Econometric Project

Apurva Shekhar | Ritu Ranjani Ravi Shankar | Suchita Negi | Vidhi Gandhi

May 28, 2020

Does Pregnancy cause Diabetes?

Agenda:

- Introduction
- Data reading and Description of some of the variables of the dataset
- Data Pre-Processing and Feature Engineering
- Exploratory Data Analysis
- Regression - Effect of pregnancies on diabetes
    - Descriptive statistics using Stargazer
- Conclusion

Introduction

- This study was carried out to investigate the significance of health-related predictors of diabetes in Pima Indian Women.
- Many types of research by the U.S health department show that Native Americans are at more risk of getting diabetes than any other racial group.
- This dataset contains patient details who are of Pima Indians/Native American origin and are females of at least 21 years old.
- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and the variables of this dataset are selected based on certain criteria from a larger dataset.
- Data Source: https://www.kaggle.com/uciml/pima-indians-diabetes-database (https://www.kaggle.com/uciml/pima-indians-diabetes-database)

Business Objective

- Identify if the number of times a female is pregnant has any impact on diabetes

Why is this question important?

- This exploration is worthy as researches have shown that women of Native American(Pima Indian Origin), White or Asian origin are at more risk of getting Diabetes called Gestational Diabetes when they are pregnant. For most of them this Diabetes converts to type 2 diabetes after the pregnancy goes away.

Method Used

Logit - Binary Logistic regression

We are not using LPM/Probit and have decided to proceed with Logit model because,

- LPM models the probabilty as linear function of X and also LPM allows the probability value to be greater than 1.
- Logit over Probit due to the convinence and ease of use.

Dataset

- Number of Rows: 768
- Number of columns: 9 (1 - dependent, 8 - independent variables)
- Variable description:
  - Pregnancies - Number of times pregnant
  - Glucose - Plasma glucose concentration - a 2 hours in an oral glucose tolerance test
  - BloodPressure - Diastolic blood pressure (mm Hg)
  - SkinThickness - Triceps skinfold thickness (mm)
  - Insulin - 2-Hour serum insulin (mu U/ml)
  - BMI - Body mass index (weight in kg/(height in m)^2)
  - Diabetes pedigree function - A function that represents how likely they are to get the disease by extrapolating from their ancestor's history.

Does the dataset contain all variables that account for diabetes ?

- Diabetes is believed to have a strong genetic link, meaning that it tends to run in families.

The major factors/risks that causes diabetes are the following :

1. Sedentary Lifestyle(Obesity or being overweight) : This factor is captured in our dataset through the variable - **BMI**
2. Gestational diabetes or giving birth to a baby : This factor is attributed to **the number of pregnancy a patient has had (Variable of Interest)**
3. strong genetic link : Captured by the variable - **Diabetespedigreefunction**
4. Aging: Increasing age is a significant risk factor for type 2 diabetes. Captured by the variable : **Age**
5. High Blood Pressure : Captured by the variable - **bloodpressure**
6. Glucose/Insulin : Insulin and other hormones control the amount of glucose in your bloodstream.People with diabetes either don't make insulin or their body's cells can no longer use their insulin. This leads to high blood sugars. Captured by variable - **Glucose and Insulin**

From the above medical domain research, we could see that, the dataset contains all major factors that could cause diabetes.

Recent studies have shown that skinfolds thickness were associated with 2·8-fold and 6·4-fold risk of developing T2DM(Type 2 Diabetes). This factor is also attributed in our dataset through **SkinThickness**

Dependent Variable and Variable of Interest:

- Null Hypothesis: Pregnancy doesn't cause Diabetes.
- Alternative Hypothesis: Pregnancy causes Diabetes.
- Dependent variable: does the person have Diabetes(=1) or not(=0)?
- Independent Variables:
  - Variable of interest: Pregnancies
  - Control Variables: BMI, Age, Glucose,DiabetesPedigreeFunction, BloodPressure, SkinThickness.

	Pregnancies <int>	Gluco... <int>	BloodPressure <int>	SkinThickness <int>	Insulin <int>	B... <dbl>	DiabetesPedig
1	6	148	72	35	0	33.6	
2	1	85	66	29	0	26.6	
3	8	183	64	0	0	23.3	
4	1	89	66	23	94	28.1	
5	0	137	40	35	168	43.1	
6	5	116	74	0	0	25.6	

6 rows | 1-8 of 10 columns

Descriptive statistics of the data

```
stargazer(pima, type="text", median=TRUE, iqr=TRUE,digits=1, title="Descriptive Statistic
s")
```

```
Descriptive Statistics
============================================================================
Statistic                 N  Mean  St. Dev. Min Pctl(25) Median Pctl(75) Max
----------------------------------------------------------------------------
Pregnancies              768  3.8    3.4     0     1       3       6      17
Glucose                  768 120.9   32.0    0     99     117    140.2   199
BloodPressure            768 69.1    19.4    0     62      72      80     122
SkinThickness            768 20.5    16.0    0     0       23      32      99
Insulin                  768 79.8   115.2    0     0      30.5   127.2   846
BMI                      768 32.0    7.9    0.0   27.3    32.0    36.6   67.1
DiabetesPedigreeFunction 768  0.5    0.3    0.1   0.2     0.4     0.6    2.4
Age                      768 33.2    11.8    21    24      29      41      81
Outcome                  768  0.3    0.5     0     0       0       1       1
----------------------------------------------------------------------------
```

Data cleaning

```
# check for NA
sapply(pima, function(x) sum(is.na(x)))
```

```
             Pregnancies                    Glucose             BloodPressure             Ski
nThickness
                       0                          0                         0
0
                 Insulin                        BMI DiabetesPedigreeFunction
Age
                       0                          0                         0
0
                 Outcome
                       0
```

- The dataset revealed many abnormal values for biological measures. Variables such as Skin Thickness and Glucose had 227 and 374 zero-values respectively.
- The fact that both measures cannot hold zero values indicated that the missing values in the dataset were represented as zero values in the dataset.
- The missing values in the dataset constituted to about 30% of the observations in the dataset.
- As removing these values would result in significant information loss, kNN imputation was performed to impute the missing values in the data set.

- Only obvious wrong values in the dataset (zero values) were imputed. Large outliers in the variables were not handled.

```
# Dealing with zeros
missing_data <- pima[,setdiff(names(pima), c('Outcome', 'Pregnancies'))]
features_miss_num <- apply(missing_data, 2, function(x) sum(x <= 0))
features_miss <- names(missing_data)[ features_miss_num > 0]

rows_miss <- apply(missing_data, 1, function(x) sum(x <= 0) >= 1)
sum(rows_miss)
```

```
[1] 376
```

```
missing_data[missing_data <= 0] <- NA
pima[, names(missing_data)] <- missing_data

# KNN imputation
orig_data <- pima
colSums(is.na(pima))
```

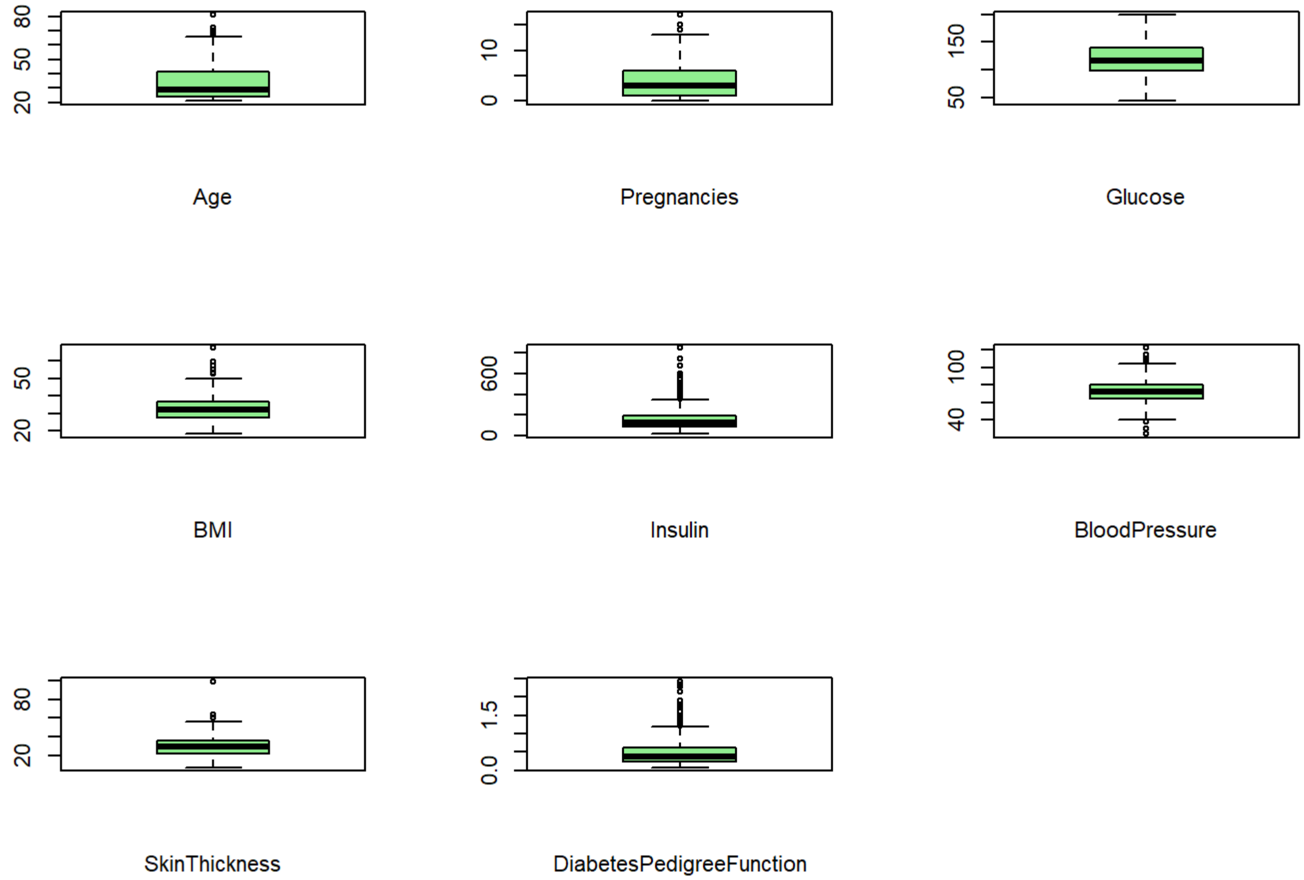
```
             Pregnancies                  Glucose            BloodPressure            Ski
nThickness
                       0                        5                       35
227
                 Insulin                      BMI DiabetesPedigreeFunction
Age
                     374                       11                        0
0
                 Outcome
                       0
```

```
pima[,c(-8,-9)] <- knnImputation(pima[,c(-8,-9)], k = 3)
sapply(pima, function(x) sum(is.na(x)))
```

```
             Pregnancies                  Glucose            BloodPressure            Ski
nThickness
                       0                        0                        0
0
                 Insulin                      BMI DiabetesPedigreeFunction
Age
                       0                        0                        0
0
                 Outcome
                       0
```

Outlier Detection

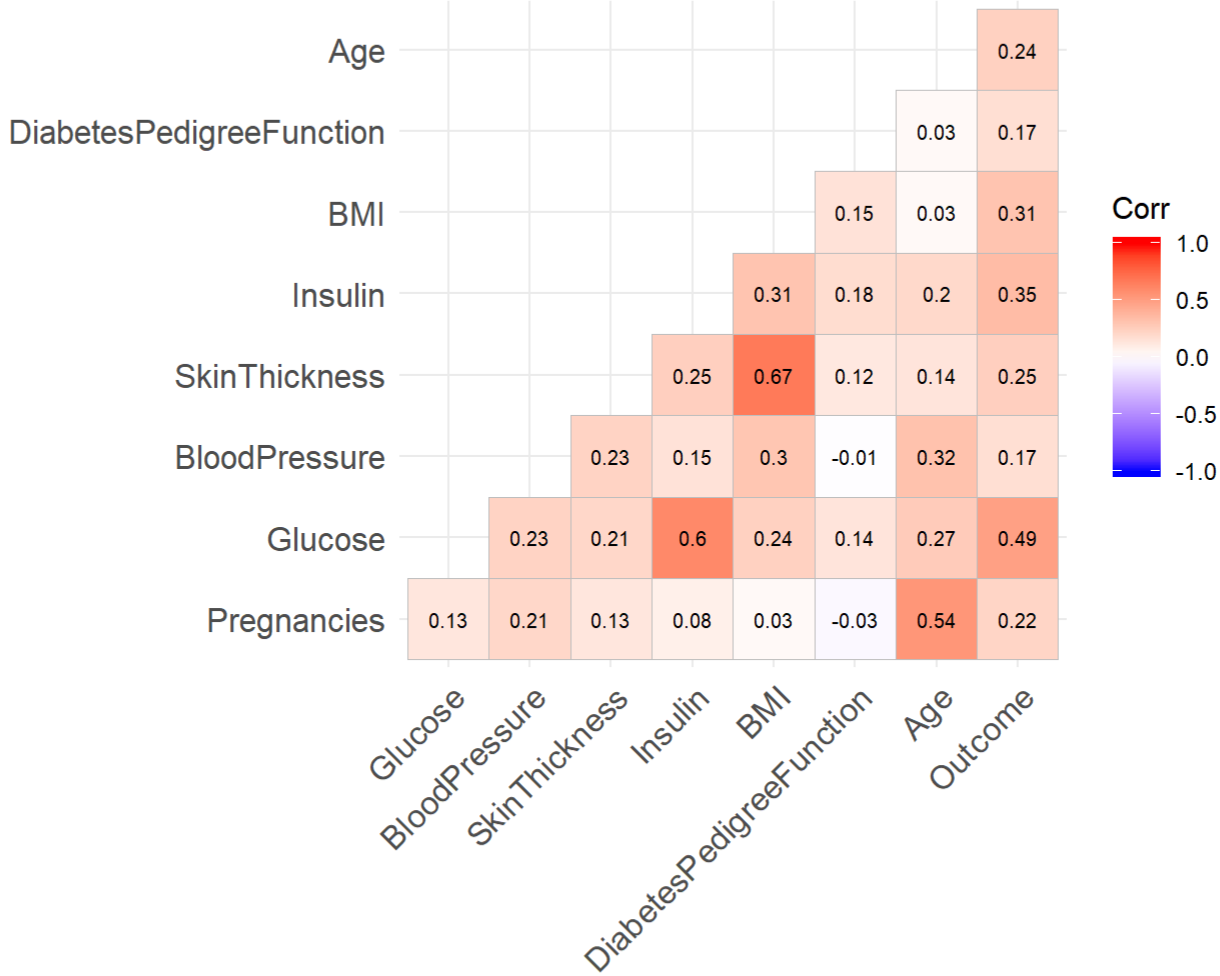
```
par(mfrow=c(3,3))
boxplot(x=pima$Age, xlab="Age",col=c('lightgreen'))
boxplot(x=pima$Pregnancies, xlab="Pregnancies",col=c('lightgreen'))
boxplot(x=pima$Glucose, xlab='Glucose',col=c('lightgreen'))
boxplot(x=pima$BMI, xlab='BMI',col=c('lightgreen'))
boxplot(x=pima$Insulin, xlab='Insulin',col=c('lightgreen'))
boxplot(x=pima$BloodPressure, xlab='BloodPressure',col=c('lightgreen'))
boxplot(x=pima$SkinThickness, xlab='SkinThickness',col=c('lightgreen'))
boxplot(x=pima$DiabetesPedigreeFunction, xlab='DiabetesPedigreeFunction',col=c('lightgree
n'))
#boxplot(x=pima$Outcome, xlab='Outcome',col=c('lightgreen'))
```



- From the above plot it is clear we have outlier values in our data. But since this is medical data, we cannot remove or normalise those outlier values.

