Project 2: Diabetes Analysis

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of the instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The dataset consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Goal:

(1) **Finding and visualising frequency of pregnancies and means of different parameters.**

```
In [14]: data = pd.read_csv('../workspace/diabetes.csv')
         data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

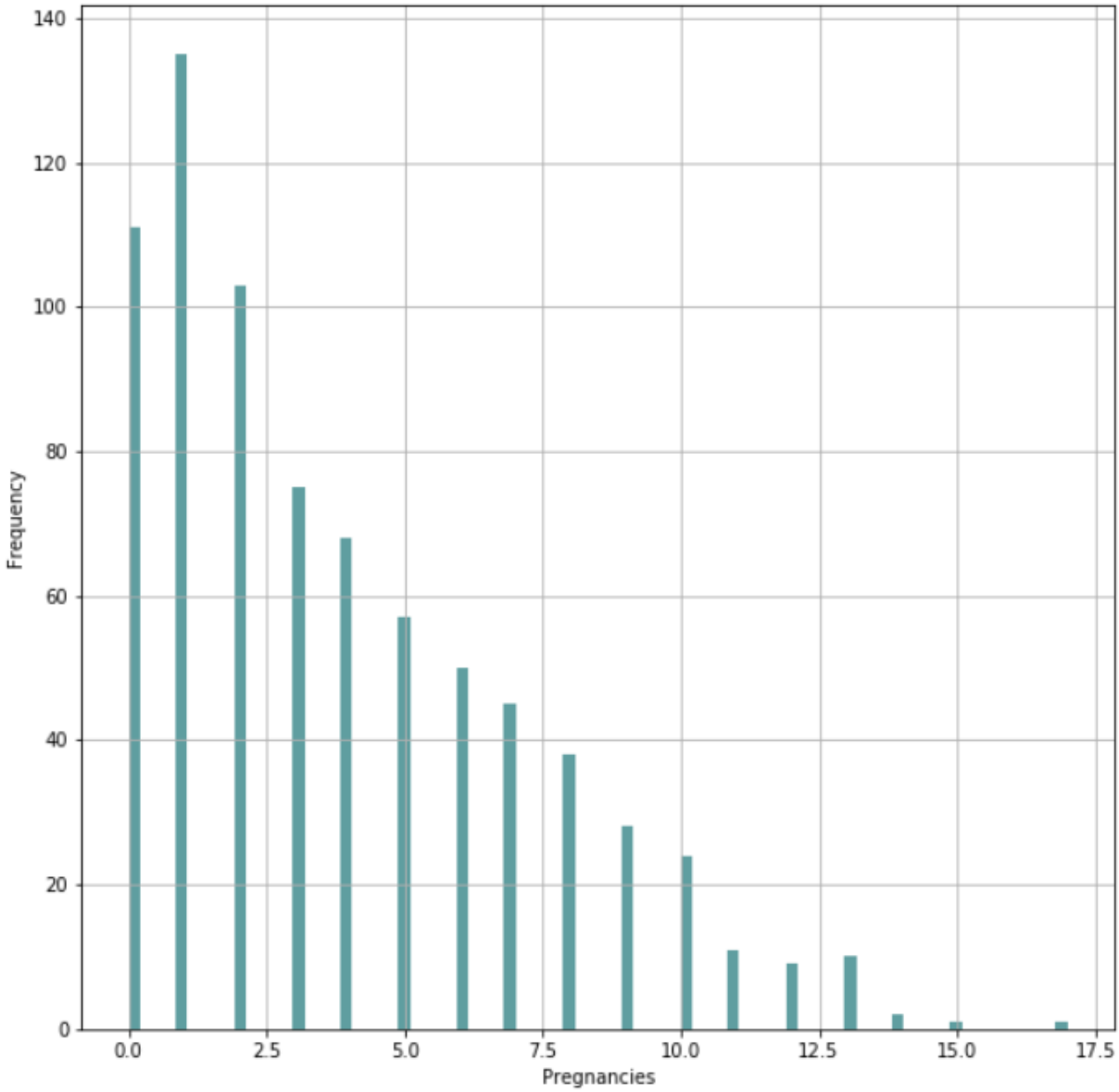
In [15]:
```
data.describe()
```

Out[15]:

	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

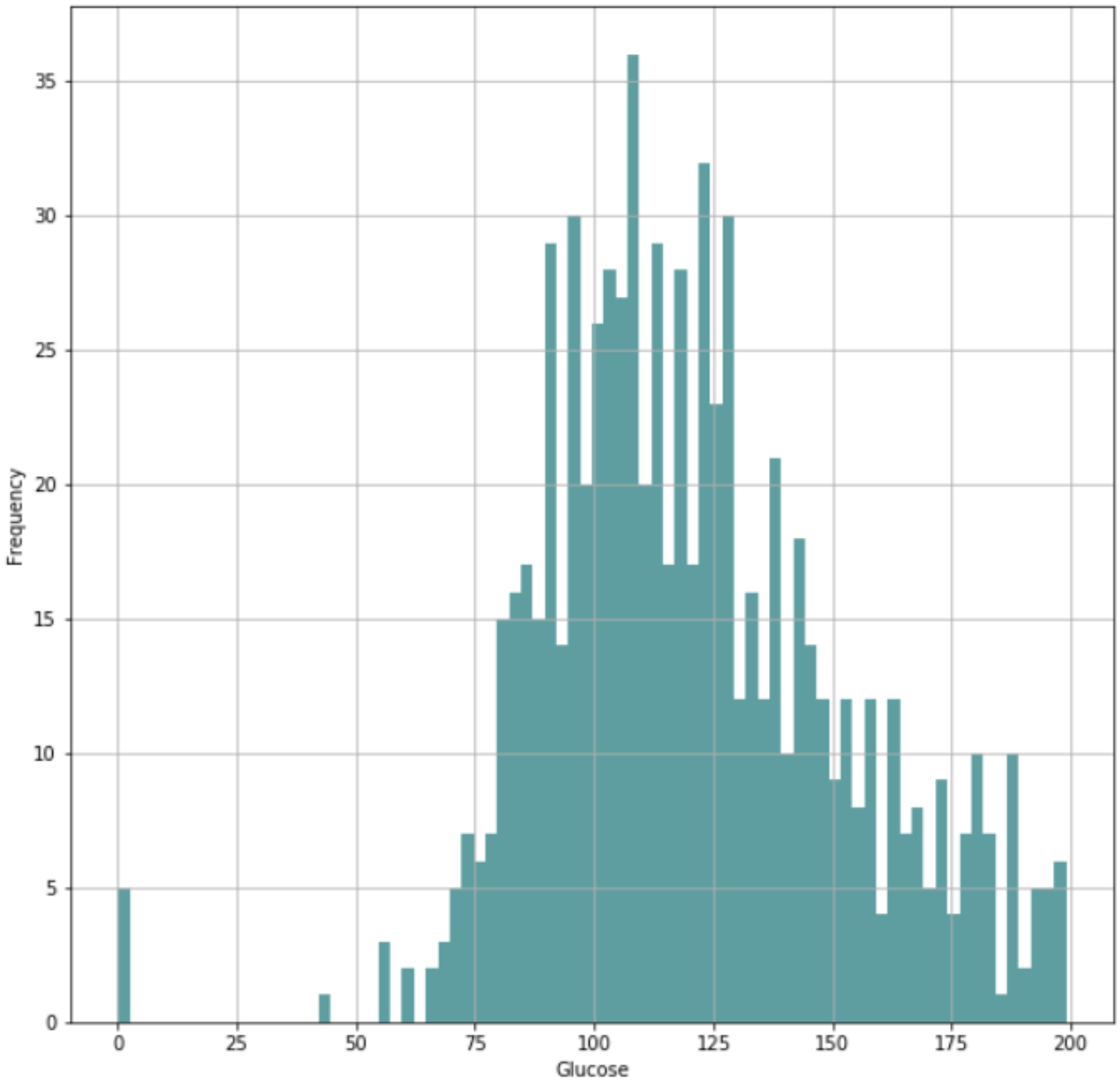
Conclusion: This shows the different columns of the dataset along with various statistical data like mean.

