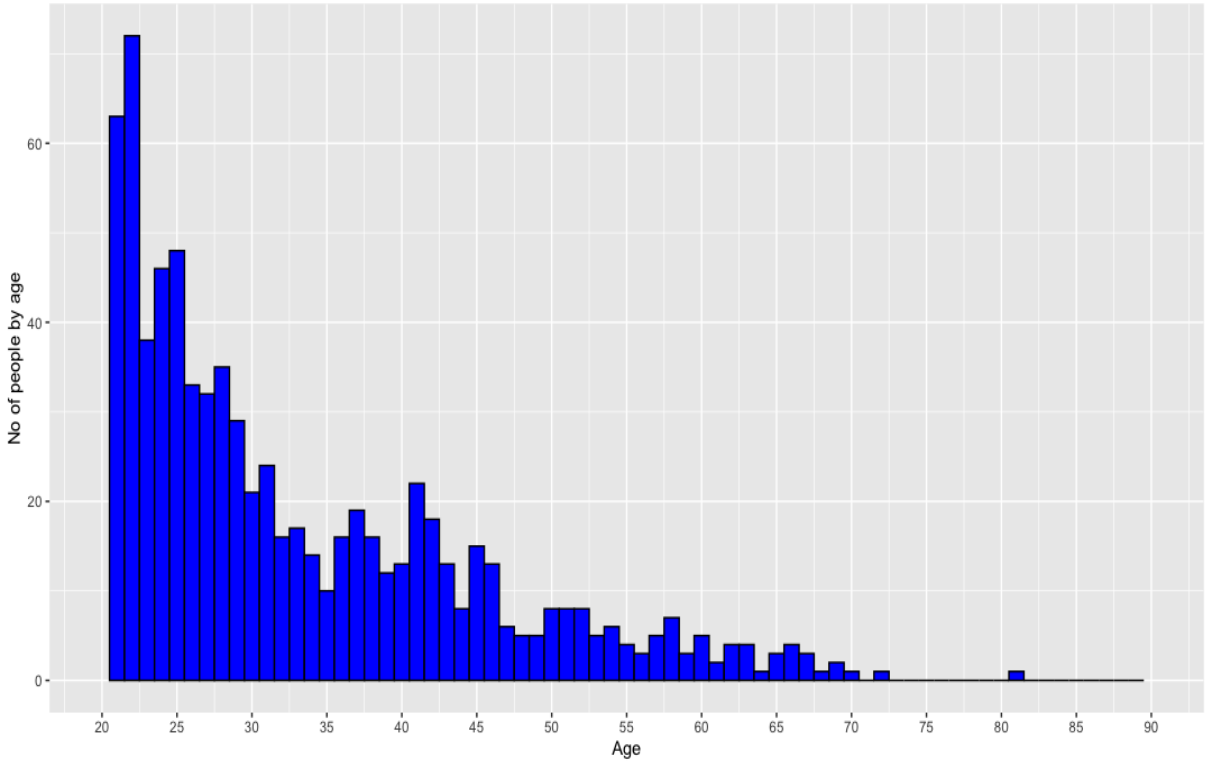
A SECONDARY DATA ANALYSIS ON PIMA INDIAN DIABETES PREDICTION

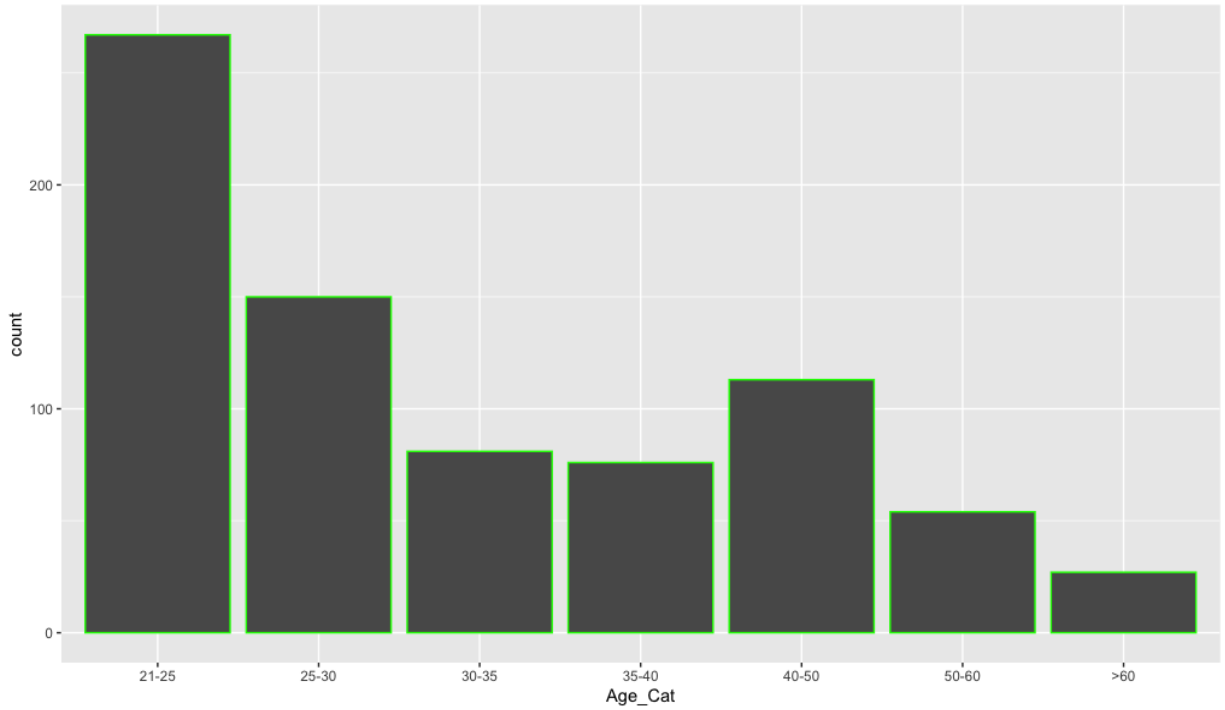
-Priyal Shah

The purpose of the study was that I wanted to understand the effects of various predictor variable on target variable on Diabetes. Diabetes is a condition in which blood sugar or blood glucose is really high leading to type 1 diabetes is a chronic condition in which the body does not make insulin or make very little insulin. The type 2 diabetes is a more common type where your body does not make or use insulin well. Without enough insulin, the glucose stays in your blood for a longer time. One can also be prediabetic that means your blood sugar is higher than normal average population but not high enough to be called diabetes. Being prediabetes, puts you at a higher risk of getting type 2 diabetes. Over 30 million of the total population is diabetic in India. India is a considerably different from that in the Western world in terms of diabetes. Diabetes in India has begun to appear much earlier in life. That means that chronic long-term complications are becoming more common. This research paper intends to analyze the data and create a model on the PIMA Indian Diabetes dataset to predict what triggers diabetes and if a particular observation is at a risk of developing diabetes, taking into consideration all the independent factors.

The dataset can be found online. This data is originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases and this dataset can be used to predict what factors contribute towards being diabetic. I will be using R Studio to perform a secondary analysis on the dataset that has been gathered by the NIDDK.