Correlation Plot

```
ggcorrplot(cor(pima, use="pairwise.complete.obs"),  hc.order=FALSE, type='lower',lab=TRUE
, lab_size=2.5)
```



From the correlation plot, we could see that -

- Pregnancy and Age are highly correlated
- Glucose and Insulin are highly correlated
- SkinThickness and BMI are highly correlated
- Glucose and Diabetes(outcome) are moderately correlated

This gives us the idea of mulitcollinearity issue, that could arise when including two highly correlated variables in our model.

Correlation is not causation. Though, the correlation plot gives us an idea of the variables that impacts the outcome variable and also are correlated with our variable of interest, to identify the control variables for logit model, we perform Exploratory Data Analysis.

T-test

Conducting t-tests enables us to identify variables that are statistically significant.

We perform two t-test :

Test 1:

T-test of all X variables against Variable of interest-Pregnancies

```
lapply(pima[,c("Glucose", "DiabetesPedigreeFunction", "BMI","Insulin",'Age','BloodPressur
e','SkinThickness')], function(x) anova(lm(x ~ pima$Pregnancies)))
```

```
$Glucose
Analysis of Variance Table

Response: x
                  Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Pregnancies   1  12017 12017.4  13.143 0.0003076 ***
Residuals        766 700368   914.3                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$DiabetesPedigreeFunction
Analysis of Variance Table

Response: x
                  Df Sum Sq  Mean Sq F value Pr(>F)
pima$Pregnancies   1  0.095 0.094622  0.8618 0.3535
Residuals        766 84.106 0.109798               

$BMI
Analysis of Variance Table

Response: x
                  Df Sum Sq Mean Sq F value Pr(>F)
pima$Pregnancies   1     23  22.836  0.4813  0.488
Residuals        766  36341  47.443               

$Insulin
Analysis of Variance Table

Response: x
                  Df  Sum Sq Mean Sq F value  Pr(>F)  
pima$Pregnancies   1   46689   46689  4.6831 0.03077 *
Residuals        766 7636888    9970                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Age
Analysis of Variance Table

Response: x
                  Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Pregnancies   1  31432 31431.8  322.54 < 2.2e-16 ***
Residuals        766  74647    97.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$BloodPressure
Analysis of Variance Table
```

```
Response: x
                  Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Pregnancies   1   5160  5160.1   36.19 2.771e-09 ***
Residuals        766 109218   142.6                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$SkinThickness
Analysis of Variance Table

Response: x
                  Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Pregnancies   1   1273 1273.31  13.519 0.0002527 ***
Residuals        766  72148   94.19                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test 2:

T-test of all X variables against Dependent Variables - Outcome

```
lapply(pima[,c("Glucose", "DiabetesPedigreeFunction", "BMI","Insulin",'Age','BloodPressure','SkinThickness')], function(x) anova(lm(x ~ pima$Outcome)))
```

```
$Glucose
Analysis of Variance Table

Response: x
              Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Outcome   1 173729  173729  247.05 < 2.2e-16 ***
Residuals    766 538657     703                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$DiabetesPedigreeFunction
Analysis of Variance Table

Response: x
              Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Outcome   1  2.545  2.5447  23.871 1.255e-06 ***
Residuals    766 81.656  0.1066                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$BMI
Analysis of Variance Table

Response: x
              Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Outcome   1   3575  3575.4  83.528 < 2.2e-16 ***
Residuals    766  32789    42.8                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Insulin
Analysis of Variance Table

Response: x
              Df  Sum Sq Mean Sq F value    Pr(>F)    
pima$Outcome   1  949697  949697  108.03 < 2.2e-16 ***
Residuals    766 6733880    8791                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Age
Analysis of Variance Table

Response: x
              Df Sum Sq Mean Sq F value   Pr(>F)    
pima$Outcome   1   6027  6026.7  46.141 2.21e-11 ***
Residuals    766 100052   130.6                     
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$BloodPressure
Analysis of Variance Table

Response: x
              Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Outcome   1   3393  3392.5  23.415 1.579e-06 ***
Residuals    766 110985   144.9                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$SkinThickness
Analysis of Variance Table

Response: x
              Df Sum Sq Mean Sq F value    Pr(>F)    
pima$Outcome   1   4601  4600.7  51.208 1.947e-12 ***
Residuals    766  68821    89.8                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

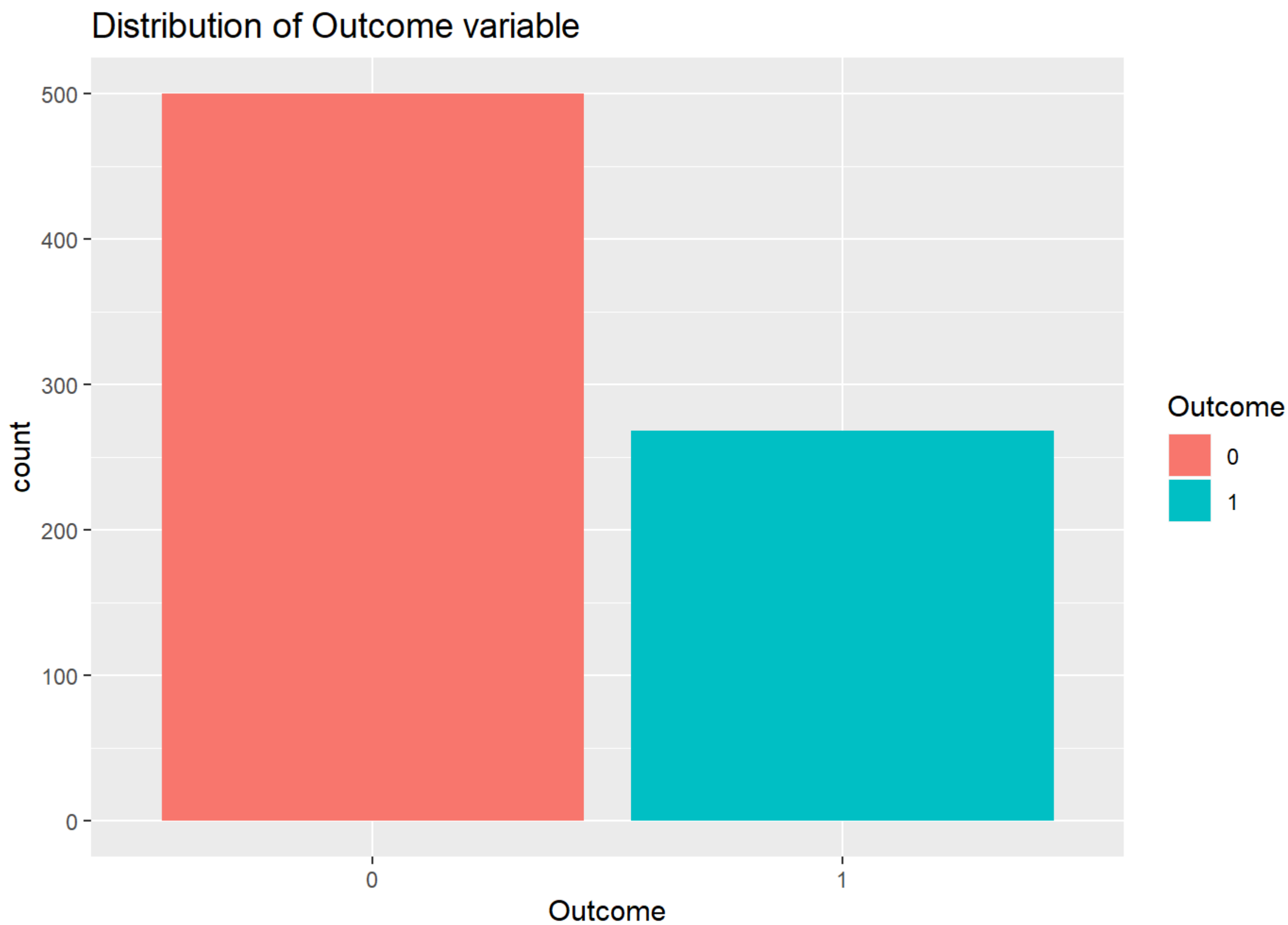
From the above t-tests, we saw that DiabetesPedigreeFunction and BMI are not signifcant variables against Pregnancies. However, both DiabetesPedigreeFunction and BMI are significant at 0.1% interval when tested against outcome(diabetes/not). Also, since few variables are highly correlated with other variables, we could conclude which variables to include in our model only after performing Exploratory Data Analysis.

Exploratory Data Analysis:

Inspecting the distribution of dependent variables and independent variables to identify control variables that also affect diabetes outcome along with pregnancies.

Distribution of Outcome variable

```
pima$Outcome <- factor(pima$Outcome)

ggplot(pima,aes(Outcome,fill = Outcome)) +
  geom_bar() +
  ggtitle("Distribution of Outcome variable")
```

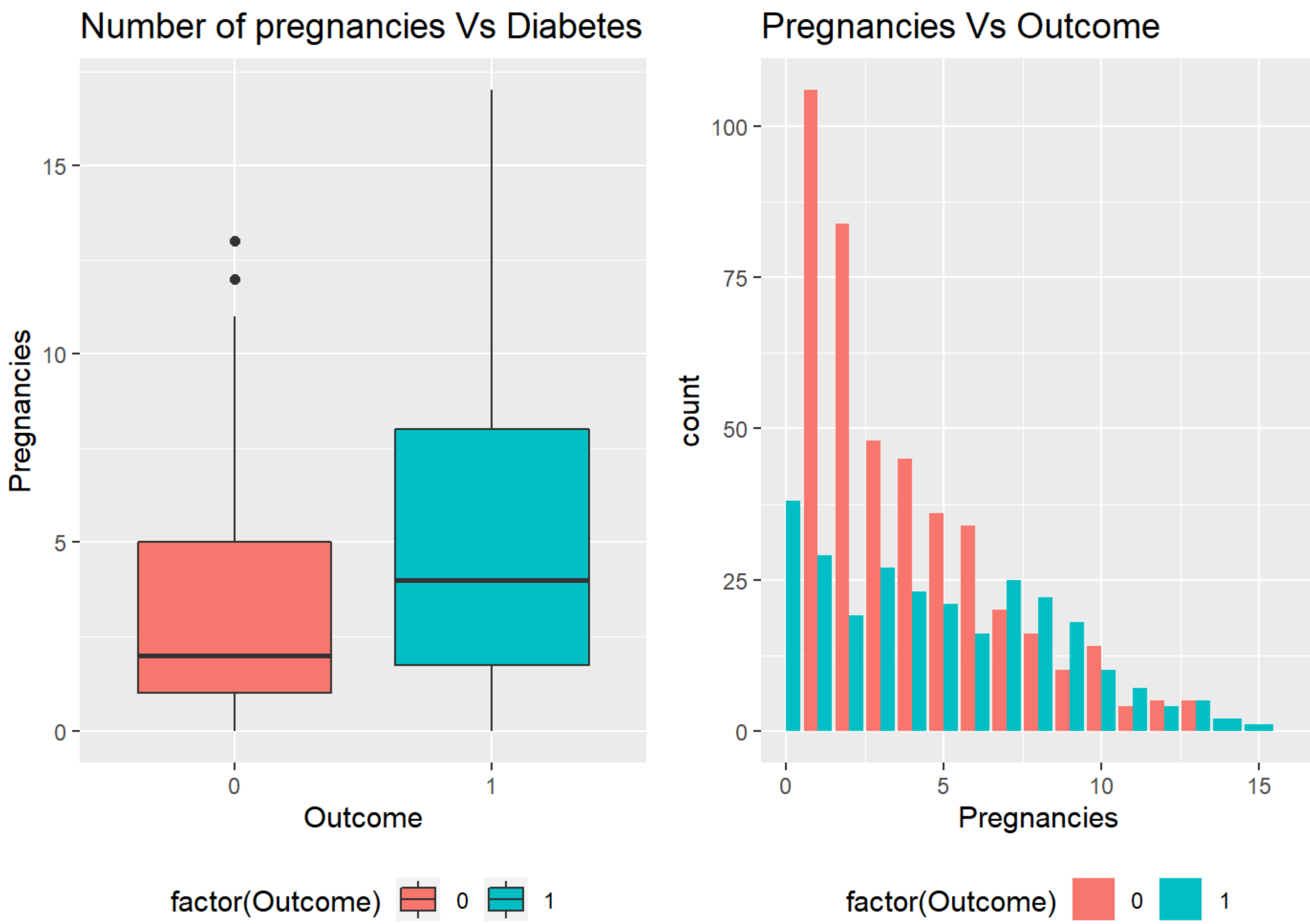


- We have approximately 500 females with no diabetes and more than 350 females with diabetes in our data.

