**Basic statistics 1**

|  |
| --- |
|  |
|  | | |

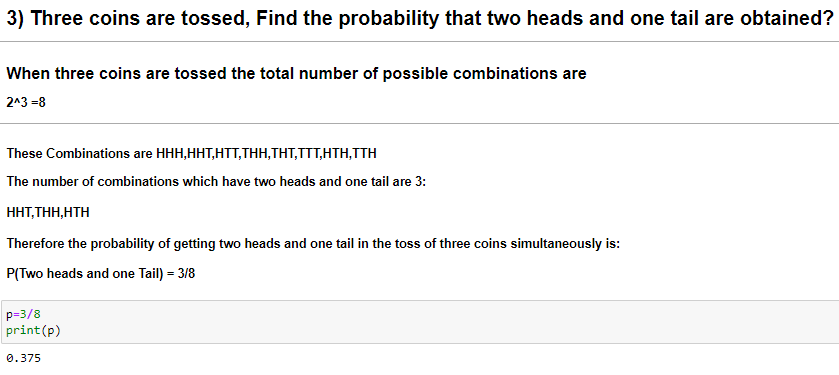
1. **Identify the following datatypes**

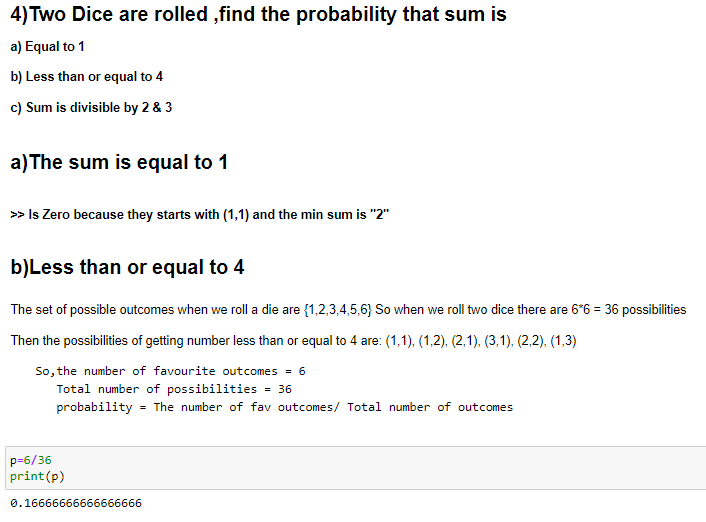
|  |  |  |
| --- | --- | --- |
| Activity | | Data Type |
| Number of beatings from Wife | | Discrete |
| Results of rolling a dice | | Ordinal |
| Weight of a person | | Ratio |
| Weight of gold | | Ratio |
| Distance between two places | | Ratio |
| Length of a leaf | | Ratio |
| Dog’s weight | | Ratio |
| Blue Colour | Nominal | |
| Number of Kids | Discrete | |
| Number of Tickets in Indian railways | Ordinal | |
| Number of times married | Discrete | |
| Gender (male or female) | Nominal | |

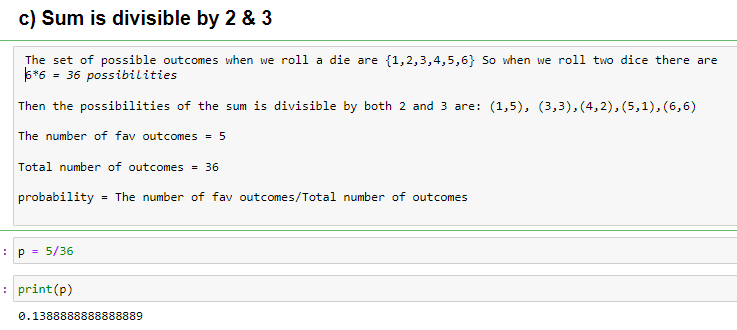
2) Identify the Data types, which were among the following