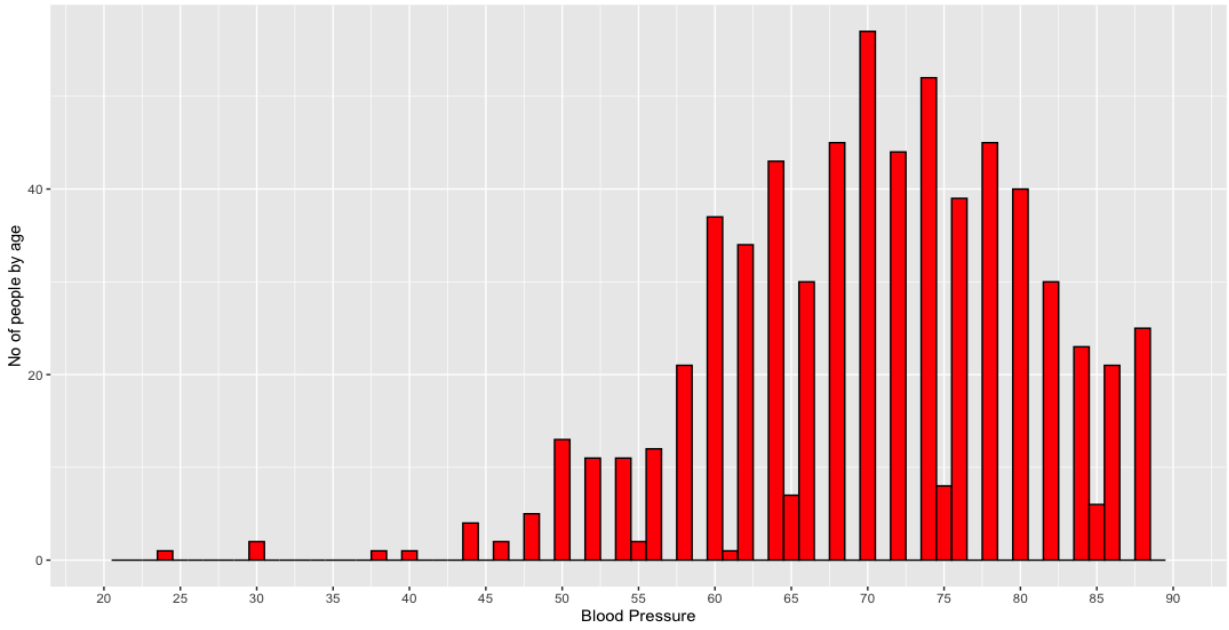
Age





The data on age shows that the mean age is 33.24 whereas the median age is 29. Looking at the graph it says that the population age is not a representative of the same as it not a normal distribution. The graph is positively skewed or is skewed towards the right showing some inconsistency and biasness. For a normal distribution mean should be more or less similar to median and our age dataset shows a that it is not normally distributed. There no abnormality in the age dataset as the minimum age is 21 and the maximum age is 81 which looks quite fair. Our majority of population in the data set falls under the age category of 21-25 followed by 25 to 30. This data analysis will further illustrate or predict how much of the young population in India is at the risk of getting diabetes or has diabetes and what health factors triggers diabetes.

Blood Pressure



The blood pressure dataset shows us that the graph is negatively skewed to confirm this the Shapiro-Wilk normality test and the moments package was used in the R studio. The results showed that the p-value < 0.05 and therefore the data does not follow normal distribution. With maximum blood pressure being 122 and the minimum being 0. The dataset shows incorrect information when it says that the minimum blood pressure of these samples in the dataset is 0. By dropping these observation from the dataset will result in loss of data(35 observation). Therefore, it is best to replace it with the median value.

Shapiro-Wilk normality test
data: db$BloodPressure
W = 0.81892, p-value < 2.2e-16

library("moments")
> skewness(db$BloodPressure)
[1] -1.840005

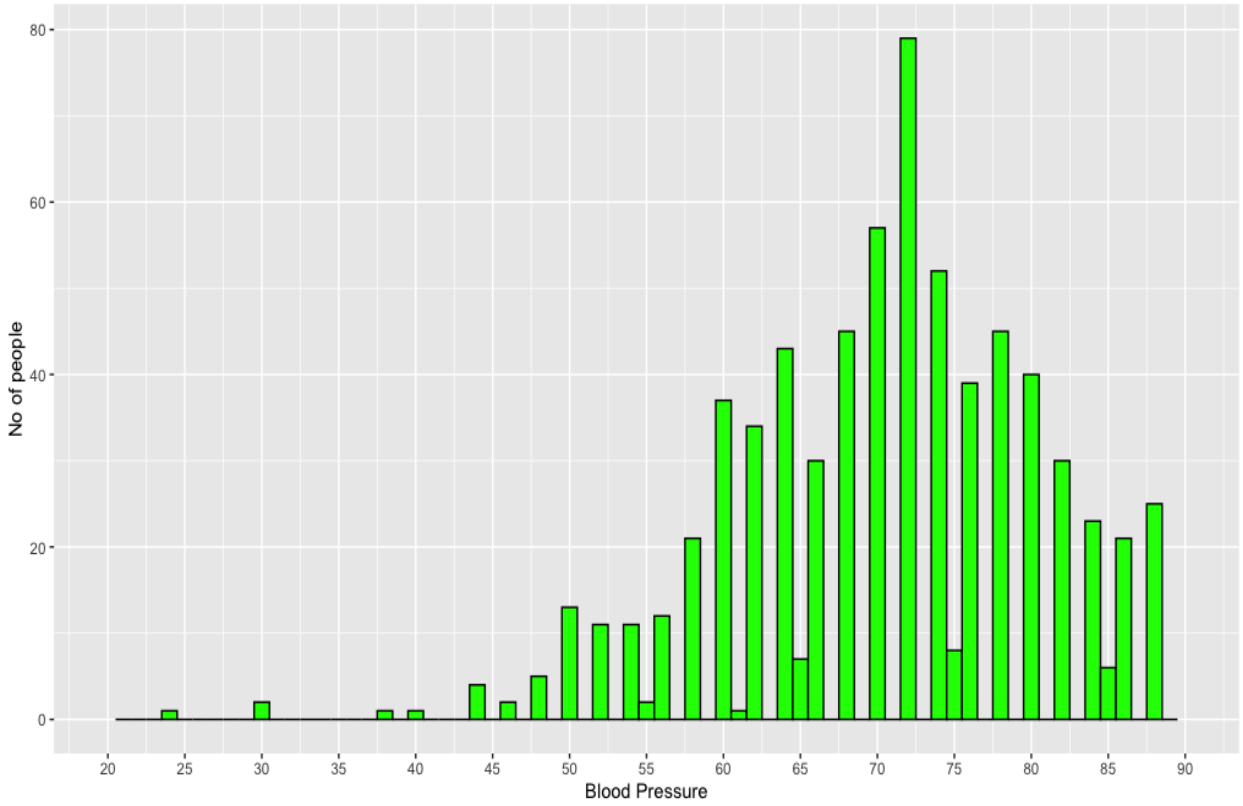
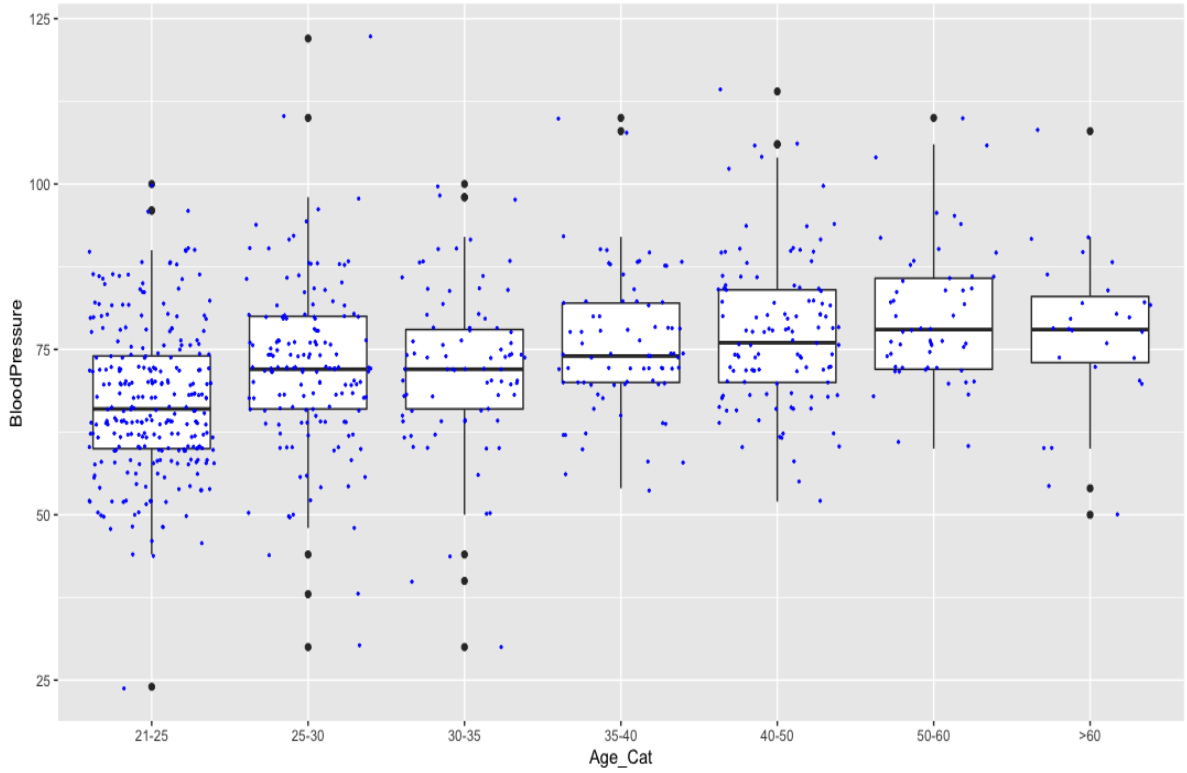


