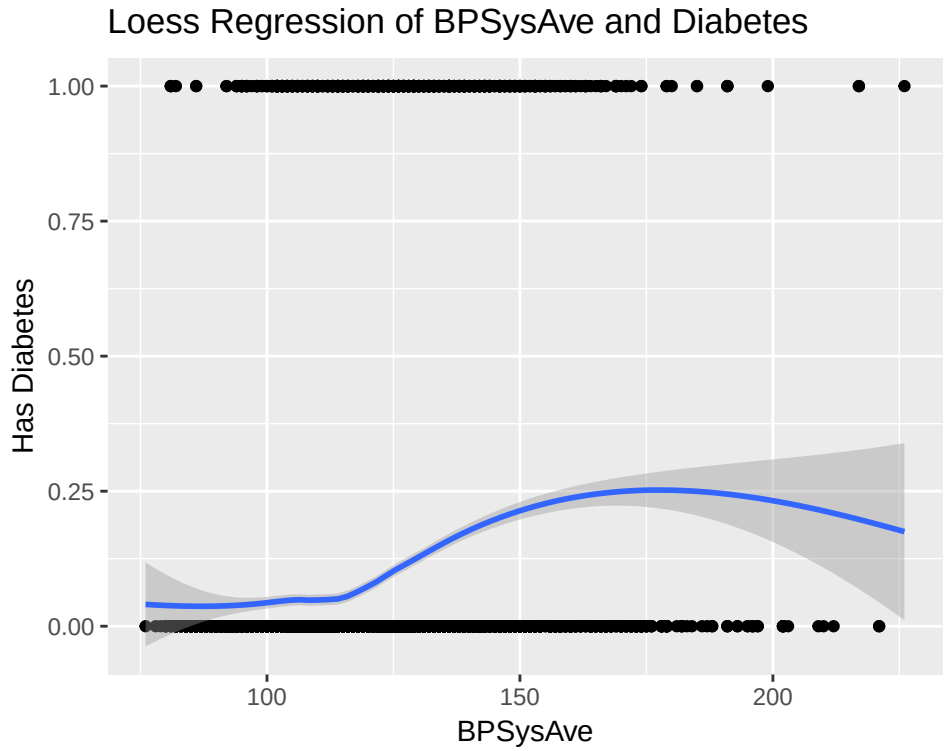
```
## Warning: Removed 1147 rows containing non-finite values (stat_boxplot).
```




```
dat.mod <- dat.mod %>% filter(!is.na(BMI))

ggplot(data = dat.mod, aes(x=BPSysAve, y=as.numeric(Diabetes)-1)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title="Loess Regression of BPSysAve and Diabetes", y="Has Diabetes")
```

```
## Warning: Removed 1147 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1147 rows containing missing values (geom_point).
```

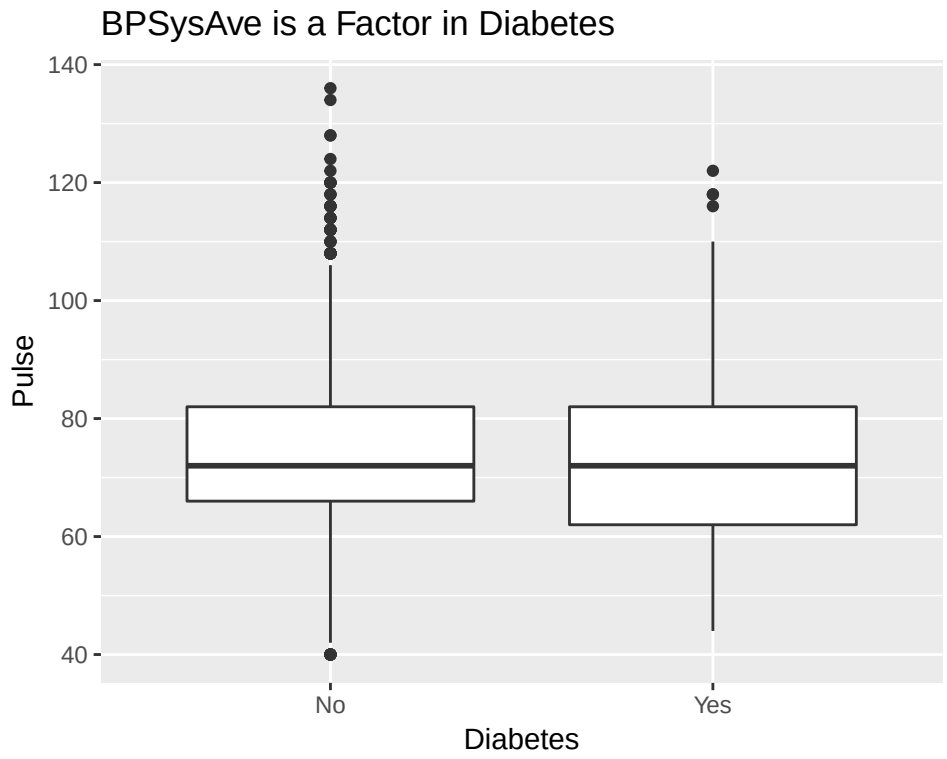


The effect of BPSysAve appears to be uncertain. We can't say much about this.

