Predicting Diabetes in PIMA Women

edX Capstone Project Submission

Kirtimay Pendse

6/23/2020

Introduction

Diabetes is a metabolic disorder defined as when one's blood glucose is too high (known as hyperglycemia) for a prolonged period of time. Glucose is an essential simple sugar widely consumed daily, and the hormone insulin helps absorbing glucose from food and transform it into energy; however, sometimes one's body doesn't make enough insulin or is unable to use it well, resulting in glucose staying in the blood stream unidgested and unable to reach the cells.[1]. This can cause health problems, especially diabetes.Around 9.5% -almost 30.5 million- of the United States population had diabetes in 2015 [2], and factors such as being overweight, being physically inactive, having a family history are linked with higher chances of developing diabetes. Due to several factors not discussed in this paper [3], diabetes is extremely prevalent in Native Americans, most notably within the Pima tribe- since the Pima tribe is a mostly homogenous group, Pima people have been the subject of several studies of diabetes.

This project is the final part of the HarvardX: PH125.9x Data Science: Capstone course[4], the last course for the Data Science Professional Certificate. This project is centered around predicting the presence of diabetes in Pima Indian women using data on factors such as age, body mass index, blood pressure etc. compiled together in the Pima Indians Diabetes dataset.

The dataset, loaded as 'pima_diabetes', is split into a training set containing 80% of the data and a test set containing 20% of the data for validation. This report is split into four sections: first, the objective and motivation behind the project is highlighted, then exploratory data analysis is conducted, following which the modeling approach to develop the diabetes prediction algorithm is presented. Finally, the modeling results are presented along with a discussion on the algorithm's performance and its limitations.

Objective

The dataset[5] is available on Kaggle and is originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, a part of the Department of Health and Human Services. The objective of this analysis is to diagnostically predict whether or not a patient is diabetic, based on select diagnostic measurements included in the dataset (such as BMI, Age, Blood Pressure). There are 786 individuals in the dataset, all of whom are females of at least 21 years of age, and of Pima Indian heritage.

[1] https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes#:~:text=Diabetes%20is%20a%20disease%20that,to%20be%

[2] Centers for Disease Control and Prevention. National diabetes statistics report, 2017. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[3] more can be found at https://care.diabetesjournals.org/content/29/8/1866

[4] https://courses.edx.org/courses/course-v1:HarvardX+PH125.9x+1T2020/course/

[5] https://www.kaggle.com/ksp585/pima-indian-diabetes-logistic-regression-with-r

Methods and Analysis

Preparing the data

First, the dataset is downloaded and split into a train set and a test set. The train set is used to create the prediction algorithm, and then the algorithm is tested on the test set for a final validation.

```
#Loading required packages
library(lubridate)
if(!require(ggthemes))
  install.packages("ggthemes", repos = "http://cran.us.r-project.org")
if(!require(scales))
  install.packages("scales", repos = "http://cran.us.r-project.org")
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
library(dplyr)
library(knitr)
library(ggplot2)
library(dslabs)
library(lubridate)
library(corrplot)
library(readr)

#Downloading the data
dl <- tempfile()
download.file("https://github.com/kirtimay/edX_Capstone/blob/master/cyo-diabetes/diabetes.csv", dl)
pima_diabetes <- read.csv("diabetes.csv", col.names=c("pregnancies","glucose","bp","skin_thickness","ins

#convert outcome to factor
pima_diabetes$outcome <- factor(pima_diabetes$outcome)
```

Description of Variables

As seen in the table, there are 9 variables in total. The response variable is 'outcome', which is a binary variable- 1 indicates that the patient is diabetic, and 0 indicates that they are not. The other 8 variables are predictors, and their descriptions are provided below.

It should be noted that the plasma glucose concentration was measured after a 2-hour glucose tolerance oral test, BMI is calculated as the patient's weight in kgs divided by their height in meters squared, and the DPF is a variable synthesizing family history of diabetes [6].

Variable	Class	Description
pregnancies	integer	No. of Pregnancies
glucose	integer	Plasma Glucose Concentration (mg/dL)
bp	integer	Diastolic BP (mm Hg)
skin_thickness	integer	Triceps Skin Thickness (mm)
insulin	integer	2 Hour Serum Insulin (uU/mL)
bmi	numeric	Body Mass Index
dpf	numeric	Diabetes Pedigree Function
age	integer	Age in Years
outcome	factor	Presence of Diabetes

[6] http://www.personal.kent.edu/~mshanker/personal/Zip_files/sar_2000.pdf

Data is split

```
set.seed(1, sample.kind="Rounding")
test_index <- createDataPartition(y = pima_diabetes$outcome, times = 1, p = 0.2, list = FALSE)
train_set <- pima_diabetes[-test_index,]
test_set <- pima_diabetes[test_index,]
```

Exploratory Analysis

For the initial data exploration, the head() function was used to get a broad understanding of the data.

pregnancies	glucose	bp	skin_thickness	insulin	bmi	dpf	age	outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0

The table above shows that there seems to be a lot of variation within all the variables, and that a value of 0 for skin_thickness and insulin seems to indicate some missing data. Summary statistics were then calculated to get a better understanding of the variables.

pregnancies	glucose	bp	skin_thickness	insulin	bmi	dpf	ag
Min. : 0.00	Min. : 0	Min. : 0.0	Min. : 0.0	Min. : 0.0	Min. : 0.0	Min. :0.078	Min.
1st Qu.: 1.00	1st Qu.: 99	1st Qu.: 62.0	1st Qu.: 0.0	1st Qu.: 0.0	1st Qu.:27.3	1st Qu.:0.244	1st Qu
Median : 3.00	Median :117	Median : 72.0	Median :23.0	Median : 30.5	Median :32.0	Median :0.372	Media
Mean : 3.85	Mean :121	Mean : 69.1	Mean :20.5	Mean : 79.8	Mean :32.0	Mean :0.472	Mean
3rd Qu.: 6.00	3rd Qu.:140	3rd Qu.: 80.0	3rd Qu.:32.0	3rd Qu.:127.2	3rd Qu.:36.6	3rd Qu.:0.626	3rd Qu
Max. :17.00	Max. :199	Max. :122.0	Max. :99.0	Max. :846.0	Max. :67.1	Max. :2.420	Max.