In [19]:
```
data_columns = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age']
for each in data_columns:
    fig1, ax1 = plt.subplots(figsize =(10,10))
    plt.hist(data[each], bins=80,color = "cadetblue")
    plt.xlabel(each)
    plt.ylabel("Frequency")
    plt.grid()
    plt.show()
```

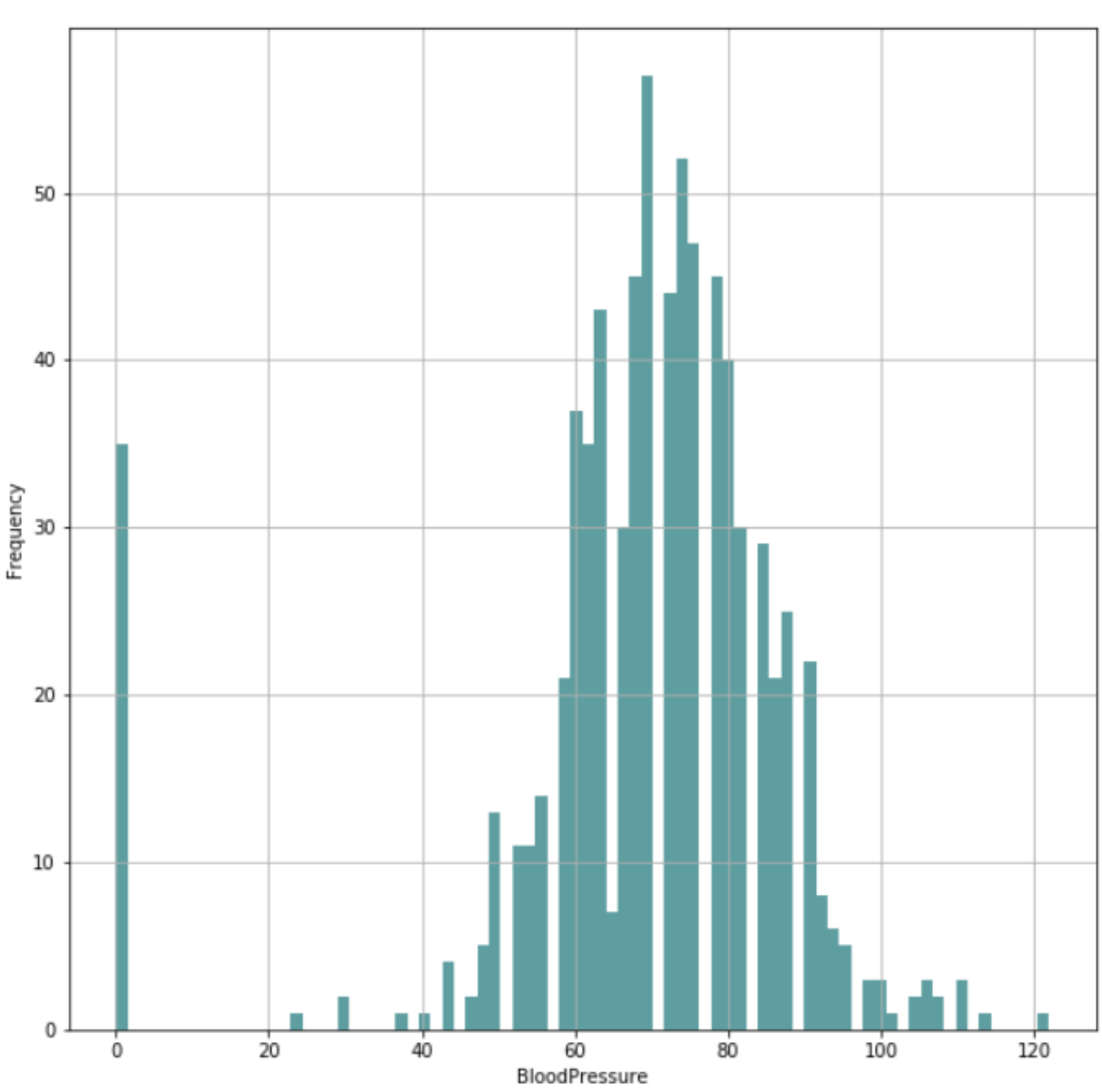


Conclusion: This graph shows the frequency of pregnancies plotted against the number of pregnancies. As we can see, the most frequent number a female was pregnant is 