**MDSC-102-Final Lab Test Report**

Name : BASIREDDY LOKANATH REDDY

Regd No : 23901

Subject Code : Inferential Statistics(lab)-MDSC-102-Lab

**Problem Statement:**

**Taken a Dataset: Smoking and Drinking dataset with Body signals.**

1. **Done a Pre-processing.**
2. **Done a visualization for the data from the dataset using mat plot library by plotting the suitable graphs for the data.**
3. **After visualization, make the inferences from the plots and checked the Relations.**
4. **Next, seen that the dataset features are normal or skewed if they are not normal then we made normal.**
5. **Taken two features and done a Z-test and T-test for that features.**

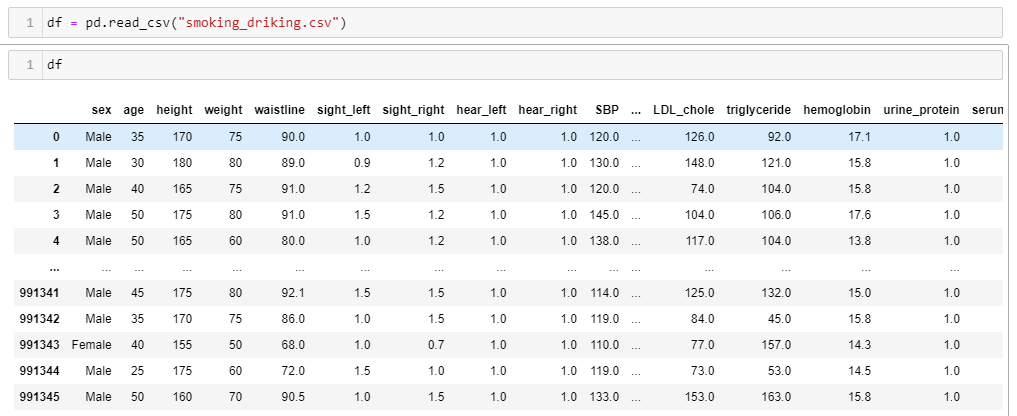
Dataset: [data set link](https://github.com/Loknath07/MDSC-102-FinalLabTest/blob/main/smoking_driking_dataset_Ver01.csv)

This dataset is someone collected from National Health Insurance Service and kept in web and I have taken from there.

The purpose of this data set is to analysis of body signal of smokers and drinkers.

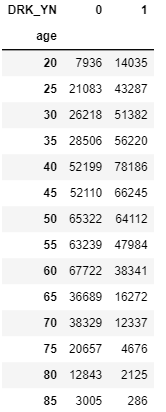
**Data Pre-processing:**

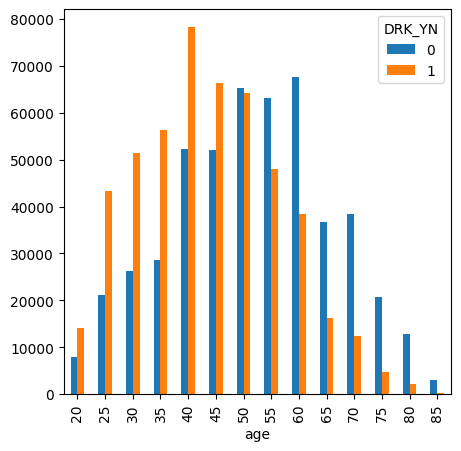
we read the data as a CSV file:



****In this dataset for features **SMK\_stat\_type\_cd** , **DRK\_YN** s and **sex** the values are not in numerical. It is easy to work with the dataset where the most of the features in the dataset are the numerical. Therefore, all that features we convert into float type.