A. Distribution of variable of interest over outcome

1. Distribution of Number of pregnancies

```
p1 <- ggplot(pima, aes(x = Outcome, y = Pregnancies,fill = factor(Outcome))) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Number of pregnancies Vs Diabetes")

p2 <- ggplot(pima,aes(x = Pregnancies,fill = factor(Outcome))) +
  geom_bar(position = "Dodge") +
  scale_x_continuous(limits = c(0,16)) +
  theme(legend.position = "bottom") +
  labs(title = "Pregnancies Vs Outcome")

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

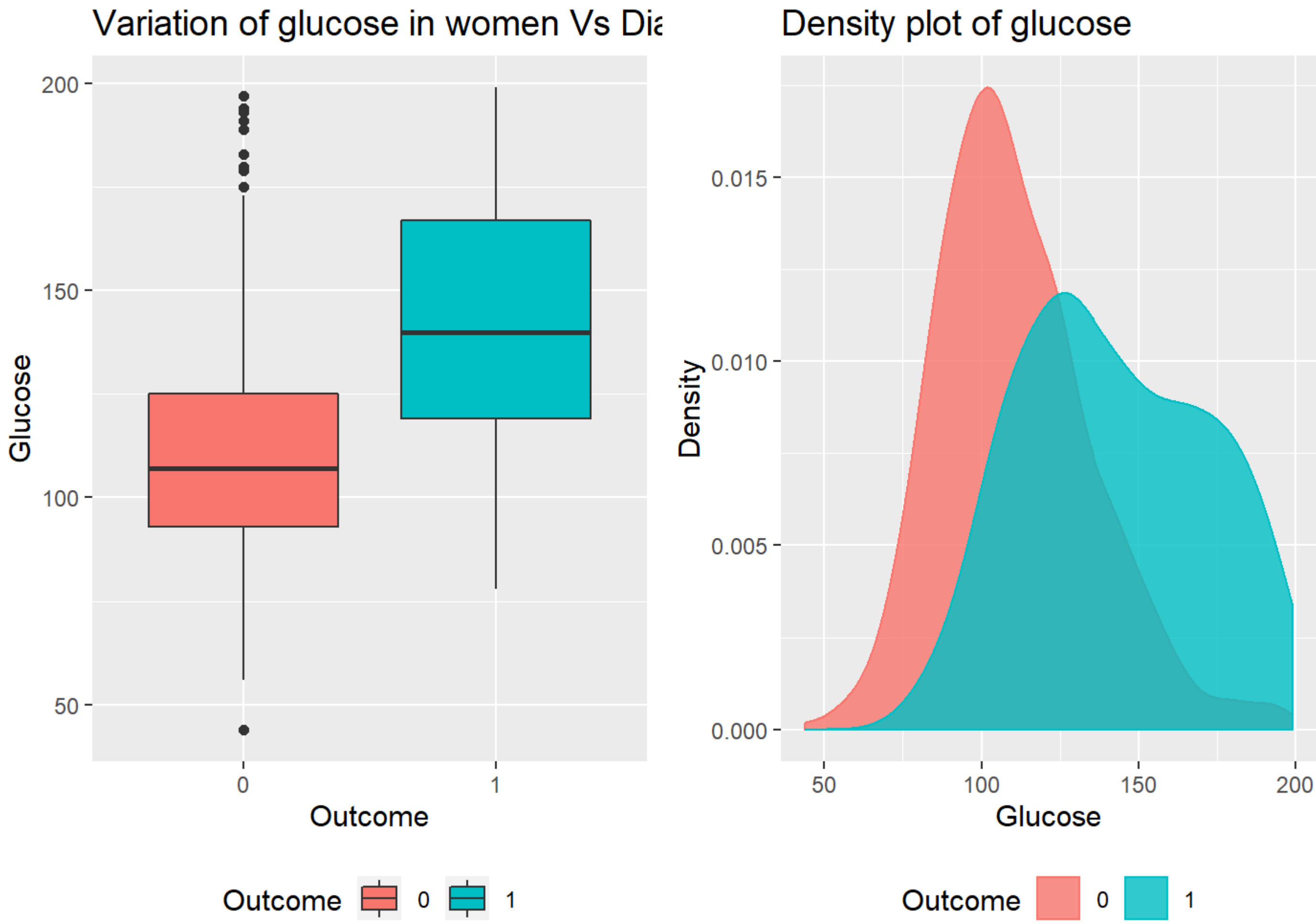


- From the plot between Pregnancies and outcome, it appears that as the number of Pregnancies increases, the risk of getting diabetes increases.

B. Distribution of control variables over outcome

1. Distribution of Glucose variable

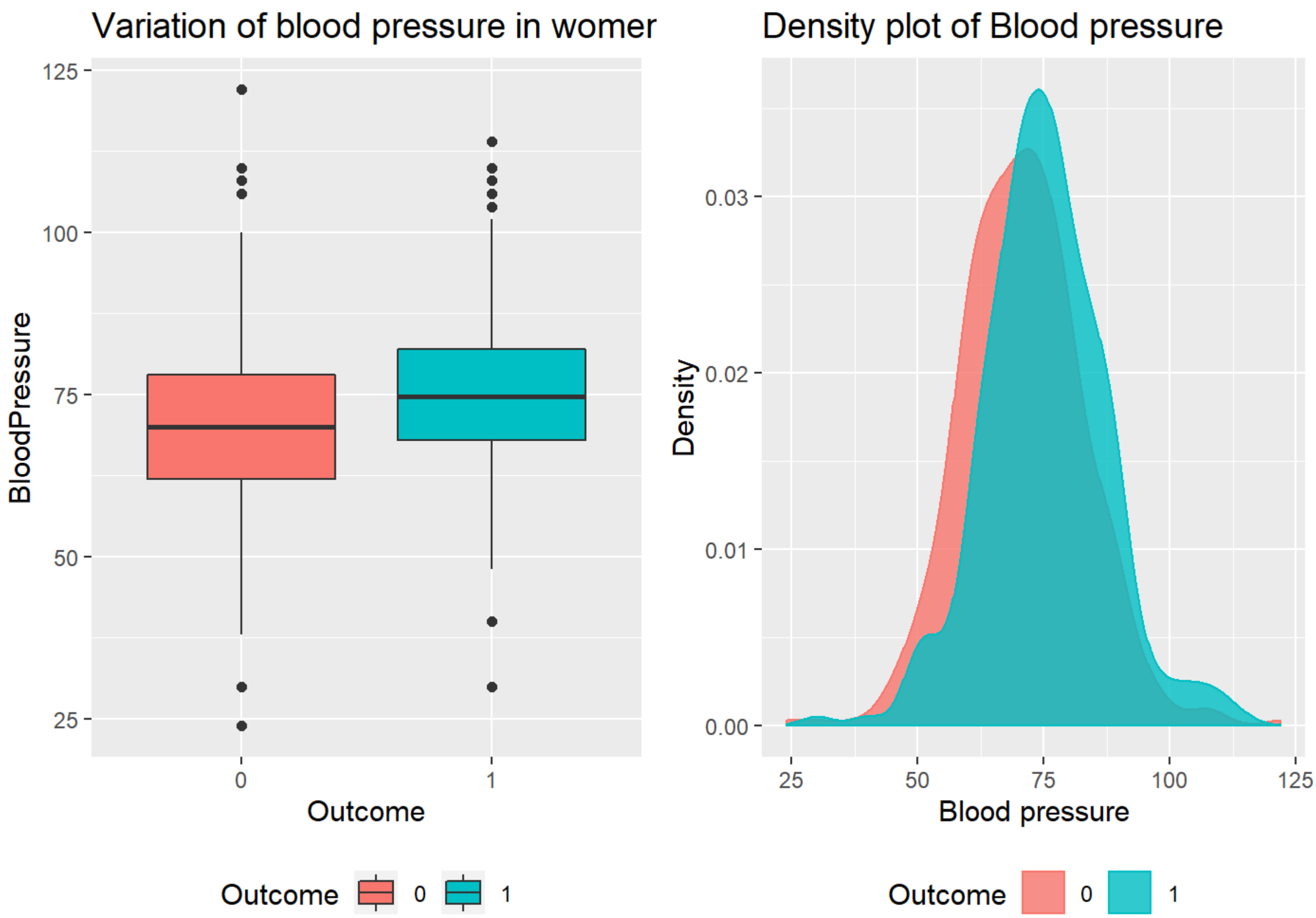
```
p1 <- ggplot(pima, aes(x = Outcome, y = Glucose,fill = Outcome)) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of glucose in women Vs Diabetes")

p2 <- ggplot(pima, aes(x = Glucose, color = Outcome, fill = Outcome)) +
  geom_density(alpha = 0.8) +
  theme(legend.position = "bottom") +
  labs(x = "Glucose", y = "Density", title = "Density plot of glucose")

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



- There's a clear difference in the amount of Glucose present in the female who have been diagnosed with Diabetes and those who haven't. Higher the glucose levels, more prone is the female to diabetes.

2. Distribution of BloodPressure variable

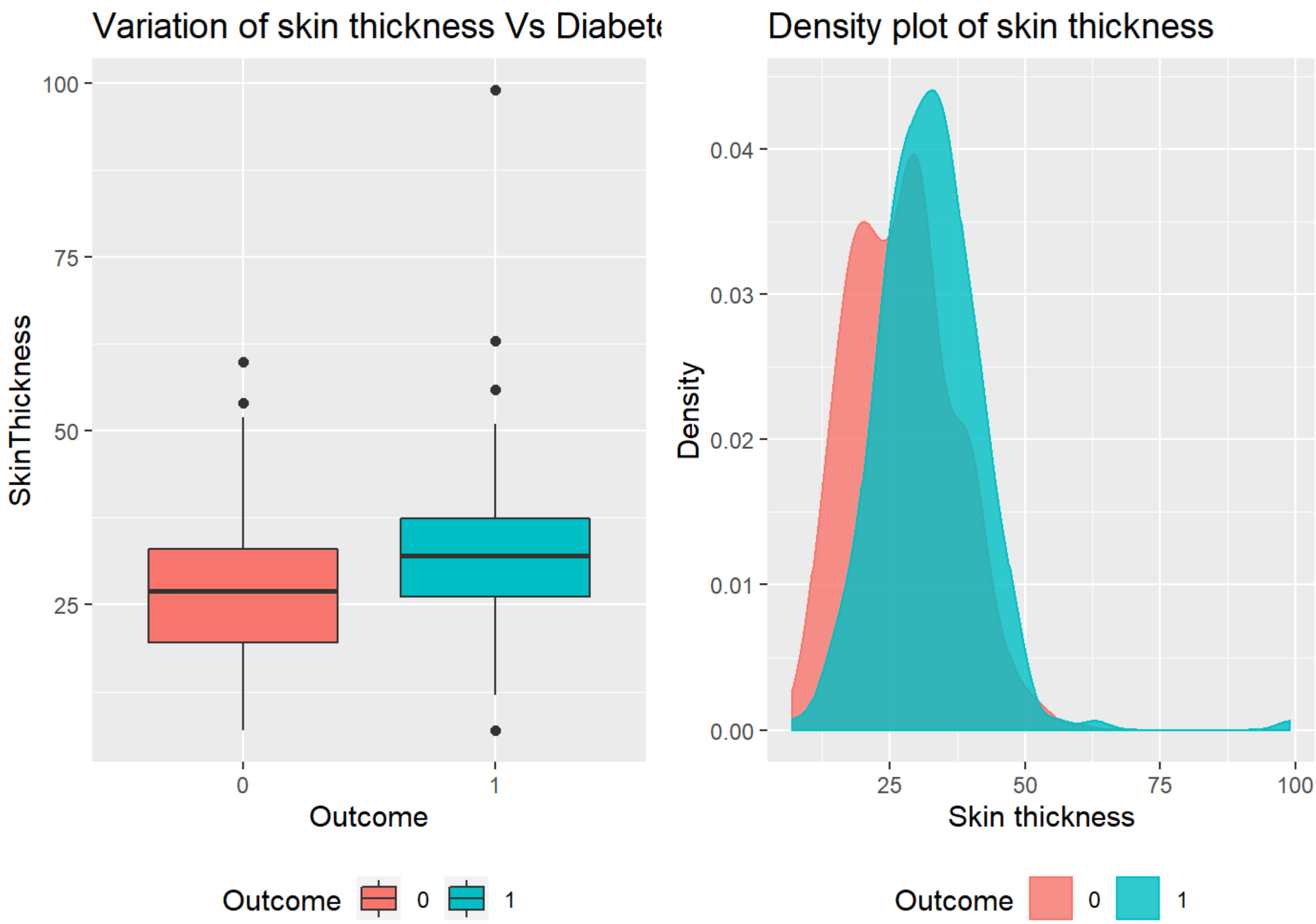
```
p1 <- ggplot(pima, aes(x = Outcome, y = BloodPressure,fill = Outcome)) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of blood pressure in women Vs Diabetes")

p2 <- ggplot(pima, aes(x = BloodPressure, color = Outcome, fill = Outcome)) +
  geom_density(alpha = 0.8) +
  theme(legend.position = "bottom") +
  labs(x = "Blood pressure", y = "Density", title = "Density plot of Blood pressure")


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



- There is no clear difference seen in the two categories of females who have and don't have Diabetes. This shows that Blood Pressure might not be a good predictor of the response variable.

3. Distribution of SkinThickness variable

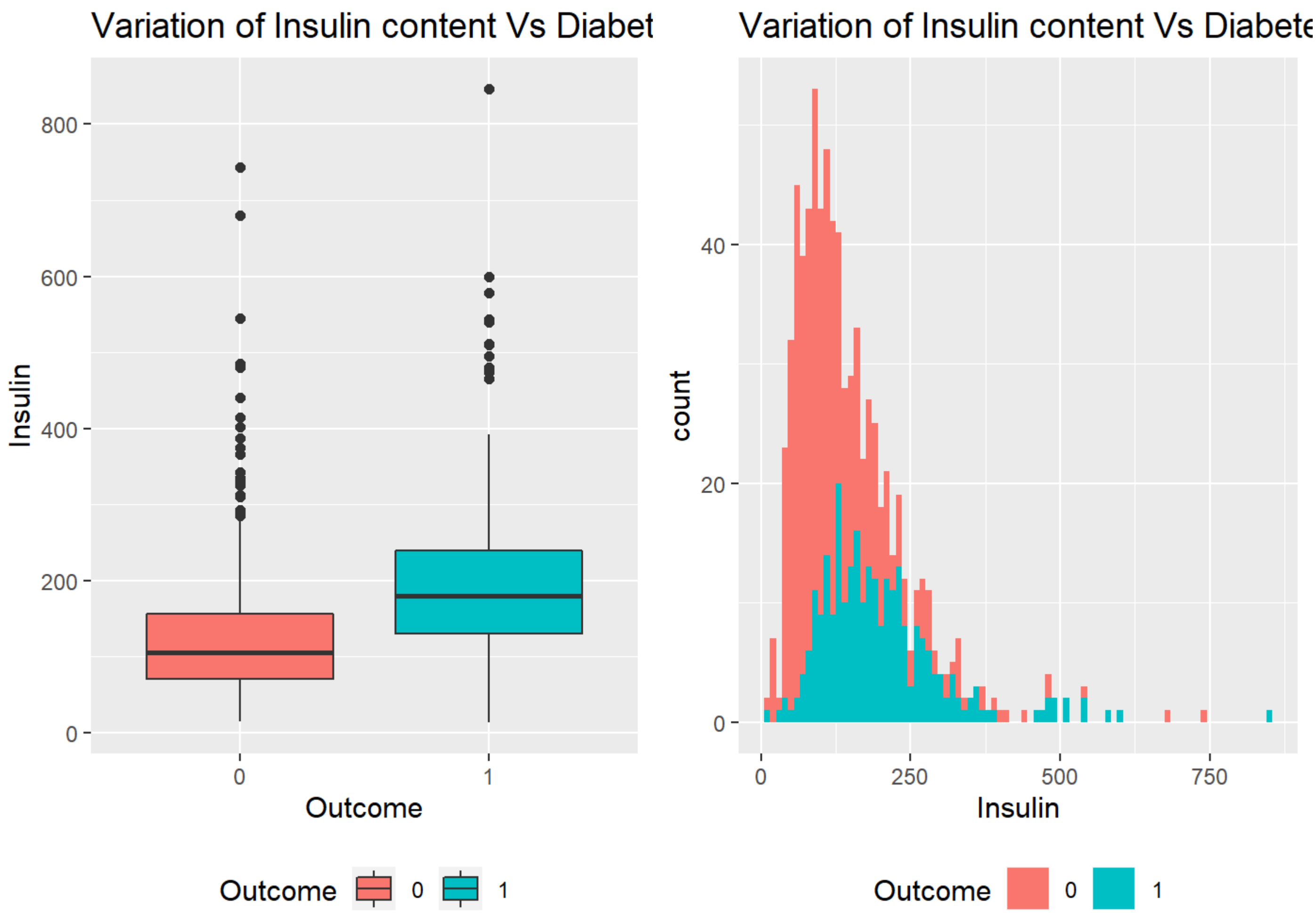
```
p1 <- ggplot(pima, aes(x = Outcome, y = SkinThickness,fill = Outcome)) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of skin thickness Vs Diabetes")

p2 <- ggplot(pima, aes(x = SkinThickness, color = Outcome, fill = Outcome)) +
  geom_density(alpha = 0.8) +
  theme(legend.position = "bottom") +
  labs(x = "Skin thickness", y = "Density", title = "Density plot of skin thickness")


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



- There is no clear difference seen in the two categories of females who have and don't have Diabetes. This shows that Skin Thickness might not be a good predictor of the response variable.