Figure 1 when the minimum value (0) is replaced with the median value (72)

After replacing the missing values with the median values it can be seen that the data is slightly less skewed.

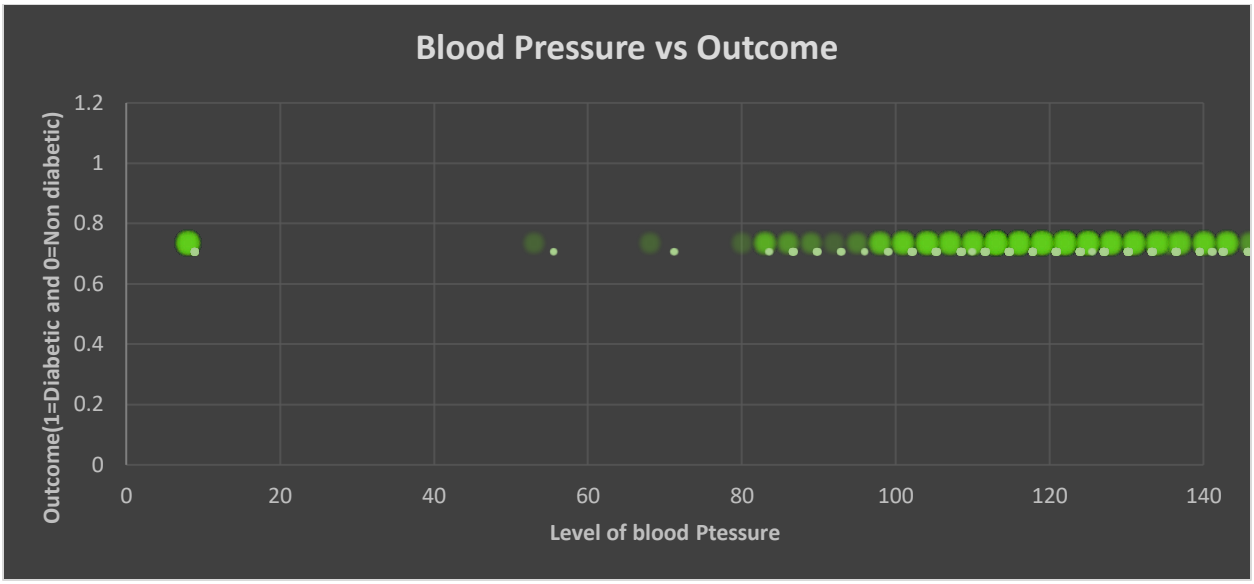
Blood pressure among the population



The box plot above and the table below shows that at the age of 21 to 25 the data is more clustered towards the median having a normal distribution. The majority of people in our sample with the age between 21 to 25 years have a normal blood pressure ranging from 60mmHg to 74mmHg with majority of them having a blood pressure of 66mmHg and few with 100mmHg. As age increases the data starts deviating away from mean showing people with higher age ted to have high blood pressure in India.

