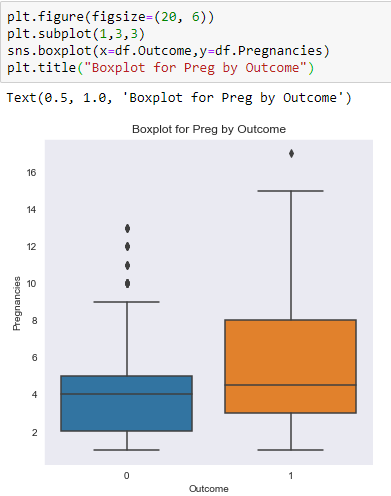
**CAPSTONE HEATHCARE PROJECT**

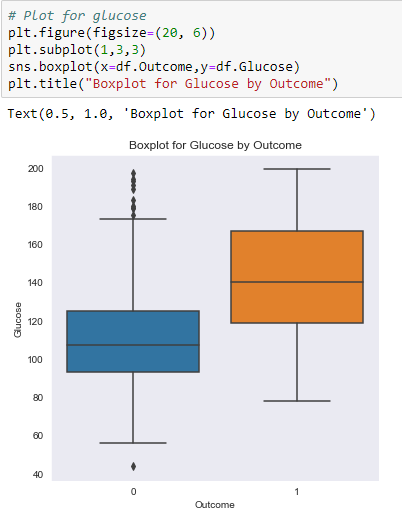
1. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of actions.
2. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

3. Perform correlation analysis. Visually explore it using a heat map.

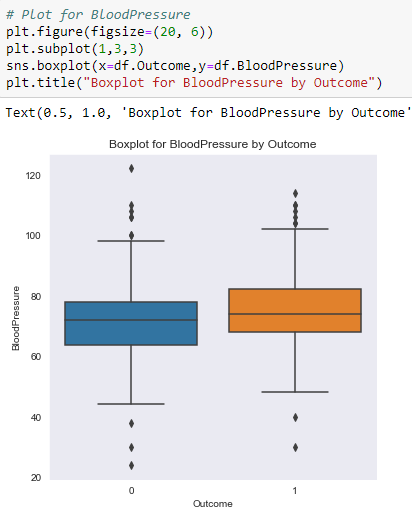
***1 - Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of actions -***



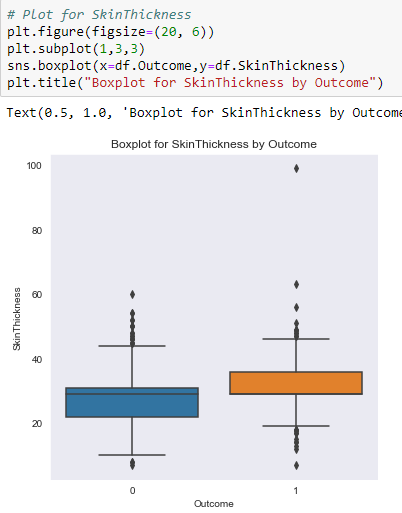
**In the boxplots, we find few outliers in both subsets. Some of the non-diabetic women have had many pregnancies. To validate this hypothesis, we need to statistically test it.**

****

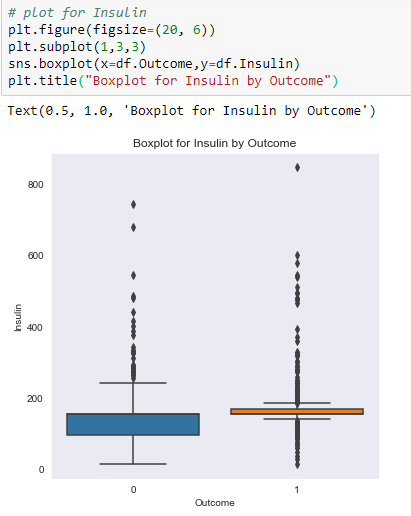
**Glucose data is slightly skewed to right the data set contains over 60% who are diabetic and its likely that their Glucose levels were higher. visually skewness seems acceptable (<2) and its also likely that confidence intervels of the means are not overlapping. So a hypothesis that Glucose is measure of outcome, is likely to be true. But needs to be statistically tested.**