4. Distribution of Insulin variable

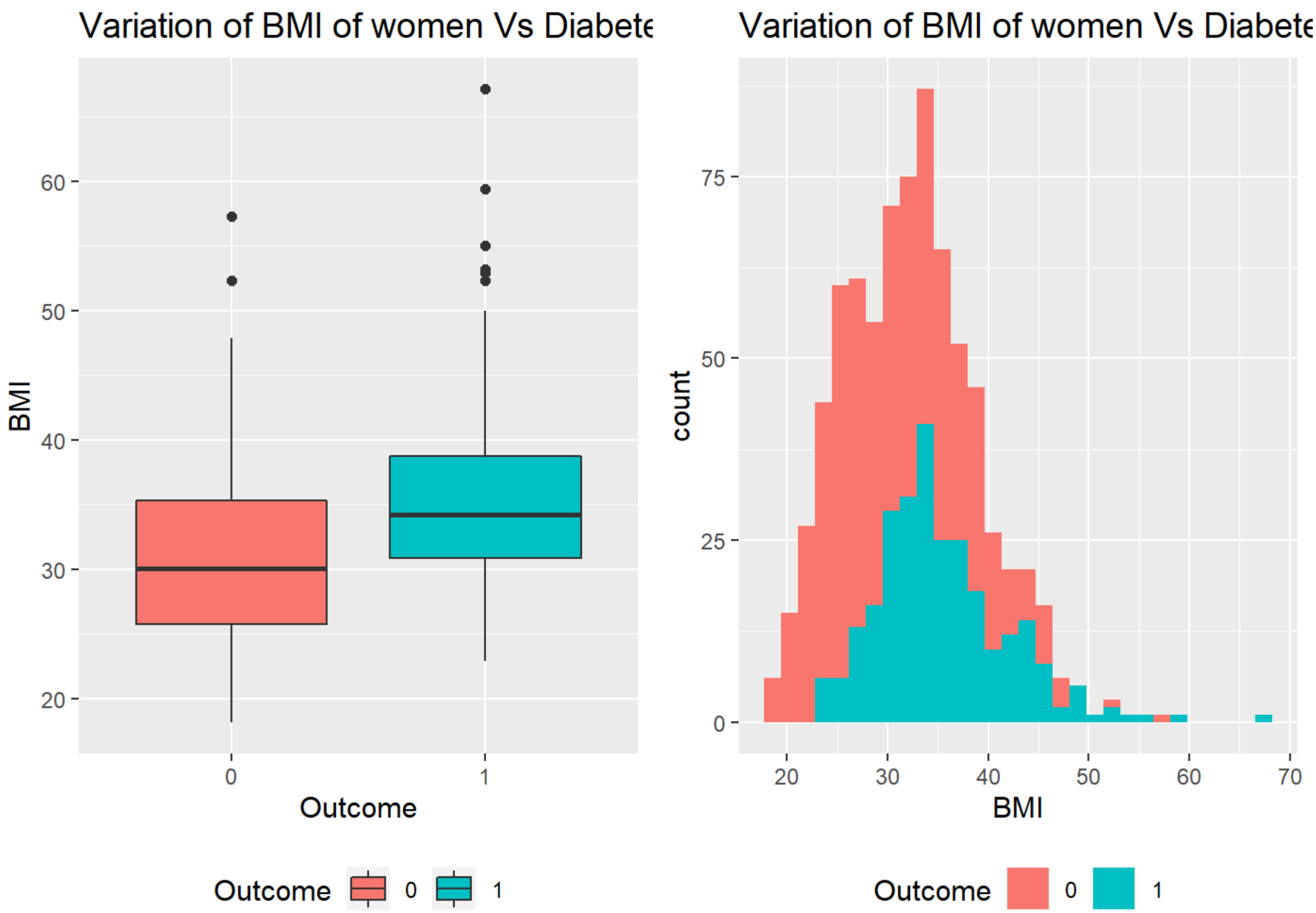
```
p1 <- ggplot(pima, aes(x = Outcome, y = Insulin,fill = Outcome)) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of Insulin content Vs Diabetes")

p2 <- ggplot(pima, aes(Insulin, fill = Outcome)) +
  geom_histogram(binwidth=10) +
  theme(legend.position = "bottom") +
  ggtitle("Variation of Insulin content Vs Diabetes")

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



- Females with higher insulin count are slightly more prone to diabetes than with lower insulin count.

5. Distribution of BMI variable

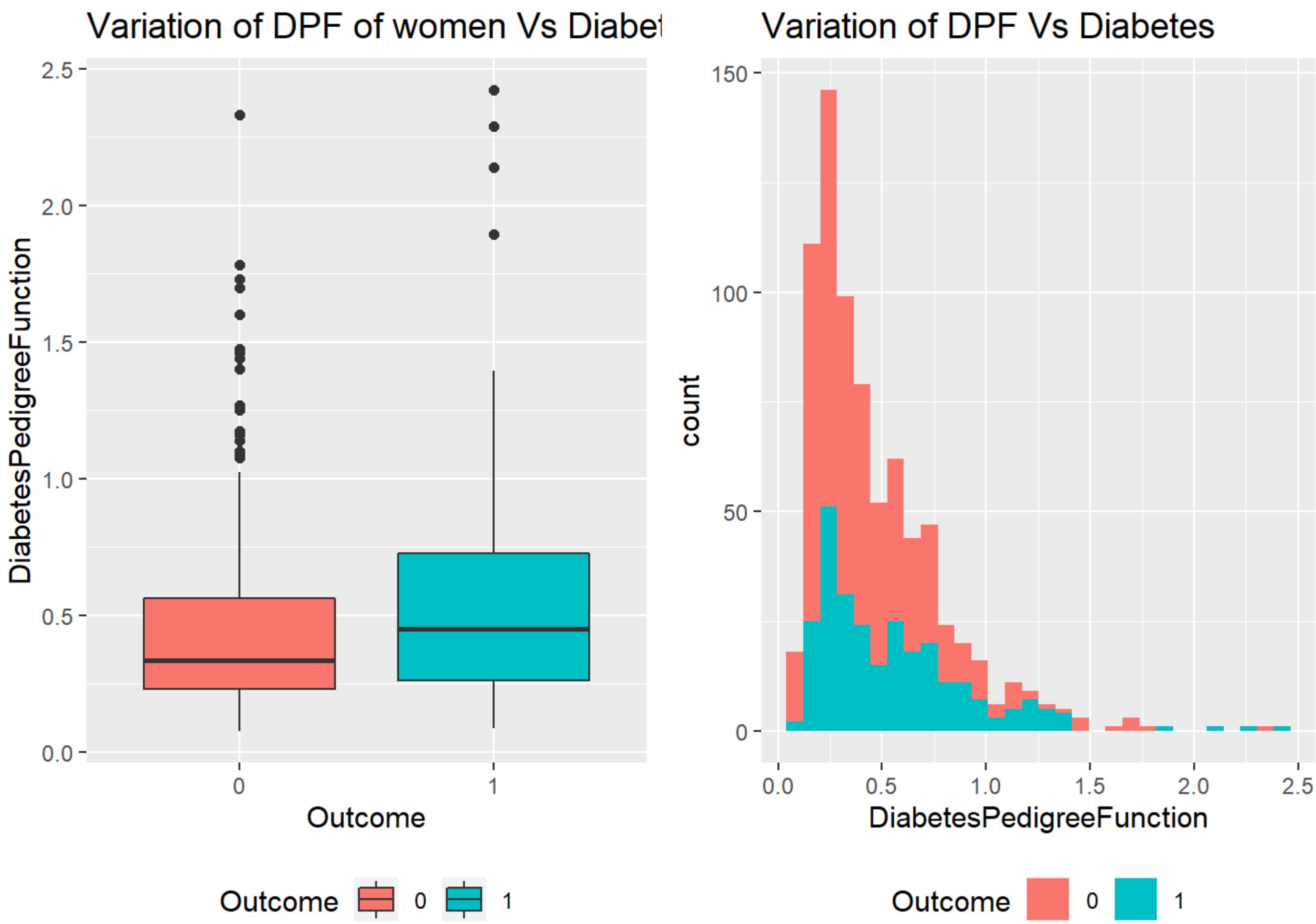
```
p1 <- ggplot(pima, aes(x = Outcome, y = BMI,fill = Outcome)) +
  geom_boxplot(binwidth = 5) +
  theme(legend.position = "bottom") +
  ggtitle("Variation of BMI of women Vs Diabetes")

p2 <- ggplot(pima, aes(BMI, fill = Outcome)) +
  geom_histogram() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of BMI of women Vs Diabetes")


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



- All the females who had Diabetes had a BMI greater than 25, which is above the normal levels. On the other hand, females who did not have Diabetes had a BMI ranging from 18 to 60. Females with low BMI are less likely to have diabetes.

6. Distribution of DiabetesPedigreeFunction variable

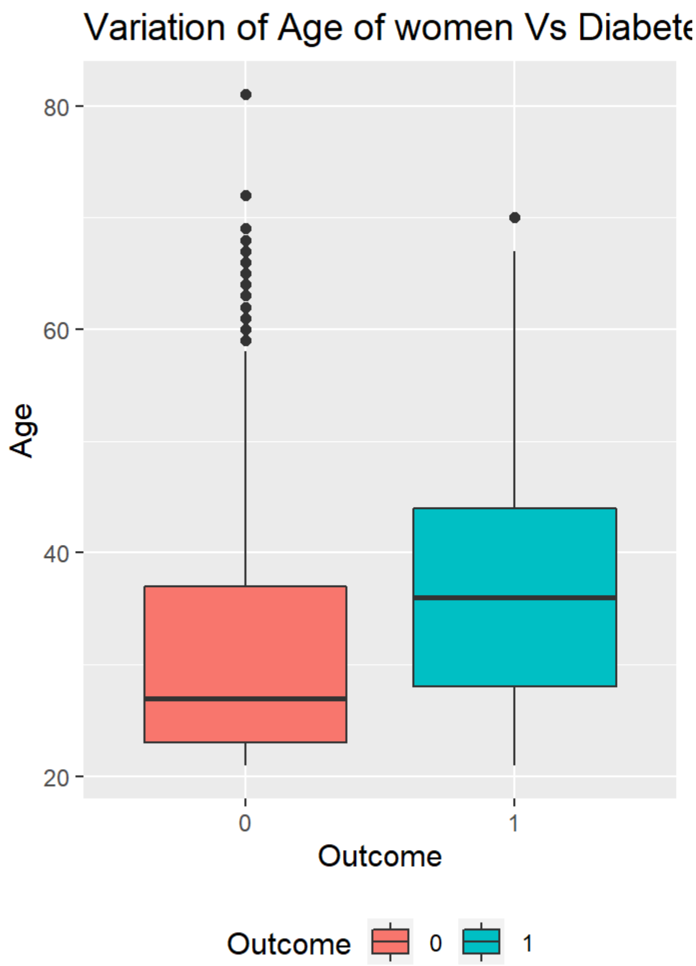
```
p1 <- ggplot(pima, aes(x = Outcome, y = DiabetesPedigreeFunction,fill = Outcome)) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of DPF of women Vs Diabetes")


p2 <- ggplot(pima, aes(DiabetesPedigreeFunction,fill = Outcome)) +
  geom_histogram() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of DPF Vs Diabetes")

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

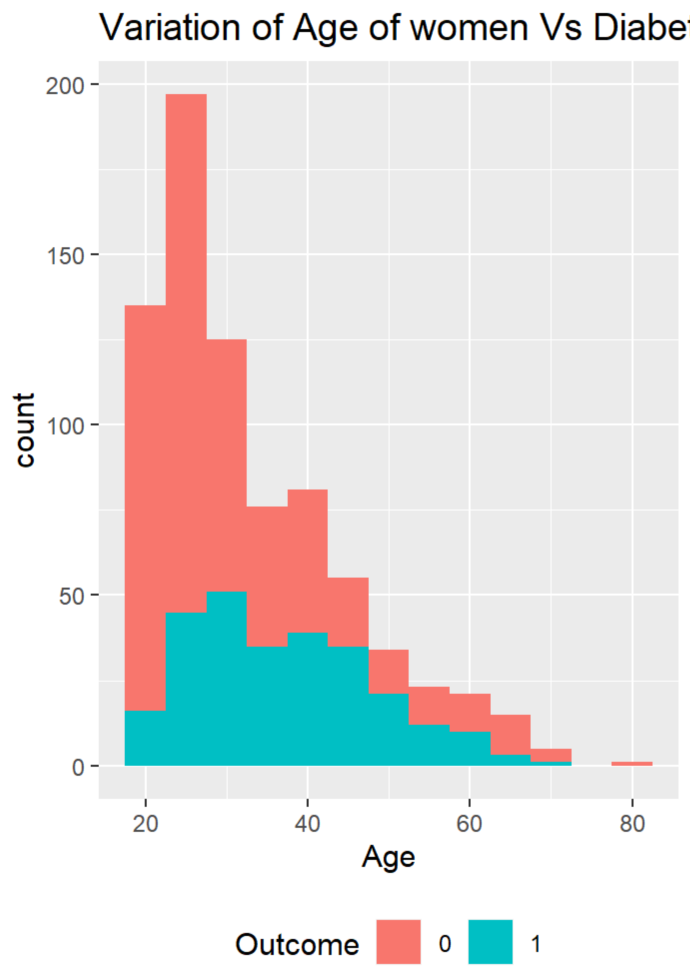


- Diabetes pedigree function is a function that scores the likelihood of diabetes based on family history. Females with higher DPF value are more likely to have diabetes than with lower values.

7. Distribution of Age variable

```
p1 <- ggplot(pima, aes(x = Outcome, y = Age,fill = Outcome)) +
  geom_boxplot() +
  theme(legend.position = "bottom") +
  ggtitle("Variation of Age of women Vs Diabetes")


p2 <- ggplot(pima, aes(Age, fill = Outcome)) +
  geom_histogram(binwidth = 5) +
  theme(legend.position = "bottom") +
  ggtitle("Variation of Age of women Vs Diabetes")