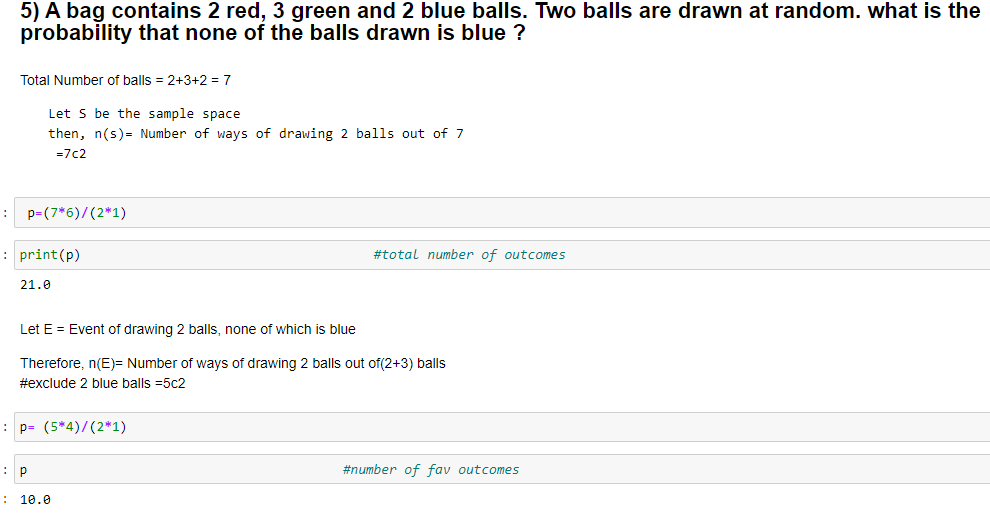
Nominal, Ordinal, Interval, Ratio.

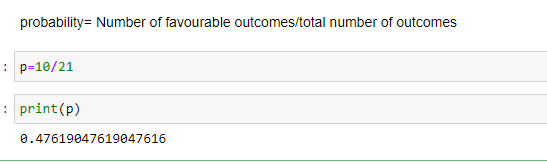
|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High school class ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Colour | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figure | Interval |
| Blood Group | Nominal |
| Time of Day | Ratio |
| Time on a clock with hands | Ratio |
| Number of children | Ordinal |
| Religious preference | Nominal |
| Barometer pressure | Ratio |
| SAT scores | Ratio |
| Years of education | Interval |