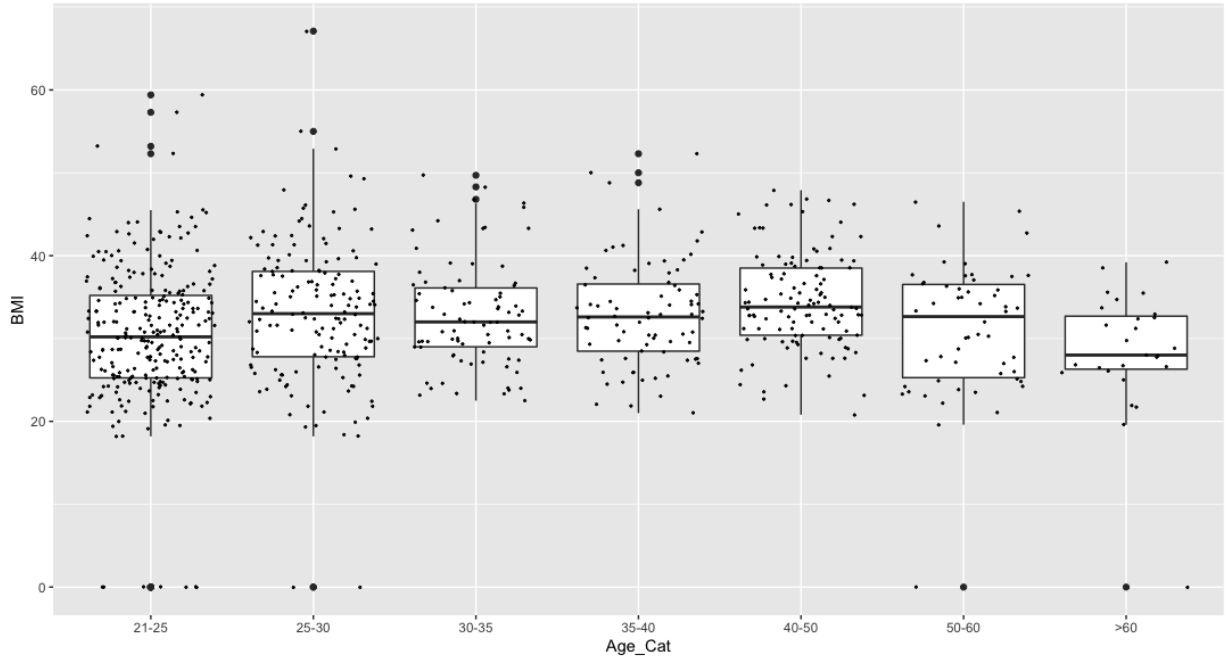
Age	Blood Pressure					
	Min	1st Qu	Median	Mean	3rd Qu	Max
21 to 25	24.00	60.00	66.00	67.32	74.00	100.00
25 to 30	30.00	66.00	72.00	72.33	80.00	122.00
30 to 35	30.00	66.00	72.00	72.37	78.00	100.00
35 to 40	54.00	70.00	74.00	75.62	82.00	110.00
40 to 50	52.00	70.00	76.00	77.34	84.00	114.00
50 to 60	60.00	72.00	78.00	80.06	85.75	110.00
>60	50.0	73.0	78.0	77.7	83.0	108.0

The table shows an interesting relationship between age and blood pressure. A direct relationship between age and blood pressure illustrates that as population age increases the average blood pressure increases. Also, as age increases the median starts deviating from mean showing that the distribution starts to skew.



If we observe the graph, we can see that diabetics generally have high blood pressure as there are many observations that have exceeded 100mmHg. The blood pressure range for diabetics ranges from 40mmHg to 115mmHg and for non-diabetic, the majority of the sample falls below 100mmHg. A |t-stat| of 0.815 against the critical value 2.576 and with a p value of 0.41 shows that blood pressure is not a significant determinant of outcome.

BMI

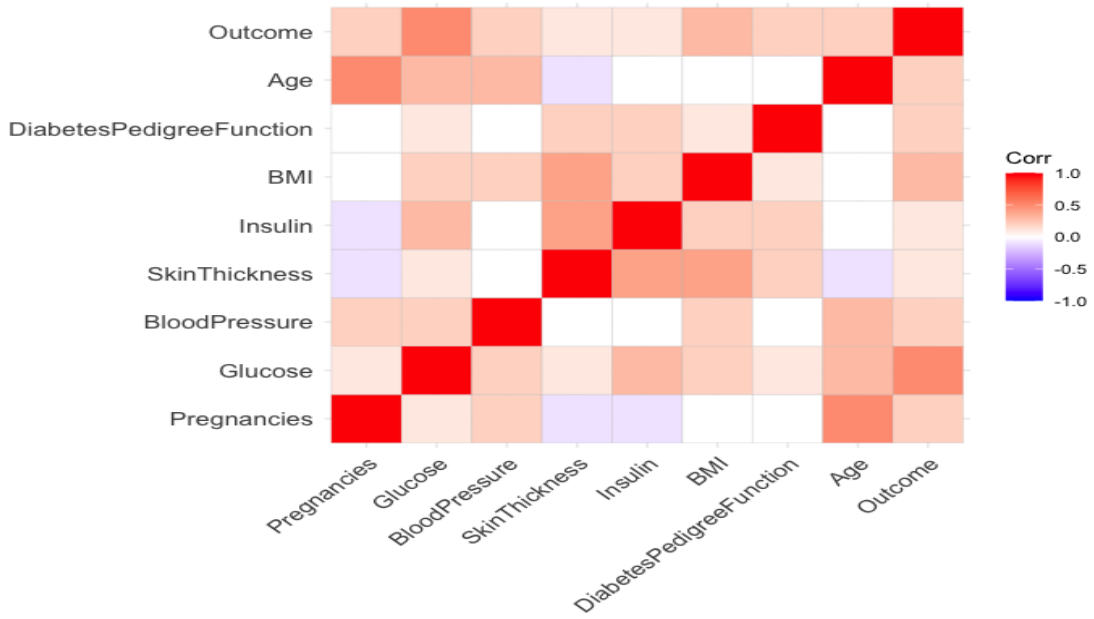


The population with the age of 21 to 25 years that is the young population has a BMI that is clustered toward its median and the majority of the data is falls in the interquartile range is between 25 to 35 but as people start aging the BMI of the elderly moves away from the median and out of the interquartile range. This also sheds light on the proven fact that as you age your lipids do not get converted into fats and this makes you put on weight easily resulting in further deviation from BMI in elderly. A further deviation from median and away from the interquartile range makes the data

skewed leading to inconsistent and biased results. For example, the boxplot for people within 50 to 60 years is negatively skewed with its median close to 75 percentile whereas the boxplot for the population with age greater than 60 is positively skewed with it's median more towards the 25 percentile.

Age_Cat	BMI					
	Min	1st Qu	Median	Mean	3rd Qu	Max
21 to 25	0.00	27.80	33.00	33.04	38.10	67.10
25 to 30	0.00	27.80	33.00	33.04	38.10	67.10
30 to 35	22.50	29.00	32.00	32.81	36.10	49.70
35 to 40	21.00	28.48	32.60	32.97	36.58	52.30
40 to 50	20.8	30.4	33.8	34.5	38.5	47.9
50 to 60	0.00	25.27	32.65	31.11	36.52	46.50
>60	0.0	26.3	28.0	28.4	32.7	39.2

Correlation Matrix



The correlation matrix shows a significant correlation between Age and Pregnancies to confirm the correlation Person's test was used. The correlation coefficient(r) was 0.544 and by the rule of thumb if r is greater than 0.70 multicollinearity might be observed. **Thus, there is no significant multicollinearity observed.**