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

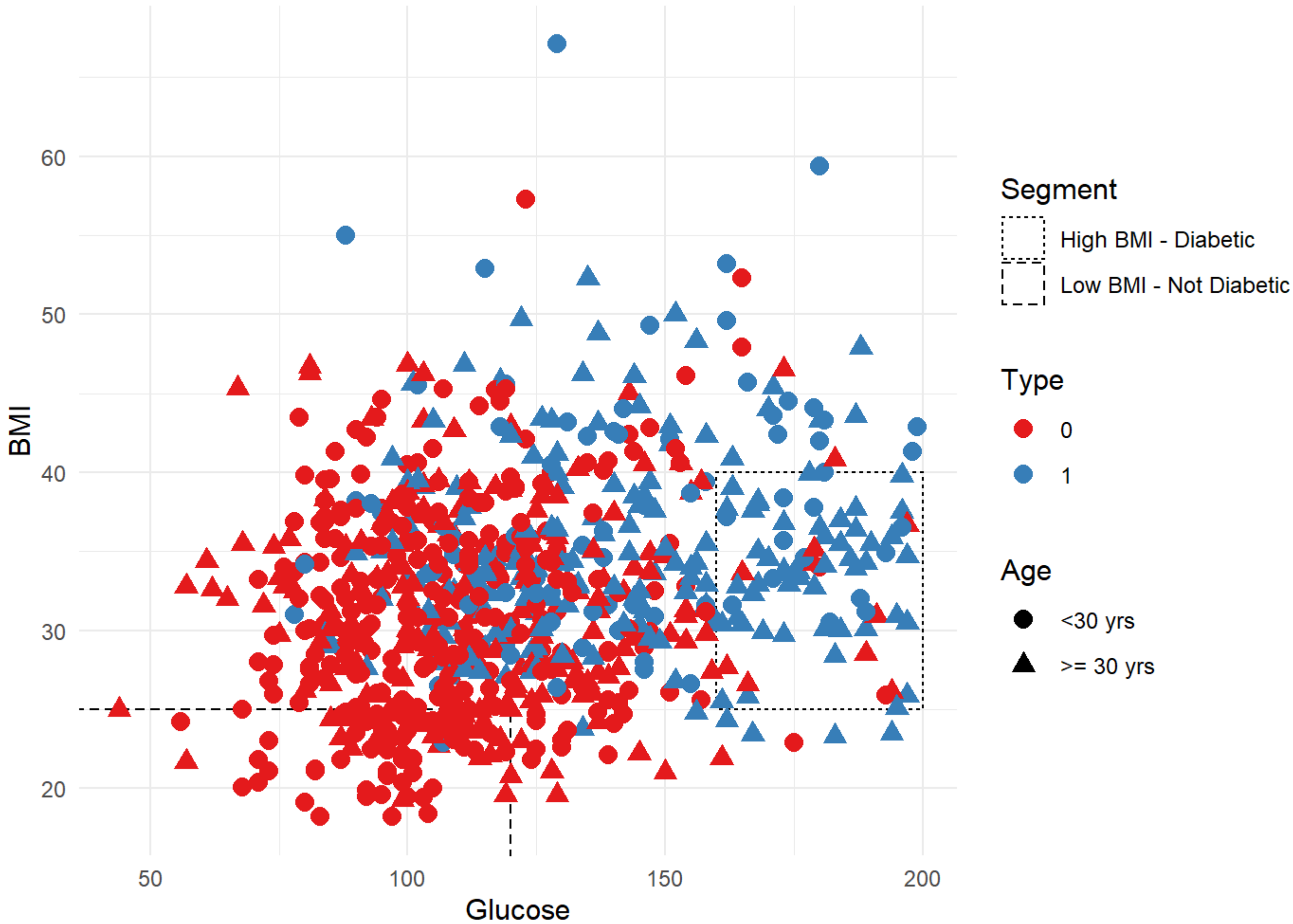




- Females over the age of 28 are more likely to have diabetes than females below the age of 28.

C. Cluster analysis of impact of BMI,Age,Glucose on Outcome

```
d<-pima
d$Age <- ifelse(d$Age < 30, "<30 yrs", ">= 30 yrs")

ggplot(d, aes(x = Glucose, y = BMI)) +
  geom_rect(aes(linetype = "High BMI - Diabetic"), xmin = 160, ymax = 40, fill = NA, xmax
= 200,
            ymin = 25, col = "black") +
  geom_rect(aes(linetype = "Low BMI - Not Diabetic"), xmin = 0, ymax = 25, fill = NA, xma
x = 120,
            ymin = 10, col = "black") +
  geom_point(aes(col = factor(Outcome), shape = factor(Age)), size = 3) +
  scale_color_brewer(name = "Type", palette = "Set1") +
  scale_shape(name = "Age") +
  scale_linetype_manual(values = c("High BMI - Diabetic" = "dotted", "Low BMI - Not Diabe
tic" = "dashed"),
                        name = "Segment") + theme_minimal()
```

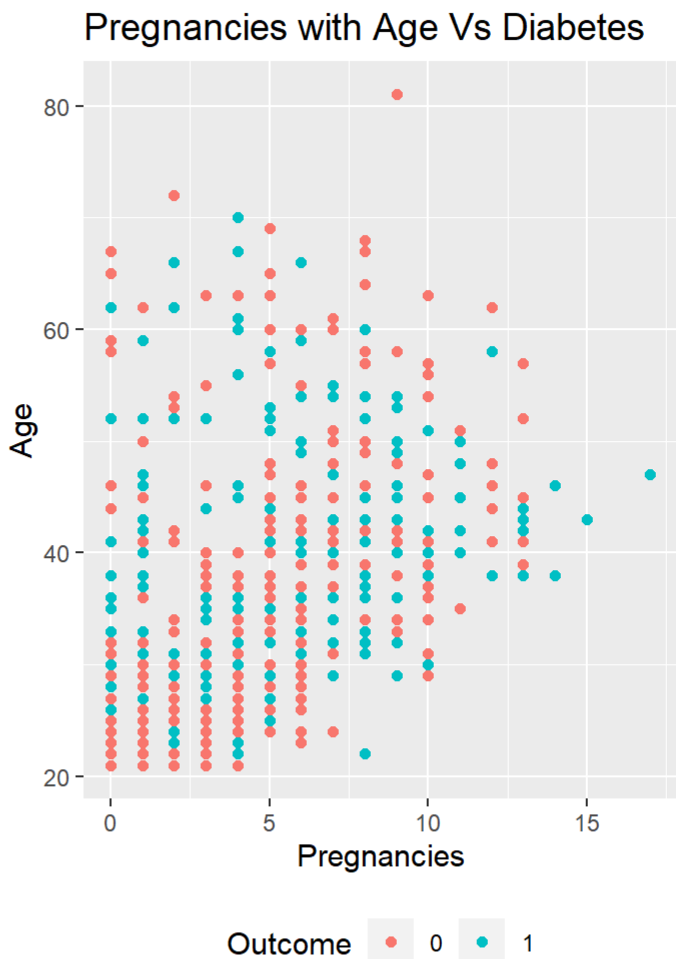


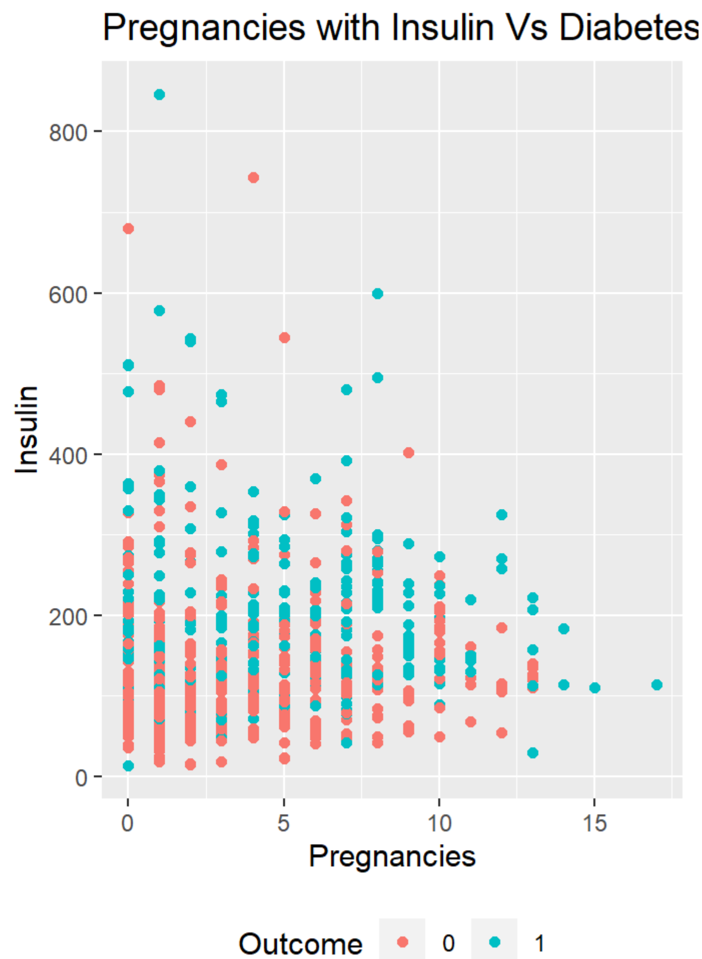
The dotted box toward the right side of the plot indicates:

- High BMI correlates with being diabetic when combined with glucose
- Females over the age of 30 are more prone to diabetes than females below 30

D. Distribution of Pregnancies and control variables Vs Outcome:

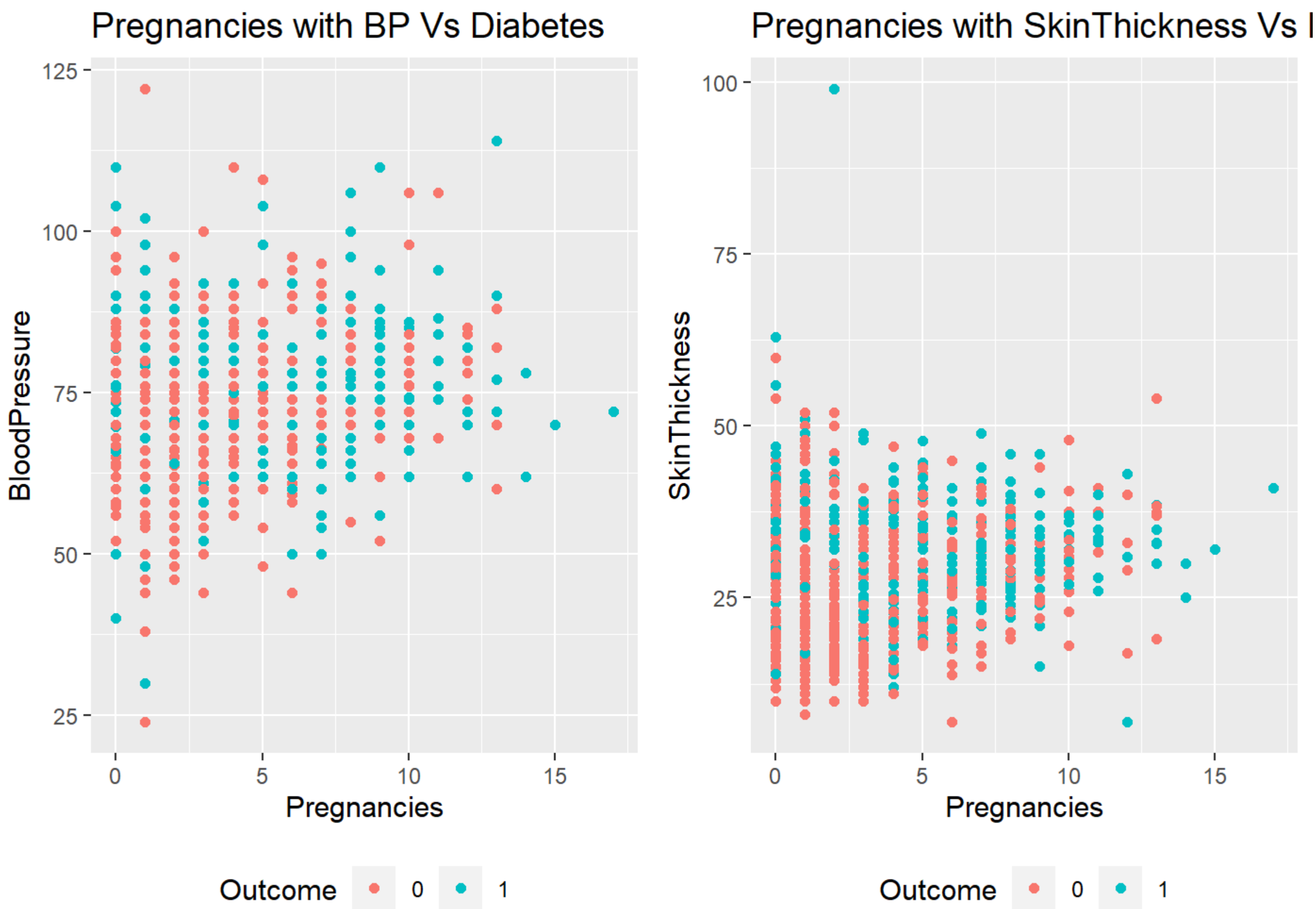
```
#Pregnancies with Age Vs Diabetes
p1 <- ggplot(pima, aes(x = Pregnancies, y = Age)) +
  geom_point(aes(color=Outcome)) +
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with Age Vs Diabetes")

#Pregnancies with Insulin Vs Diabetes
p2 <- ggplot(pima,aes(x=Pregnancies,y=Insulin))+
  geom_point(aes(color=Outcome))+
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with Insulin Vs Diabetes")

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





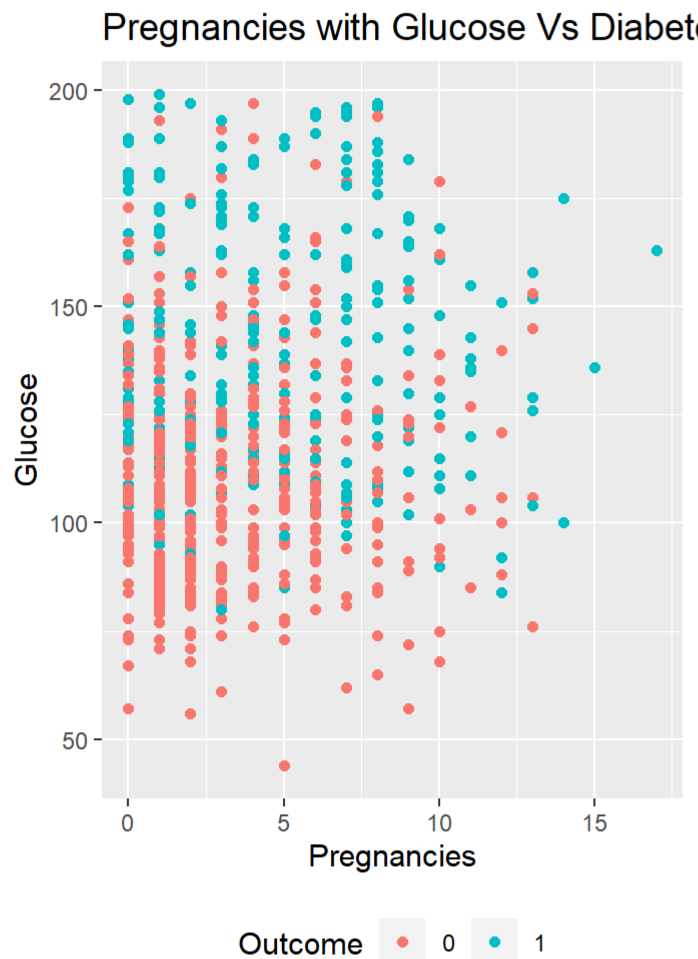
- No clear boundary can be drawn that separates Non-diabetic and Diabetic women based on Number of Pregnancies vs Age
- Non-diabetic women seemed to have lower levels of Insulin as opposed to Diabetic women who recorded low to high levels of Insulin. There is no clear pattern observed with increase in number of pregnancies and insulin count on Diabetes outcome.