In the summary statistics presented above, it's observed that the mean number of pregnancies is 3.85, which seems pretty high at first glance but is consistent with previous findings on Native American pregnancy rates and statistics [7]. The maximum value is 17, which is significantly higher than the 75th percentile value of 6. The mean glucose level is 121 mg/dL, which is towards the high end of the normal 70 to 130 mg/dL range [8] and the mean diastolic blood pressure is 69.1, which is well within a normal range. An average skin thickness of 20.5mm is within a normal range [9], and an average insulin of 127.2 μU/mL is within the normal range for an oral test conducted 2 hours after administration of glucose [10]. Interestingly, the mean BMI value of 32 seems to be very high, as the normal range of BMI is 18 to 24, and while a value of 67.1 is extremely high (the max value), it doesn't seem to be an outlier as BMIs have been measured in three figures before. Some concern arose here as the minimum value for glucose, bp, skin_thickness, and bmi are 0, which are not possible and there maybe some missing data to address before any modeling is done. [11]

```
##    pregnancies        glucose             bp skin_thickness        insulin
##              0              0              0              0              0
##            bmi            dpf            age        outcome
##              0              0              0              0
```

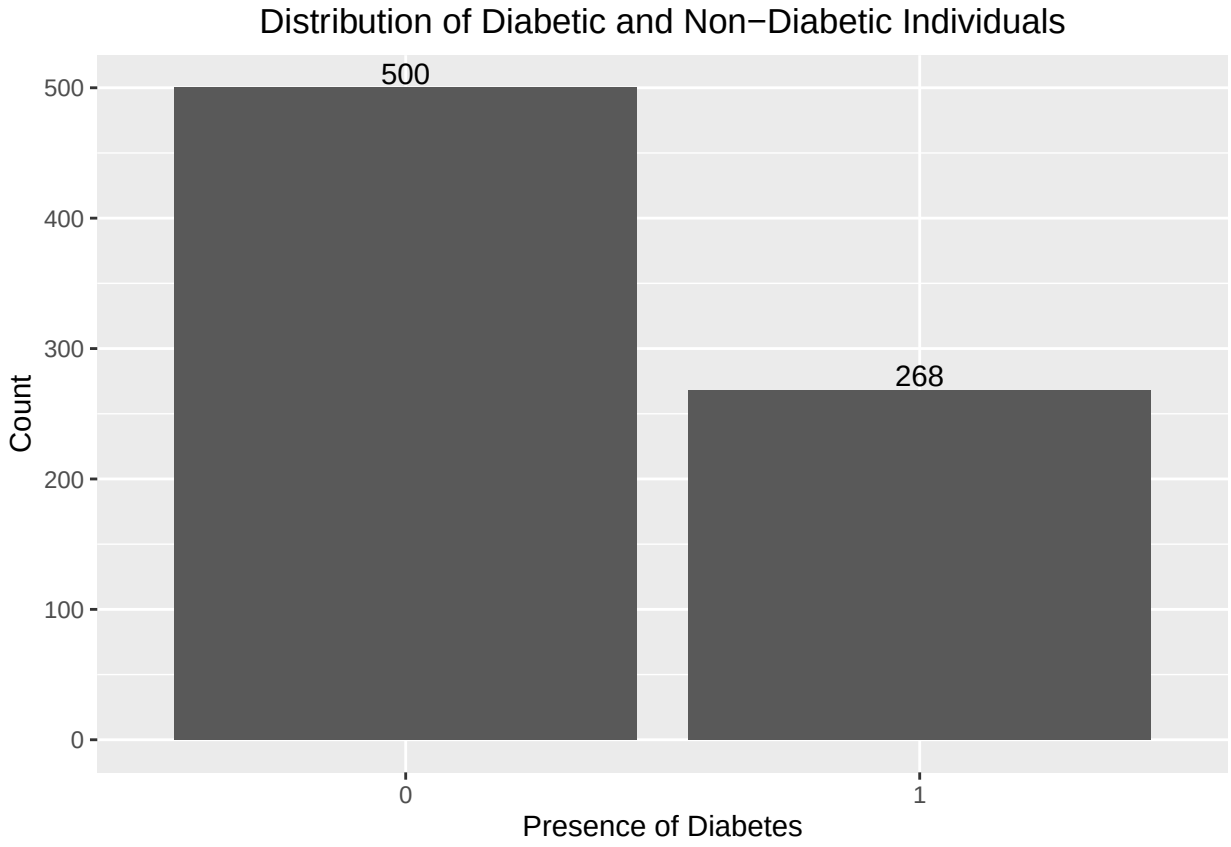
[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909384/
[8] https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5083983/
[10] https://emedicine.medscape.com/article/2089224-overview
[11] The minimum value for insulin is 0 as well, but in cases of Type 1 diabetes, it is possible that the human body doesn't produce any insulin at all.

