Analysis of Diabetes Dataset

1. Research Question

Is there significant relationship between glucose and the likelihood of developing diabetes and how does this relationship vary across different age groups ?

2. Exploratory Data analysis

2.1 Dataset Statistics

[5]:

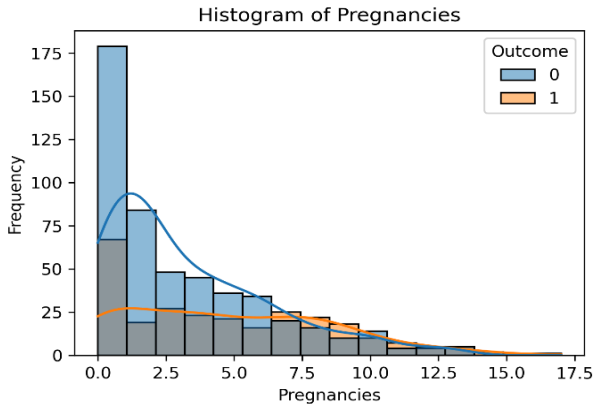
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

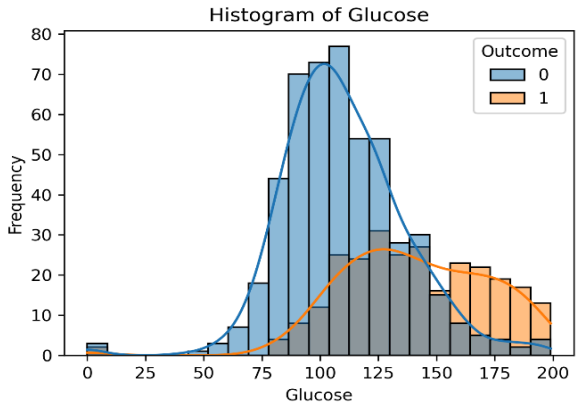
For each feature in the dataset the above statistics were measured,

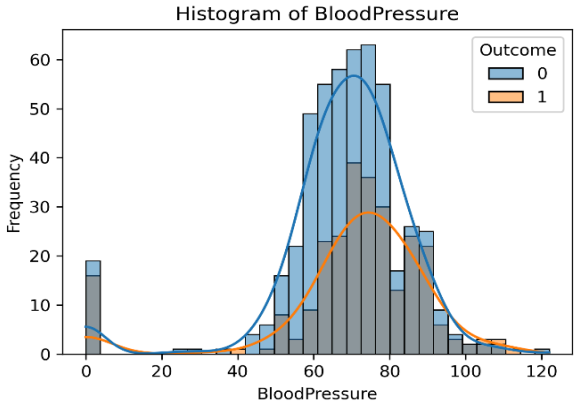
- Mean
- Standard Deviation
- Max & Min
- 25%, 50% and 75% percentile