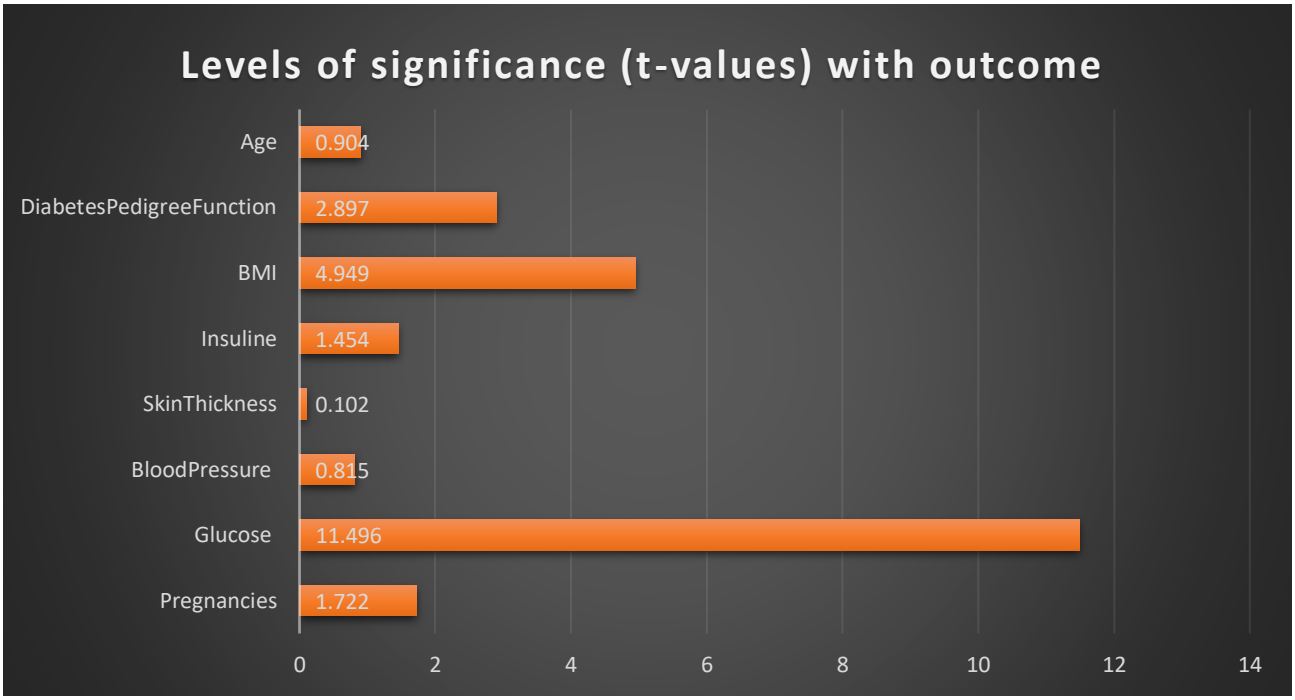
```
> cor.test(db$Age,db$Pregnancies, method="pearson")

        Pearson's product-moment correlation

data:  db$Age and db$Pregnancies
t = 17.959, df = 766, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4925652 0.5922775
sample estimates:
      cor
0.5443412

>
```

The significance table

The significance table shows weather number of pregnancies, levels of glucose, levels of blood pressure, skintickness, insuline, BMI, diabetes pedigree function and age are significant determinant of outcome. The dataset mentions outcome in form of a dummy variable 0 and 1 where 0 = non-diabetic and 1= diabetic. The highlighted variables are significant determinant of the outcome.

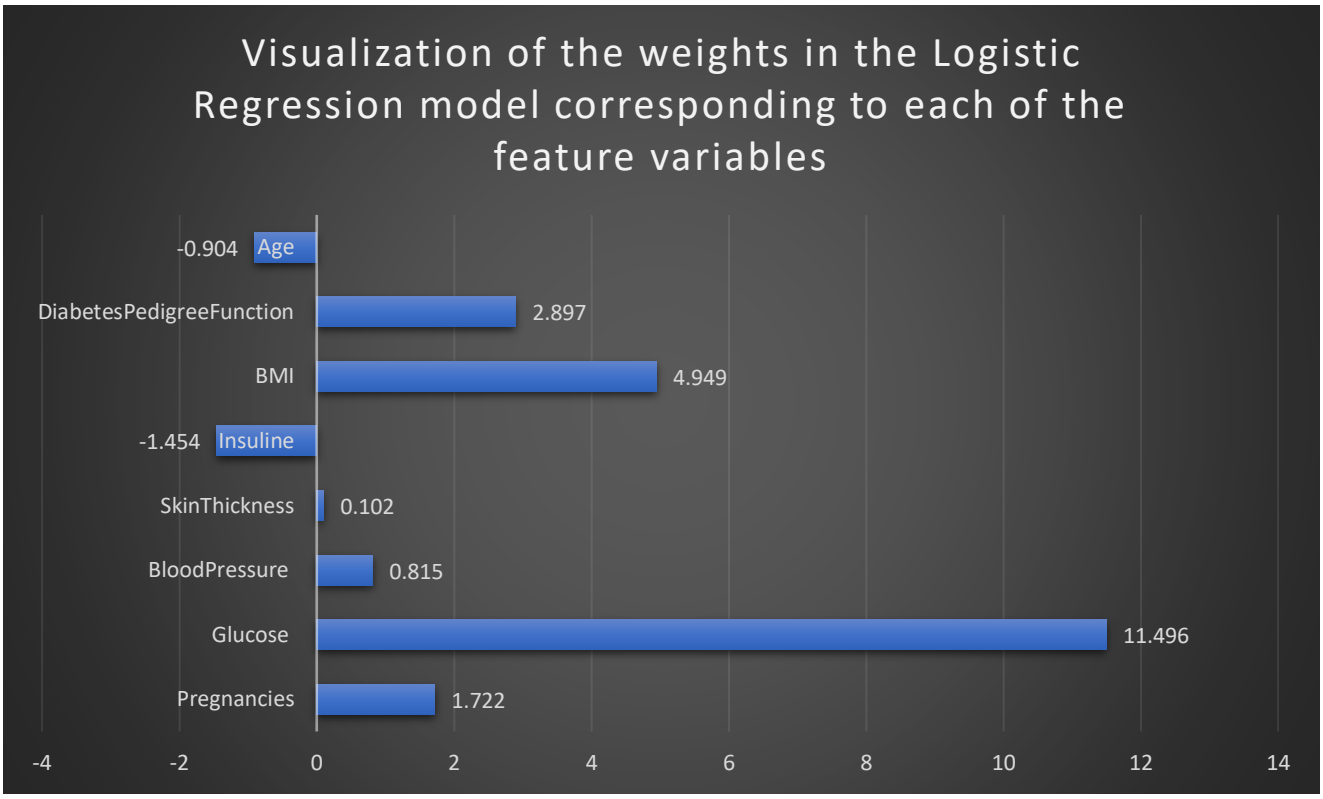
Column1	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-0.6439462	0.1952358	-3.298	0.00102 **
Pregnancies	0.009706	0.005635	1.722	0.0854
Glucose	0.005948	0.0005174	**11.496**	**< 2e-16 *** **
BloodPressure	-0.0010727	0.0013159	-0.815	0.41523
SkinThickness	0.0001118	0.0010978	0.102	0.91888
Insuline	-0.0002209	0.0001519	-1.454	0.14627
BMI	0.0105188	0.0021255	**4.949**	**9.21e-07 *** **
DiabetesPedigreeFunction	0.1313419	0.0453395	**2.897**	**0.00388 ** **
Age	-0.0066693	0.0073816	-0.904	0.36654
Age_Cat25-30	0.0762076	0.0556298	1.37	0.17112
Age_Cat30-35	0.2550904	0.0891347	**2.862**	**0.00433 ** **
Age_Cat35-40	0.2202671	0.1232195	**1.788**	**0.07424 . **
Age_Cat40-50	0.3657313	0.1656971	**2.207**	**0.02760 * **
Age_Cat50-60	0.3970433	0.2436981	1.629	0.10368
Age_Cat>60	0.2093276	0.3257491	0.643	0.52068

T test the level of significance the t values and p values will be used. From the above table it can be seen that **BMI, Glucose, Diabetes Pedigree Function and Age_cat 30-35, Age_cat 35-40, Age_cat 40-50** are significant determinants of outcome at 99% significance level (t*=2.58)



The bar graph shows that the variables like glucose, BMI, Diabetes Pedigree Function, Insuline, Skin thickness, BloodPressure and Pregnancies are all significant determinant of the Outcome.