```
#Pregnancies with BP Vs Diabetes
p1 <- ggplot(pima,aes(x=Pregnancies,y=BloodPressure))+
  geom_point(aes(color=Outcome))+
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with BP Vs Diabetes")

#Pregnancies with SkinThickness Vs Diabetes
p2 <- ggplot(pima,aes(x=Pregnancies,y=SkinThickness))+
  geom_point(aes(color=Outcome))+
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with SkinThickness Vs Diabetes")

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



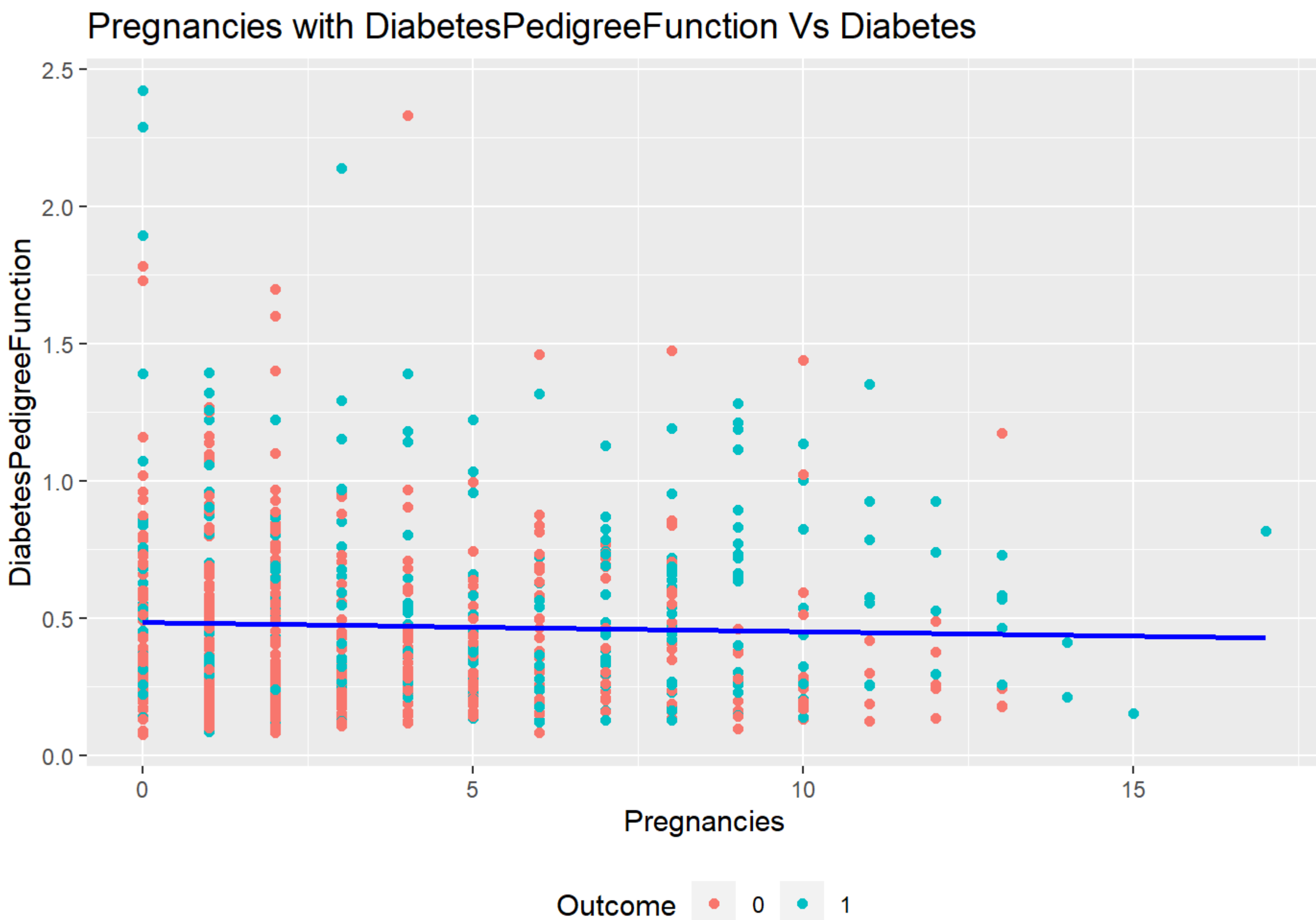
- Women who have Diabetes can't be differentiated from those who don't have based on BP values
- Women with low values of Skin Thickness are less prone to have Diabetes. However, there is no significant impact of SkinThickness along with increase in pregnancies on the Outcome

```
#Pregnancies with BMI Vs Diabetes
p1 <- ggplot(pima,aes(x=Pregnancies,y=BMI))+
  geom_point(aes(color=Outcome))+
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with BMI Vs Diabetes")

#Pregnancies with Glucose Vs Diabetes
p2 <- ggplot(pima,aes(x=Pregnancies,y=Glucose))+
  geom_point(aes(color=Outcome))+
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with Glucose Vs Diabetes")

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



- Pima females with fewer number of pregnancy, had diabetes at larger ranges of BMI(above 30). As the number of pregnancies increased, Pima female with lower BMI were more prone to diabetes. For females with diabetes, there were more outliers in which a female who had very few pregnancies had very high BMIs. For females who didnt have diabetes, the female who had about 6-8 pregnancies seemed to have relatively high BMIs.
- Higher the glucose count, more prone is the female to getting diabetes. As the number of pregnancies increase, we see that even at comparatively lower glucose levels, pima indian women are more prone to getting diabetes.

```
#Pregnancies with DiabetesPedigreeFunction Vs Diabetes
ggplot(pima,aes(x=Pregnancies,y=DiabetesPedigreeFunction))+
  geom_point(aes(color=Outcome))+
  theme(legend.position = "bottom") +
  ggtitle("Pregnancies with DiabetesPedigreeFunction Vs Diabetes")+stat_smooth(method="l
m",se=FALSE, col = "blue" )
```



- Pima females with high diabetespedigree function are more prone to getting diabetes. However, we also see from above graph that, even if the likelihood of getting diabetes is low for few patients, because they have more pregnancy count(pregnancy count around 6-9) they are more prone to getting diabetes.

From our Exploratory Analysis, we see that :

- Variables that have substantial effect on Outcome:
    - Glucose
    - DiabetesPedigreeFunction
    - BMI
    - Insulin
    - Age
- Variables that have substantial effect on both Pregnancies and Outcome:
    - Glucose
    - DiabetesPedigreeFunction
    - BMI

We believe through the correlation plot, t-test and exploratory analysis, combined with our intuition from survey, variables - Age, Glucose,BMI,DiabetesPedigreeFunction should be included as control variables for our variable of interest - Pregnancies in the model.

Regression : Effect of Pregnancies on Diabetes

We perform logit regression to observe the causal effect of pregnancies on Diabetes along with other variables.

We run different model by including the control variables one by one.

Adding variable ‘Age’ in the base model along with the variable of interest.

```
l1 = glm(Outcome~Pregnancies, family=binomial, x=TRUE,  data=pima)
l2 = glm(Outcome~Pregnancies+Age, family=binomial, x=TRUE,  data=pima)


stargazer(l1, l2, se=list(NULL, NULL),
          column.labels=c("logit-1", "logit-2"),
          title="Logit- Pregnancy and Diabetes", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```

Logit- Pregnancy and Diabetes
================================================
                      Dependent variable:
                  ------------------------------
                            Outcome
                      logit-1        logit-2
                        (1)            (2)
------------------------------------------------
Pregnancies          0.137***       0.082**
                      (0.023)       (0.027)

Age                                 0.030***
                                    (0.008)

Constant             -1.177***     -1.963***
                      (0.123)       (0.241)

------------------------------------------------
Observations           768            768
Log Likelihood       -478.105       -470.549
Akaike Inf. Crit.     960.210       947.098
================================================
Note:               *p<0.05; **p<0.01; ***p<0.001
```

```
## For l1 model
pseudoR2=(l1$null.deviance-l1$deviance)/l1$null.deviance
print(paste("pseudo_R2 for model 1",pseudoR2)) # 0.0375185004267976
```

```
[1] "pseudo_R2 for model 1 0.0375185004267976"
```

```
## For l2 model
pseudoR2=(l2$null.deviance-l2$deviance)/l2$null.deviance
print(paste("pseudo_R2 for model 2",pseudoR2)) #0.0527293813429582
```

```
[1] "pseudo_R2 for model 2 0.0527293813429582"
```

```
## Calculates marginal effect of the regressors
fm1a=maBina(l1, x.mean=TRUE, rev.dum=TRUE, digits=3)
fm2a=maBina(l2, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, se=list(NULL, NULL),
          title="Logit- Marginal Effects", type="text",
          column.labels=c("logit-1", "logit-2"),
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```


Logit- Marginal Effects
================================================
                      Dependent variable:
                  ------------------------------
                            Outcome
                      logit-1        logit-2
                        (1)            (2)
------------------------------------------------
Pregnancies          0.031***       0.018**
                      (0.005)       (0.006)

Age                                 0.007***
                                    (0.002)

Constant             -0.265***     -0.441***
                      (0.024)       (0.050)


------------------------------------------------
Observations            768           768
Akaike Inf. Crit.     960.210       947.098
================================================
Note:               *p<0.05; **p<0.01; ***p<0.001
```

1. Adding ‘Age’ in the model increased log likelihood of the second model from -478.105 to -470.549.
2. Omitted variable bias corrected- By adding statistically significant variable ‘Age’,the coefficient of VOI went down from 0.031 to 0.018. Therefore, adding the variable ‘age’ corrected ‘upward omitted variable bais’.

3. The AIC(Akaike Inf.Crit) has reduced from 960.210 to 947.098, showing that the model is getting better with an addition of the variable. The lower the AIC value the better.

Pseudo_r2: Pseudo_R2 increased from 0.03751 to 0.05272. This shows that adding the variabale 'age' increased the explanatory power of the model from 3.7% to 5.2%.

Adding variable 'BMI' in the model

```
l3 = glm(Outcome~Pregnancies+Age+BMI, family=binomial, x=TRUE, data=pima)

stargazer(l1, l2, l3, se=list(NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3"),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Pregnancy and Diabetes
=================================================
                        Dependent variable:
                  -------------------------------
                              Outcome
                   logit-1    logit-2   logit-3
                     (1)        (2)       (3)
-------------------------------------------------
Pregnancies        0.137***   0.082**   0.090**
                   (0.023)    (0.027)   (0.028)

Age                           0.030*** 0.033***
                              (0.008)   (0.008)

BMI                                    0.110***
                                        (0.013)

Constant          -1.177***  -1.963*** -5.730***
                   (0.123)    (0.241)   (0.540)


-------------------------------------------------
Observations         768        768       768
Log Likelihood     -478.105   -470.549 -430.165
Akaike Inf. Crit.  960.210    947.098   868.330
=================================================
Note:               *p<0.05; **p<0.01; ***p<0.001
```

```
pseudoR2=(l3$null.deviance-l3$deviance)/l3$null.deviance
print(paste("pseudo_R2 for model 3",pseudoR2))
```

```
[1] "pseudo_R2 for model 3 0.134027341631049"
```

```
fm3a=maBina(l3, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, fm3a, se=list(NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3"),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Marginal Effects
=================================================
                        Dependent variable:
                  -------------------------------
                              Outcome
                   logit-1    logit-2    logit-3
                     (1)        (2)        (3)
-------------------------------------------------
Pregnancies        0.031***   0.018**    0.020**
                   (0.005)    (0.006)    (0.006)

Age                           0.007***  0.007***
                              (0.002)    (0.002)

BMI                                     0.024***
                                         (0.003)

Constant          -0.265***  -0.441***  -1.252***
                   (0.024)    (0.050)    (0.110)

-------------------------------------------------
Observations         768        768        768
Akaike Inf. Crit.  960.210    947.098    868.330
=================================================
Note:               *p<0.05; **p<0.01; ***p<0.001
```

1. Adding 'BMI' in the model increased log likelihood of the third model from -470.549 to -430.165.
2. Omitted variable bias corrected- By adding statistically significant variable 'BMI',the coefficient of VOI went up from 0.018 to 0.020. Therefore, adding the variable 'BMI' corrected 'downward omitted variable bais'.
3. The AIC(Akaike Inf.Crit) has reduced from 947.098 to 868.330, showing that the model is getting better with an addition of the variable. The lower the AIC value the better.

Pseudo_r2: Pseudo_R2 increased from 0.05272 to 0.13. This shows that adding the variabale 'BMI' increased the explanatory power of the model from 5.2% to 13%.

Adding variable 'Glucose' in the model

```
l4 = glm(Outcome~Pregnancies+Age+BMI+Glucose, family=binomial, x=TRUE, data=pima)


stargazer(l1, l2, l3,l4, se=list(NULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", "logit-4"),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Pregnancy and Diabetes
=========================================================
                             Dependent variable:
                  ---------------------------------------
                                   Outcome
                   logit-1   logit-2   logit-3   logit-4
                     (1)       (2)       (3)       (4)
---------------------------------------------------------
Pregnancies       0.137***   0.082**   0.090**  0.116***
                   (0.023)   (0.027)   (0.028)   (0.032)

Age                         0.030***  0.033***    0.011
                             (0.008)   (0.008)   (0.009)

BMI                                   0.110***  0.093***
                                       (0.013)   (0.015)

Glucose                                         0.036***
                                                 (0.004)

Constant          -1.177*** -1.963*** -5.730*** -9.122***
                   (0.123)   (0.241)   (0.540)   (0.715)

---------------------------------------------------------
Observations         768       768       768       768
Log Likelihood    -478.105  -470.549  -430.165  -361.777
Akaike Inf. Crit.  960.210   947.098   868.330   733.554
=========================================================
Note:                       *p<0.05; **p<0.01; ***p<0.001
```

```
## For l2 model
pseudoR2=(l4$null.deviance-l4$deviance)/l4$null.deviance
print(paste("pseudo_R2 for model 4",pseudoR2))
```

```
[1] "pseudo_R2 for model 4 0.271700038302199"
```

```
fm4a=maBina(l4, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, fm3a, fm4a, se=list(NULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4'),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Marginal Effects
=========================================================
                             Dependent variable:
                  ---------------------------------------
                                   Outcome
                   logit-1   logit-2   logit-3   logit-4
                     (1)       (2)       (3)       (4)
---------------------------------------------------------
Pregnancies       0.031***   0.018**   0.020**  0.024***
                   (0.005)   (0.006)   (0.006)   (0.007)

Age                         0.007***  0.007***    0.002
                             (0.002)   (0.002)   (0.002)

BMI                                   0.024***  0.019***
                                       (0.003)   (0.003)

Glucose                                         0.008***
                                                 (0.001)

Constant          -0.265*** -0.441*** -1.252*** -1.906***
                   (0.024)   (0.050)   (0.110)   (0.137)

---------------------------------------------------------
Observations         768       768       768       768
Akaike Inf. Crit.  960.210   947.098   868.330   733.554
=========================================================
Note:                        *p<0.05; **p<0.01; ***p<0.001
```

1. Adding ‘Glucose’ in the model increased log likelihood of the fourth model from -430.165 to -361.777.
2. Omitted variable bias corrected- By adding statistically significant variable ‘Glucose’,the coefficient of VOI went up from 0.020 to 0.024. Therefore, adding the variable ‘Glucose’ corrected ‘downward omitted variable bais’.
3. The AIC(Akaike Inf.Crit) has reduced from 868.330 to 733.554, showing that the model is getting better with an addition of the variable. The lower the AIC value the better.

Pseudo_r2: Pseudo_R2 increased from 0.13 to 0.2717. This shows that adding the variabale ‘Glucose level’ increased the explanatory power of the model from 13% to 27%.

Adding variable ‘DiabetesPedigreeFunction’ in the model

