Project 2 - Diabetes Dataset

Matthew Roland, Jean Jimenez, Kelly Eng

2023-10-03

For this part of project 2, I will be working with the Diabetes dataset uploaded by Folorunsho Atanda on BlackBoard (https://bbhosted.cuny.edu/webapps/discussionboard/do/message?action=list__messages&course__id=__2303106__1&nav=discussion__board__entry&conf__id=__2781274__1&forum__id=__3717486__1&message__id=__69324188__1)

To begin, I will read the dataset into R from my Github repo, add an ID column, clean the data such that the predictors of diabetes are grouped under one column, and then perform simple, descriptive analyses to determine which factors are most greatly associated with diabetes outcomes

Loading the dataset & Adding an ID Variable

```r
diabetes <- read.csv("https://raw.githubusercontent.com/Mattr5541/DATA-607/main/Project%202/diabetes.cs

##Now, I want to add an ID variable, so we can more easily keep track of observations
diabetes <- diabetes %>% mutate(ID = row_number())

##Now, I would like to reorder the columns so that ID appears first
diabetes <- diabetes %>% select(ID, everything())
```

Converting this dataframe to a long format

```r
diabetes_long <- diabetes %>% pivot_longer(cols = c(Pregnancies:BMI, Age), names_to = "Predictors", valu

diabetes_long <- diabetes_long %>% select(ID, Predictors, Values, DiabetesPedigreeFunction, Outcome)
```

Determining mean values for preictors based on outcomes

```r
##Next, I want to split this dataframe into two: one for negative (0) outcomes and one for positive (1)

diabetes_pos <- diabetes_long %>% filter(Outcome == 1)
diabetes_neg <- diabetes_long %>% filter(Outcome == 0)
```

First, I will determine the average pedigree level for individuals with diabetes versus those without diabetes, to determine if this is an appropriate marker for predicting diabetes

```r
print((paste0("Number of people without diabetes: ", nrow(diabetes_neg))))
```

```
## [1] "Number of people without diabetes: 3500"
```

```r
print((paste0("Number of people with diabetes: ", nrow(diabetes_pos))))
```

```
## [1] "Number of people with diabetes: 1876"
```

As we can see, the total number of individuals without diabetes is larger than the number of individuals with diabetes. From the perspective of generalizing to the population, this makes sense

```r
mean_pedigree_neg <- diabetes_neg %>% summarize(mean_pedigree = mean(DiabetesPedigreeFunction))
print(mean_pedigree_neg)
```

```
## # A tibble: 1 x 1
##   mean_pedigree
##           <dbl>
## 1         0.430
```

```r
mean_pedigree_pos <- diabetes_pos %>% summarize(mean_pedigree = mean(DiabetesPedigreeFunction))
print(mean_pedigree_pos)
```

```
## # A tibble: 1 x 1
##   mean_pedigree
##           <dbl>
## 1         0.550
```