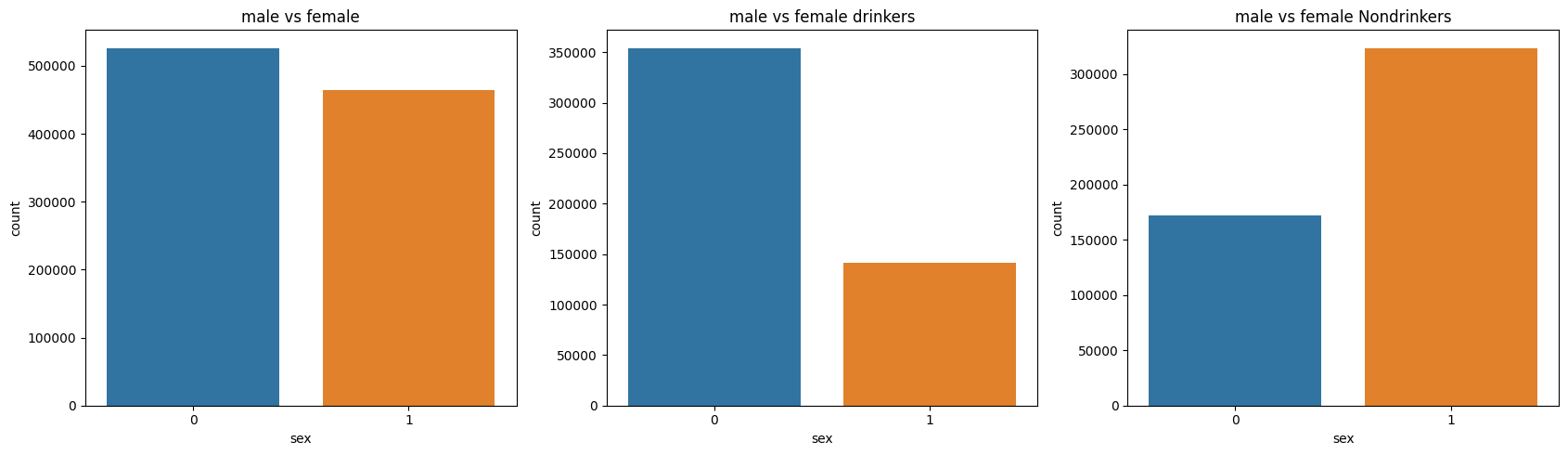
We converted the values of above mentioned features into float.

Next, done some basic thing that seeing is there any missing values, seeing the statistical data and information of the data (where we get mean, median, std, datatype and etc.)