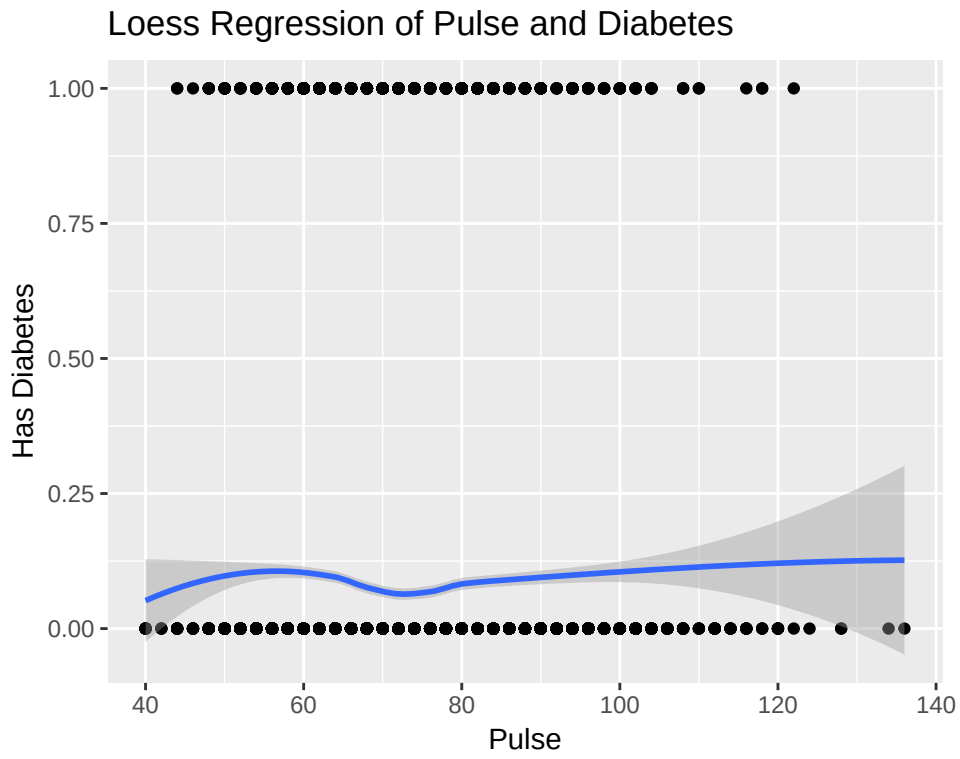
Pulse vs. Diabetes

```
ggplot(dat.mod, aes(x=Diabetes, y=Pulse)) +
  geom_boxplot() +
  labs(title="BPSysAve is a Factor in Diabetes")
```

```
## Warning: Removed 1136 rows containing non-finite values (stat_boxplot).
```




```
dat.mod <- dat.mod %>% filter(!is.na(Pulse))

ggplot(data = dat.mod, aes(x=Pulse, y=as.numeric(Diabetes)-1)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title="Loess Regression of Pulse and Diabetes", y="Has Diabetes")
```



Pulse doesn't appear to have any effect on diabetes.

c) Fit a model and provide the summary

Of the plots shown above, BMI seems to be the biggest factor so we'll fit a model to it.

```
model <- glm(Diabetes ~ BMI, family=binomial(link='logit'), data=dat.mod)
summary(model)
```

```
##
## Call:
## glm(formula = Diabetes ~ BMI, family = binomial(link = "logit"),
##     data = dat.mod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1878  -0.4403  -0.3579  -0.2943   2.6947
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.04949    0.15922  -31.71   <2e-16 ***
## BMI          0.09115    0.00492   18.53   <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: 5006.5  on 8492  degrees of freedom
## Residual deviance: 4658.5  on 8491  degrees of freedom
```

```
## AIC: 4662.5
##
## Number of Fisher Scoring iterations: 5
```

d) Your conclusion

We have a statisticially significant coefficient with BMI with a very low p-value. Therefore we can say that BMI absolutely has a positivie impact on Diabetes.