****

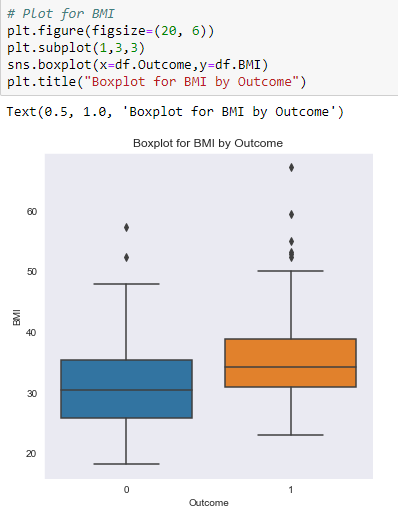
**Few outliers in the data. Its likely that some people have low and some have high BloodPressure. So the association between diabetic (Outcome) and BloodPressure needs to be statistically tested.**

****

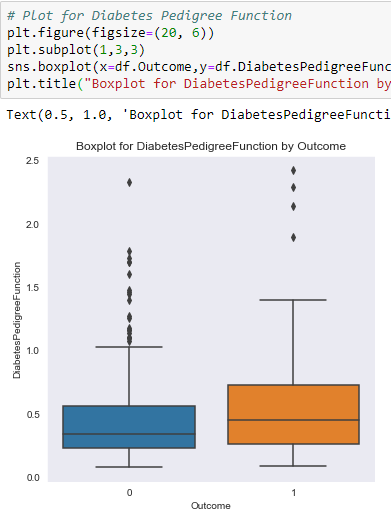
**People who are not diabetic have lower skin thickness. This is a hypothesis has to be validated. As non-diabetic data is skewed but diabetic samples seems to be normally distributed.**

****

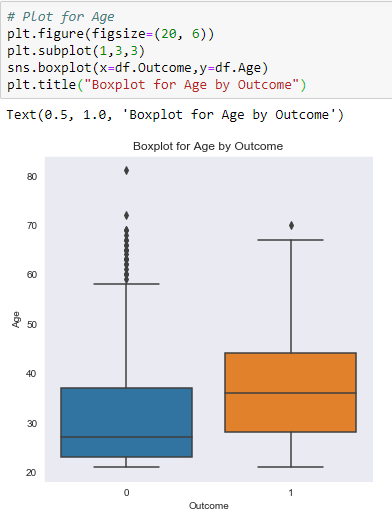
**There are Outliers in this data. These Outliers are concern for us and most of them with higher insulin values are also diabetic.**

****

**BMI also has outliers. People on the higher side of BMI have diabetes.**

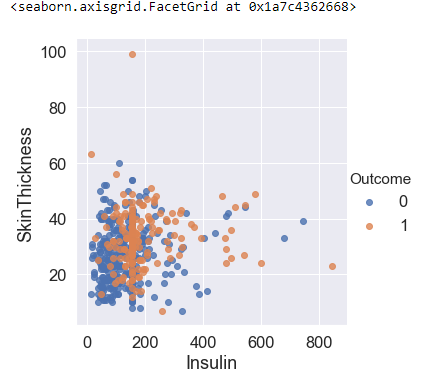
****

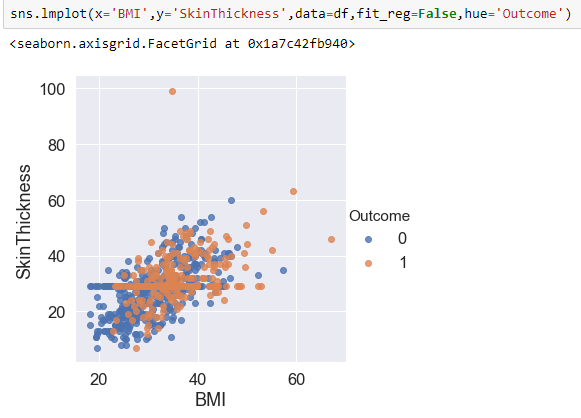
**As Diabetes Pedigree Function increases, there is a strong likelihood of being diabetic, but needs statistical evaluation.**