**Visualization:**

**Drinkers and Non-Drinkers according to the age:**

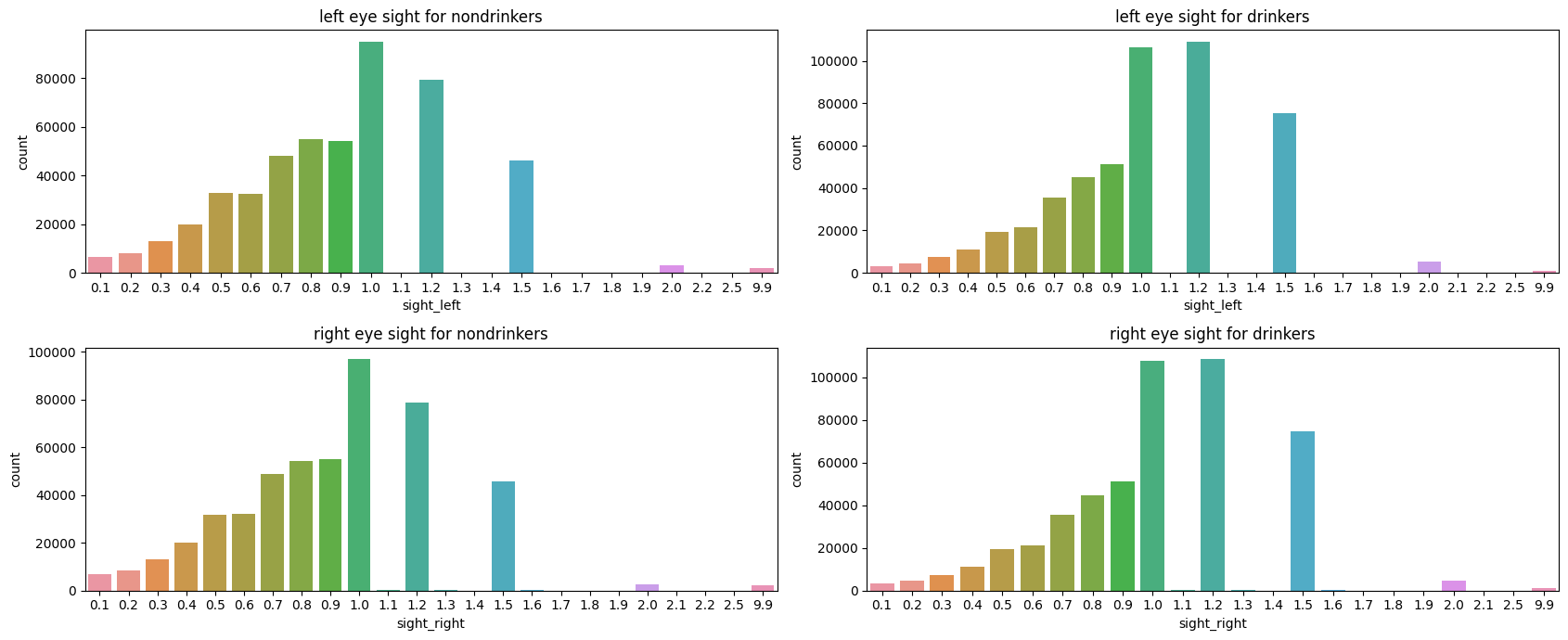
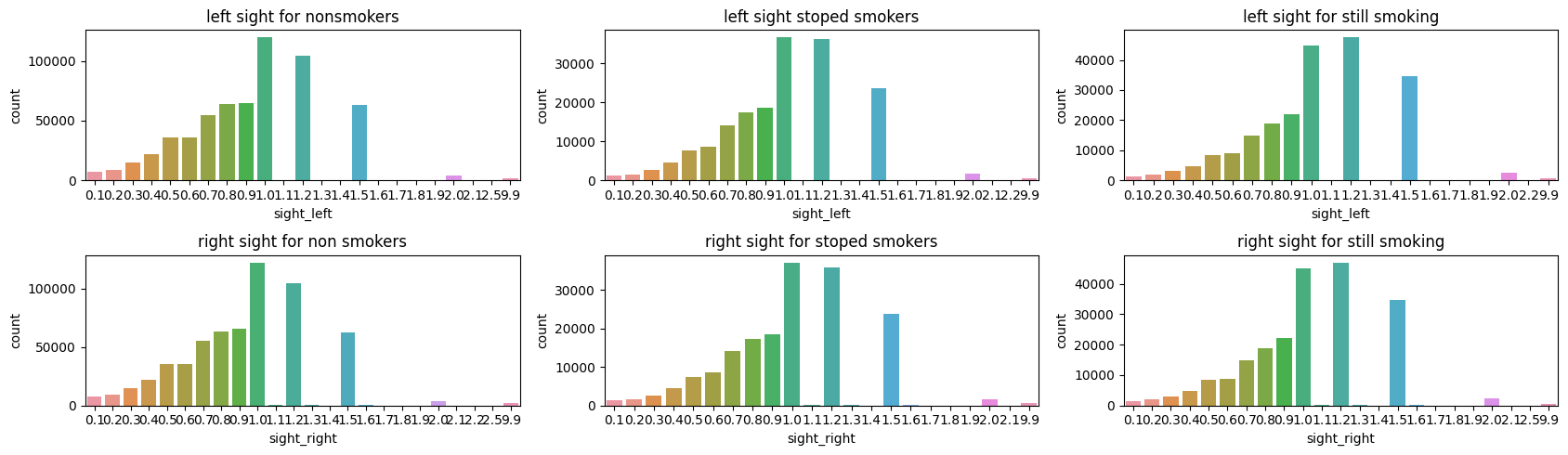
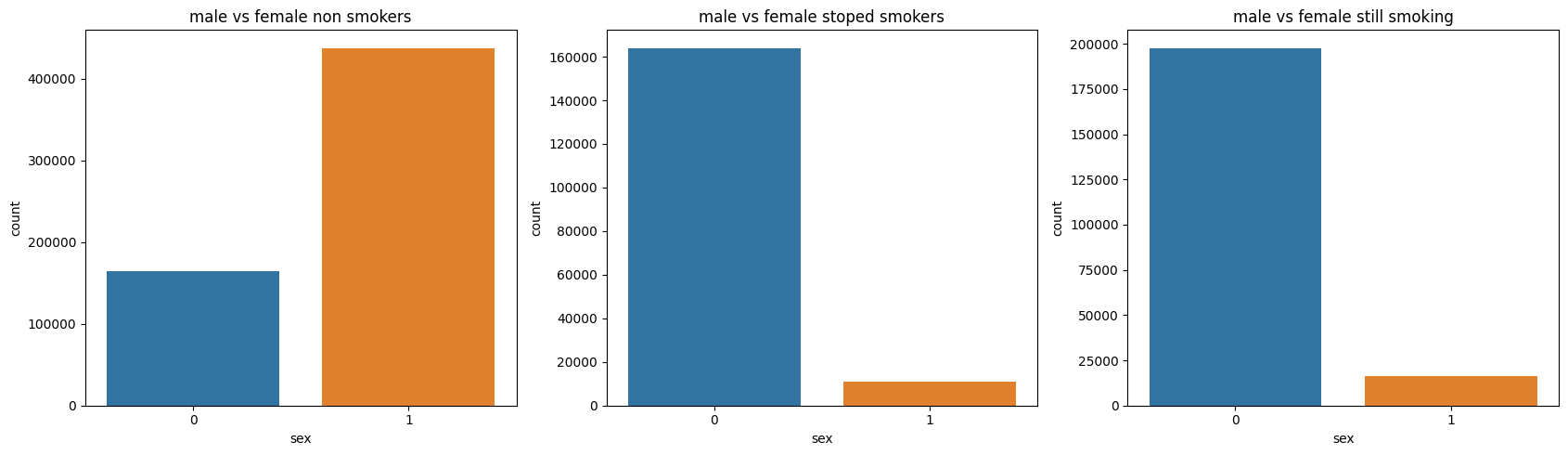
The above plotted graph is the drinkers in the dataset where 0(blue bar) mention the non-drinkers and the 1(Orange bar) mention the drinkers. We can see that the number of drinkers decreases with age from 40 to 85 and increase with age from 20 to 40. The number of non-drinkers increases with age from 20 to 60 and decreases with age from 60 to 85. The largest number of drinkers are in the 35-45 age range. The largest number of non-drinkers are in the 40-50 age range. These inference are written by see in above plot as in numbers we can see beside the plot table for clarity.

