Final Project

Jed Brough

11/18/2020

Roughly 30 percent of the world’s estimated warehoused data comes from the healthcare industry. There are such large amounts of data to be searched through and so many insights that we could get if we were able to make sense of all the data. I was able to find a dataset on diabetes that provides some interesting opportunities. We know that there are several factors that contribute to diabetes, some of them are genetic and some of them are influenced by lifestyle. I want to use the data to be able to see which variables are most correlated with diabetes and which variables could be affected by lifestyle changes to prevent or manage diabetes. If diabetes could be prevented or managed primarily through lifestyle changes, that could potentially save money and increase the quality of life for many individuals.

For this project, I used a dataset that looked at health factors in women among the Pima Indian tribe. The dataset was downloaded from Kaggle.com. The dataset has 768 observations with 9 variables to work with. I removed the variable Skin Thickness because it had a low correlation with the other variables, most importantly, a low correlation with the response variable Outcome. I removed values with 0 blood pressure because 0 blood pressure indicates either death or an error.

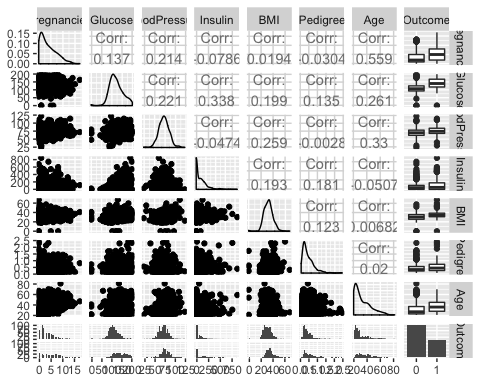
Below are the first 10 observations in the dataset followed by the structure of the dataset that includes each variable and it’s data type.

## Pregnancies Glucose BloodPressure Insulin BMI DiabetesPedigreeFunction  
## 1 6 148 72 0 33.6 0.627  
## 2 1 85 66 0 26.6 0.351  
## 3 8 183 64 0 23.3 0.672  
## 4 1 89 66 94 28.1 0.167  
## 5 0 137 40 168 43.1 2.288  
## 6 5 116 74 0 25.6 0.201  
## Age Outcome  
## 1 50 1  
## 2 31 0  
## 3 32 1  
## 4 21 0  
## 5 33 1  
## 6 30 0