Plots

Outcome Variable

Diabetes

```
ggplot(pima_diabetes,aes(outcome)) +
  geom_bar() +
  ggtitle("Distribution of Diabetic and Non-Diabetic Individuals") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("Count") +
  geom_text(stat='count', aes(label=..count..), vjust=-0.2)
```

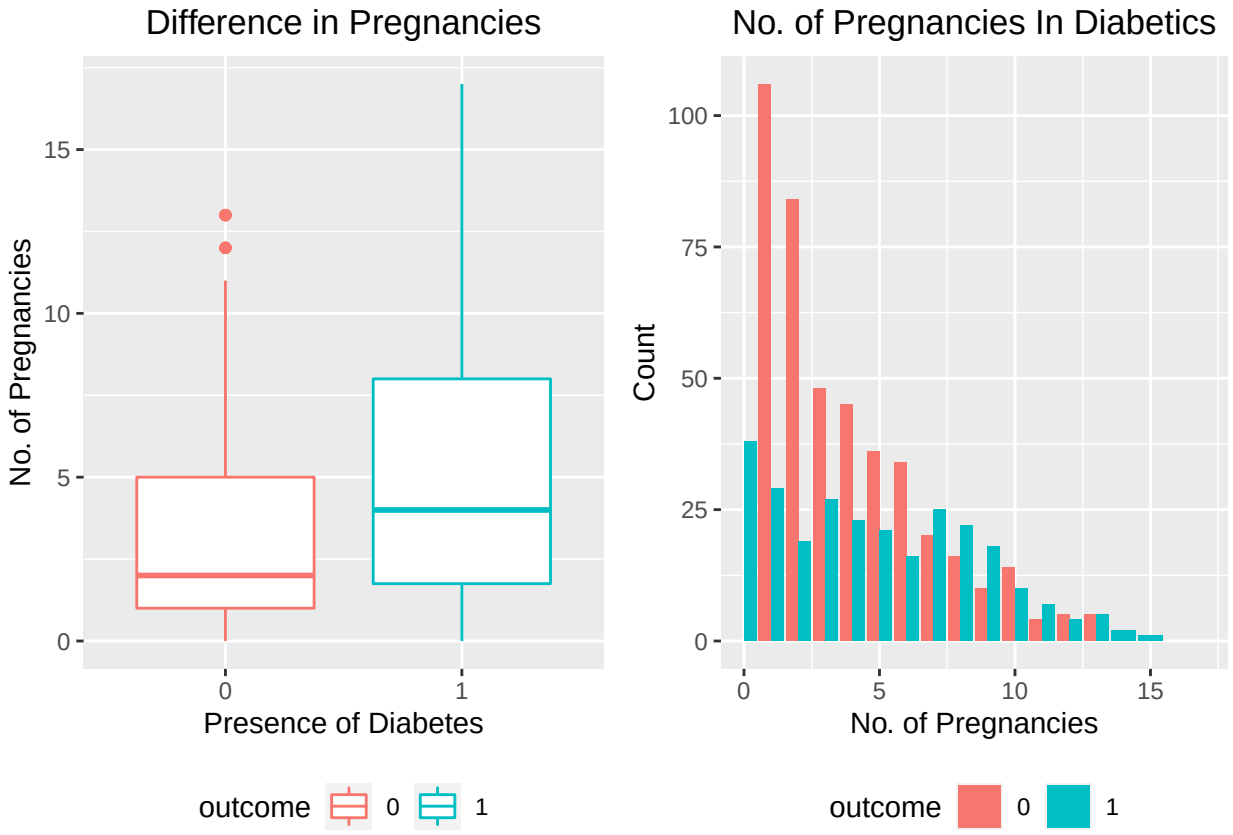


Predictor Variables

Pregnancies

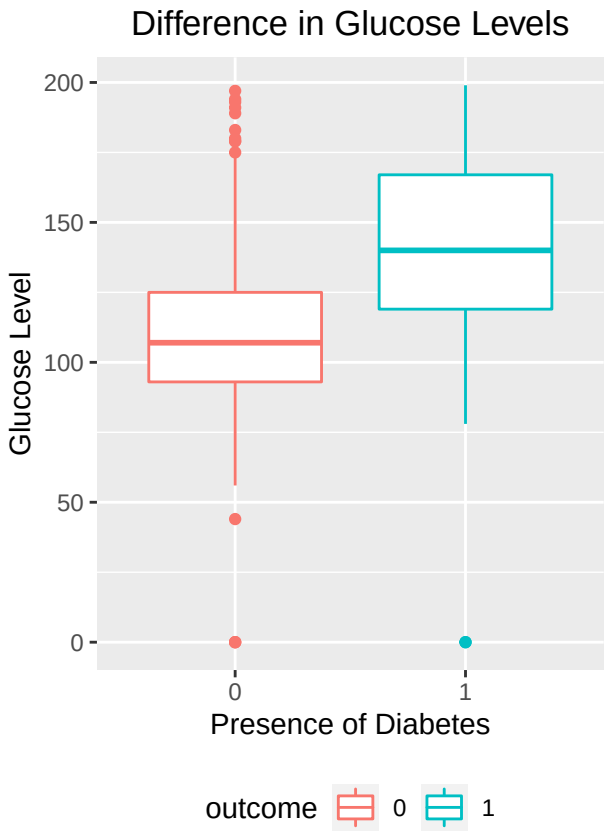
```
p1 <- ggplot(pima_diabetes, aes(x = outcome, y = pregnancies, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in Pregnancies") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("No. of Pregnancies") +
  theme(legend.position = "bottom")

p2 <- ggplot(pima_diabetes,aes(x = pregnancies, fill=outcome)) +
  geom_bar(position = "Dodge") +
  scale_x_continuous(limits = c(0,17)) +
```