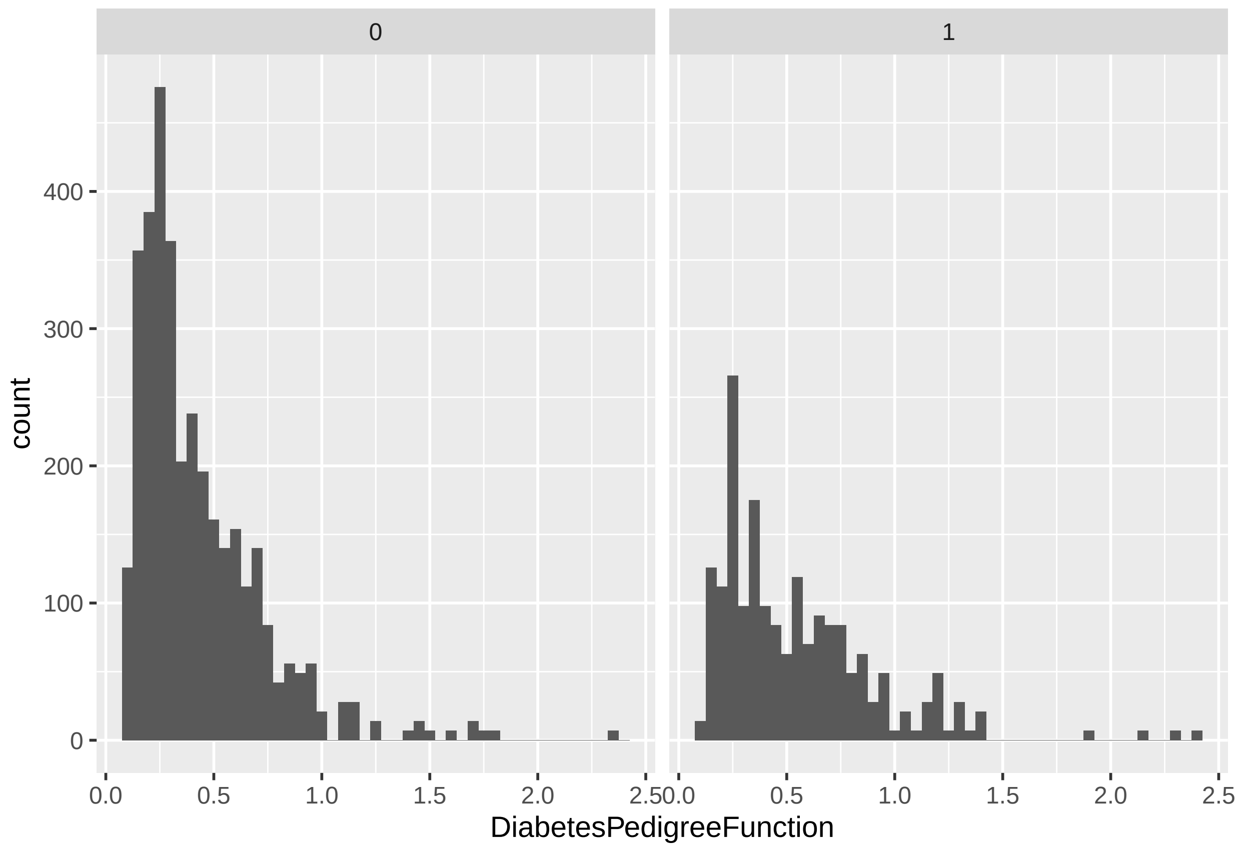
```r
##Median Values
median_pedigree_neg <- diabetes_neg %>% summarize(median_pedigree = median(DiabetesPedigreeFunction))
print(median_pedigree_neg)
```

```
## # A tibble: 1 x 1
##   median_pedigree
##             <dbl>
## 1           0.336
```

```r
median_pedigree_pos <- diabetes_pos %>% summarize(median_pedigree = median(DiabetesPedigreeFunction))
print(median_pedigree_pos)
```

```
## # A tibble: 1 x 1
##   median_pedigree
##             <dbl>
## 1           0.449
```

```r
ggplot(diabetes_long, aes(x = DiabetesPedigreeFunction, y = after_stat(count))) + geom_histogram(binwid
```



All of this shows that the diabetes pedigree function, on average, is marginally higher for individuals diagnosed with diabetes than those who were not diagnosed with diabetes. The graphs show that the distributions for both categories are heavily positively skewed (which is sensible for this type of indicator), and the values tend to cluster closer to 1.0 when people are positive for diabetes than when they test negative for diabetes. Because of this heavy skew, I opted to determine the median values, as well. As we can see, the difference median differences between pedigrees diagnosed with diabetes versus not diagnosed with diabetes is substantially different. However, it may not serve as the most reliable indicator for predicting diabetes, as both the mean and median values for those diagnosed with diabetes are only around .5. It is important to note that this measure, of course, is only a marker for risk factors that lead to diabetes, however, and not of diabetes development, itself.

Now, I will find the averages for each predictor:

```r
unique(diabetes_long$Predictors)
```

```
## [1] "Pregnancies"   "Glucose"       "BloodPressure" "SkinThickness"
## [5] "Insulin"       "BMI"           "Age"
```

```r
diabetes_means_neg <- diabetes_neg %>% group_by(Predictors) %>% summarize(Group_means = mean(Values))

diabetes_means_pos <- diabetes_pos %>% group_by(Predictors) %>% summarize(Group_means = mean(Values))

print(diabetes_means_neg)
```

```
## # A tibble: 7 x 2
##   Predictors    Group_means
##   <chr>               <dbl>
## 1 Age                 31.2
## 2 BMI                 30.3
## 3 BloodPressure       68.2
## 4 Glucose            110.
## 5 Insulin             68.8
## 6 Pregnancies          3.30
## 7 SkinThickness       19.7
```

```
print(diabetes_means_pos)
```