1 followed by 0. As the number of pregnancies increases, the number of females being pregnant that number of times decreases.

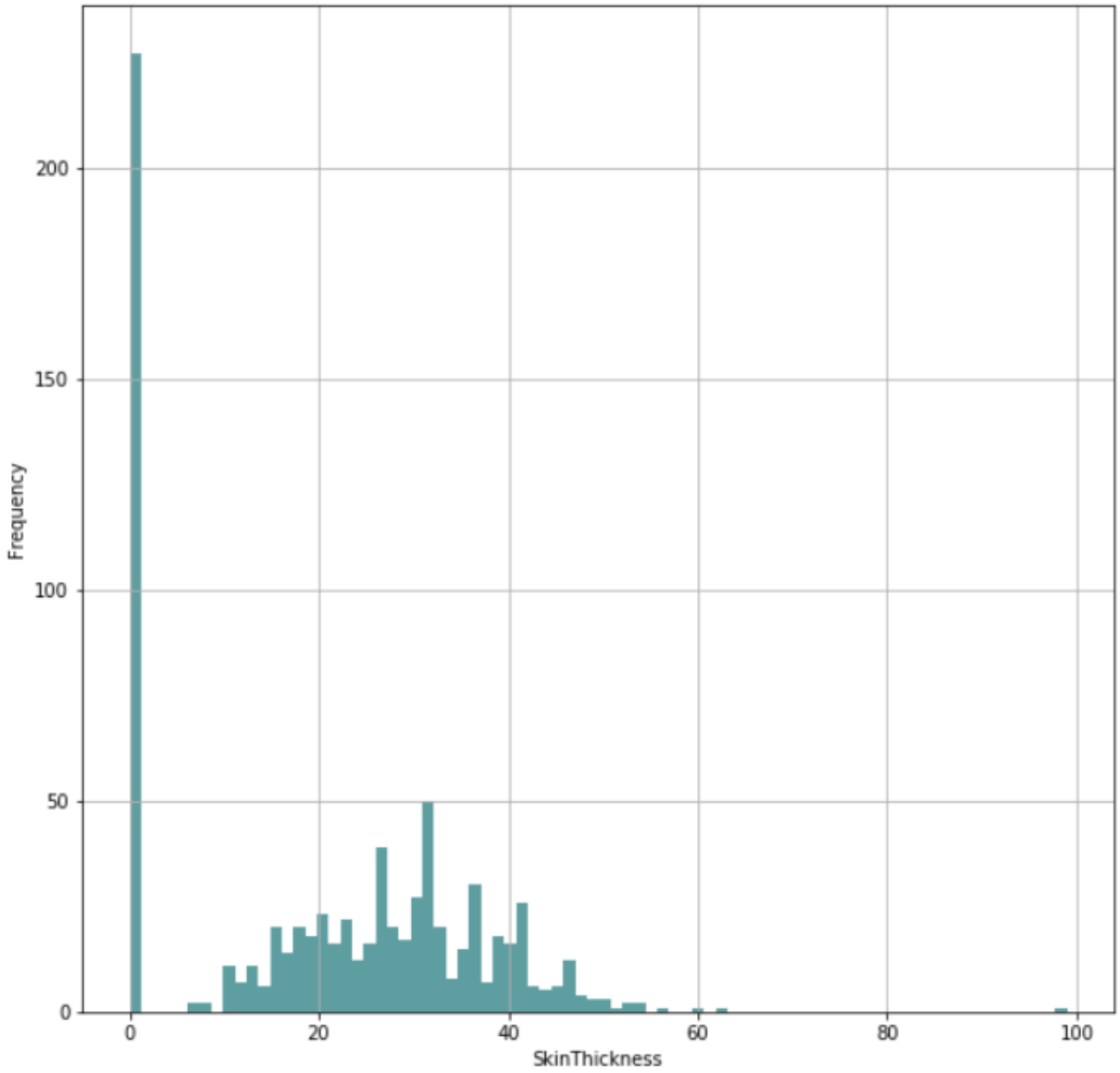
(2) Finding and visualising the frequencies of other parameters.



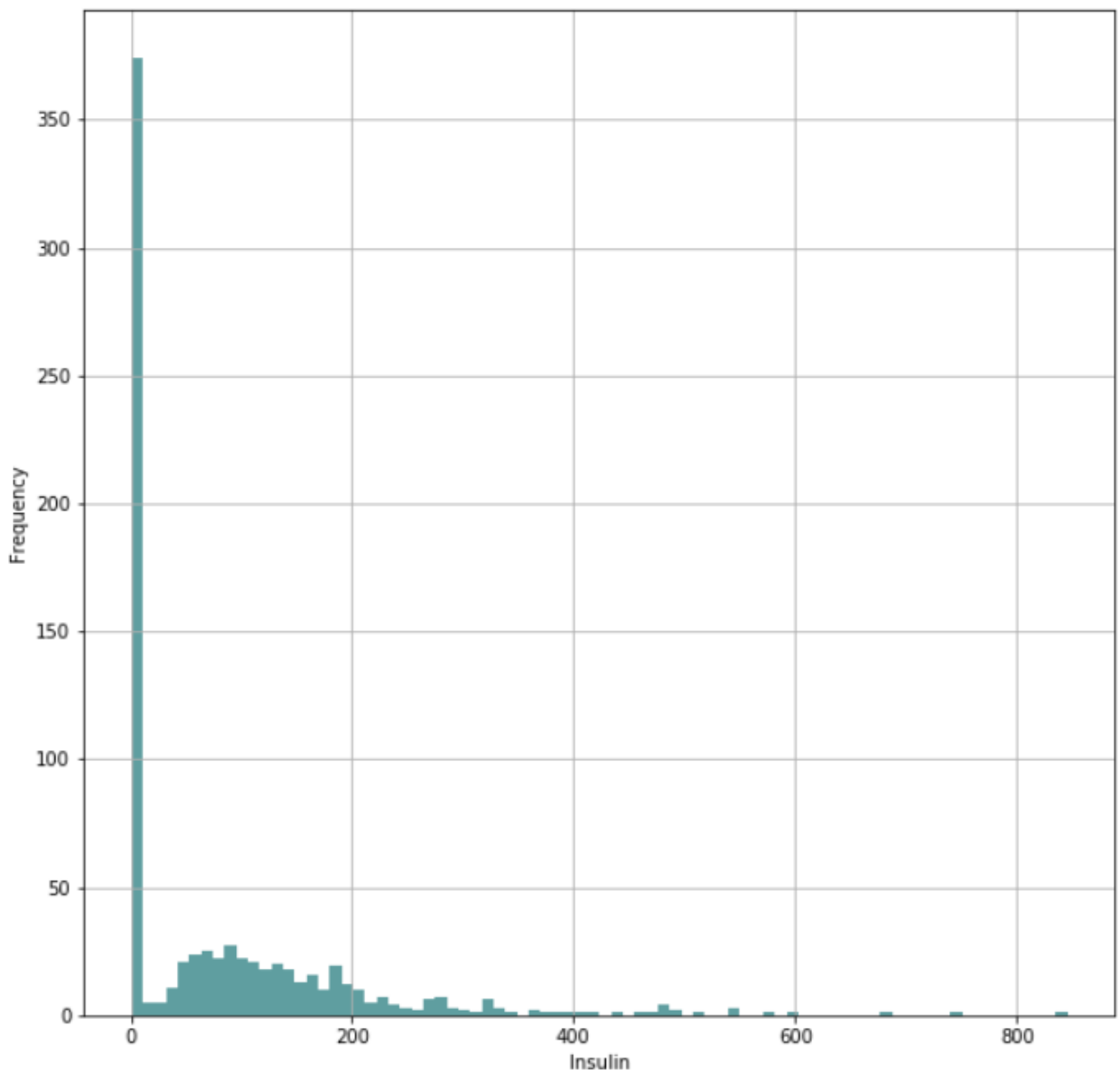
<u>Conclusion:</u> This graph visualisation shows the distribution of Glucose and its frequency. We can infer from the graph that blood glucose level in the range of 100-125 is the most frequent level with highest frequency in