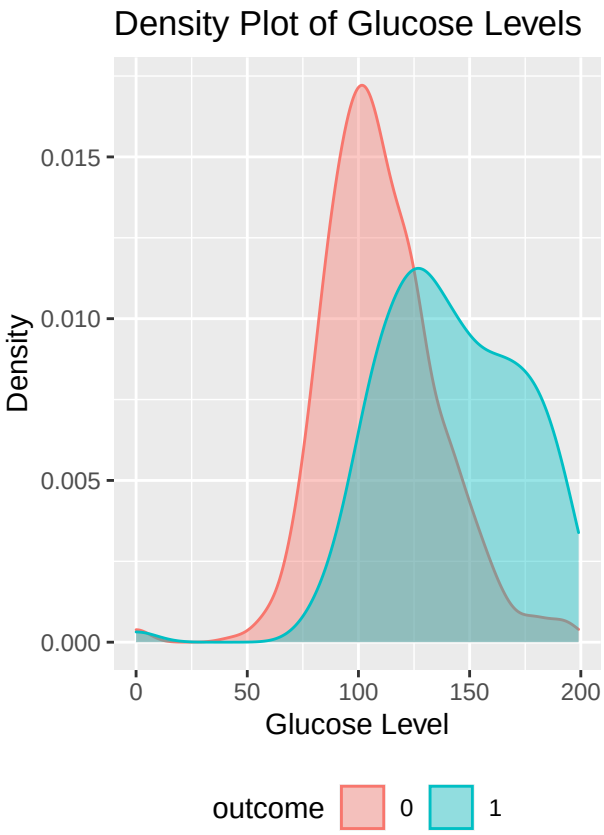
```
  labs(title = "No. of Pregnancies In Diabetics") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("No. of Pregnancies") +
  ylab("Count") +
  theme(legend.position = "bottom")

gridExtra::grid.arrange(p1, p2, ncol = 2)
```



Glucose

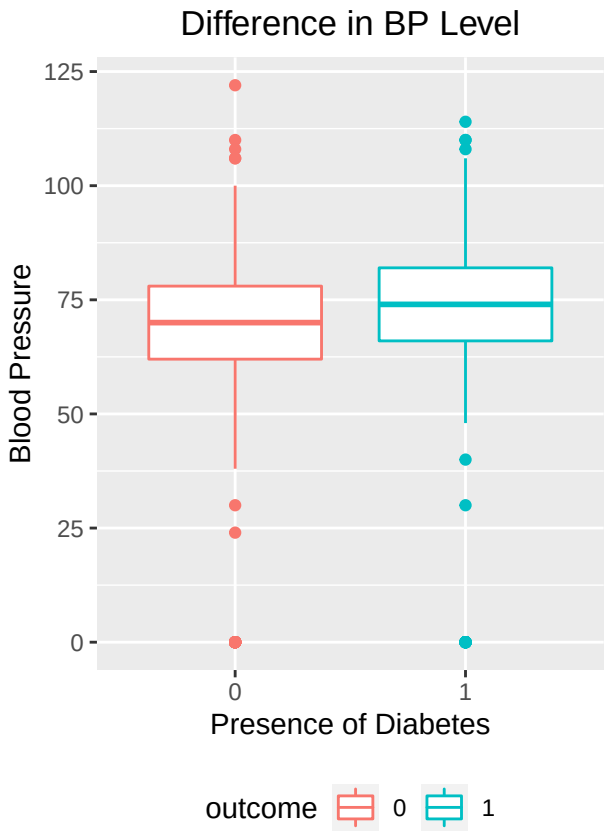
```
p3 <- ggplot(pima_diabetes, aes(x = outcome, y=glucose, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in Glucose Levels") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("Glucose Level") +
  theme(legend.position = "bottom")

p4 <- ggplot(pima_diabetes, aes(x = glucose, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "Glucose Level", y = "Density", title = "Density Plot of Glucose Levels")

gridExtra::grid.arrange(p3, p4, ncol = 2)
```