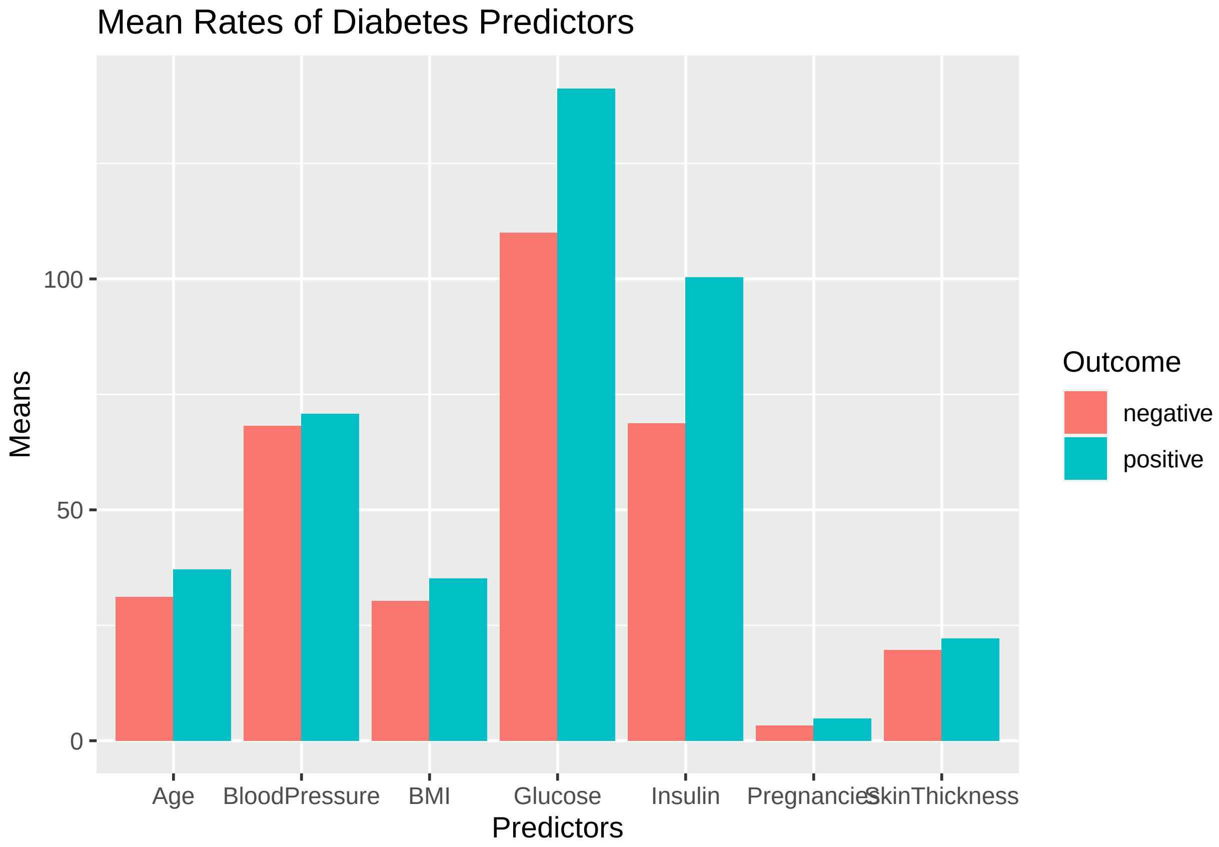
```
## # A tibble: 7 x 2
##   Predictors    Group_means
##   <chr>               <dbl>
## 1 Age                 37.1
## 2 BMI                 35.1
## 3 BloodPressure       70.8
## 4 Glucose            141.
## 5 Insulin            100.
## 6 Pregnancies          4.87
## 7 SkinThickness       22.2
```

We can see that–as expected–all categories are elevated in the positive group when compared to the negative group. Let’s investigate this further by using graphical comparisons

```
##I want to first combine these mean comparisons into a single dataframe
diabetes_means_neg <- data_frame(diabetes_means_neg)
```

```
## Warning: ‘data_frame()‘ was deprecated in tibble 1.1.0.
## i Please use ‘tibble()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
diabetes_means_pos <- data_frame(diabetes_means_pos)

diabetes_means_neg <- diabetes_means_neg %>% mutate(Outcome = 0)
diabetes_means_pos <- diabetes_means_pos %>% mutate(Outcome = 1)

diabetes_mean_append <- rbind(diabetes_means_neg, diabetes_means_pos)

diabetes_mean_append <- diabetes_mean_append %>% mutate(Outcome = recode(Outcome, '0' = 'negative', '1'

ggplot(diabetes_mean_append, aes(x = Predictors, y = Group_means, fill = Outcome)) + geom_bar(stat= "id
```