this group being greater than 35. We also notice 5 '0' values which indicate that these values were not collected properly and resulted in an improbable value.



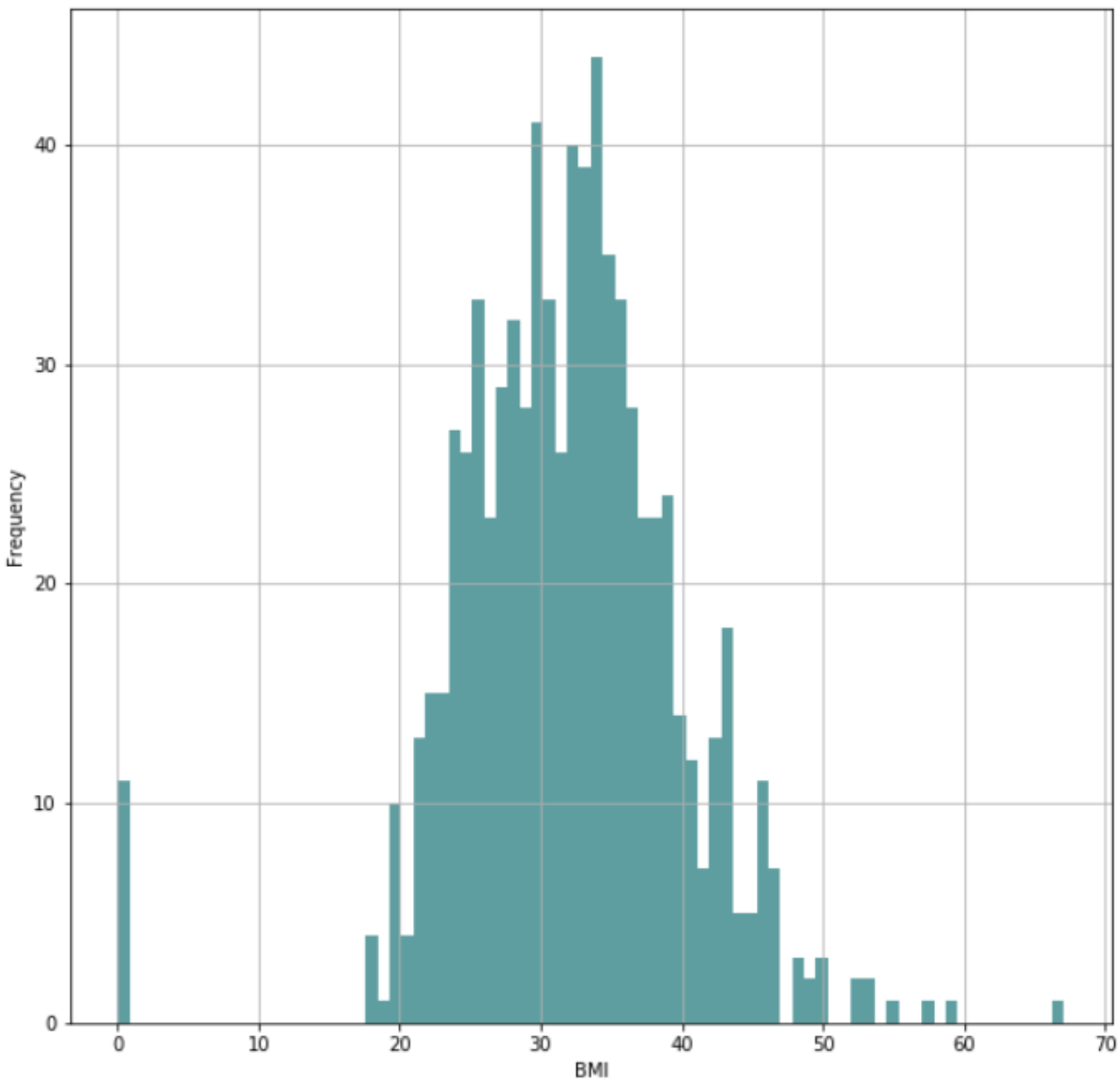
Conclusion: This graph visualisation shows the distribution of given blood pressure. It shows that majority of females have blood group less than 80, with very few (less than 5) having values as low as 20. We again notice discrepancy in data collection as we notice some values of blood pressure to be 0.