```
l5 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction, family=binomial, x
=TRUE, data=pima)


stargazer(l1, l2, l3,l4,l5, se=list(NULL, NULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", "logit-4", "logit-5"),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Pregnancy and Diabetes
==========================================================================
                                        Dependent variable:
                         -------------------------------------------------
                                              Outcome
                          logit-1   logit-2   logit-3   logit-4   logit-5
                            (1)       (2)       (3)       (4)       (5)
--------------------------------------------------------------------------
Pregnancies              0.137***   0.082**   0.090**  0.116***  0.123***
                          (0.023)   (0.027)   (0.028)   (0.032)   (0.032)

Age                                0.030***  0.033***    0.011     0.011
                                    (0.008)   (0.008)   (0.009)   (0.009)

BMI                                          0.110***  0.093***  0.090***
                                              (0.013)   (0.015)   (0.015)

Glucose                                                0.036***  0.036***
                                                        (0.004)   (0.004)

DiabetesPedigreeFunction                                          0.877**
                                                                  (0.295)

Constant                 -1.177*** -1.963*** -5.730*** -9.122*** -9.393***
                          (0.123)   (0.241)   (0.540)   (0.715)   (0.734)


--------------------------------------------------------------------------
Observations                768       768       768       768       768
Log Likelihood           -478.105  -470.549  -430.165  -361.777  -357.261
Akaike Inf. Crit.         960.210   947.098   868.330   733.554   726.521
==========================================================================
Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

```
## For l2 model
pseudoR2=(l5$null.deviance-l5$deviance)/l5$null.deviance
print(paste("pseudo_R2 for model 5",pseudoR2))
```

```
[1] "pseudo_R2 for model 5 0.280792200627194"
```

```
fm5a=maBina(l5, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, fm3a, fm4a, fm5a, se=list(NULL, NULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4', 'logit-5'),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Marginal Effects
===========================================================================
                                       Dependent variable:
                         --------------------------------------------------
                                             Outcome
                          logit-1   logit-2   logit-3   logit-4   logit-5
                            (1)       (2)       (3)       (4)       (5)
---------------------------------------------------------------------------
Pregnancies              0.031***   0.018**   0.020**  0.024***  0.026***
                          (0.005)   (0.006)   (0.006)   (0.007)   (0.007)

Age                                0.007***  0.007***    0.002     0.002
                                    (0.002)   (0.002)   (0.002)   (0.002)

BMI                                          0.024***  0.019***  0.019***
                                              (0.003)   (0.003)   (0.003)

Glucose                                                0.008***  0.008***
                                                        (0.001)   (0.001)

DiabetesPedigreeFunction                                          0.183**
                                                                  (0.062)

Constant                 -0.265*** -0.441*** -1.252*** -1.906*** -1.962***
                          (0.024)   (0.050)   (0.110)   (0.137)   (0.142)

---------------------------------------------------------------------------
Observations                768       768       768       768       768
Akaike Inf. Crit.         960.210   947.098   868.330   733.554   726.521
===========================================================================
Note:                                         *p<0.05; **p<0.01; ***p<0.001
```

1. Adding 'DiabetesPedigreeFunction' in the model increased log likelihood of the fifth model from -361.777 to -357.261
2. Omitted variable bias corrected- By adding statistically significant variable 'DiabetesPedigreeFunction',the coefficient of VOI went up from 0.024 to 0.026. Therefore, adding the variable 'DiabetesPedigreeFunction' corrected 'downward omitted variable bais'.

3. The AIC(Akaike Inf.Crit) has reduced from 733.554 to 726.521, showing that the model is getting better with an addition of the variable. The lower the AIC value the better.

Pseudo_r2: Pseudo_R2 increased from 0.2717 to 0.280. This shows that adding the variabale 'Glucose level' increased the explanatory power of the model from 27.17% to 28%.

Adding variable 'BloodPressure' in the model

```
l6 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction+BloodPressure, fami
ly=binomial, x=TRUE, data=pima)


stargazer(l1, l2, l3,l4,l5,l6, se=list(NULL, NULL, NULL, NULL, NULL,NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", "logit-4", "logit-5", "logit-
6"),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Pregnancy and Diabetes
=====================================================================================
                                          Dependent variable:
                         ------------------------------------------------------------
                                                Outcome
                          logit-1   logit-2   logit-3   logit-4   logit-5   logit-6
                            (1)       (2)       (3)       (4)       (5)       (6)
-------------------------------------------------------------------------------------
Pregnancies              0.137***   0.082**   0.090**  0.116***  0.123***  0.125***
                          (0.023)   (0.027)   (0.028)   (0.032)   (0.032)   (0.032)

Age                                0.030***  0.033***    0.011     0.011     0.013
                                    (0.008)   (0.008)   (0.009)   (0.009)   (0.009)

BMI                                          0.110***  0.093***  0.090***  0.094***
                                              (0.013)   (0.015)   (0.015)   (0.015)

Glucose                                                0.036***  0.036***  0.036***
                                                        (0.004)   (0.004)   (0.004)

DiabetesPedigreeFunction                                           0.877**   0.864**
                                                                  (0.295)   (0.297)

BloodPressure                                                               -0.009
                                                                            (0.009)

Constant                 -1.177*** -1.963*** -5.730*** -9.122*** -9.393*** -9.044***
                          (0.123)   (0.241)   (0.540)   (0.715)   (0.734)   (0.803)

-------------------------------------------------------------------------------------
Observations                768       768       768       768       768       768
Log Likelihood           -478.105  -470.549  -430.165  -361.777  -357.261  -356.736
Akaike Inf. Crit.         960.210   947.098   868.330   733.554   726.521   727.473
=====================================================================================
Note:                                                 *p<0.05; **p<0.01; ***p<0.001
```

```
## For l2 model
pseudoR2=(l6$null.deviance-l6$deviance)/l6$null.deviance
print(paste("pseudo_R2 for model 6",pseudoR2))
```

```
[1] "pseudo_R2 for model 6 0.281847690693421"
```

```
fm6a=maBina(l6, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, fm3a, fm4a, fm5a, fm6a,se=list(NULL, NULL, NULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4', 'logit-5', 'logit-
6'),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Marginal Effects
=====================================================================================
                                              Dependent variable:
                         ------------------------------------------------------------
                                                    Outcome
                          logit-1   logit-2   logit-3   logit-4   logit-5   logit-6
                            (1)       (2)       (3)       (4)       (5)       (6)
-------------------------------------------------------------------------------------
Pregnancies              0.031***   0.018**   0.020**  0.024***  0.026***  0.026***
                          (0.005)   (0.006)   (0.006)   (0.007)   (0.007)   (0.007)

Age                                0.007***  0.007***   0.002     0.002     0.003
                                    (0.002)   (0.002)   (0.002)   (0.002)   (0.002)

BMI                                          0.024***  0.019***  0.019***  0.020***
                                              (0.003)   (0.003)   (0.003)   (0.003)

Glucose                                                0.008***  0.008***  0.008***
                                                        (0.001)   (0.001)   (0.001)

DiabetesPedigreeFunction                                          0.183**   0.180**
                                                                  (0.062)   (0.062)

BloodPressure                                                               -0.002
                                                                            (0.002)

Constant                 -0.265*** -0.441*** -1.252*** -1.906*** -1.962*** -1.888***
                          (0.024)   (0.050)   (0.110)   (0.137)   (0.142)   (0.158)

-------------------------------------------------------------------------------------
Observations                768       768       768       768       768       768
Akaike Inf. Crit.         960.210   947.098   868.330   733.554   726.521   727.473
=====================================================================================
Note:                                                *p<0.05; **p<0.01; ***p<0.001
```

1. Adding 'BloodPressure' in the model increased log likelihood of the sixth model from -357.261 to -356.736.
2. Omitted variable bias already corrected in Model 5.

3. The AIC(Akaike Inf.Crit) has increased from 726.521 to 727.473, showing that the addition of the new variable is actually puling the model down. The new variable - 'BloodPressure' does not help in explaining our causal relationship.

Pseudo_r2: Pseudo_R2 increased from 0.280 to 0.281. This shows that adding the variabale 'BloodPressure' increased the explanatory power of the model from 28% to 28.1%. We see miniscule change in pseudo_r2 implementing that BloodPResuure does not contribute to effect the chance of getting diabetes if you are pregnant.

Adding variable 'Skin Thickness' in the model

```
l7 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction+BloodPressure+SkinT
hickness, family=binomial, x=TRUE, data=pima)


stargazer(l1, l2, l3,l4,l5,l6, l7, se=list(NULL, NULL, NULL, NULL, NULL,NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", "logit-4", "logit-5", "logit-
6", "logit-7"),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

Logit- Pregnancy and Diabetes

	Dependent variable:						
				Outcome			
	logit-1	logit-2	logit-3	logit-4	logit-5	logit-6	logit-7
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pregnancies	0.137***	0.082**	0.090**	0.116***	0.123***	0.125***	0.125***
	(0.023)	(0.027)	(0.028)	(0.032)	(0.032)	(0.032)	(0.032)
Age		0.030***	0.033***	0.011	0.011	0.013	0.013
		(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
BMI			0.110***	0.093***	0.090***	0.094***	0.096***
			(0.013)	(0.015)	(0.015)	(0.015)	(0.020)
Glucose				0.036***	0.036***	0.036***	0.036***
				(0.004)	(0.004)	(0.004)	(0.004)
DiabetesPedigreeFunction					0.877**	0.864**	0.865**
					(0.295)	(0.297)	(0.297)
BloodPressure						-0.009	-0.009
						(0.009)	(0.009)
SkinThickness							-0.001
							(0.013)

```
Constant                   -1.177*** -1.963*** -5.730*** -9.122*** -9.393*** -9.044*** -9.0
51***
                            (0.123)   (0.241)   (0.540)   (0.715)   (0.734)   (0.803)   (0.
806)

-----------------------------------------------------------------------------------------
-----
Observations                  768       768       768       768       768       768       7
68
Log Likelihood             -478.105  -470.549  -430.165  -361.777  -357.261  -356.736  -35
6.731
Akaike Inf. Crit.           960.210   947.098   868.330   733.554   726.521   727.473   72
9.462
=========================================================================================
=====
Note:                                                          *p<0.05; **p<0.01; ***p<
0.001
```

```
## For l2 model
pseudoR2=(l7$null.deviance-l7$deviance)/l7$null.deviance
print(paste("pseudo_R2 for model 7",pseudoR2))
```

```
[1] "pseudo_R2 for model 7 0.28185814829639"
```

```
fm7a=maBina(l7, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, fm3a, fm4a, fm5a, fm6a,fm7a,se=list(NULL, NULL, NULL, NULL, NULL, N
ULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4', 'logit-5', 'logit-
6', 'logit-7'),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Marginal Effects
==========================================================================================
=====
                                                 Dependent variable:
                         -----------------------------------------------------------------
-----
                                                                   Outcome
                          logit-1   logit-2   logit-3   logit-4   logit-5   logit-6   log
it-7
                            (1)       (2)       (3)       (4)       (5)       (6)
(7)
------------------------------------------------------------------------------------------
-----
Pregnancies              0.031***   0.018**   0.020**  0.024***  0.026***  0.026***  0.02
6***
                          (0.005)   (0.006)   (0.006)   (0.007)   (0.007)   (0.007)   (0.
007)

Age                                0.007***  0.007***    0.002     0.002     0.003     0.
003
                                    (0.002)   (0.002)   (0.002)   (0.002)   (0.002)   (0.
002)

BMI                                          0.024***  0.019***  0.019***  0.020***  0.02
0***
                                              (0.003)   (0.003)   (0.003)   (0.003)   (0.
004)

Glucose                                                0.008***  0.008***  0.008***  0.00
8***
                                                        (0.001)   (0.001)   (0.001)   (0.
001)

DiabetesPedigreeFunction                                           0.183**   0.180**   0.1
81**
                                                                  (0.062)   (0.062)   (0.
062)

BloodPressure                                                               -0.002    -0.
002
                                                                            (0.002)   (0.
002)

SkinThickness                                                                           0.
000
                                                                                      (0.
003)
```

```
Constant                    -0.265*** -0.441*** -1.252*** -1.906*** -1.962*** -1.888*** -1.8
89***
                             (0.024)   (0.050)   (0.110)   (0.137)   (0.142)   (0.158)   (0.
159)

-----------------------------------------------------------------------------------------
-----
Observations                  768       768       768       768       768       768       7
68
Akaike Inf. Crit.            960.210   947.098   868.330   733.554   726.521   727.473   72
9.462
=========================================================================================
=====
Note:                                                          *p<0.05; **p<0.01; ***p<
0.001
```

1. Adding 'SkinThickness' in the model increased log likelihood of the seventh model from -356.736 to -356.731. It's a trivial change in the model as we see that SkinThickness is not a statistically significant model.
2. Omitted variable bias already corrected in Model 5.
3. The AIC(Akaike Inf.Crit) has increased from 727.473 to 729.462, showing that the addition of the new variable is actually puling the model further down. The new variable - 'SkinThickness' does not as well help in explaining our causal relationship.

Pseudo_r2: Pseudo_R2 did not increase much. This shows that adding the variabale 'SkinThickness' does not effect the chances of getting diabetes if you are pregnant.

Adding variable 'Insulin' in the model

```
l8 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction+BloodPressure+SkinT
hickness+Insulin, family=binomial, x=TRUE, data=pima)



stargazer(l1, l2, l3,l4,l5,l6, l7, l8,se=list(NULL, NULL, NULL, NULL, NULL,NULL, NULL, NU
LL),
          column.labels=c("logit-1", "logit-2", "logit-3", "logit-4", "logit-5", "logit-
6", "logit-7", "logit-8"),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