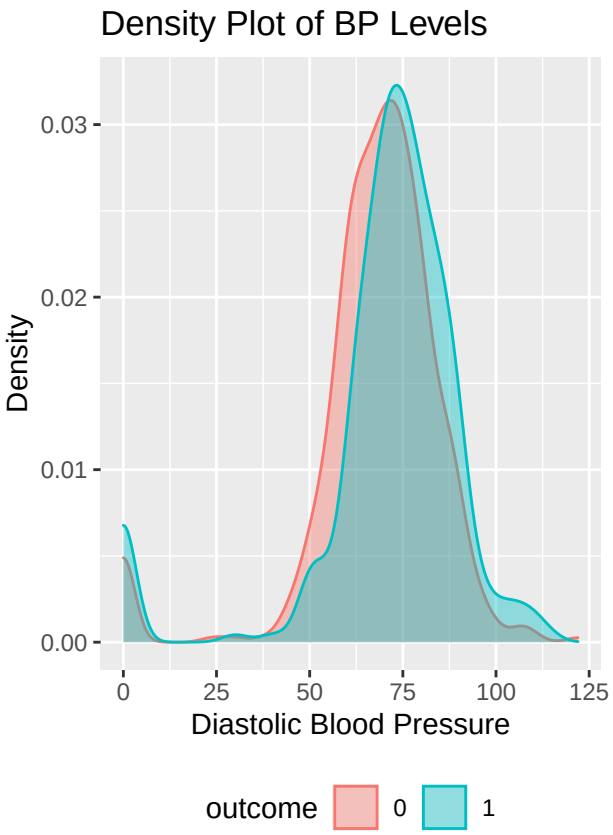




Blood Pressure

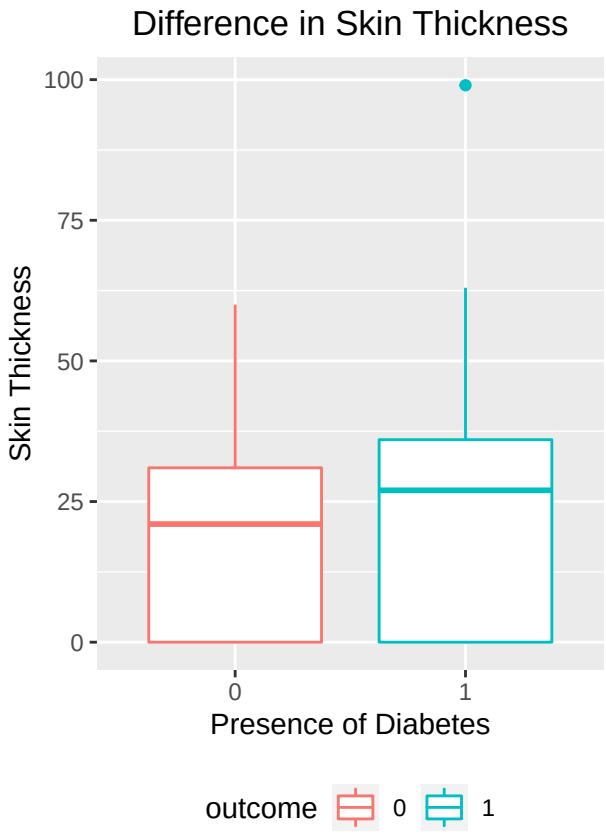
```
p5 <- ggplot(pima_diabetes, aes(x = outcome, y=bp, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in BP Level") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("Blood Pressure") +
  theme(legend.position = "bottom")

p6 <- ggplot(pima_diabetes, aes(x = bp, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "Diastolic Blood Pressure", y = "Density", title = "Density Plot of BP Levels")

gridExtra::grid.arrange(p5, p6, ncol = 2)
```

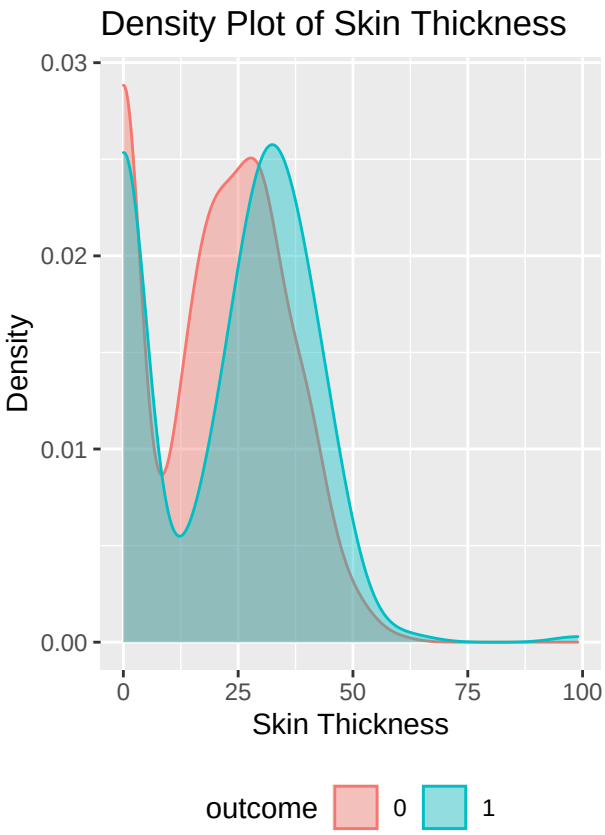




Skin Thickness

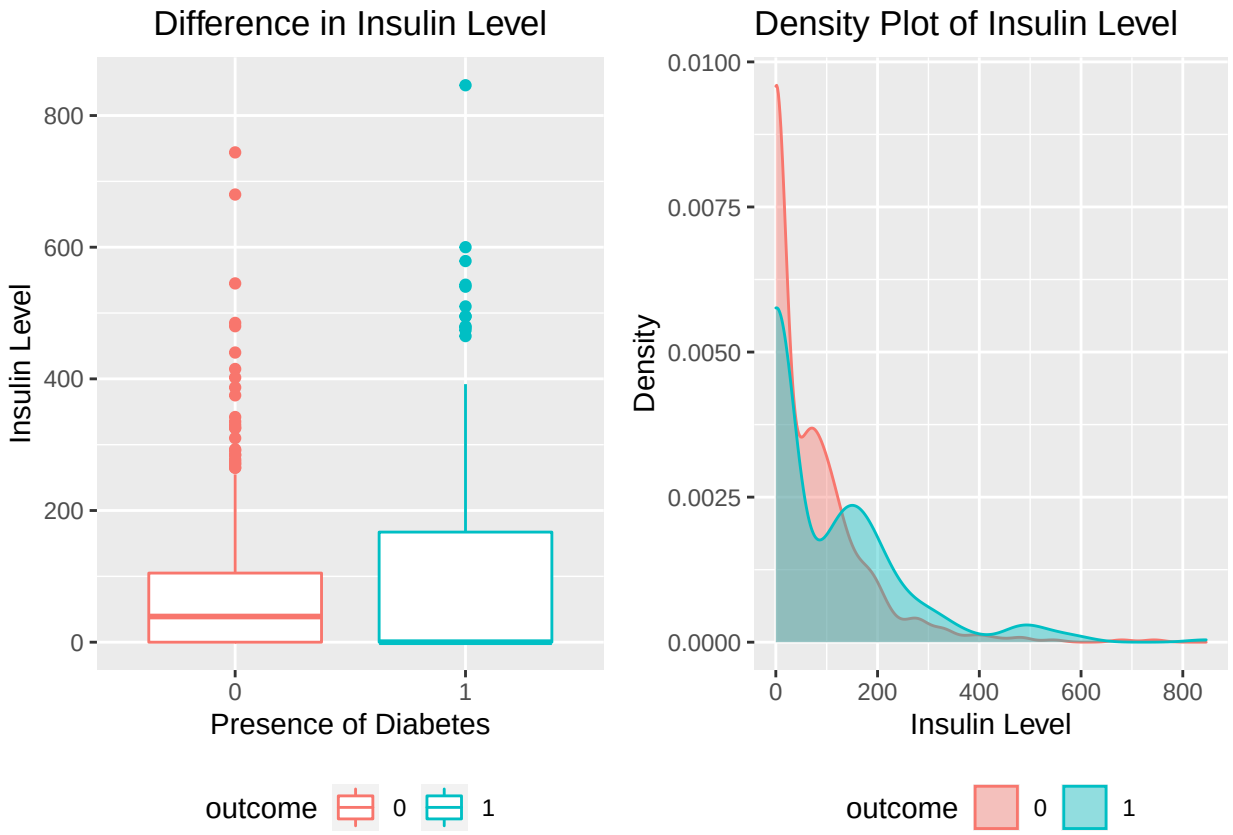
```
p7 <- ggplot(pima_diabetes, aes(x = outcome, y=skin_thickness, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in Skin Thickness") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("Skin Thickness") +
  theme(legend.position = "bottom")

p8 <- ggplot(pima_diabetes, aes(x = skin_thickness, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "Skin Thickness", y = "Density", title = "Density Plot of Skin Thickness")

gridExtra::grid.arrange(p7, p8, ncol = 2)
```





