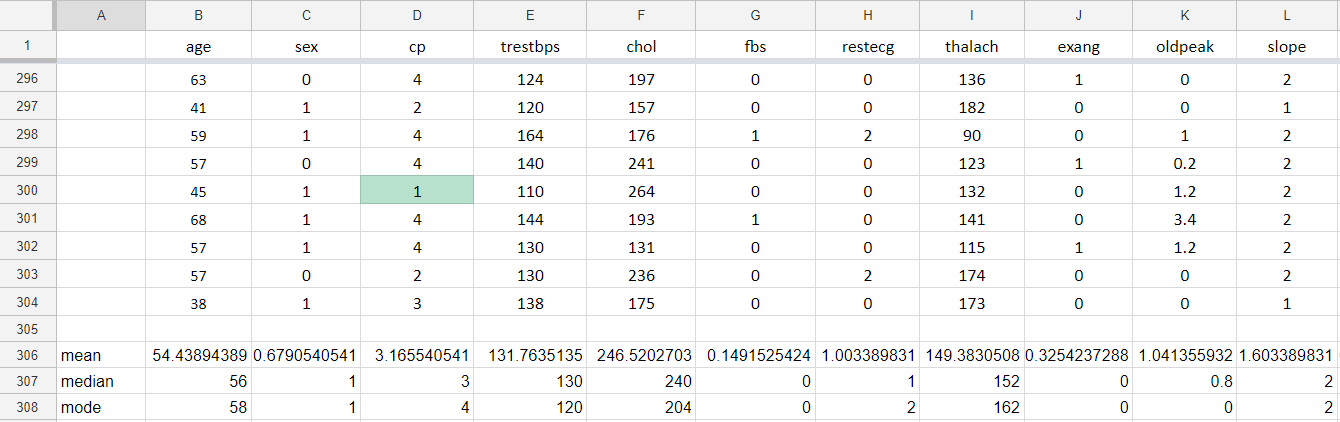
**EXPERIMENT- 2 (PART A)**

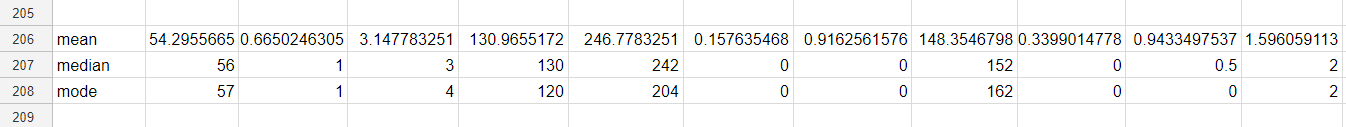
**DATA-PREPROCESSING USING EXCEL:**

**Original Dataset:**

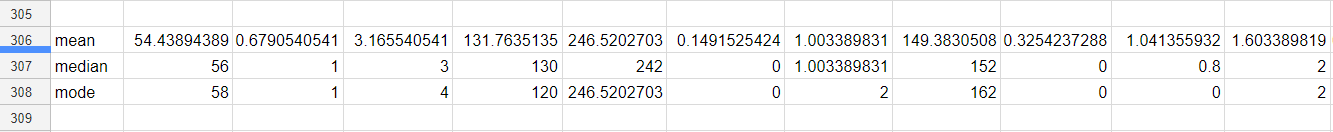
****

**HANDLING MISSING VALUES:-**

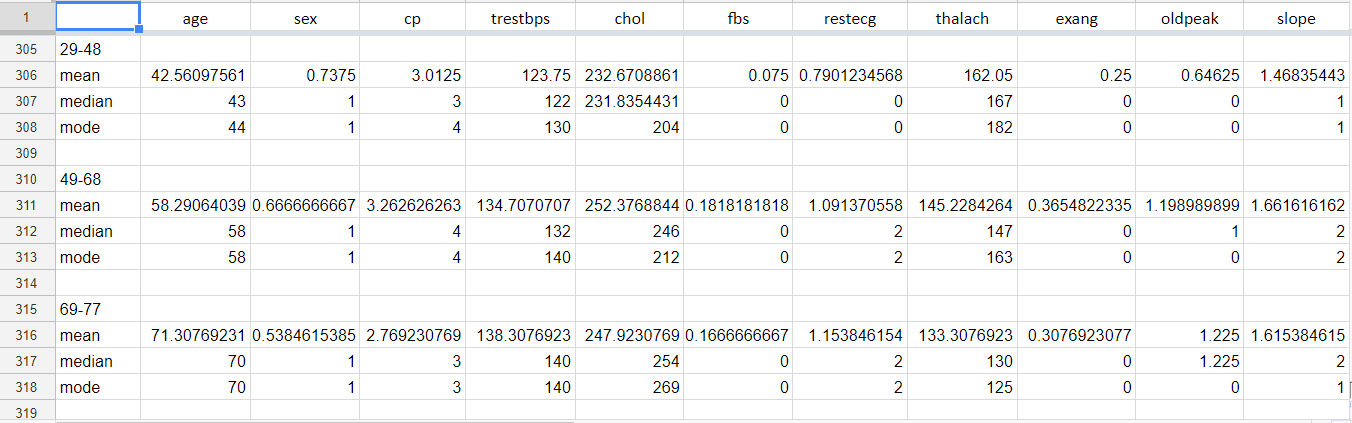
**1) AFTER DELETING THE ROWS WHICH CONTAIN MISSING VALUES :-**

****

**2) AFTER FILLING WITH MEAN VALUES:**

****

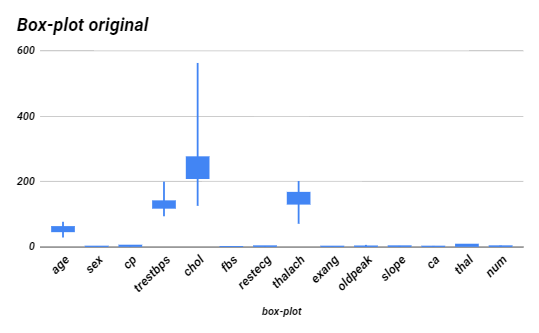
**3) AFTER FILLING WITH CLASS MEAN VALUES :-**

****

Since replacing missing values with mean gives the results similar or approximation to original dataset , we choose **mean substitution for missing values.**

**DETECT NOISY DATA:-**

**1) BOXPLOT :**

****

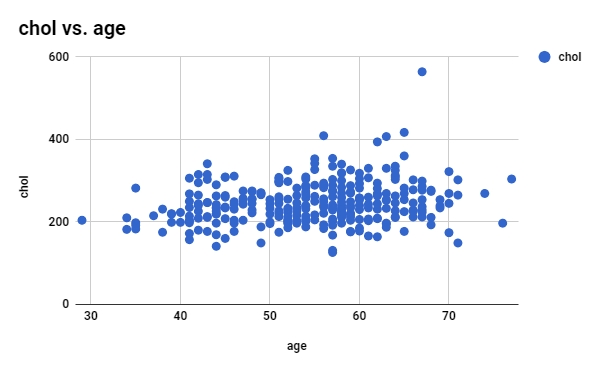
This plotting shows the noisy data and all the attributes that consist of outliers. This can be removed using clustering i.e detecting all the outliers and replacing them with the median value of the respective column.