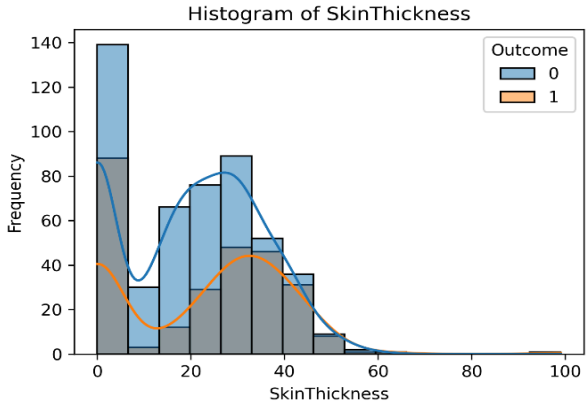
2.2 Visualizations

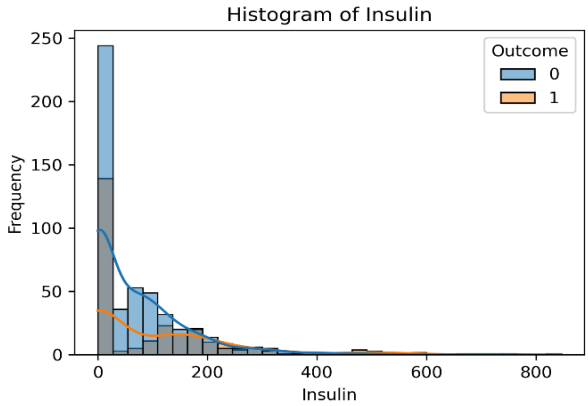
Lets take a look at the distribution of each feature in the dataset .

