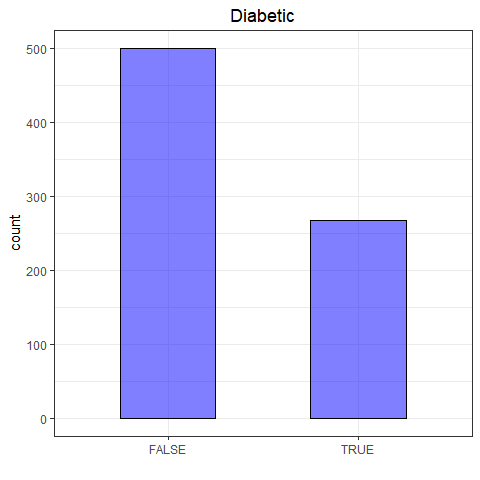
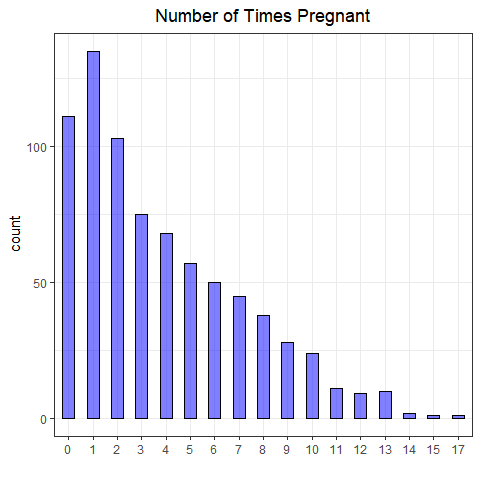
Exploratory Data Analysis - Pima Data Set

27th March 2018

# **Univariate Analysis**

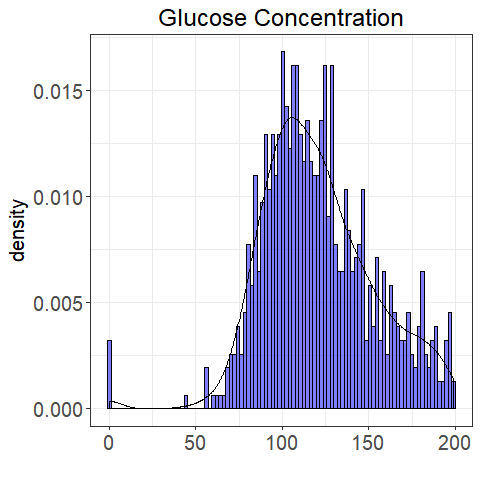


A smaller percentage (35%) of the pregnant women tested were diagnosed with diabetes.



## Count Percentage  
## 1 135 18  
## 0 111 14  
## 2 103 13  
## 3 75 10  
## 4 68 9  
## 5 57 7  
## 6 50 7  
## 7 45 6  
## 8 38 5  
## 9 28 4  
## 10 24 3  
## 11 11 1  
## 13 10 1  
## 12 9 1  
## 14 2 0  
## 15 1 0  
## 17 1 0  
## NA 0 0

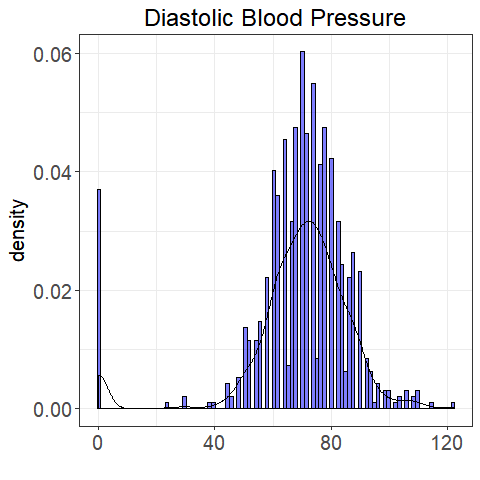
55% of the women had 3 or fewer children.