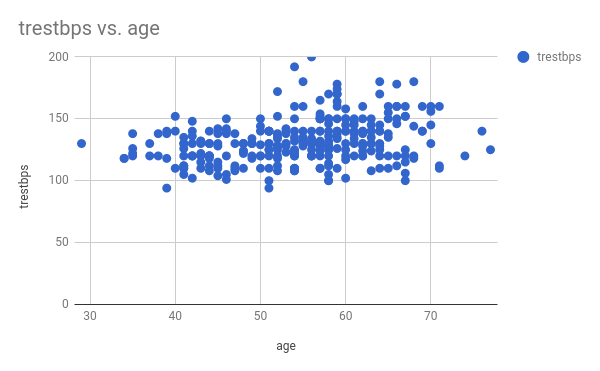
**For the Data set, if missing values are filled with mean of actual data set than the answers of both the dataset are matching.**

**DATA VISUALIZATION:**

**SCATTER PLOT FOR VARIOUS ATTRIBUTES:**

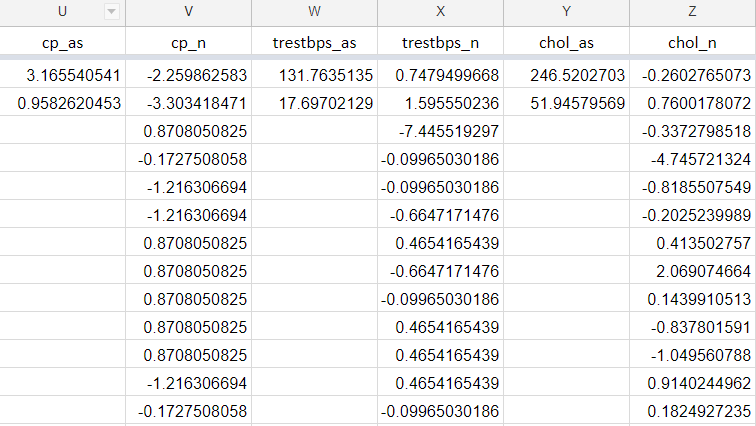
Here , we have certain outliers which can be detected using interquartile range formulae and can be removed by data cleansing technique.