**Male vs Female: All, Drinkers and Non-drinkers**

By seeing the above plot, we can say that in this data set there 5,00,000 males and 4,50,000 females are there. there are more male drinkers than female drinkers. there are less male drinkers than female drinkers.

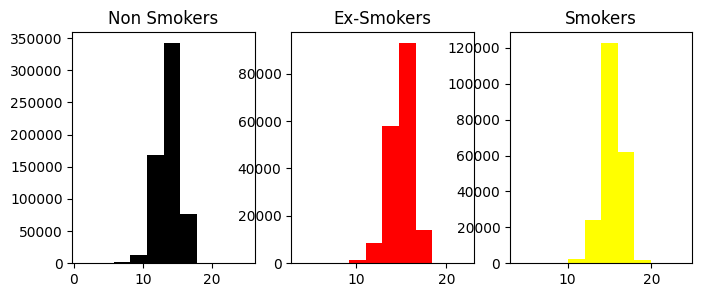
**Male vs Female: Non-smokers, stopped Smokers and still Smoking people**

The plot is given below for smoking description in thatthere are more females as non-smokers as compared to males. almost all the females that who have smoking habit the are stopped. Males only 80 percent people stopped smoking in middle. A small proportion of women are smokers, whereas among men, there is a relatively equal distribution, with most being either current smokers or individuals who have smoked in the past.

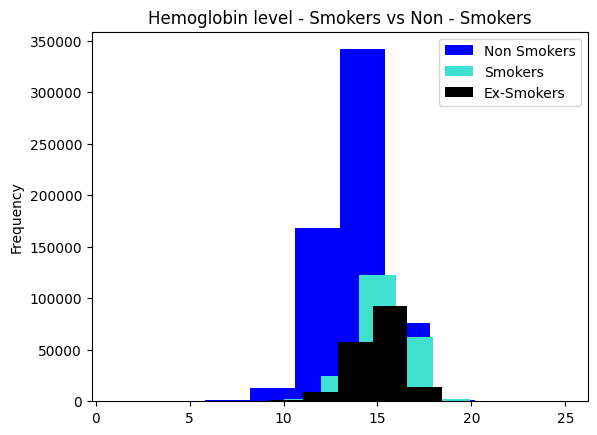
******Comparing:**

From the above 1st plot comparing people who drink and non-drinkers there is no big difference for eye sight for both and right and left. So we can say that by drinking there is no problem for eye sight. From 2nd plot there is no much relationship for smoker and non-smokers and eye sight.

