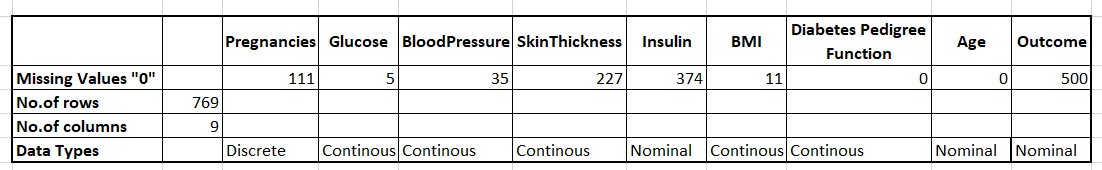
**Mandatory project**

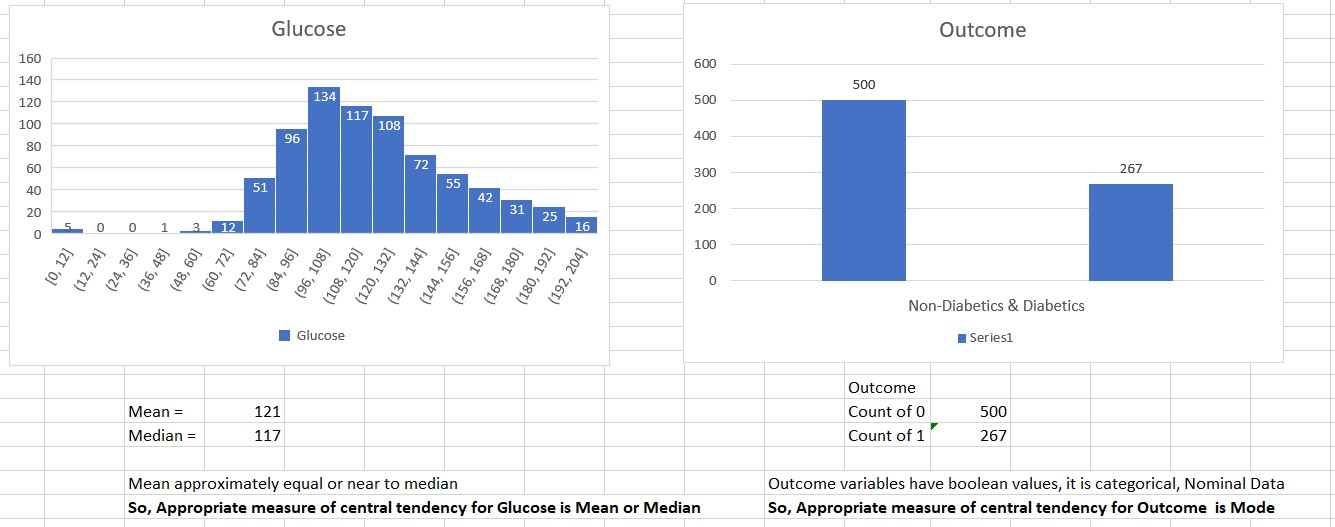
**Pima Indians Diabetes Database Descriptive Statistics Project**  
Predict the onset of diabetes based on diagnostic measures. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset.

**Columns of the dataset  
Pregnancies**: Number of times pregnant  
**Glucose**: Plasma glucose concentration 2 hours in an oral glucose tolerance test  
**Blood Pressure**: Diastolic blood pressure (mm Hg)  
**Skin Thickness**: Triceps skin fold thickness (mm  
**Insulin**: 2-Hour serum insulin (mu U/ml)  
**BMI**: Body mass index (weight in kg/(height in m)^2)   
**Diabetes Pedigree Function**: Diabetes pedigree function  
**Age**: Age (years)  
**Outcome**: Class variable (0 or 1) 0 mean non-diabetic and 1 means diabetic

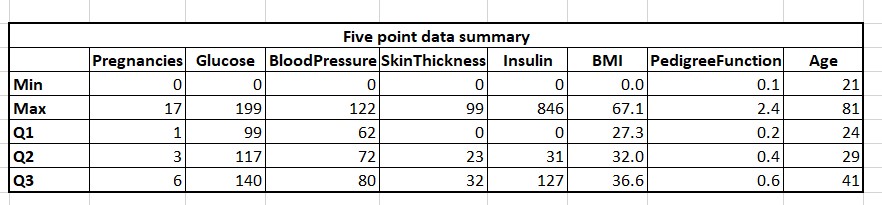
1. Please do the basic exploration of data and explain missing values, number of rows and columns and data types in statistical term. Hint: be cautious with missing values. Missing values do not mean NA only. They may enter also in a different format. Sometimes they may represent as ?, #, etc or also with extreme values. Please report information as required.