Glucose Concentration

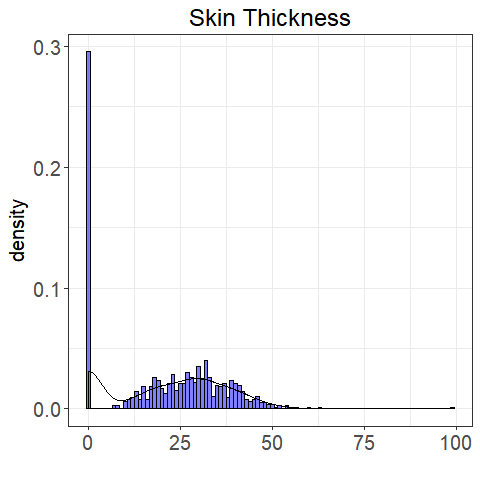
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 99 | 117 | 140 | 199 | 121 | 32 | 768 | 0 |

The maximum glucose concentration was 199 units



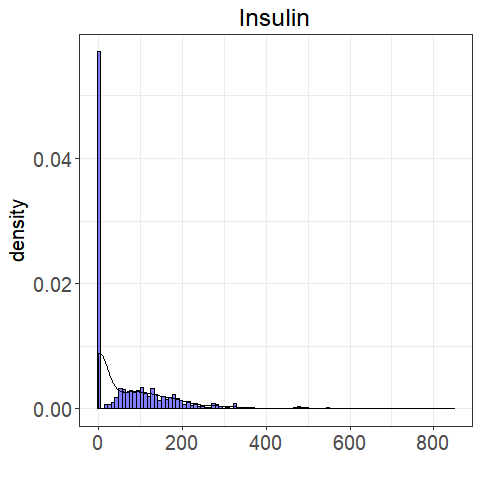
Diastolic Blood Pressure

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 62 | 72 | 80 | 122 | 69 | 19 | 768 | 0 |
|  |  |  |  |  |  |  |  |  |  |



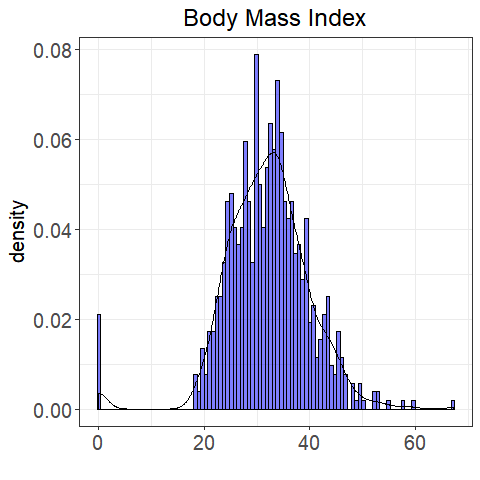
Skin Thickness

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 0 | 23 | 32 | 99 | 21 | 16 | 768 | 0 |
|  |  |  |  |  |  |  |  |  |  |



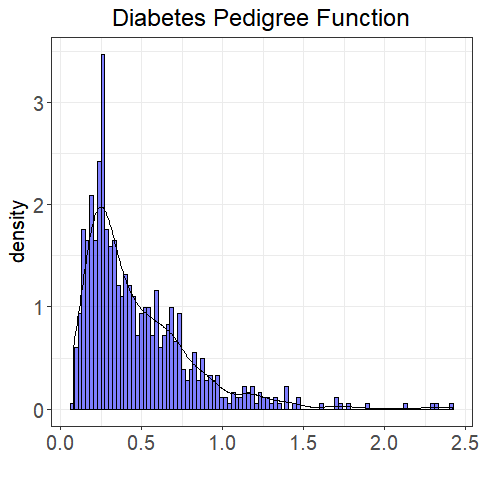
Insulin

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 0 | 30 | 127 | 846 | 80 | 115 | 768 | 0 |
|  |  |  |  |  |  |  |  |  |  |



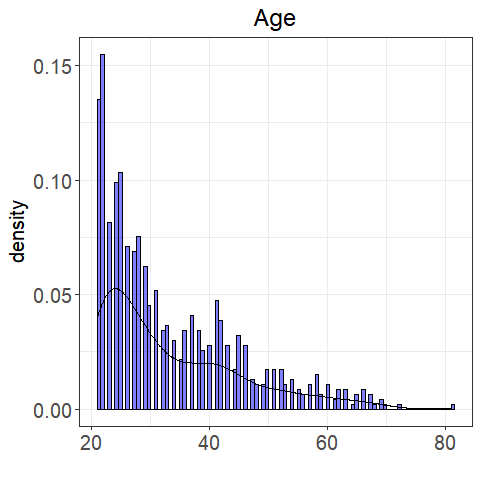
Body Mass Index

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 27 | 32 | 37 | 67 | 32 | 8 | 768 | 0 |