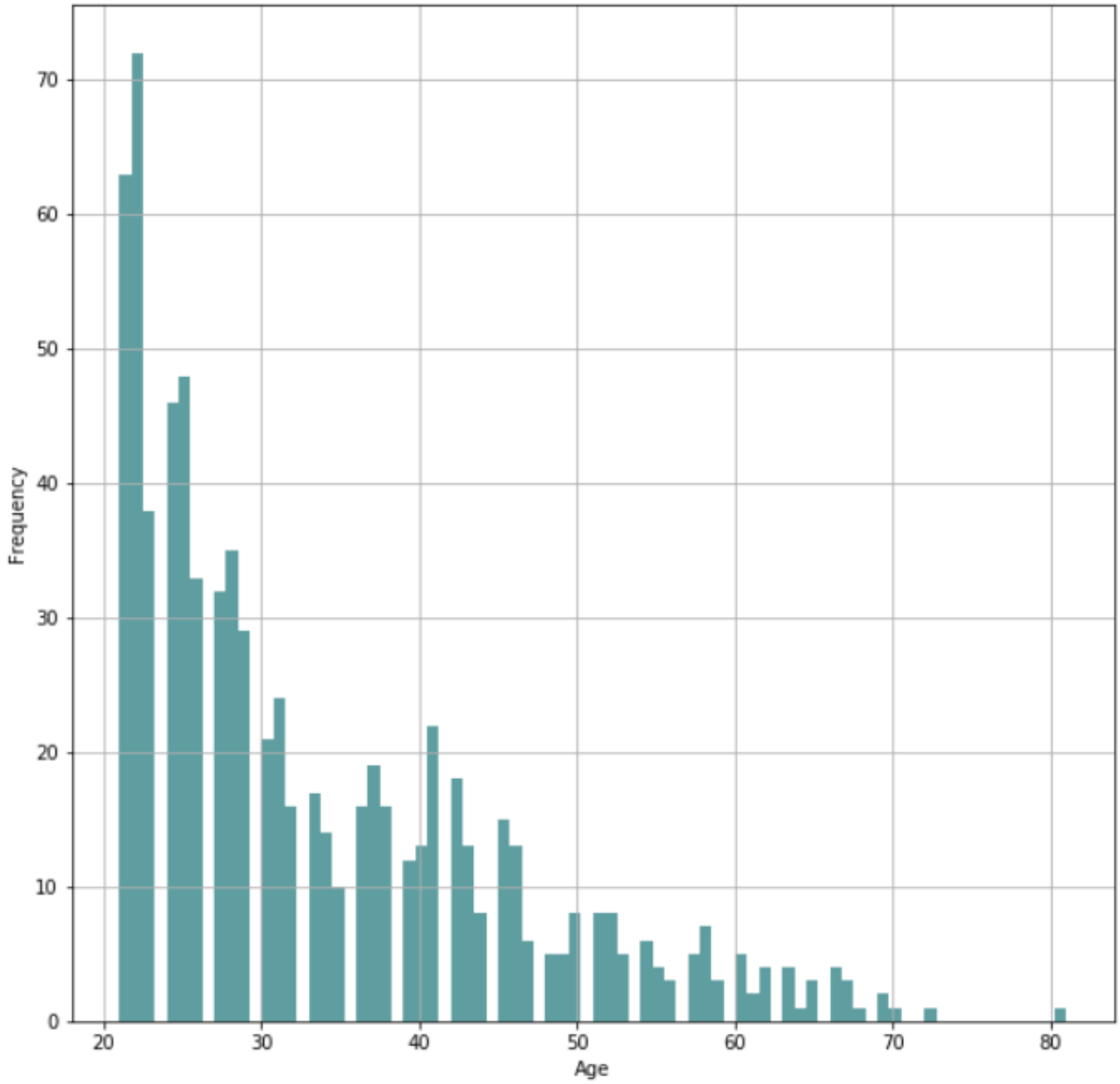
Conclusion: This graph shows the skin thickness plotted against the frequency. This graph shows that the most common skin thickness among the given females lies between 20 and 40. Here again we can notice the discrepancy in data collection as skin thickness can't be 0.