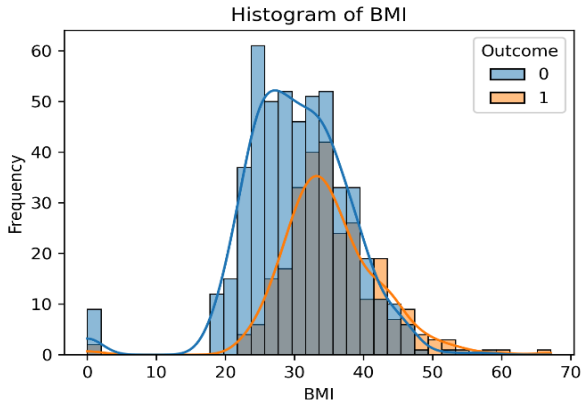


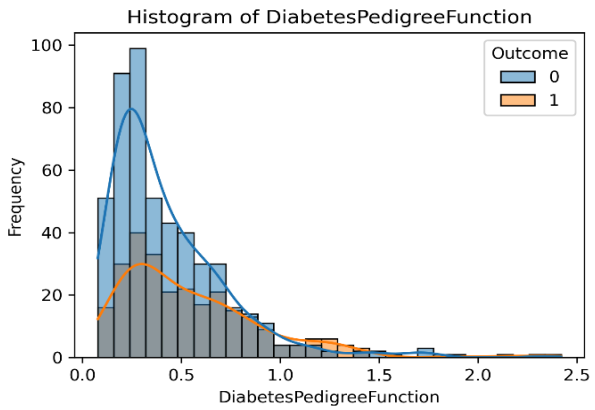


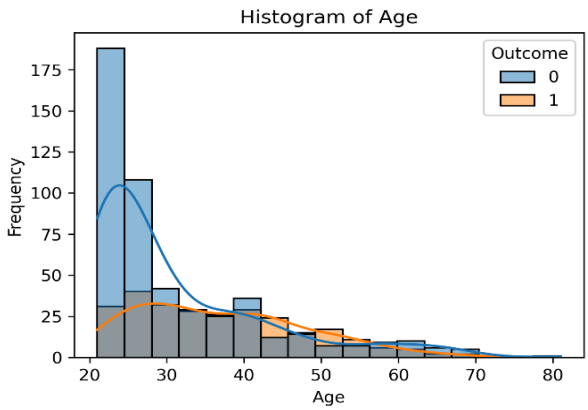






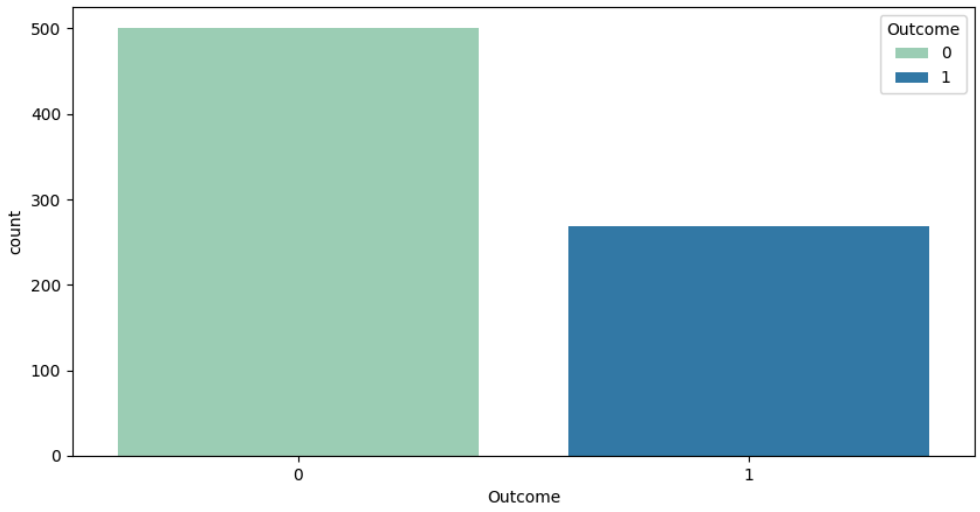






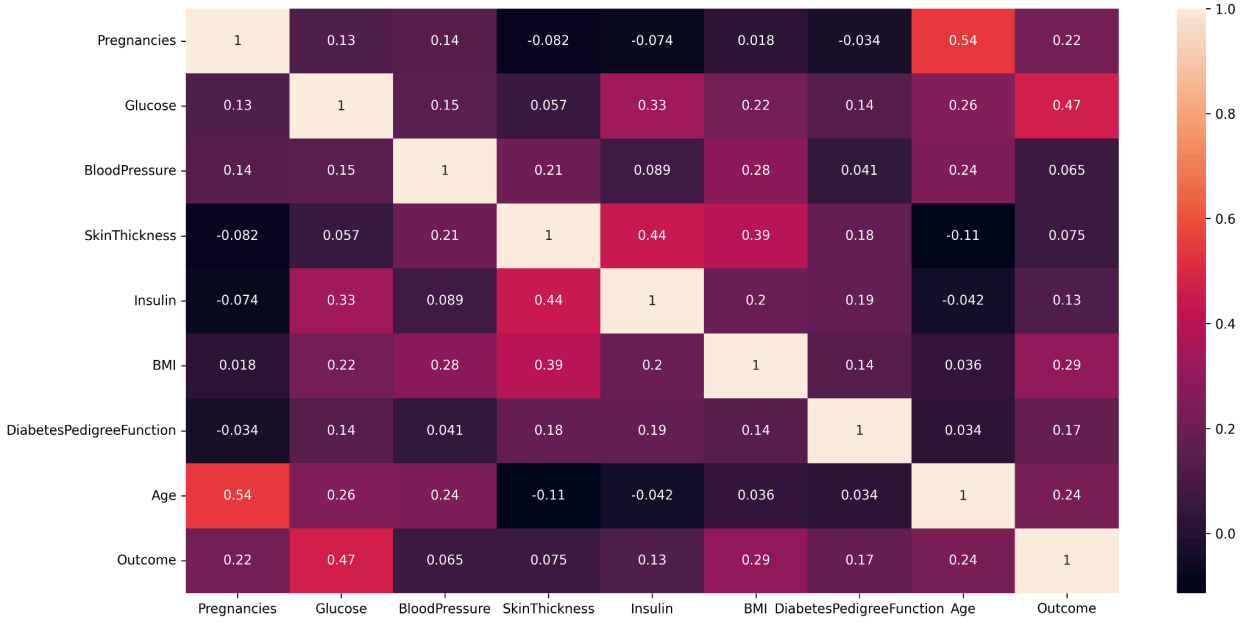
From the above chart we can also interpret the skewness as well as the kurtosis of the features.

Now we will check if the classes are evenly distributed.