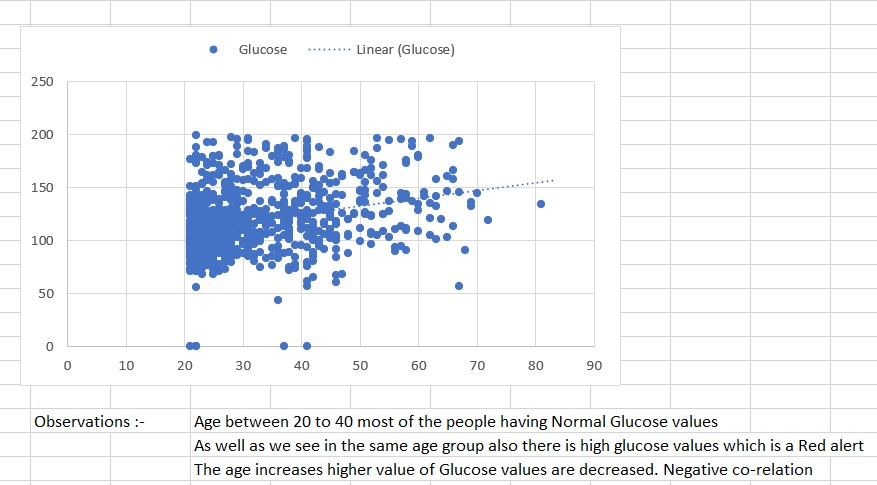
1. Calculate appropriate measures of central tendency for Glucose and outcome column only?



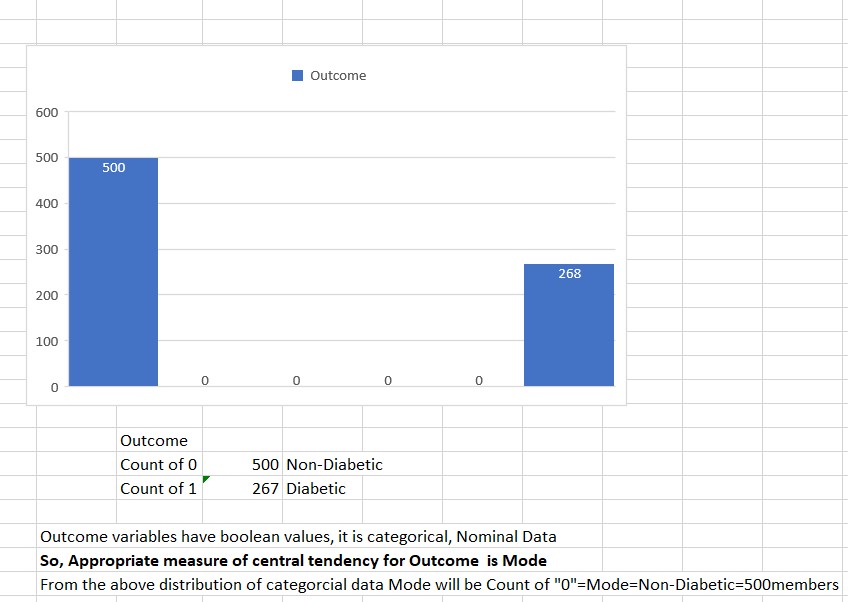
1. Please provide 5 points data summaries for required columns?