**Hemoglobin levels for Non-Smokers, Ex-Smokers and still Smokers:**

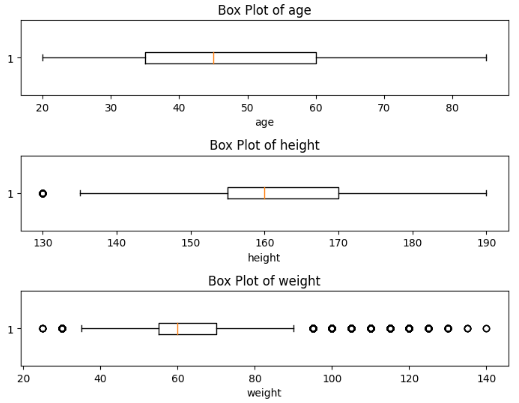
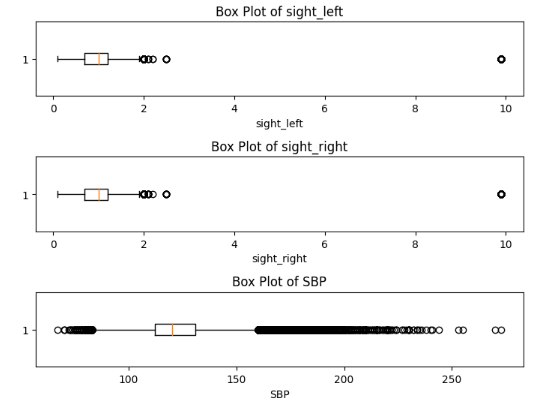


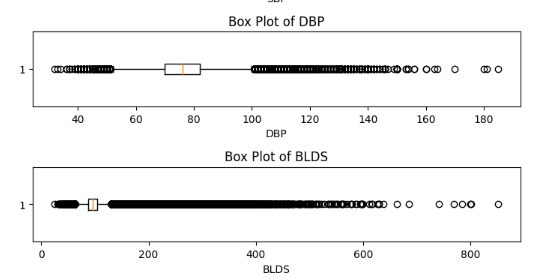
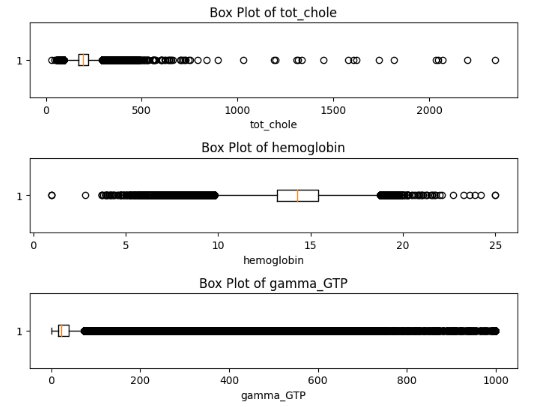
* Non-smokers have higher hemoglobin levels than smokers and ex-smokers.
* Ex-smokers have higher hemoglobin levels than smokers.
* Smoking is associated with lower hemoglobin levels.
* From the above plot we can say that quitting smoking can lead to an increase in hemoglobin levels.



* We can see clearly by keeping all the three graphs in one graph, where Non-smokers have higher hemoglobin levels than smokers and ex-smokers.

**Checking the data is in normal or not in normal and if it is not converted into normal:**

**Done a Boxplot:**



**Inferences for above box plots:**

**Age:** The box plot is skewed to the right, with a few outliers on the right side. The median is around 40 years old. This suggests that the data is not normally distributed.

**Height:** Box plot skewed right, with outliers. Median height is around 170 cm. Not normally distributed.

**Weight:** Box plot skewed right, with outliers. Median weight is around 70 kg. Not normally distributed.

**Sight left:** Box plot skewed right, with outliers. Median sight left is around 1.00. Not normally distributed.

**Sight right:** Box plot skewed right, with outliers. Median sight right is around 1.00. Not normally distributed.

**SBP:** Box plot skewed right, with outliers. Median SBP is around 150 mmHg. Not normally distributed.

**DBP:** Box plot skewed right, with outliers. Median DBP is around 80 mmHg. Not normally distributed.

**BLDS:** Box plot skewed right, with outliers. Median BLDS is around 300 mg/dL. Not normally distributed.

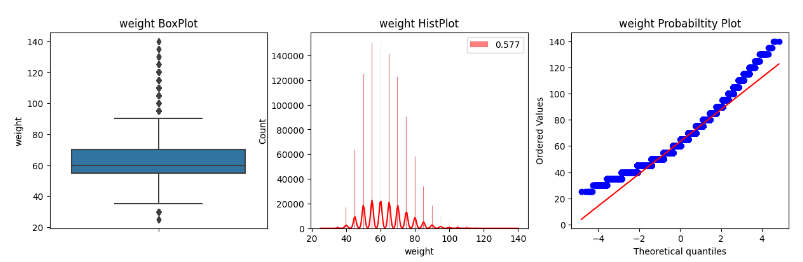
**Total cholesterol:** Box plot is skewed right, with outliers. Median total cholesterol is around 500 mg/dL. Not normally distributed.