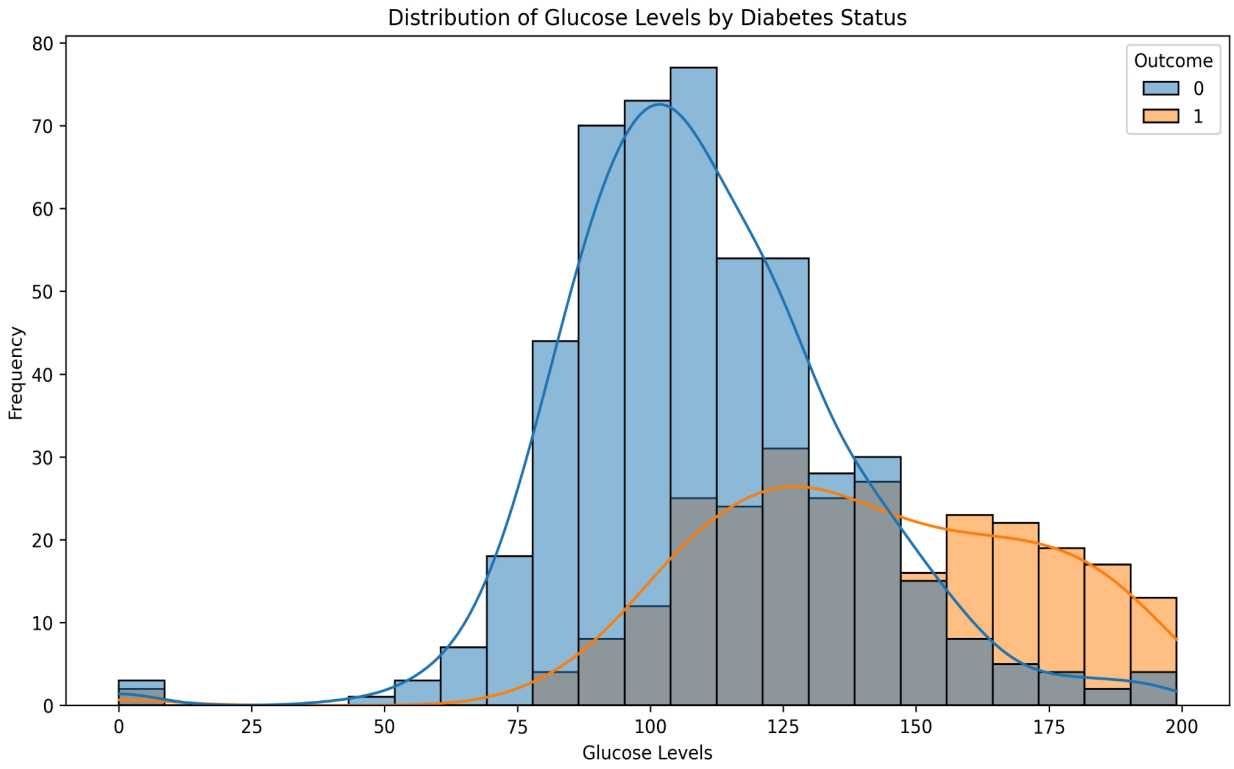


From the above bar chart we can conclude that there are more instances of 0 (non-diabetes) than 1 (diabetes).

Now lets look at the correlation matrix which will help us understand the relationship between each features better.



Visualizing the Glucose level distribution by diabetes status.



From the above plot we can infer that the glucose level among diabetic patients are more than in non-diabetic patients.

3. Hypothesis Testing

- **Null Hypothesis (H0) :** There is no significant difference in glucose levels between diabetic and non-diabetic patients.
- **Alternative Hypothesis (H1) :** There is significant difference in glucose levels between diabetic and non-diabetic patients.

We will perform **Independent T-Test** for our hypothesis testing .