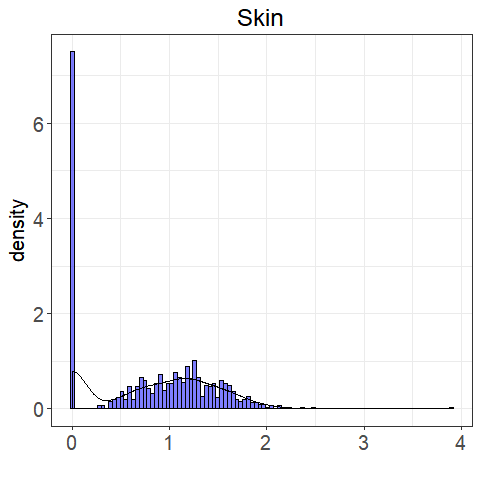
Diabetes Pedigree Function

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 768 | 0 |



Age

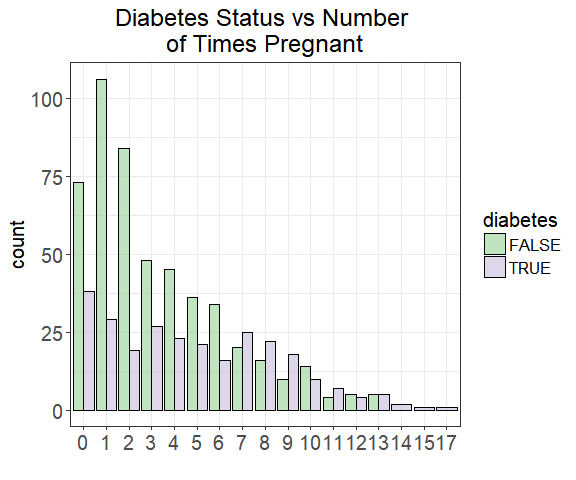
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 21 | 24 | 29 | 41 | 81 | 33 | 12 | 768 | 0 |



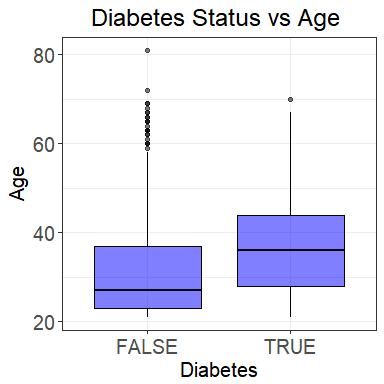
Skin

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|  | 0 | 0 | 1 | 1 | 4 | 1 | 1 | 768 | 0 |

# **Bivariate Analysis**

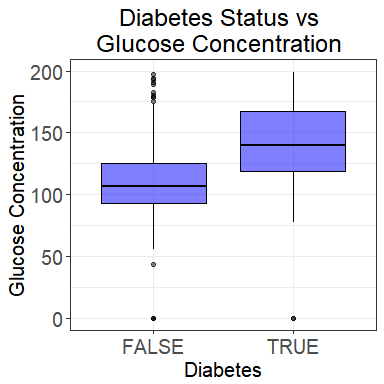


##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 768   
##   
##   
## | Diabetes Status   
## Pregnancies | FALSE | TRUE | Row Total |   
## -------------|-----------|-----------|-----------|  
## 0 | 73 | 38 | 111 |   
## -------------|-----------|-----------|-----------|  
## 1 | 106 | 29 | 135 |   
## -------------|-----------|-----------|-----------|  
## 2 | 84 | 19 | 103 |   
## -------------|-----------|-----------|-----------|  
## 3 | 48 | 27 | 75 |   
## -------------|-----------|-----------|-----------|  
## 4 | 45 | 23 | 68 |   
## -------------|-----------|-----------|-----------|  
## 5 | 36 | 21 | 57 |   
## -------------|-----------|-----------|-----------|  
## 6 | 34 | 16 | 50 |   
## -------------|-----------|-----------|-----------|  
## 7 | 20 | 25 | 45 |   
## -------------|-----------|-----------|-----------|  
## 8 | 16 | 22 | 38 |   
## -------------|-----------|-----------|-----------|  
## 9 | 10 | 18 | 28 |   
## -------------|-----------|-----------|-----------|  
## 10 | 14 | 10 | 24 |   
## -------------|-----------|-----------|-----------|  
## 11 | 4 | 7 | 11 |   
## -------------|-----------|-----------|-----------|  
## 12 | 5 | 4 | 9 |   
## -------------|-----------|-----------|-----------|  
## 13 | 5 | 5 | 10 |   
## -------------|-----------|-----------|-----------|  
## 14 | 0 | 2 | 2 |   
## -------------|-----------|-----------|-----------|  
## 15 | 0 | 1 | 1 |   
## -------------|-----------|-----------|-----------|  
## 17 | 0 | 1 | 1 |   
## -------------|-----------|-----------|-----------|  
## Column Total | 500 | 268 | 768 |   
## -------------|-----------|-----------|-----------|  
##   
##