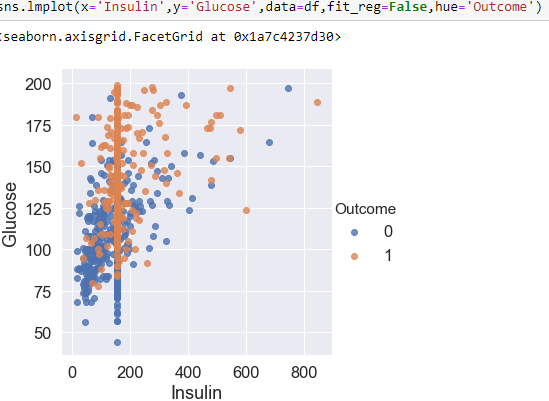
****

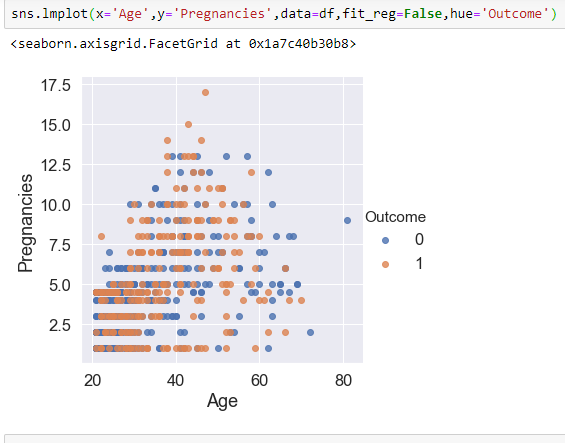
**Age data also has outliers. The more the age goes up, the more people are prone to diabetes**

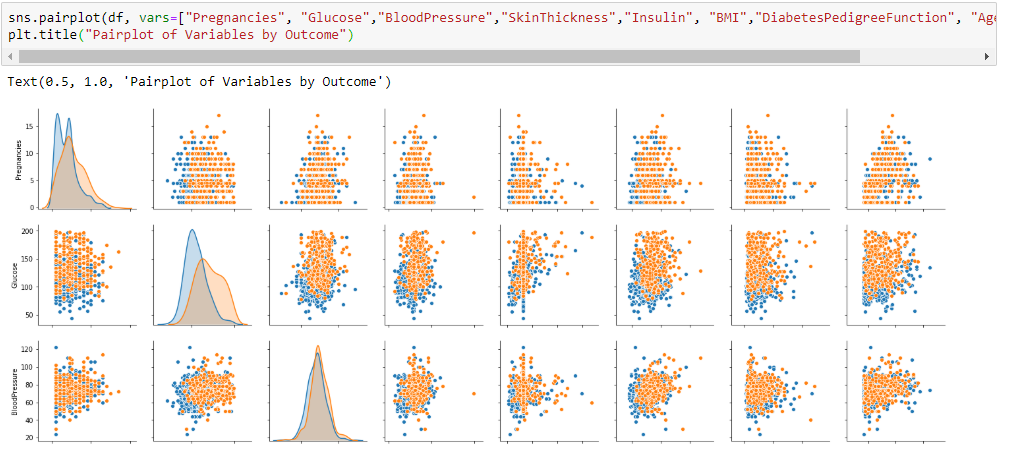
***2. Create scatter charts between the pair of variables to understand the relationships. Describe your findings -***