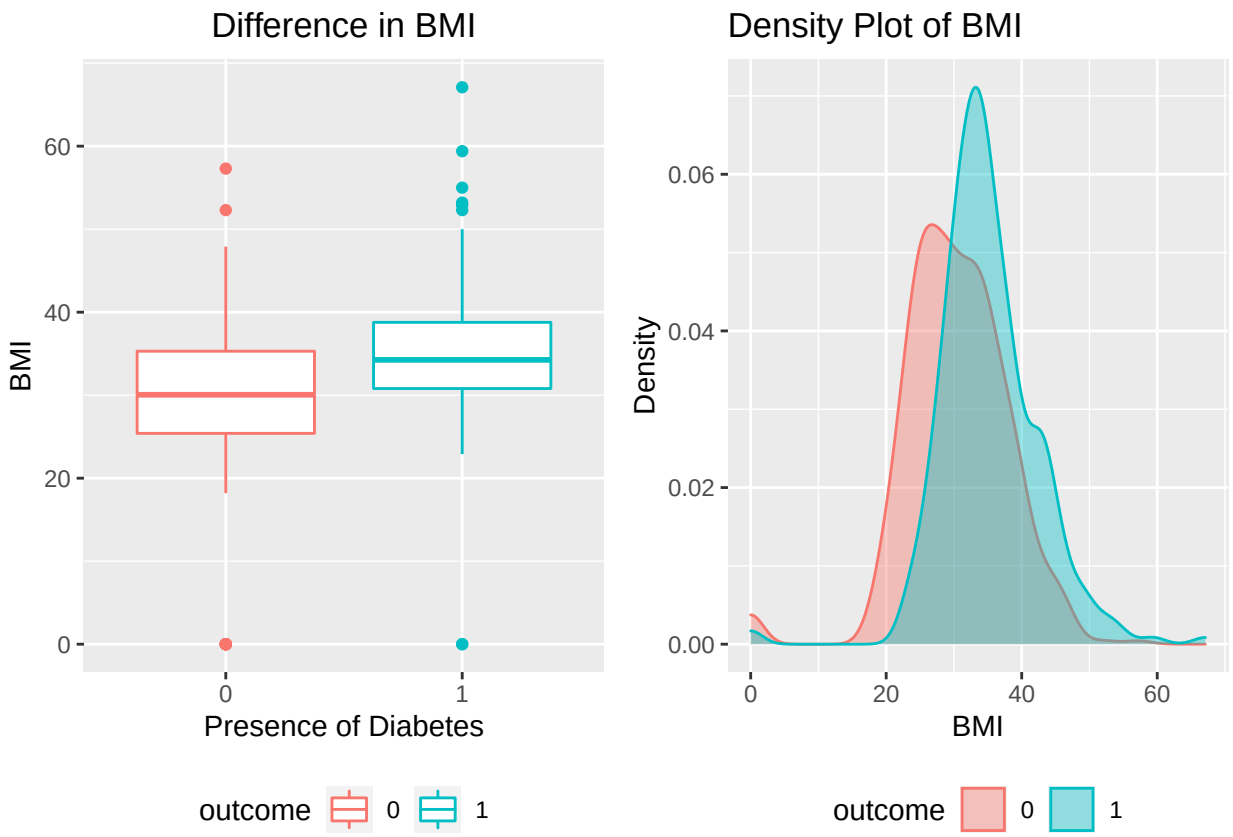
Insulin

```
p9 <- ggplot(pima_diabetes, aes(x = outcome, y=insulin, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in Insulin Level") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("Insulin Level") +
  theme(legend.position = "bottom")

p10 <- ggplot(pima_diabetes, aes(x = insulin, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "Insulin Level", y = "Density", title = "Density Plot of Insulin Level")

gridExtra::grid.arrange(p9, p10, ncol = 2)
```