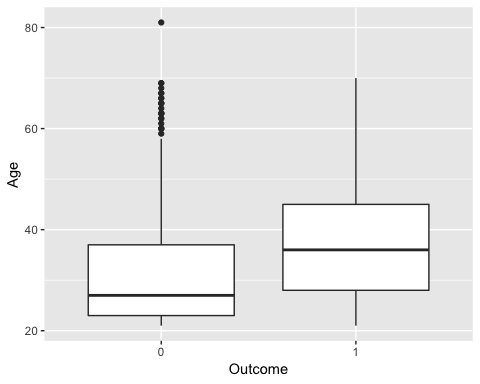
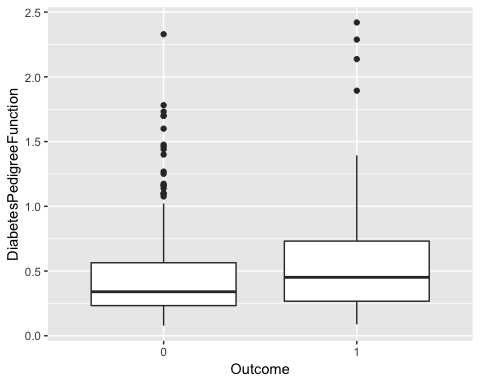
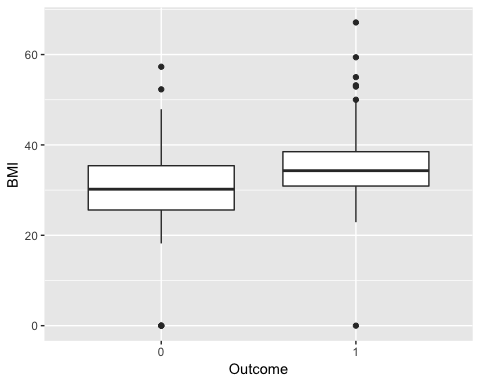
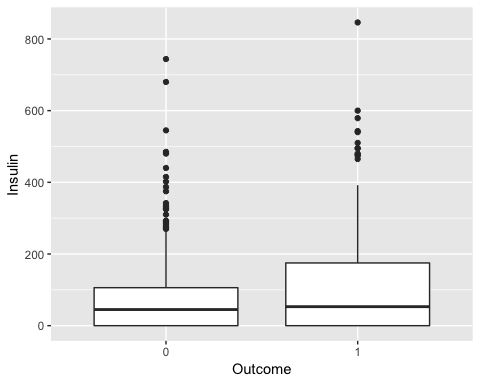
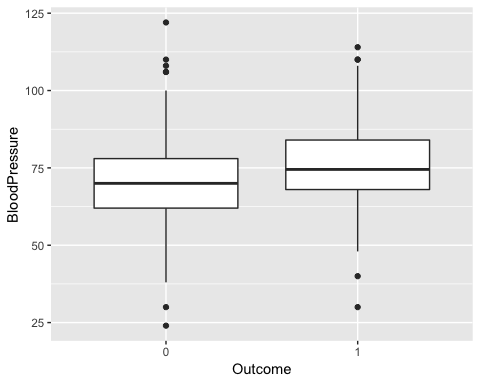
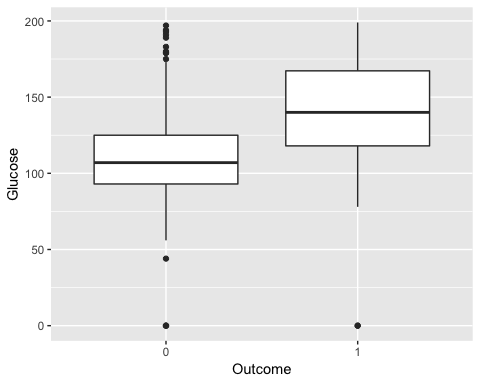
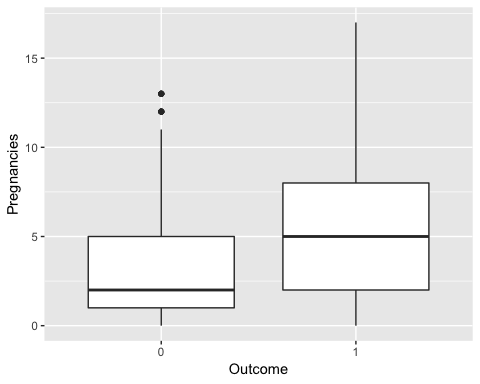
## 'data.frame': 733 obs. of 8 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 2 8 4 ...  
## $ Glucose : int 148 85 183 89 137 116 78 197 125 110 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 70 96 92 ...  
## $ Insulin : int 0 0 0 94 168 0 88 543 0 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 30.5 0 37.6 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 53 54 30 ...  
## $ Outcome : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 2 2 1 ...

Outcome is the response variable that indicates a diabetes diagnosis. 1 indicates a diagnosis of diabetes and 0 indicates a diagnosis of no diabetes. The remaining variables will be used as independent variables.

Below is a plot that shows scatterplots and correlation coefficients of each of the variable combinations and their relationship with the response variable.

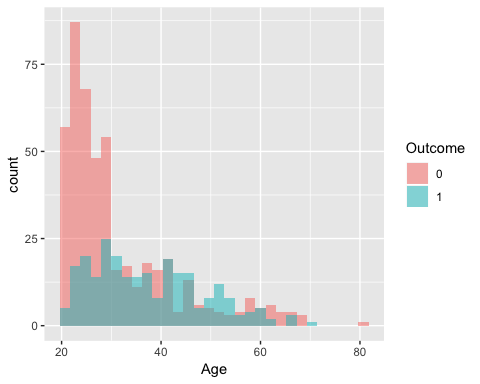
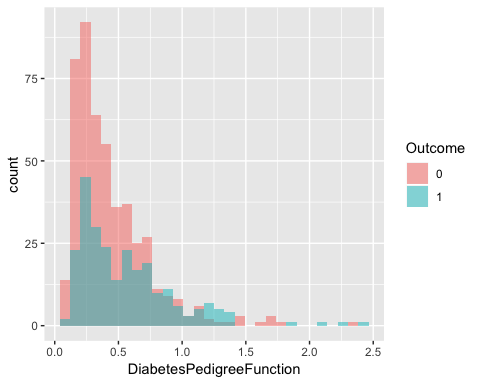
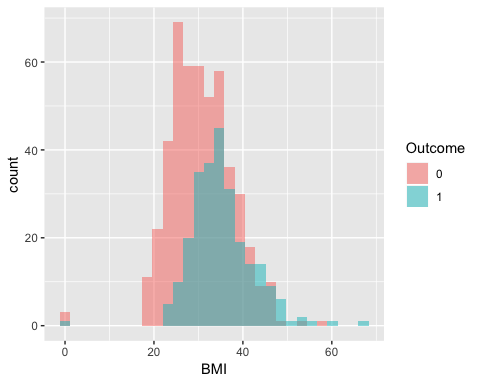
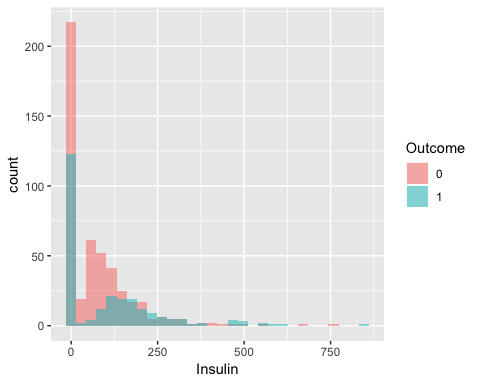
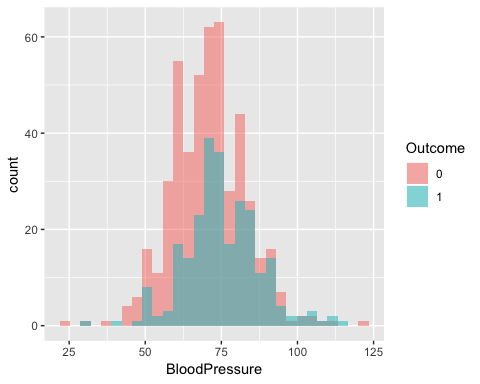
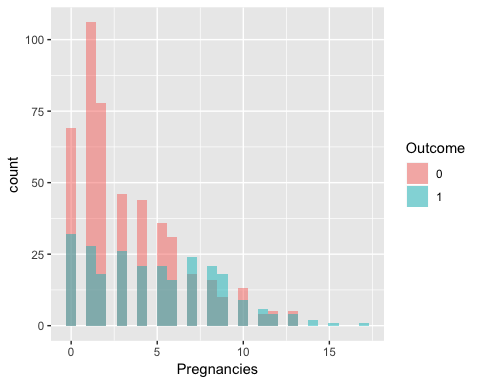