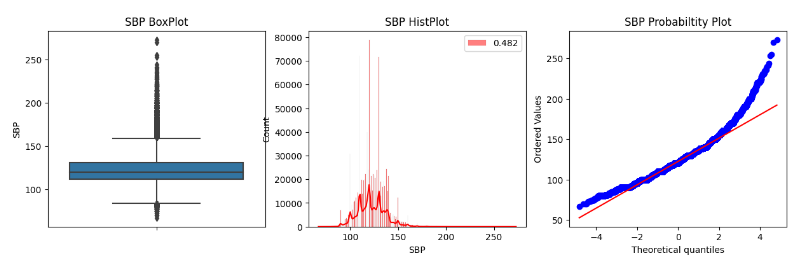
**Hemoglobin:** Box plot is skewed right, with outliers. Median hemoglobin is around 15 g/dL. Not normally distributed.

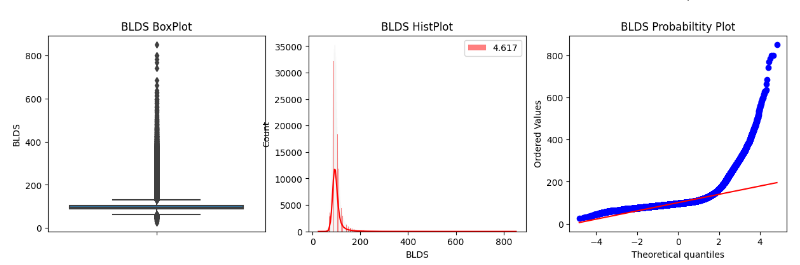
**Gamma GTP:** Box plot is skewed right, with outliers. Median gamma-glutamic transferase (GGT) is around 200 U/L. Not normally distributed.

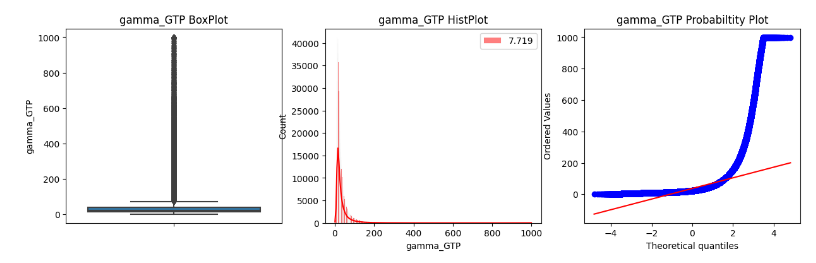
**Plotted box plot, Histogram plot and probability plot for the above features to see clearly weather they are in normal or not normal:**

Taken some features that are not formal from the given dataset:

**1) Weight**

**2) SBP **

**3) BLDS**

**4) GTP **

# Since the above features are not in normal we convert them into normal by using” Boxcox”.

# 

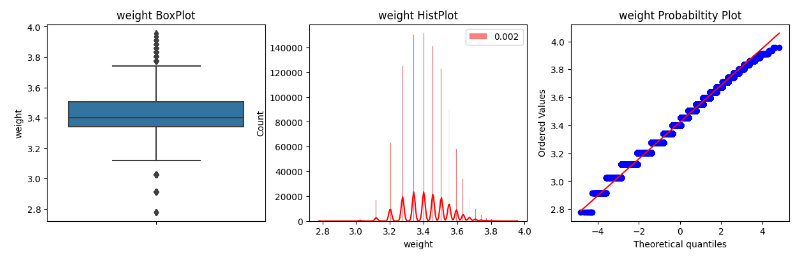
# After converting them into normal the by changing the skewness in data points

# 

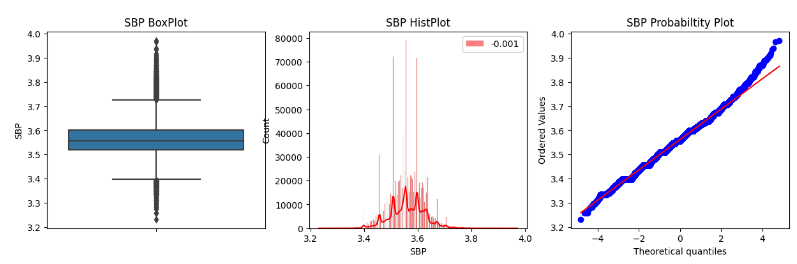
# We can observe that there is a change in the skewness of the data points after transformation.