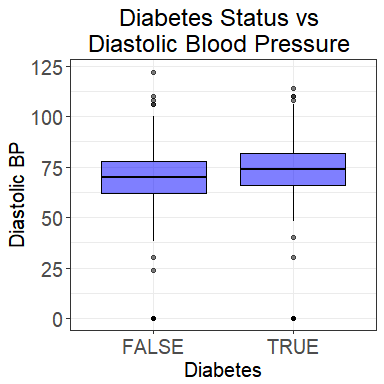


## Group.1 age  
## 1 FALSE 31.19000  
## 2 TRUE 37.06716

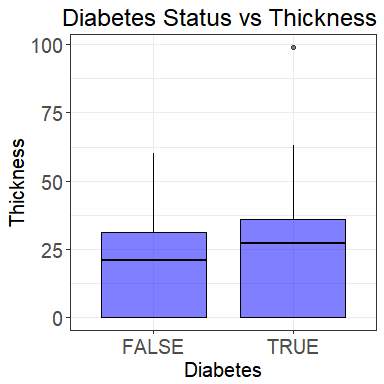
On average, older women (women aged 37) were diagnosed with diabetes and those aged 31 were not diabetic



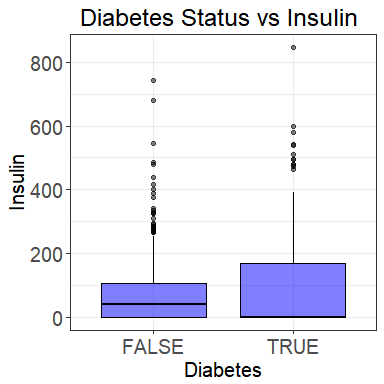
## Group.1 glucose\_conc  
## 1 FALSE 109.9800  
## 2 TRUE 141.2575



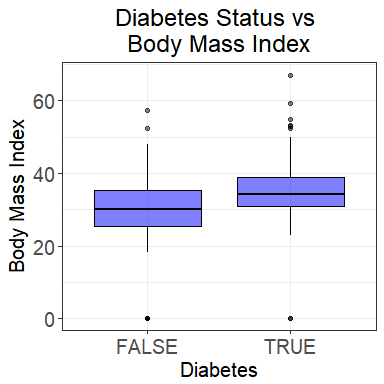
## Group.1 diastolic\_bp  
## 1 FALSE 68.18400  
## 2 TRUE 70.82463



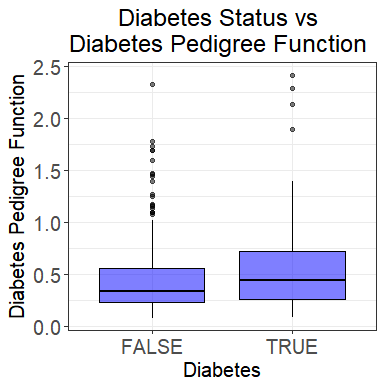
## Group.1 thickness  
## 1 FALSE 19.66400  
## 2 TRUE 22.16418