## Boxplots: each variable against Outcome



## 

## Histograms of each variable.

Histograms are color coded. Blue represents the data of the diabetic individuals and pink of the non-diabetic individuals.

## Logistic Regression Model

I fit a logistic regression model to determine which variables are most significant in predicting diabetes. I started by splitting the data into train and test sets with an 80-20 ratio. I then used the train set to create a logistic regression model.

##   
## Call:  
## glm(formula = Outcome ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4190 -0.7187 -0.4220 0.7526 2.9074   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.4287838 0.9280708 -9.082 < 2e-16 \*\*\*  
## Pregnancies 0.1113029 0.0375878 2.961 0.00307 \*\*   
## Glucose 0.0332979 0.0042821 7.776 7.48e-15 \*\*\*  
## BloodPressure -0.0140356 0.0099300 -1.413 0.15752   
## Insulin -0.0009186 0.0009872 -0.930 0.35213   
## BMI 0.0877235 0.0175333 5.003 5.64e-07 \*\*\*  
## DiabetesPedigreeFunction 0.8030325 0.3441471 2.333 0.01963 \*   
## Age 0.0245600 0.0109942 2.234 0.02549 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 694.00 on 549 degrees of freedom  
## Residual deviance: 515.04 on 542 degrees of freedom  
## AIC: 531.04  
##   
## Number of Fisher Scoring iterations: 5

Based on the regression summary output, we can see that pregnancies, glucose, BMI, and the diabetes pedigree function are all significant in predicting Outcome.

I used the test set to calculate predictions using the logistic model. I created a confusion matrix to compare the model predictions against the actual results and calculated the accuracy of the model

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## 0 333 38  
## 1 83 96

## [1] 0.78

We can see that the model is 79.09% accurate.

## Implications

Type 2 diabetes can primarily be managed by making lifestyle changes in diet and activity levels. BMI and glucose are two variables that are tied to a person’s lifestyle, and can be altered by diet and activity. Healthcare providers can improve the quality of life of an individual and reduce the cost of treatment if they are able to manage diabetes through diet and exercise alone. While a family history of diabetes is indicative of diabetes in an individual, we can see that both BMI and glucose were more significant than pedigree in predicting diabetes.

## Limitations

While type 2 diabetes can be managed by diet and exercise, type 1 can only be treated through insulin shots. Unfortunately, the outcome variable does not distinguish between type 1 and type 2 diabetes. This makes it difficult to know if variables like glucose and BMI are significant for only type 2 or for both types of diabetes. Another limitation of this analysis is that it implies correlation and not causation. While the study implies that glucose levels and BMI can be used to manage diabetes, it could be possible that diabetes leads to significant glucose levels and BMI. Instead of using diet and exercise to manage diabetes, diabetes could be affecting the glucose and BMI of the individuals.