Logit- Pregnancy and Diabetes

	Dependent variable:							
	Outcome							
	logit-1	logit-2	logit-3	logit-4	logit-5	logit-6	logit-7	logit-8
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pregnancies	0.137***	0.082**	0.090**	0.116***	0.123***	0.125***	0.125***	0.125***
	(0.023)	(0.027)	(0.028)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
Age		0.030***	0.033***	0.011	0.011	0.013	0.013	0.013
		(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)
BMI			0.110***	0.093***	0.090***	0.094***	0.096***	0.095***
			(0.013)	(0.015)	(0.015)	(0.015)	(0.020)	(0.020)
Glucose				0.036***	0.036***	0.036***	0.036***	0.036***
				(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
DiabetesPedigreeFunction					0.877**	0.864**	0.865**	0.861**
					(0.295)	(0.297)	(0.297)	(0.298)
BloodPressure						-0.009	-0.009	-0.009
						(0.009)	(0.009)	(0.009)
SkinThickness							-0.001	-0.001
							(0.013)	(0.013)

```
Insulin
0.0002

(0.001)

Constant                 -1.177*** -1.963*** -5.730*** -9.122*** -9.393*** -9.044*** -9.0
51*** -9.019***
                          (0.123)   (0.241)   (0.540)   (0.715)   (0.734)   (0.803)   (0.
806)   (0.831)

------------------------------------------------------------------------------------------
---------------
Observations                768       768       768       768       768       768       7
68       768
Log Likelihood           -478.105  -470.549  -430.165  -361.777  -357.261  -356.736  -35
6.731  -356.719
Akaike Inf. Crit.         960.210   947.098   868.330   733.554   726.521   727.473   72
9.462   731.438
==========================================================================================
===============
Note:                                                                       *p<0.05; **p<
0.01; ***p<0.001
```

```
## For l2 model
pseudoR2=(l8$null.deviance-l8$deviance)/l8$null.deviance
print(paste("pseudo_R2 for model 8",pseudoR2))
```

```
[1] "pseudo_R2 for model 8 0.281882709782331"
```

```
fm8a=maBina(l8, x.mean=TRUE, rev.dum=TRUE, digits=3)
stargazer(fm1a, fm2a, fm3a, fm4a, fm5a, fm6a,fm7a, fm8a,se=list(NULL, NULL, NULL, NULL, N
ULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4', 'logit-5', 'logit-
6', 'logit-7', 'logit-8'),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

Logit- Marginal Effects

	Dependent variable:							
	Outcome							
	logit-1	logit-2	logit-3	logit-4	logit-5	logit-6	logit-7	logit-8
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pregnancies	0.031***	0.018**	0.020**	0.024***	0.026***	0.026***	0.026***	0.026***
	(0.005)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Age		0.007***	0.007***	0.002	0.002	0.003	0.003	0.003
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
BMI			0.024***	0.019***	0.019***	0.020***	0.020***	0.020***
			(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Glucose				0.008***	0.008***	0.008***	0.008***	0.008***
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
DiabetesPedigreeFunction					0.183**	0.180**	0.181**	0.180**
					(0.062)	(0.062)	(0.062)	(0.062)
BloodPressure						-0.002	-0.002	-0.002
						(0.002)	(0.002)	(0.002)
SkinThickness							0.000	0.000
							(0.003)	(0.003)

```
Insulin
0.000

(0.000)

Constant                   -0.265*** -0.441*** -1.252*** -1.906*** -1.962*** -1.888*** -1.8
89*** -1.882***
                            (0.024)   (0.050)   (0.110)   (0.137)   (0.142)   (0.158)   (0.
159)   (0.165)

-----------------------------------------------------------------------------------------
---------------
Observations                  768       768       768       768       768       768       7
68       768
Akaike Inf. Crit.           960.210   947.098   868.330   733.554   726.521   727.473   72
9.462   731.438
=========================================================================================
===============
Note:                                                                      *p<0.05; **p<
0.01; ***p<0.001
```

1. Adding 'Insulin' in the model increased log likelihood of the eigth model from -356.731 to -356.719. It's a trivial change in the model as we see that Insulin is not a statistically significant model.
2. Omitted variable bias already corrected in Model 5.
3. The AIC(Akaike Inf.Crit) has increased from 729.462 to 731.438 , showing that the addition of the new variable is actually puling the model further down. The new variable - 'Insulin' does not as well help in explaining our causal relationship.

Pseudo_r2: Pseudo_R2 did not increase much. This shows that adding the variabale 'Insulin' does not effect the chances of getting diabetes if you are pregnant.

Best Model:

*The best model is **Model 5.***

- Omitted variable bias has been corrected by this model. The slope coefficient did not change after Model 5. This shows that all the relevant control variables have been accounted for in this model.
- In this model , we achieved E(Beta1(hat))=Beta1 = 0.026***
- Adding more variables after this model did not help in increasing pseudo_r2 much.
- Loglikelihood : *-357.261 and Model 8- -356.719*
- Goodness of fit : AIC value(Akaike Inf. Crit.) - Lower value of AIC means the better the model. *AIC : model 5 - 726.521*

Table for all logit models

```
l1 = glm(Outcome~Pregnancies, family=binomial, x=TRUE,  data=pima)
l2 = glm(Outcome~Pregnancies+Age, family=binomial, x=TRUE,  data=pima)
l3 = glm(Outcome~Pregnancies+Age+BMI, family=binomial, x=TRUE, data=pima)
l4 = glm(Outcome~Pregnancies+Age+BMI+Glucose, family=binomial, x=TRUE, data=pima)
l5 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction, family=binomial, x
=TRUE, data=pima)
l6 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction+BloodPressure, fami
ly=binomial, x=TRUE, data=pima)
l7 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction+BloodPressure+SkinT
hickness, family=binomial, x=TRUE, data=pima)
l8 = glm(Outcome~Pregnancies+Age+BMI+Glucose+DiabetesPedigreeFunction+BloodPressure+SkinT
hickness+Insulin, family=binomial, x=TRUE, data=pima)



stargazer(l1, l2, l3, l4, l5, l6,l7,l8,  se=list(NULL, NULL, NULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4', 'logit-5', 'logit-
6','logit-7', 'logit-8'),
          title='Logit-  Pregnancy and Diabetes', type='text',
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

Logit- Pregnancy and Diabetes

	Dependent variable:							
	Outcome							
	logit-1	logit-2	logit-3	logit-4	logit-5	logit-6	logit-7	logit-8
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pregnancies	0.137***	0.082**	0.090**	0.116***	0.123***	0.125***	0.125***	0.125***
	(0.023)	(0.027)	(0.028)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
Age		0.030***	0.033***	0.011	0.011	0.013	0.013	0.013
		(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)
BMI			0.110***	0.093***	0.090***	0.094***	0.096***	0.095***
			(0.013)	(0.015)	(0.015)	(0.015)	(0.020)	(0.020)
Glucose				0.036***	0.036***	0.036***	0.036***	0.036***
				(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
DiabetesPedigreeFunction					0.877**	0.864**	0.865**	0.861**
					(0.295)	(0.297)	(0.297)	(0.298)
BloodPressure						-0.009	-0.009	-0.009
						(0.009)	(0.009)	(0.009)
SkinThickness							-0.001	-0.001
							(0.013)	(0.013)

```
Insulin
0.0002

(0.001)

Constant                 -1.177*** -1.963*** -5.730*** -9.122*** -9.393*** -9.044*** -9.0
51*** -9.019***
                          (0.123)   (0.241)   (0.540)   (0.715)   (0.734)   (0.803)   (0.
806)   (0.831)

-----------------------------------------------------------------------------------------
---------------
Observations                768       768       768       768       768       768       7
68       768
Log Likelihood           -478.105  -470.549  -430.165  -361.777  -357.261  -356.736  -35
6.731  -356.719
Akaike Inf. Crit.         960.210   947.098   868.330   733.554   726.521   727.473   72
9.462   731.438
=========================================================================================
===============
Note:                                                                    *p<0.05; **p<
0.01; ***p<0.001
```

```
stargazer(fm1a, fm2a, fm3a, fm4a, fm5a, fm6a,fm7a, fm8a,se=list(NULL, NULL, NULL, NULL, N
ULL, NULL, NULL, NULL),
          column.labels=c("logit-1", "logit-2", "logit-3", 'logit-4', 'logit-5', 'logit-
6', 'logit-7', 'logit-8'),
          title="Logit- Marginal Effects", type="text",
          star.cutoffs = c(0.05,0.01,0.001), df=FALSE, digits=3)
```

```
Logit- Marginal Effects
===============================================================================
===============
                                                      Dependent variable:
                         ------------------------------------------------------
---------------
                                                            Outcome
                          logit-1   logit-2   logit-3   logit-4   logit-5   logit-6   log
it-7   logit-8
                            (1)       (2)       (3)       (4)       (5)       (6)
(7)       (8)
-------------------------------------------------------------------------------
---------------
Pregnancies              0.031***   0.018**   0.020**  0.024***  0.026***  0.026***  0.02
6***  0.026***
                          (0.005)   (0.006)   (0.006)   (0.007)   (0.007)   (0.007)   (0.
007)   (0.007)

Age                                0.007***  0.007***   0.002     0.002     0.003     0.
003     0.003
                                    (0.002)   (0.002)   (0.002)   (0.002)   (0.002)   (0.
002)   (0.002)

BMI                                          0.024***  0.019***  0.019***  0.020***  0.02
0***  0.020***
                                              (0.003)   (0.003)   (0.003)   (0.003)   (0.
004)   (0.004)

Glucose                                                0.008***  0.008***  0.008***  0.00
8***  0.008***
                                                        (0.001)   (0.001)   (0.001)   (0.
001)   (0.001)

DiabetesPedigreeFunction                                          0.183**   0.180**   0.1
81**   0.180**
                                                                  (0.062)   (0.062)   (0.
062)   (0.062)

BloodPressure                                                               -0.002    -0.
002    -0.002
                                                                            (0.002)   (0.
002)   (0.002)

SkinThickness                                                                          0.
000     0.000
                                                                                      (0.
003)   (0.003)
```

```
Insulin
0.000

(0.000)

Constant                  -0.265*** -0.441*** -1.252*** -1.906*** -1.962*** -1.888*** -1.8
89*** -1.882***
                           (0.024)   (0.050)   (0.110)   (0.137)   (0.142)   (0.158)   (0.
159)   (0.165)

-----------------------------------------------------------------------------------------
---------------
Observations                 768       768       768       768       768       768       7
68       768
Akaike Inf. Crit.          960.210   947.098   868.330   733.554   726.521   727.473   72
9.462   731.438
=========================================================================================
===============
Note:                                                                       *p<0.05; **p<
0.01; ***p<0.001
```

Interpretation of each model:

- *Logit 1*:
  - Increasing the numnber of pregnancies by 1, will increase the chance of getting diabetes by 0.03%.
- *Logit 2*:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.018%.
  - Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.007%.
- *Logit 3*:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.020%.
  - Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.007%.
  - Keeping all other variables constant, increasing the BMI by 10%, will increase the chance of getting diabetes, on an average, by 0.24%.
- *Logit 4*:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.024%.
  - Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.002%.
  - Keeping all other variables constant, increasing the BMI by 10%, will increase the chance of getting diabetes, on an average, by 0.19%.

- Keeping all other variables constant, increasing glucose level by 10%, will increase the chance of getting diabetes, on an average, by 0.08%.

- **Logit 5**:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.026%.
  - Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.002%.
  - Keeping all other variables constant, increasing the BMI by 10%, will increase the chance of getting diabetes, on an average, by 0.19%.
  - Keeping all other variables constant, increasing glucose level by 10%, will increase the chance of getting diabetes, on an average, by 0.08%.
  - Keeping all other variables constant, increasing DiabetesPedigreeFunction level by 10%, will increase the chance of getting diabetes, on an average, by 1.83%.
- *Logit 6*:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.026%.
  - Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.003%.
  - Keeping all other variables constant, increasing the BMI by 10%, will increase the chance of getting diabetes, on an average, by 0.20%.
  - Keeping all other variables constant, increasing glucose level by 10%, will increase the chance of getting diabetes, on an average, by 0.08%.
  - Keeping all other variables constant, increasing DiabetesPedigreeFunction level by 10%, will increase the chance of getting diabetes, on an average, by 1.80%.
  - Keeping all other variables constant, increasing BloodPressure level by 1 mm Hg, will decrease the chance of getting diabetes, on an average, by 0.002%. This variable is not statistically significant.
- *Logit 7*:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.026%.
  - Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.003%.
  - Keeping all other variables constant, increasing the BMI by 10%, will increase the chance of getting diabetes, on an average, by 0.20%.
  - Keeping all other variables constant, increasing glucose level by 10%, will increase the chance of getting diabetes, on an average, by 0.08%.
  - Keeping all other variables constant, increasing DiabetesPedigreeFunction level by 10%, will increase the chance of getting diabetes, on an average, by 1.81%.
  - Keeping all other variables constant, increasing BloodPressure level by 1 mm Hg, will decrease the chance of getting diabetes, on an average, by 0.002%. This variable is not statistically significant.
  - Keeping all other variables constant, increasing SkinThickness by 1 mm, will not increase the chance of getting diabetes. This variable is not statistically significant.
- *Logit 8*:
  - Keeping all other variables constant, increasing the number of preganancies by 1, will increase the chance of getting diabetes, on an average, by 0.026%.

- Keeping all other variables constant, increasing the age by 1 year, will increase the chance of getting diabetes, on an average, by 0.003%.
- Keeping all other variables constant, increasing the BMI by 10%, will increase the chance of getting diabetes, on an average, by 0.20%.
- Keeping all other variables constant, increasing glucose level by 10%, will increase the chance of getting diabetes, on an average, by 0.08%.
- Keeping all other variables constant, increasing DiabetesPedigreeFunction level by 10%, will increase the chance of getting diabetes, on an average, by 1.80%.
- Keeping all other variables constant, increasing BloodPressure level by 1 mm Hg, will decrease the chance of getting diabetes, on an average, by 0.002%. This variable is not statistically significant.
- Keeping all other variables constant, increasing SkinThickness by 1 mm, will not increase the chance of getting diabetes. This variable is not statistically significant.
- Keeping all other variables constant, increasing Insulin by 1 mm U/ml, will not increase the chance of getting diabetes. This variable is not statistically significant.

Note : *Though Age, Insulin, Skinthickness and BloodPressure are also major causes for diabetes, we see that these variables are not statistically significant in our models because of the multicolinearity that exists between: Glucose-Insulin, Pregnancies-Age, SkinThickness-BMI.*

Wald and Chi-Square Test

Chi-Square, DF and Pr > ChiSq –

- The null hypothesis is that all of the regression coefficients in the model are equal to zero.
- The small p-value from the all three tests would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero.

```
library(car)

Anova(l5, type="II", test="Wald")
```

	Df <dbl>	Chisq <dbl>	Pr(>Chisq) <dbl>
Pregnancies	1	14.579660	1.343570e-04
Age	1	1.345819	2.460103e-01
BMI	1	37.140318	1.099272e-09
Glucose	1	103.042887	3.279710e-24
DiabetesPedigreeFunction	1	8.817016	2.984343e-03

5 rows

- Although the Age variable is insignificant in this model, we include age as it is one of the major causes of the person getting diabetes and Age and Pregnancies are also related.

```
anova(l5,
      update(l5, ~1),    # update here produces null model for comparison
      test="Chisq")
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	762	714.5214	NA	NA	NA
2	767	993.4839	-5	-278.9625	3.325487e-58

2 rows

- We are testing the probability (PR>ChiSq) of observing a Chi-Square statistic as extreme as, or more so, than the observed one under the null hypothesis; The DF defines the distribution of the Chi-Square test statistics and is defined by the number of predictors in the model.
- Typically, PR>ChiSq is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01.
- The small p-value from the all three tests would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero.

Validity & limitations

- **Threats to Internal Validity:**
  - Omitted variable Bias : Added control variables till the estimate $\beta 1$ was consistent.
  - Misspecification of the functional form : Dependent variable is binary, so we used Logit.
  - Measurement errors : Could be present, not enough information to account for it.
  - Missing data and sample selection : Imputed missing values using KNN imputation methods after checking for outliers in the variables of interest.
- **Threats to External Validity:**
  - Differences in populations : Estimated results for PIMA indian women might not hold true for females of all races.
  - Differences in settings : Diagnostic and treatment procedures may vary.

Conclusion & Recommendation:

This study concluded that Pima women with greater number of pregnancies are at higher risk of getting diabetes. However, we saw that hereditary is also more likely to contribute to early onset of diabetes in the Pimas offspring generation.

Recommendation:

- Early diagnosis of diabetes onset in pregnant PIMA women.
- Continuing genetic research to prevent disease and reduce its complications. Especially the gestational diabetes and control of gestational diabetes to avoid complications of fetal uterine growth.