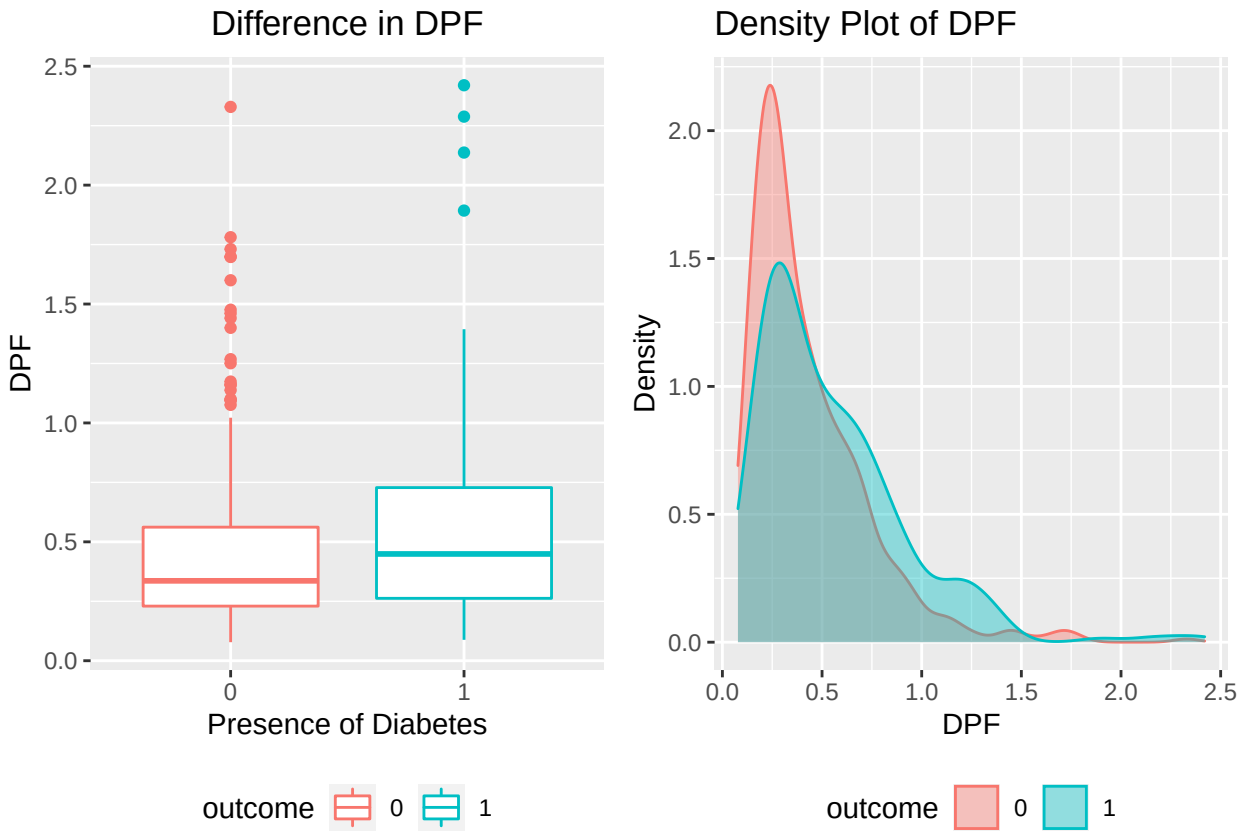
BMI

```
p11 <- ggplot(pima_diabetes, aes(x = outcome, y=bmi, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in BMI") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("BMI") +
  theme(legend.position = "bottom")

p12 <- ggplot(pima_diabetes, aes(x = bmi, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "BMI", y = "Density", title = "Density Plot of BMI")

gridExtra::grid.arrange(p11, p12, ncol = 2)
```



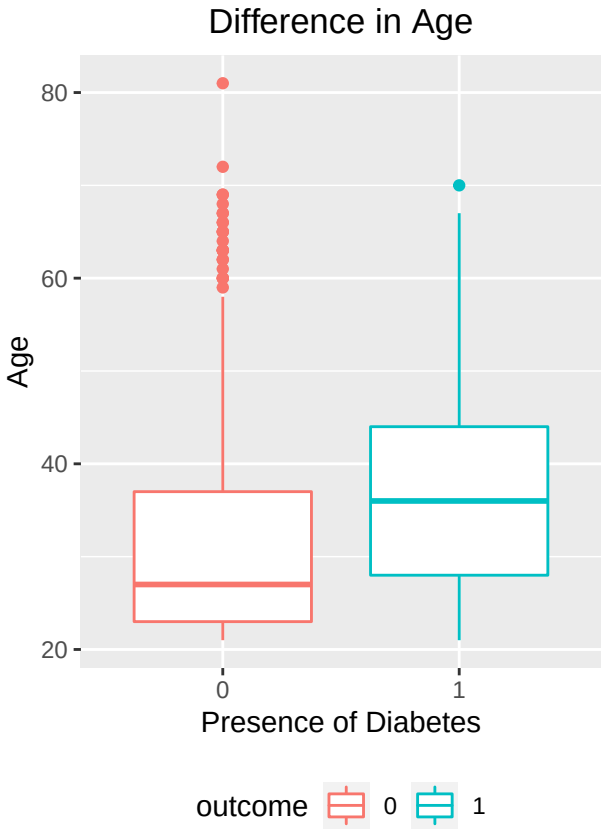
DPF

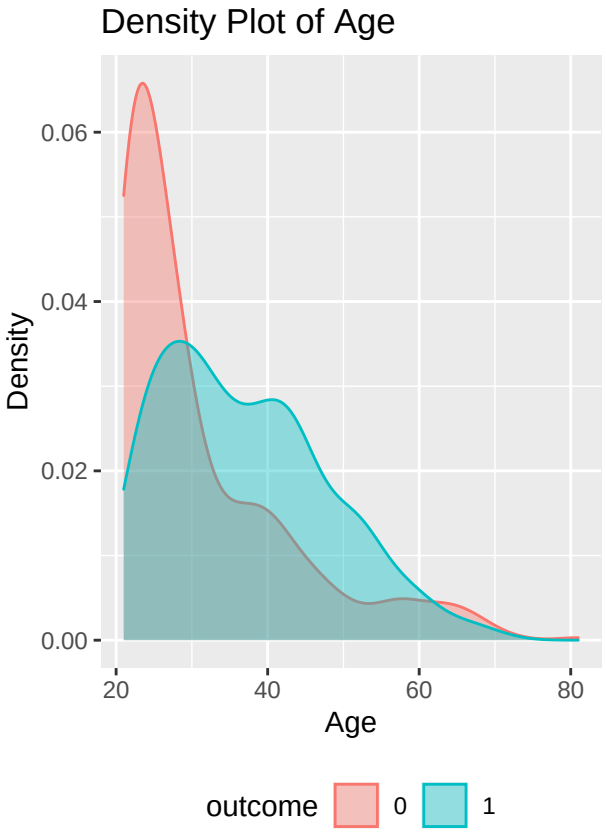
```
p13 <- ggplot(pima_diabetes, aes(x = outcome, y=dpf, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in DPF") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("DPF") +
  theme(legend.position = "bottom")

p14 <- ggplot(pima_diabetes, aes(x = dpf, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "DPF", y = "Density", title = "Density Plot of DPF")

gridExtra::grid.arrange(p13, p14, ncol = 2)
```