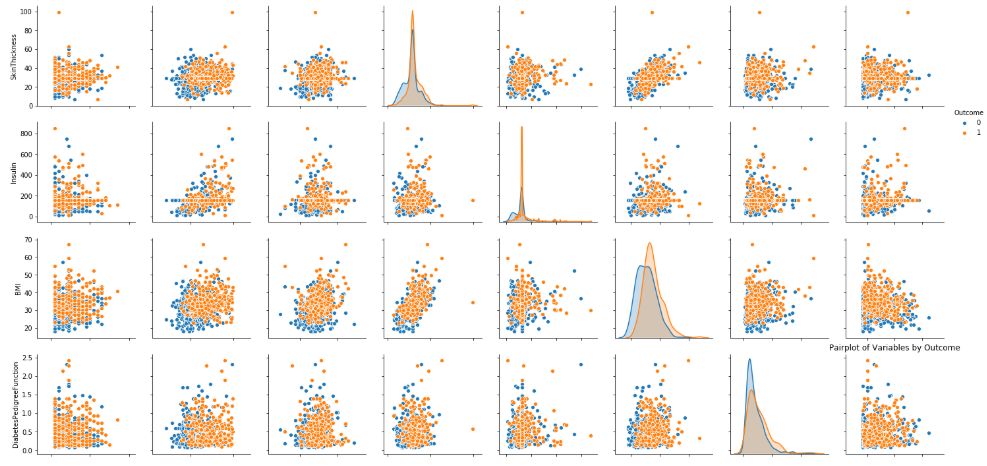


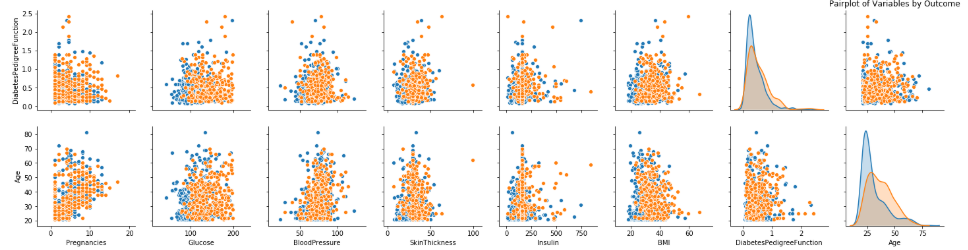












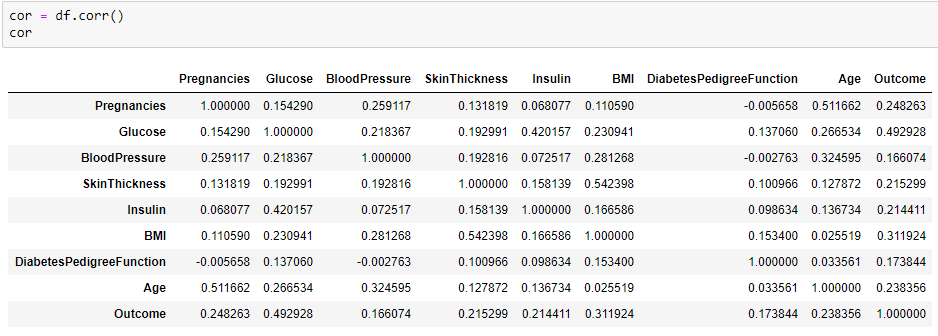
**The plot shows that there is some relationship between parameters. Outcome is added as hue. We see that blue and orange dots are overlapped. Also,**

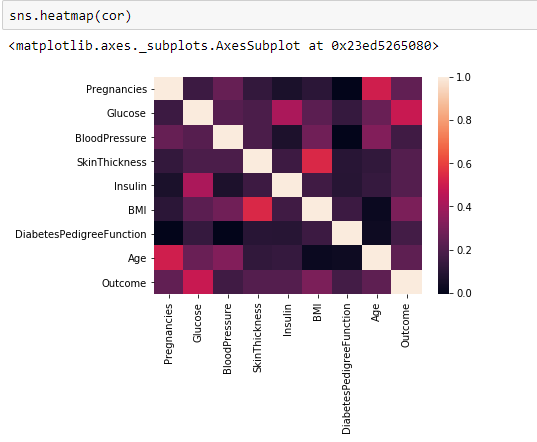
Pregnancies and age have a linear line.

BloodPressure and age have little relation. Most of the aged people have BloodPressure.

Insulin and Glucose have some relation.

***3. Perform correlation analysis. Visually explore it using a heat map -***

****

****

**The correlation plot shows the relation between the parameters.**

Glucose, Age, BMI and Pregnancies are the most correlated parameters with the Outcome

Insulin and DiabetesPedigreeFunction have little correlation with the outcome.

BloodPressure and SkinThickness have tiny correlation with the outcome.

There is a little correlation between Age and Pregnancies, Insulin and Skin Thickness, BMI and Skin Thickness, Insulin and Glucose.

***Conclusion***

1. *The dataset have nine attributes in which there are eight independent variables (Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age) and one dependent variable (Outcome).*
2. *BMI and DiabetesPedigreeFunction are a float data type and other parameters are integer data type.*
3. *The parameters do not contain any null values (missing values). However, this cannot be true. As Insulin, SkinThickness, BloodPressure, BMI, Glucose have zero values.*
4. *The Outcome parameter shows that there are 500 non-diabetic and 268 diabetic people. It means that 65.1% people are diabetic and 34.9% people are non-diabetic.*
5. *The parameters Glucose, BloodPressure, BMI are normally distributed. Pregnancies, Insulin, Age and DiabetesPedigreeFunction are rightly skewed.*
6. *The missing values '0' is replaced by the mean of the parameter to explore the dataset.*
7. *BloodPressure, SkinThickness, Insulin, BMI have outliers.*
8. *There are no convincing relationships between the parameters. Pregnancies and age have some kind of a linear line. BloodPressure and age have little relation. Most of the aged people have BloodPressure. Insulin and Glucose have some relation.*
9. *Glucose, Age BMI and Pregnancies are the most Correlated features with the Outcome. Insulin and DiabetesPedigreeFunction have little correlation with the outcome. BloodPressure and SkinThickness have tiny correlation with the outcome.*
10. *Age and Pregnancies, Insulin and Skin Thickness, BMI and Skin Thickness, Insulin and Glucose are little correlated.*
11. *After Pregnancy people have more chance of diabetes.*
12. *People with high Glucose, BloodPressure and Insulin levels are more likely to have diabetes.*