# Box plot, Histogram plot and Probability plot for the features after changing the skew:

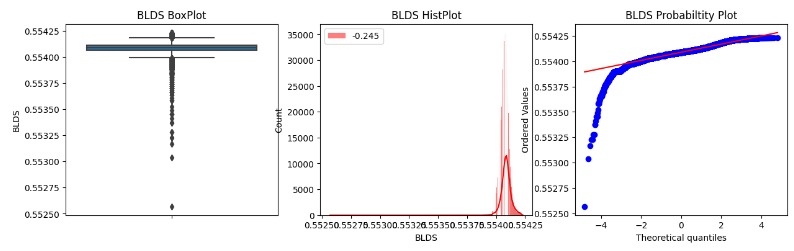
**1) Weight**

****

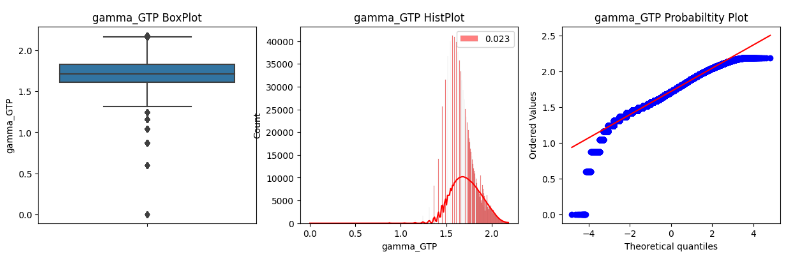
**2) SBP**

****

**3) BLDS**

****

**4) GTP**

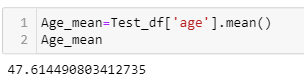
****

By comparing the 4 graphs after and before doing skew we can clearly observe that data is normalized.

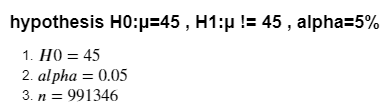
# Testing Hypothesis:

Let us choose 2 features **Age** and **Height** from dataset and done testing hypothesis.

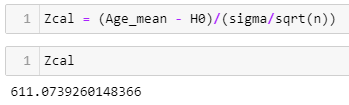
**Z-test:**

****

From the data set taken age feature and calculated mean for that, and hypothesis is:



We performed the z-test using its formula, the z value is calculated:

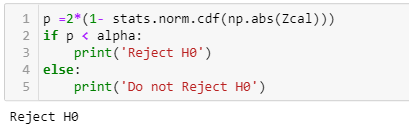


Z-calculated value is 611.07

We find a p value and we check p value is > or < alpha,

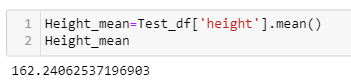
If p value is < alpha, then reject H0

If p value is > alpha, then do not reject H0

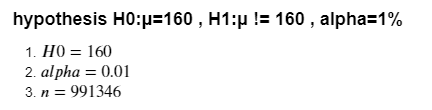


It is rejected since the mean of age is 47.614490803412735 but we have taken 45.

**t-test:**



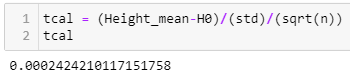
From the data set taken Height feature and calculated mean for that, and hypothesis is:



Since we don’t know variance first we find a variance:



We performed the t-test using its formula, the t value is calculated:

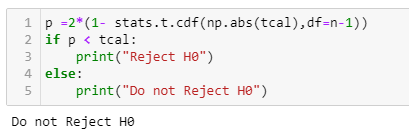


t-calculated value is 0.00024

We find a p value and we check p value is > or < alpha,

If p value is < alpha, then reject H0

If p value is >alpha, then do not reject H0



It is not rejected since the mean of height is 162 cm and we have taken nearer to that 160.

-------------------------------------- Thank You-----------------------------------