Age

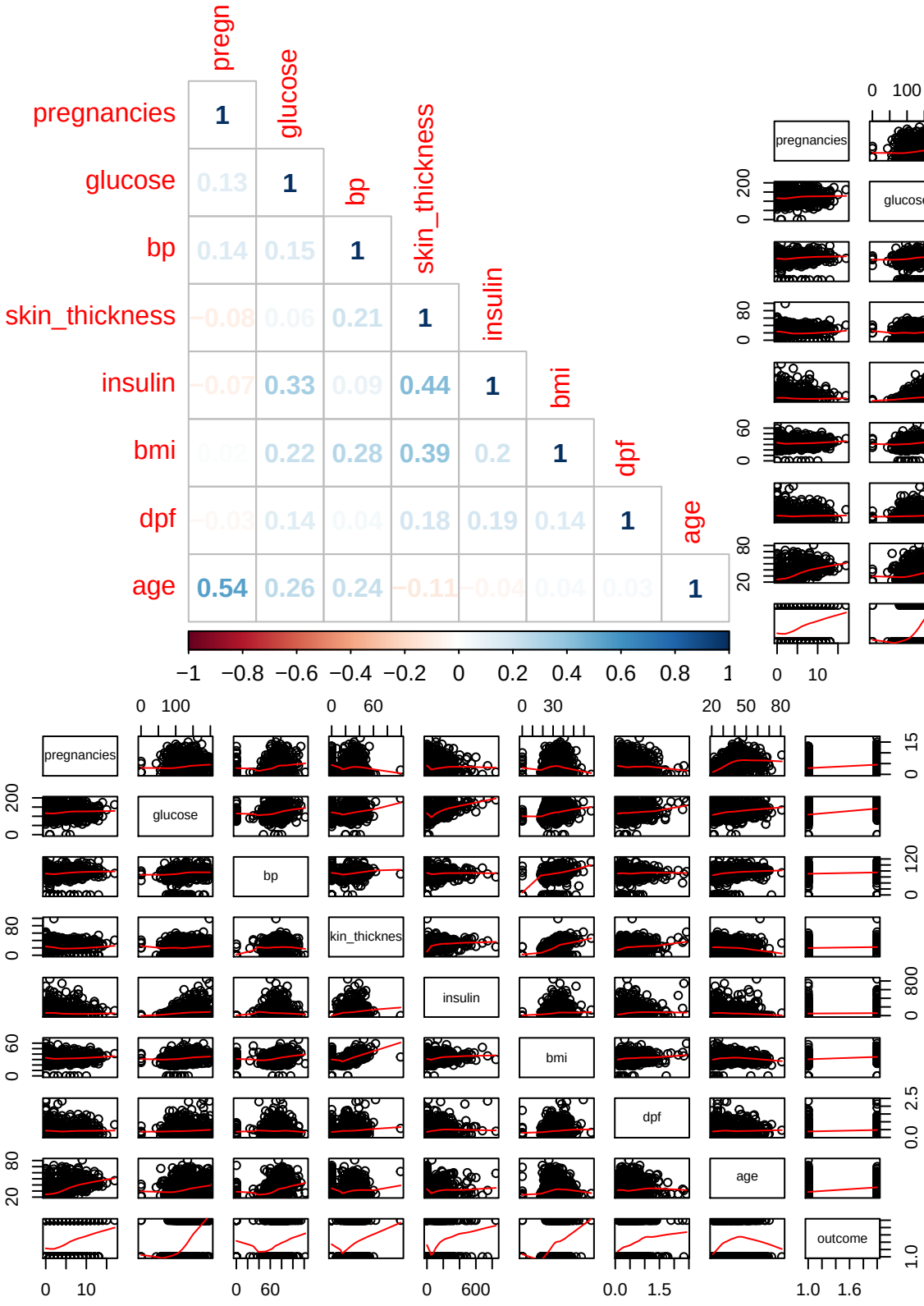
```
p15 <- ggplot(pima_diabetes, aes(x = outcome, y=age, color=outcome)) +
  geom_boxplot() +
  ggtitle("Difference in Age") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Presence of Diabetes") +
  ylab("Age") +
  theme(legend.position = "bottom")

p16 <- ggplot(pima_diabetes, aes(x = age, color = outcome, fill = outcome)) +
  geom_density(alpha = 0.4) +
  theme(legend.position = "bottom") +
  labs(x = "Age", y = "Density", title = "Density Plot of Age")

gridExtra::grid.arrange(p15, p16, ncol = 2)
```





Correlation Matrix

	pregnancies	glucose	bp	skin_thickness	insulin	bmi	dpf	age
pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54
glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26
bp	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24
skin_thickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11
insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04
bmi	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04
dpf	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03
age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00


pregn
glucose
bp
skin_thickness
insulin
bmi
dpf
age
pregnancies
1
glucose
0.13
1
bp
0.14
0.15
1
skin_thickness
−0.08
0.06
0.21
1
insulin
−0.07
0.33
0.09
0.44
1
bmi
0.02
0.22
0.28
0.39
0.2
1
dpf
−0.03
0.14
0.04
0.18
0.19
0.14
1
age
0.54
0.26
0.24
−0.11
−0.04
0.04
0.03
1
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1
pregnancies
glucose
bp
kin_thicknes
insulin
bmi
dpf
age
outcome