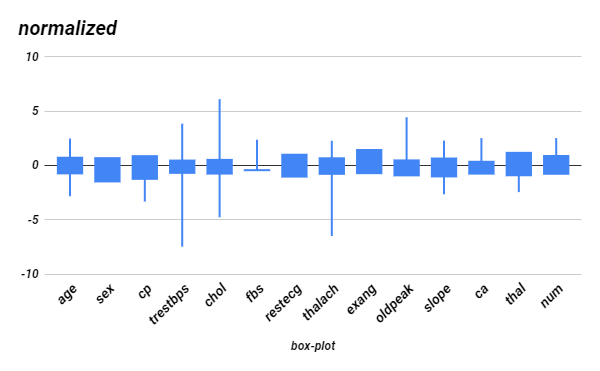
****

****

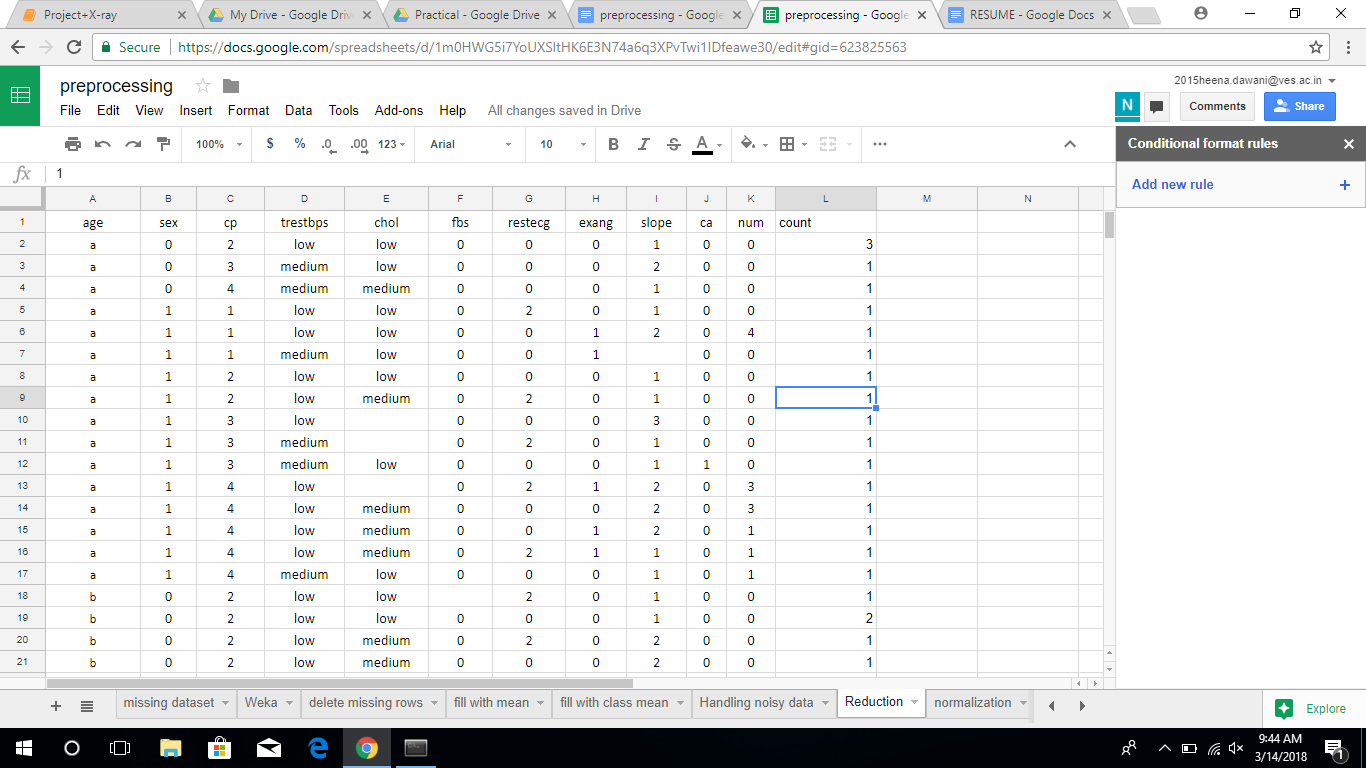
**From the scatterplots of following attributes such as Cholesterol vs age and trestbps vs age we can infer that the maximum number of people suffer from cholesterol and trestbps is in the range of 50-70.**

**NORMALIZATION:**

****

****

**Data Reduction**

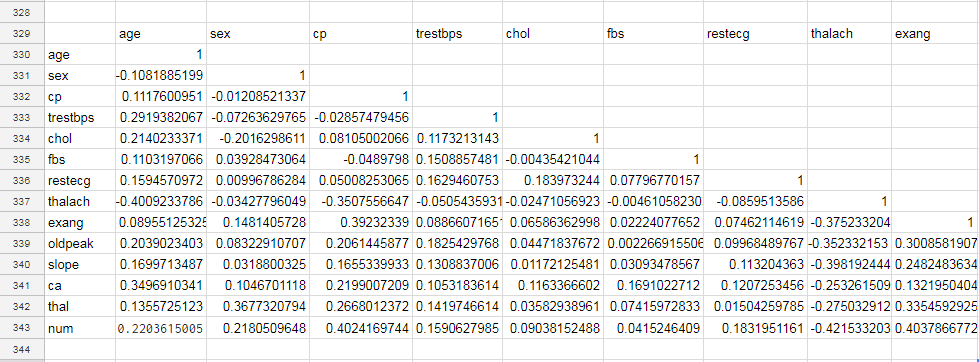
****

In data reduction, the columns which does not contribute to any conclusion or were irrelevant were deleted.

Later the numeric data was converted in categorical data and the redundant rows were deleted.

Around 20 rows were deleted.

**Correlation Analysis:**

****

a)Age is correlated to chol and trestbps i.e resting blood pressure.

b)sex is correlated to cp i.e chest pain type.

c) cp is correlated with thalach i.e maximum heart rate achieved.