Conclusion: This plot shows the insulin levels with the frequency. Majority of the females have insulin levels less than or equal to 200 with most frequent insulin level being 0. The highest insulin level goes over 800.

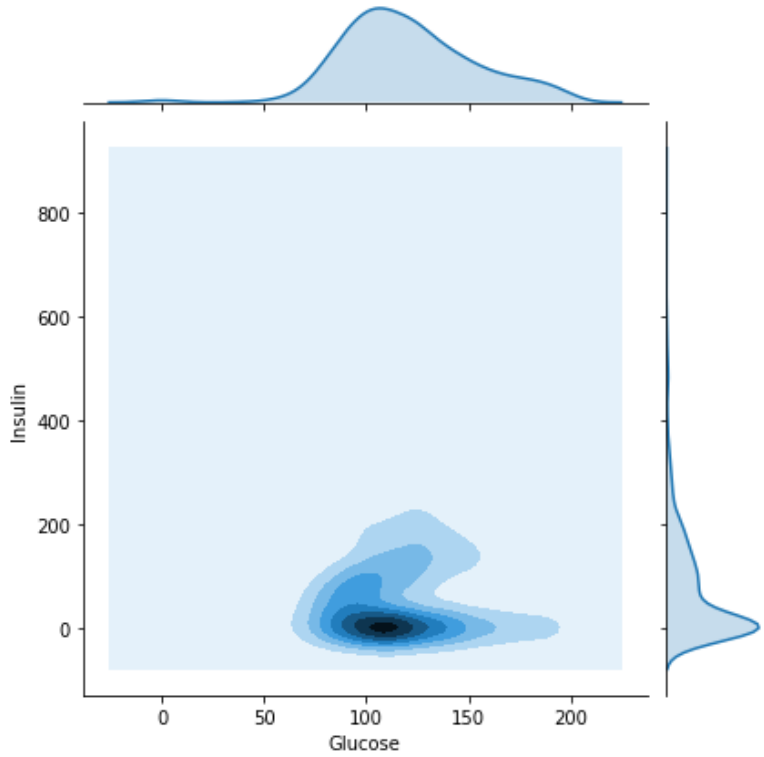


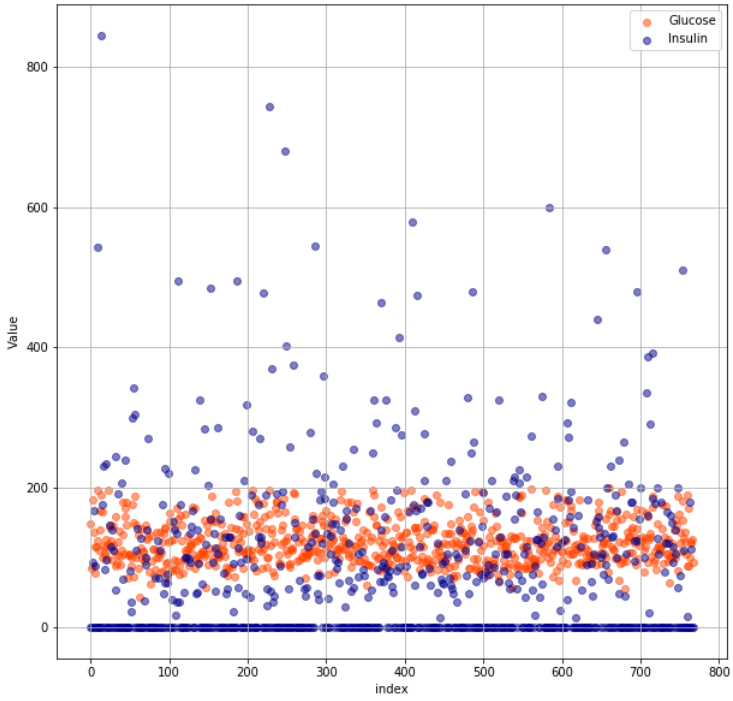
Conclusion: This graph shows the distribution of BMI. As we can see that the most frequent BMI lies in the range 30-40. From this we can infer that the given group is overweight/obese.



Conclusion: From this graph, we can infer that majority of the females in the given dataset are in the age group of 20-30.