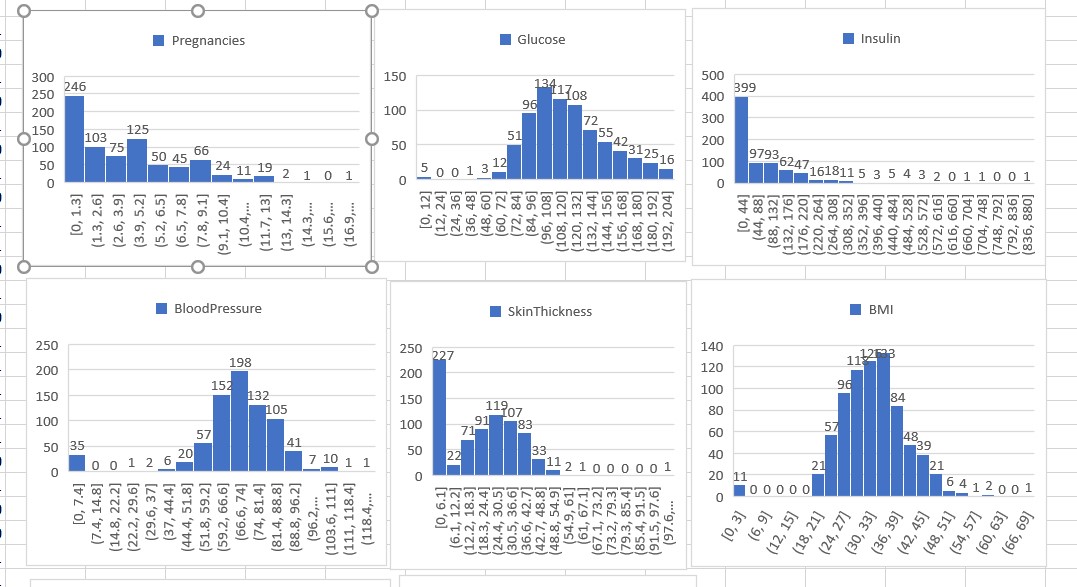
1. Please create an appropriate plot to examine the relationship between Age and Glucose.

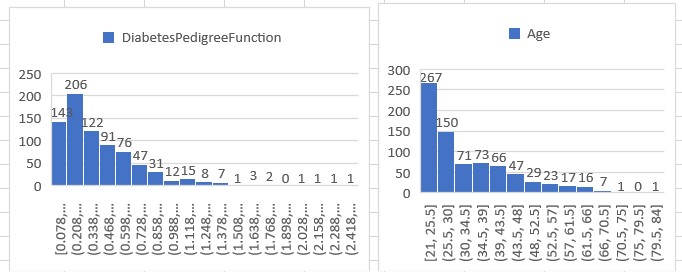


1. Please create an appropriate plot to see the distribution of Outcome variable?



1. Please examine the distribution of numerical data and explain which variable normally distributed and which variable seems to be skewed. Please also tell the direction of skewness.





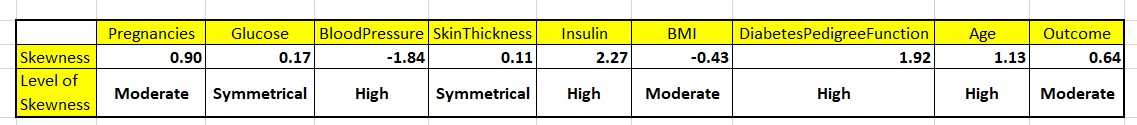
Normally Distributed : Blood Pressure, Glucose

Right Skewed : BMI, Skin Thickness, Pregnancies, Age, Diabetes Pedigree

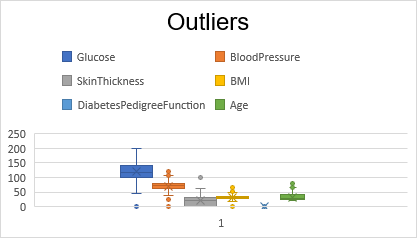
-function, Insulin

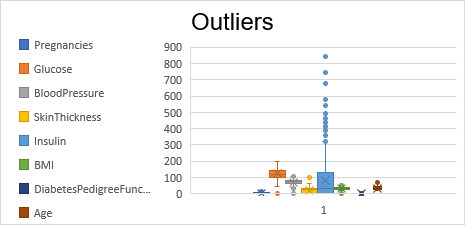
Left Skewed : Glucose

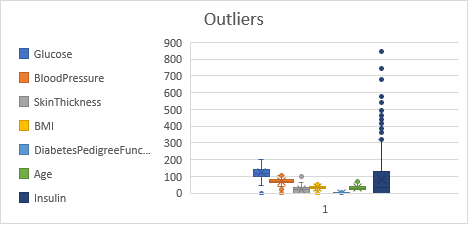
1. Please calculate the skewness value and divide variables into symmetrical, moderately skewed and highly skewed.



1. Please create appropriate plot to examine the outliers of these variables. Please name the variables which have outliers.





z

We have outliers more in Insulin, Pregnancies, Glucose

1. What should be the measures of central tendency and dispersion for skewed data? Please report appropriate measures of central tendency for skewed data only.

For skewed data

Best measure for central tendency would be median

Best measure for dispersion would be interquartile range or median absolute deviation

1. What should be the measures of central tendency and dispersion for Normally distributed data?  Please report appropriate measures of central tendency for normal data only.

For normal distribution

Best measure for CT would be mean or median u can use any one

Best measure for dispersion would be standard deviation