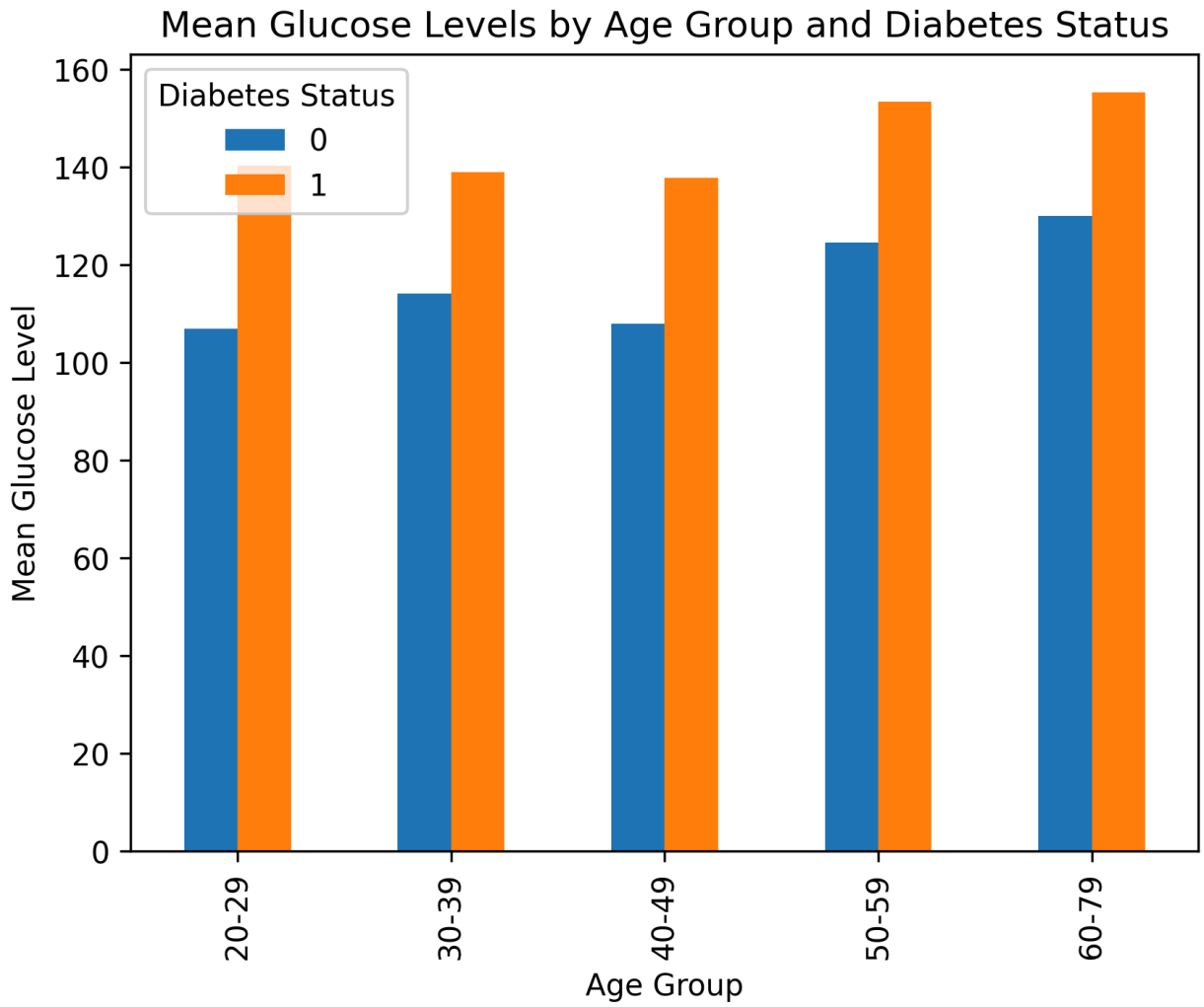
```
T-statistic: -14.60, p-value: 8.94e-43
```

From the results of the T-test we can conclude that:

- We can conclude there is significant difference between the two groups from the negative T value

- We can also reject the null hypothesis since the p-value is significantly lower than 0.05

To verify this lets take a look at the **Mean Glucose Level Distribution** by **Age Group** and **Diabetic Status.**



Conclusion

Observing the above chart we can conclude that our alternate hypothesis was correct because, we can see that across all age groups the diabetic people have higher glucose levels.