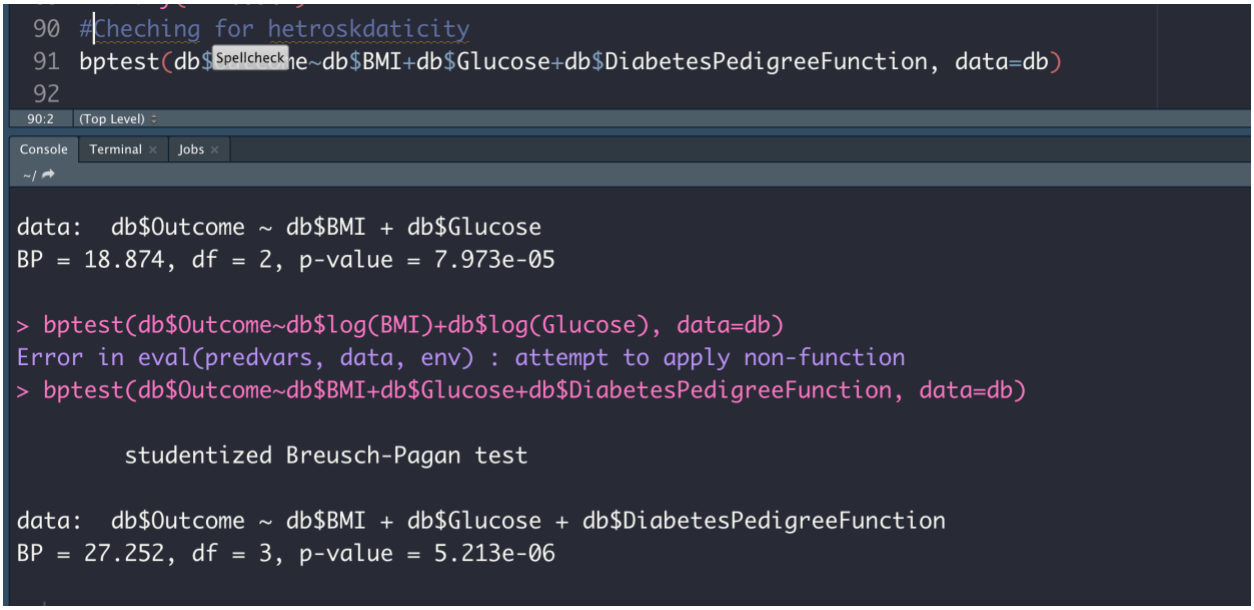
The graphs below talk more about the significant variables and their relationship with the variable outcome



From the above figure, we can draw the following conclusions.

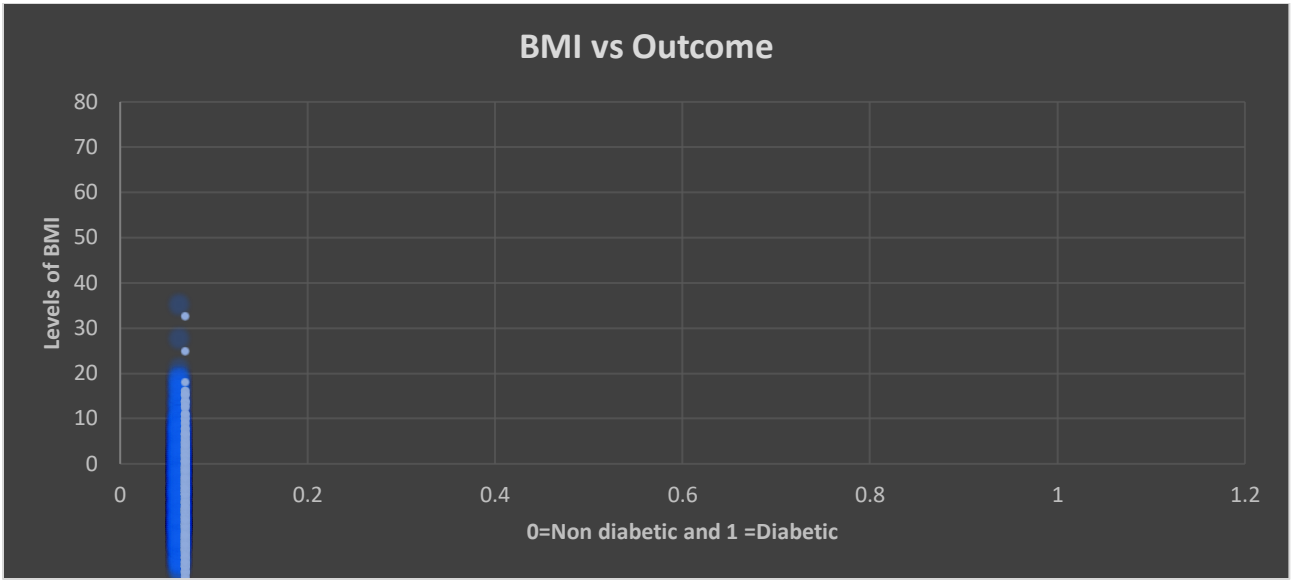
1. Glucose level, BMI, pregnancies and diabetes pedigree function have significant effect on the model, especially glucose level and BMI. It is good to see our model match with real world data.
2. Blood pressure has a negative influence on the outcome, i.e. higher blood pressure is correlated with a person being diabetic. (also, note that blood pressure is more important as a feature than age, because the magnitude is higher for blood pressure).
3. Although age was more correlated than BMI to the output variables the model is based more on BMI.