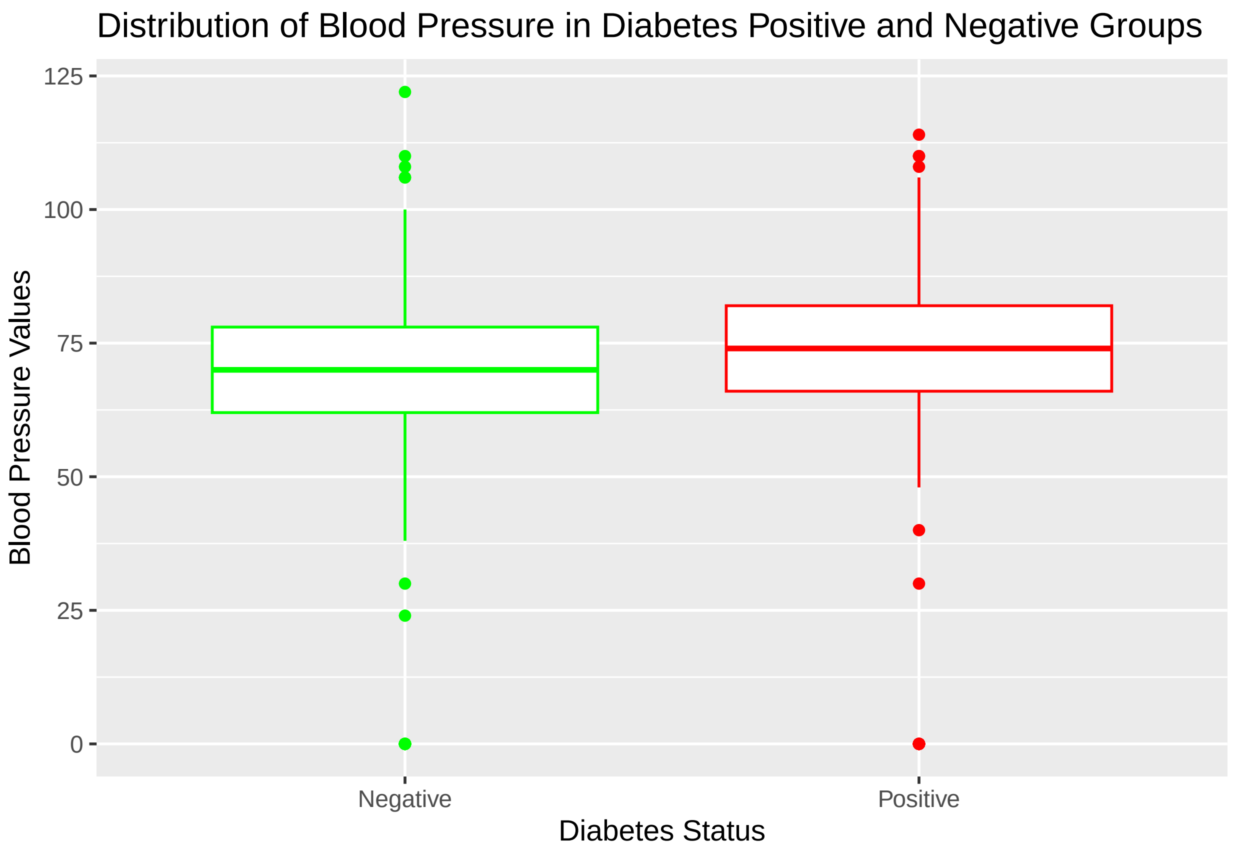


Thus, as one would assume, insulin and glucose appear to be the best predictors of having diabetes, such that highly elevated insulin and glucose levels tend to correspond with the presence of diabetes.

Now I want to see the differences in the distribution of blood pressure between the two groups.

```r
bloodPressure_pos <- diabetes_pos %>%
  filter(Predictors == "BloodPressure") %>%
  mutate(Status = "Positive")

bloodPressure_neg <- diabetes_neg %>%
  filter(Predictors == "BloodPressure")

# create an empty ggplot object
bp = ggplot() +
  ggtitle("Distribution of Blood Pressure in Diabetes Positive and Negative Groups") +
  xlab("Diabetes Status") +
  ylab("Blood Pressure Values")

# Add boxplot for Positive group
bp = bp +
  geom_boxplot(data = bloodPressure_pos, aes(x = "Positive", y = Values), colour='red')

# Add boxplot for Negative group
bp = bp +
  geom_boxplot(data = bloodPressure_neg, aes(x = "Negative", y = Values), colour='green')
```

```
# Display the plot
print(bp)
```



The distributions of blood pressure between the diabetes negative and positive group is similar. However, the mean blood pressure of the diabetes group was slightly above the mean for the negative group. This may indicate that diabetes patients have higher blood pressure (or the other way around).