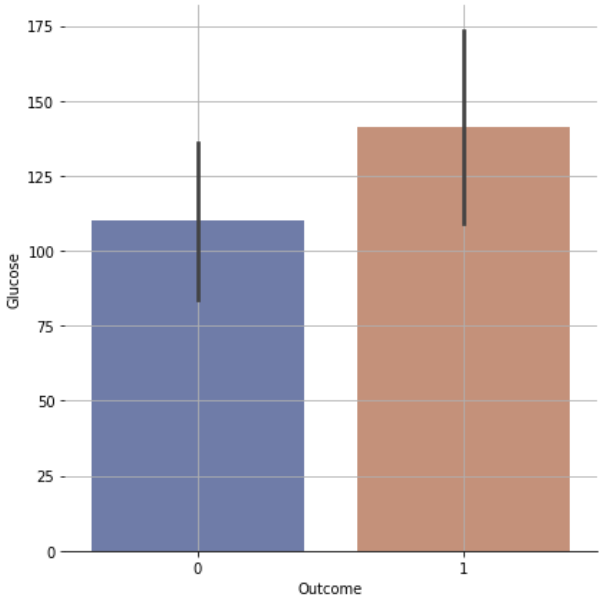
(3) **Finding and visualising the level of insulin corresponding to glucose.**

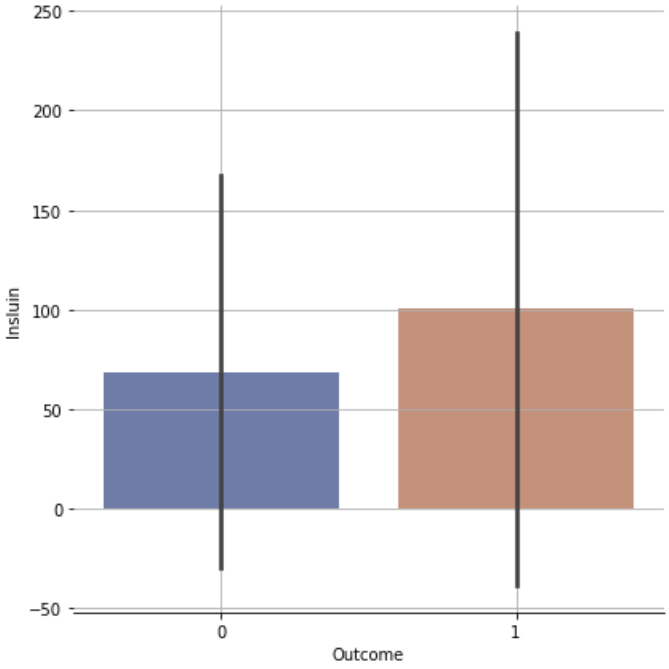




Conclusion: From the above two graphs, we can see that the concentration of the graph lies around 100-150 for glucose vs 0-200 for insulin levels.

(4) **Finding and visualising the level of different parameters for probable outcomes where 1 means probable to be pregnant and 0 means not probable to be pregnant.**



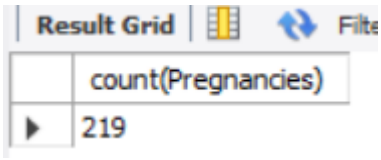


<u>Conclusion:</u> From these two graphs, we can see that with the increase in insulin and glucose levels, the probable outcome tends to 1, i.e., increase in chance of being pregnant. This is evident from the fact that the number of outcomes '1' is more in higher spectrum of glucose levels (>125) and insulin (>50).

SQL QUERIES:

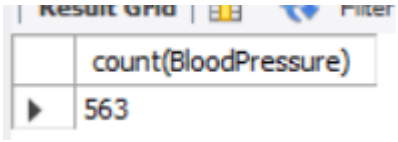
1. **To find the number of women with pregnancies more than 5:**

```
select count(Pregnancies) from diabetes
where Pregnancies>5;
```



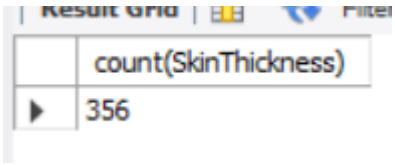
2. **To find number of women with blood pressure less than 80:**

```
select count(BloodPressure) from diabetes
where BloodPressure<80;
```



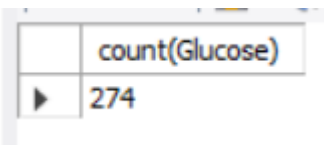
3. **To find the number of women with skin thickness between 20-40:**

```
select count(SkinThickness) from diabetes
where SkinThickness between 20 and 40;
```



4. **To find the number of women with glucose between 100-125:**

```
select count(Glucose) from diabetes
where Glucose between 100 and 125;
```



5. **To find BMI with highest frequency:**

```
select BMI, count(BMI) as BMI_frequency from diabetes
group by BMI order by BMI_frequency desc;
```

Result Grid | Filter R

BMI	BMI_frequency
32	13
31.6	12
31.2	12
0	11
32.4	10
33.3	10
30.1	9
32.9	9
32.8	9
30.8	9
33.6	8