Checking for heteroskedasticity for between outcome and the significant variables



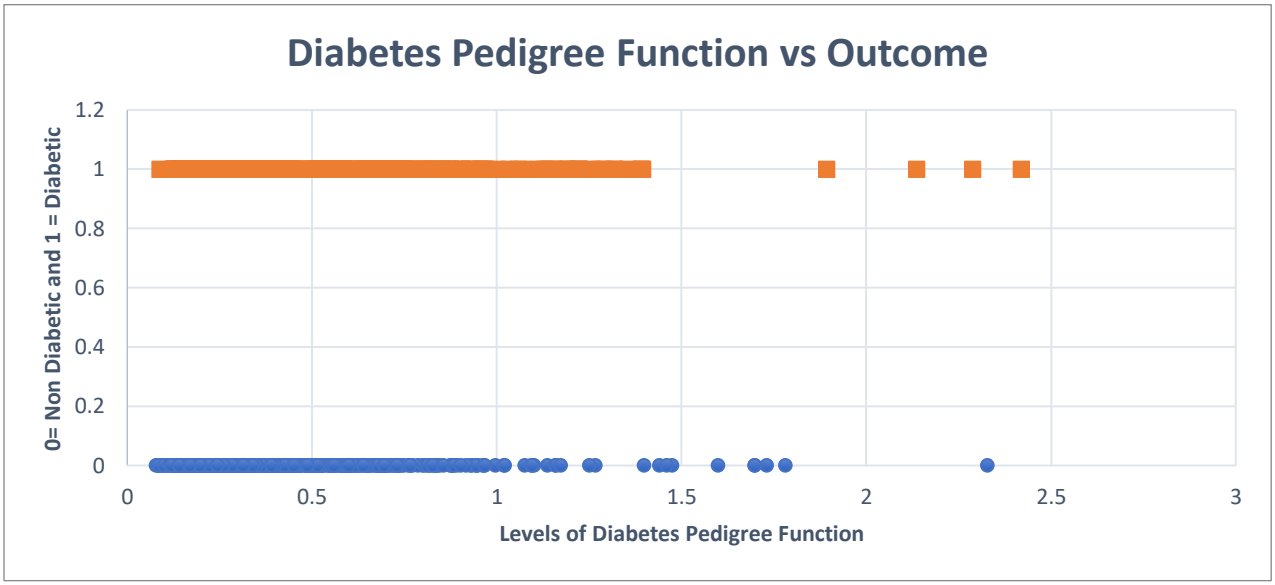
There is a less strong evidence of heteroskedasticity between the dependent and the independent variables. Using a log- log transformation might further reduce it. (I cannot log my variables in regression it shows error.)

BMI VS Outcome



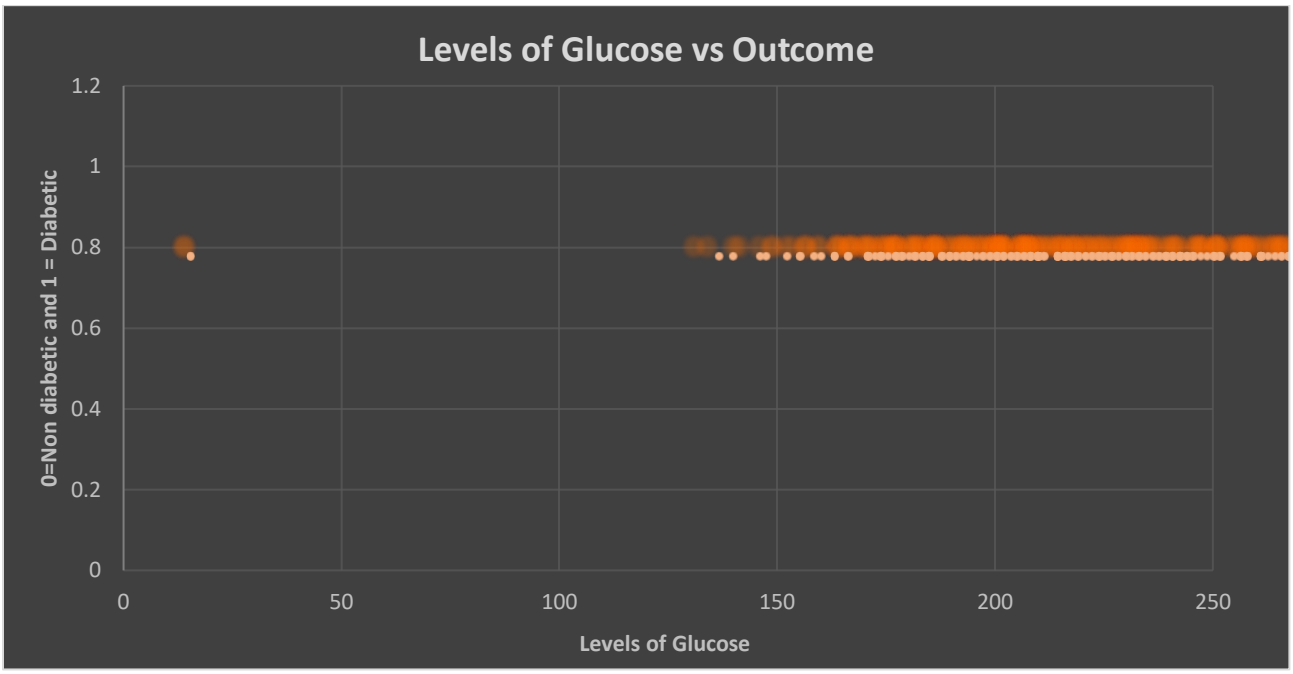
The graph above shows that people with lower BMI level that is from 15 to 50 are less prone to being diabetic whereas people with higher level of BMI 20 to 70 are more prone to being diabetes. The data also shows that as age increases BMI also increases and that leads to being at risk of getting diabetes. The data shows a positive correlation between the BMI and diabetes and also illustrates that people with diabetes generally have a higher BMI.

Diabetes Pedigree Function VS Outcome

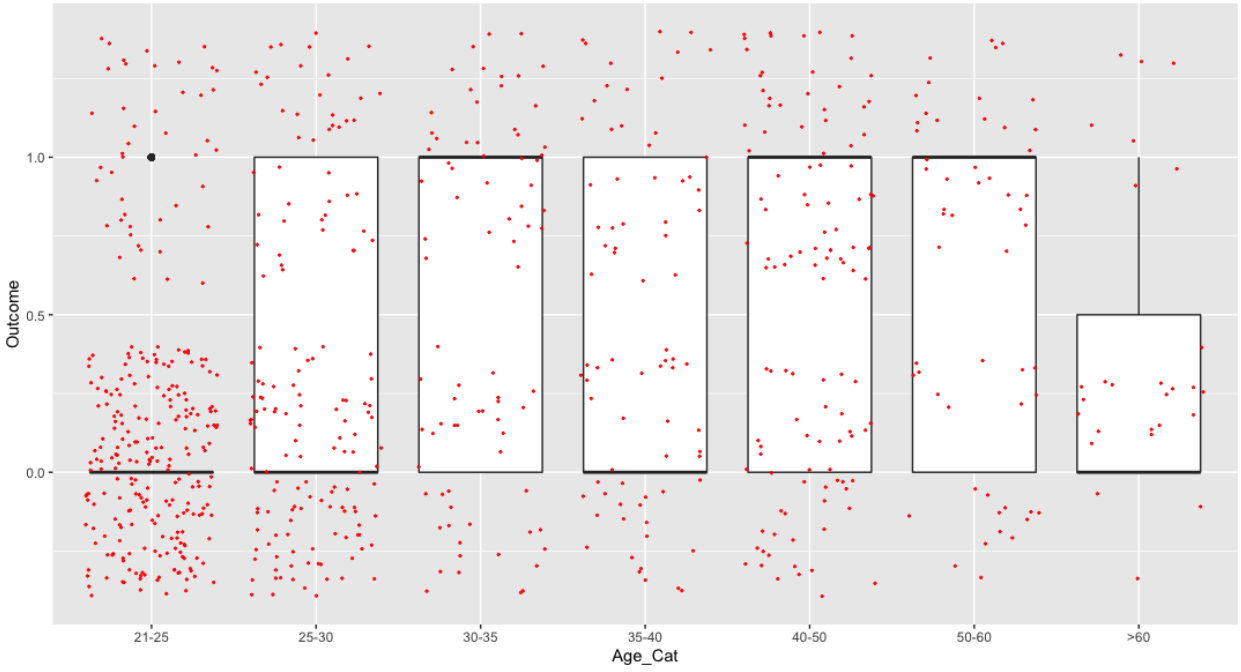


The Diabetes Pedigree Function is a significant determinant of whether or not you will be diabetic. The graph above shows that diabetics seems to have a higher diabetes pedigree function as compared to the non-diabetic. People with higher levels of Diabetes Pedigree Function ae more likely to be

diabetic. With the data that we have we can see that the majority of the diabetic have a diabetes pedigree function ranging from 0 to 1.4. Also, there are few outliers to the data that shew that Diabetes Pedigree function can go up to 2.5 in case of diabetics