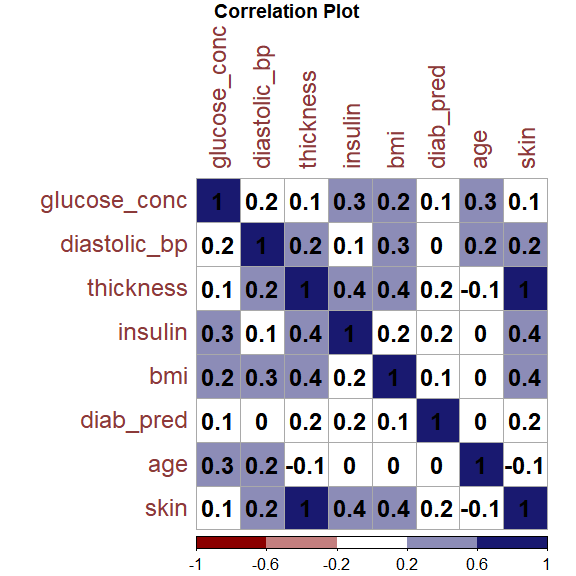
## Group.1 insulin  
## 1 FALSE 68.7920  
## 2 TRUE 100.3358



## Group.1 bmi  
## 1 FALSE 30.30420  
## 2 TRUE 35.14254

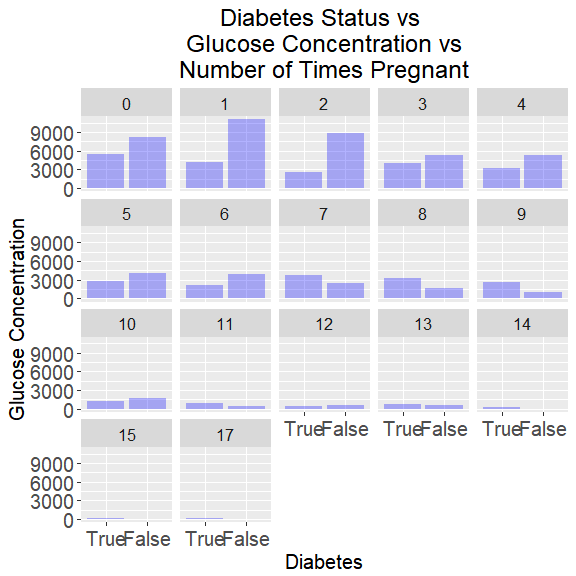


## Group.1 diab\_pred  
## 1 FALSE 0.429734  
## 2 TRUE 0.550500

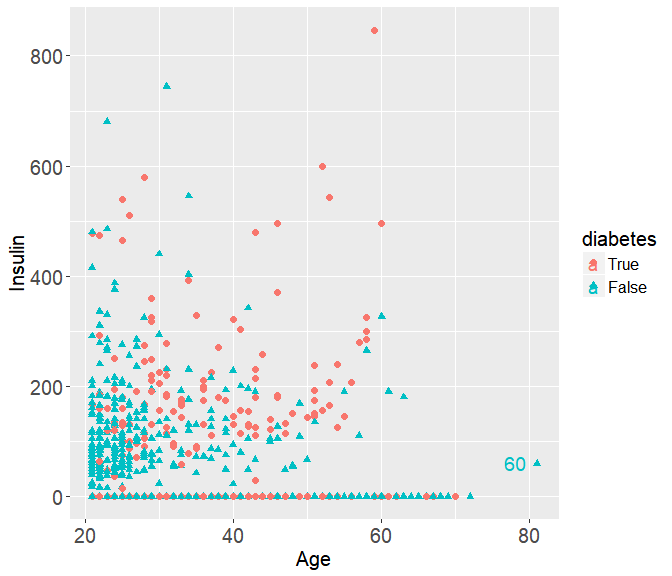


* As the age of the patient increases, their glucose concentration increases. However, the correlation is a weak positive correlation
* Insulin level, body mass index and diabetes pedigree function have no correlation with the age of the patient
* None of the variables have a strong correlation with each other except thickness and skin that are perfectly positively correlated. Investigating further, it was realised that skin and thickness were perfectly correlated because they one of them was the corresponding unit equivalent of the other.

# **Multivariate Analysis**



This shows the **aggregate glucose concentration** for the two diabetic stata of patients grouped into the number of pregnancies they had had eg: the aggregate glucose concentration for patients who didn't have diabetes but had been pregnant once is 11,051 units



This scatterplot represents the distribution of the **diabetic status** of patients across **age** and **insulin level** with the **60 representing the insulin level** of the patient who doesn't have diabetes and is above 80 years old