Level of glucose is a significant determinant of whether you are diabetic or not. The p value and the t-stats have shown that glucose is a significant determent of outcome. A lower p value shows that there are less chances of data being random. People with diabetes generally have a high level of glucose as compared to people without diabetes. The levels of glucose in diabetics ranges from 55 to 200



Age_Cat	BMI					
	Min	1st Qu	Median	Mean	3rd Qu	Max
21 to 25	0.0000	0.0000	0.0000	0.1685	0.0000	1.0000
25 to 30	0.0	0.0	0.0	0.3	1.0	1.0
30 to 35	0.0000	0.0000	1.0000	0.5062	1.0000	1.0000
35 to 40	0.0000	0.0000	0.0000	0.4605	1.0000	1.0000
40 to 50	0.0000	0.0000	1.0000	0.5664	1.0000	1.0000
50 to 60	0.0000	0.0000	1.0000	0.5741	1.0000	1.0000
>60	0.0000	0.0000	0.0000	0.2593	0.5000	1.0000

The boxplot and the chart above shows that the minimum value is equal to the 1st quartile and the maximum is equal to 3rd quartile and thus these boxplots do not have whiskers. Since the minimum, 1st quartile, median and 3rd quartile of the age group 21 to 25 is 0 therefore the boxplot is a flat line. The majority of population with the age 21 to 25 are healthy or non-diabetic this is shown by a cluster of data towards the median with few outliers towards one. As your age increases the data gets cluttered towards the maximum value this is not been evidently because the data set has more population with the age between 21 to 25.

```
> summary(db$Age_Cat)
  <21 21-25 25-30 30-35 35-40 40-50 50-60   >60
    0   267   150    81    76   113    54    27
```

Conclusion

One of the most important factors that determines the onset of diabetes is Glucose. The other major actors that contribute towards diabetes are BMI and Age. Other factors like Diabetes Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness and Insulin also contributes to the prediction of diabetes but their contribution is not much.

As we can see, the results derived from regression makes sense as one of the first things that actually is contributes towards a high-risk diabetic patient is the Glucose level. An increased BMI might also contribute towards a risk of developing Type II Diabetes. Normally, especially in case of Type II Diabetes, there is a high risk of developing as the age of a person increases (given other factors).

Information of the variables used:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/(height in m)2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- **Age:** Age (years)
- **Outcome:** Class variable (0 if non-diabetic, 1 if diabetic)

Reference

https://www.kaggle.com/ was used to extract the dataset.

Code to the Analysis

```r
db<- read.csv(file.choose())
head(db)
tail(db)
summary(db)
str(db)
#Categorizing age
db$Age_Cat <- ifelse(db$Age < 21, "<21",
                     ifelse((db$Age>=21) & (db$Age<=25), "21-25",
                            ifelse((db$Age>25) & (db$Age<=30), "25-30",
                                   ifelse((db$Age>30) & (db$Age<=35), "30-35",
                                          ifelse((db$Age>35) & (db$Age<=40), "35-40",
                                                 ifelse((db$Age>40) & (db$Age<=50), "40-50",
                                                        ifelse((db$Age>50) & (db$Age<=60), "50-60",">60")))))))
db$Age_Cat <- factor(db$Age_Cat, levels = c('<21','21-25','25-30','30-35','35-40','40-50','50-60','>60'))
table(db$Age_Cat)

install.packages("ggplot2")
library(ggplot2)
#graphing age
ggplot(aes(x = Age), data=db) +
  geom_histogram(binwidth=1, color='black', fill = "blue") +
  scale_x_continuous(limits=c(20,90), breaks=seq(20,90,5)) +
  xlab("Age") +
  ylab("No of people by age")

#graphing age categorically
ggplot(aes(x = Age_Cat), data = db) +
  geom_bar(col='green')
min(db$Age)
max(db$Age)
```

```r
#graphing bloodpressure
ggplot(aes(x =BloodPressure), data=db) +
  geom_histogram(binwidth=1, color='black', fill = "red") +
  scale_x_continuous(limits=c(20,90), breaks=seq(20,90,5)) +
  xlab("Blood Pressure") +
  ylab("No of people by age")
#testing normality using Shapiro-Wilk normality test
shapiro.test(db$BloodPressure)
#using the moments package to find the skewness
install.packages("moments")
library("moments")
skewness(db$BloodPressure)
#understanding the min and max level of bp
min(db$BloodPressure)
max(db$BloodPressure)
#using table to understand the number of people with 0 bloodpressure
table(db$BloodPressure)
#replacing the 0 in the bloodpressurs with the median value
db$BloodPressure[db$BloodPressure==0]<- 72
table(db$BloodPressure)
#graphing blood pressure with 0 replaced to the median value
ggplot(aes(x =BloodPressure), data=db) +
  geom_histogram(binwidth=1, color='black', fill = "green") +
  scale_x_continuous(limits=c(20,90), breaks=seq(20,90,5)) +
  xlab("Blood Pressure") +
  ylab("No of people")
#Age vs Blood Pressure
ggplot(aes(x=Age_Cat, y = BloodPressure), data = db) +
  geom_boxplot()+geom_jitter(color="blue", size=0.4, alpha=0.9)
#Summary of Blood Pressure and Age Categorically
by(db$BloodPressure, db$Age_Cat, summary)
#Age vs BMI
ggplot(aes(x=Age_Cat, y = BMI), data = db) +
  geom_boxplot()+geom_jitter(color="black", size=0.4, alpha=0.9)
#Summary of BMI and Age Categorically
by(db$BMI, db$Age_Cat, summary)
#Correlation between diffrent variables
db_cor <- round(cor(db[1:9]),1)
db_cor
#plotting a correlation matrix
install.packages("ggcorrplot")
library("ggcorrplot")
ggcorrplot(db_cor)
#regression of outcome on all x values
Regression1<-(lm(db$Outcome~db$Pregnancies+db$Glucose+db$BloodPressure+db$SkinThickness+db$Insulin+db$BMI+db$DiabetesPedigreeFunctio
summary(Regression1)
#plotting age_cat and outcome
ggplot(aes(x=Age_Cat, y = Outcome), data = db) +
geom_boxplot()+geom_jitter(color="red", size=0.4, alpha=0.9)
```