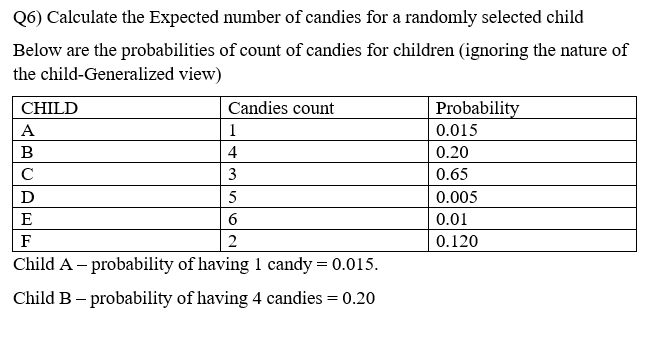


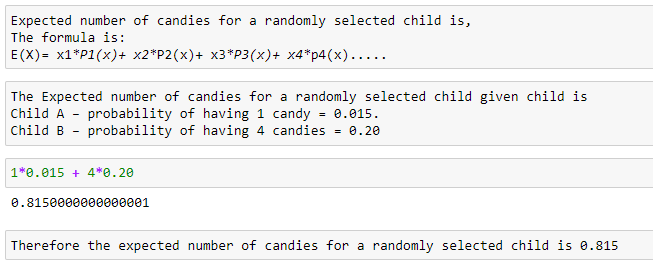


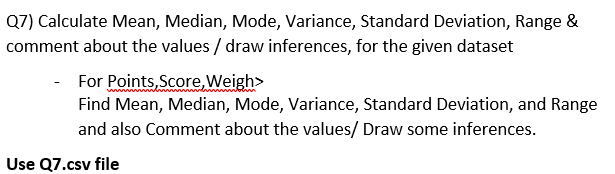


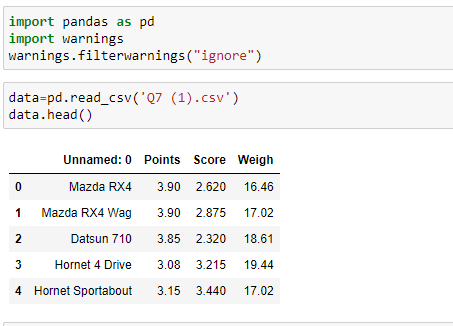


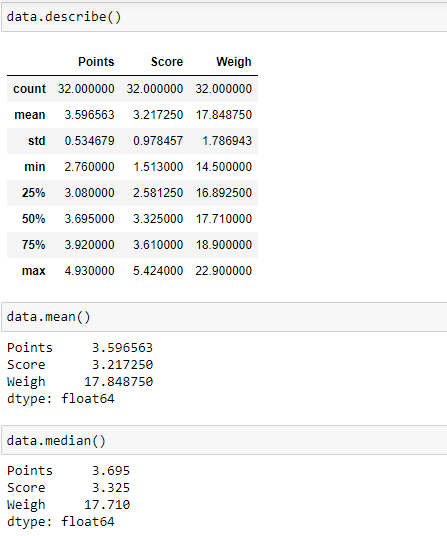


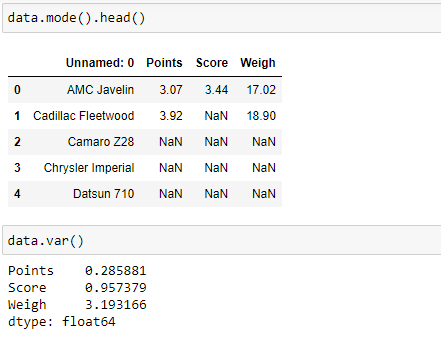




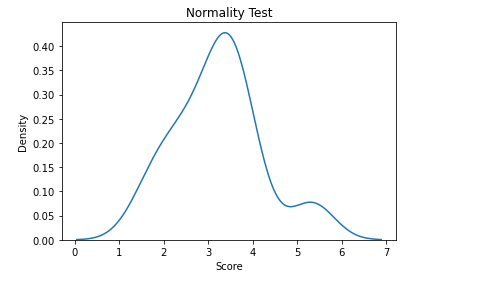


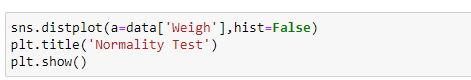


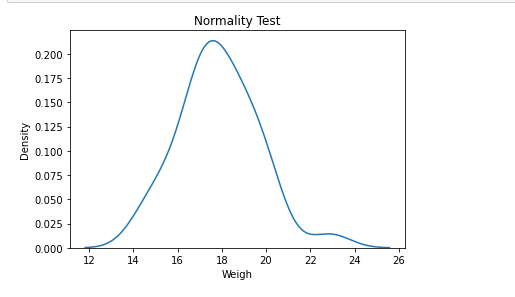


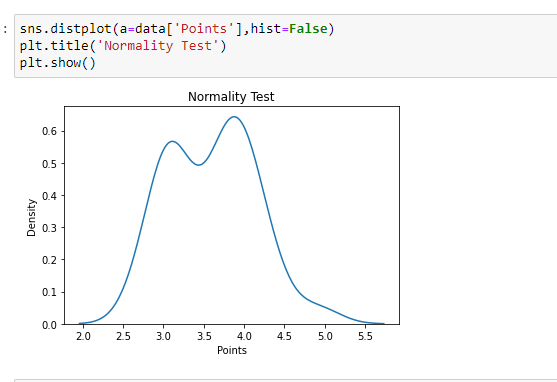


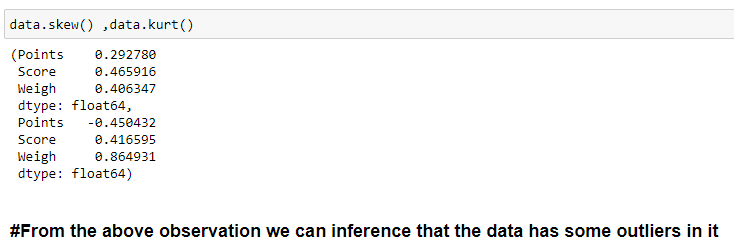




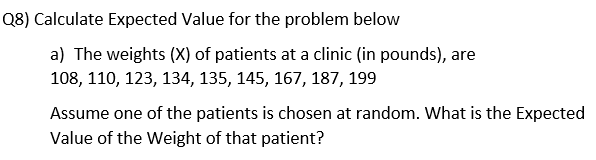


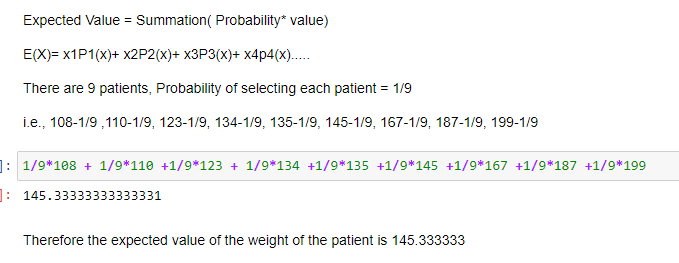


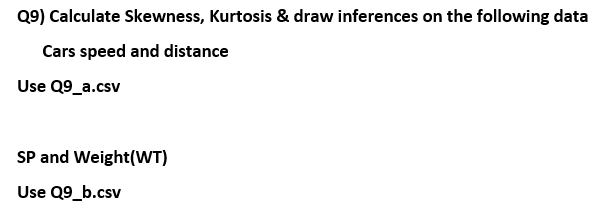


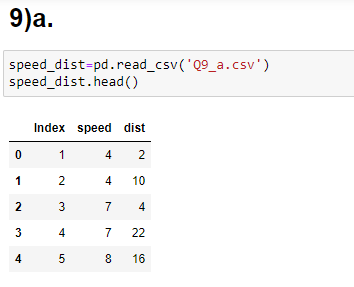


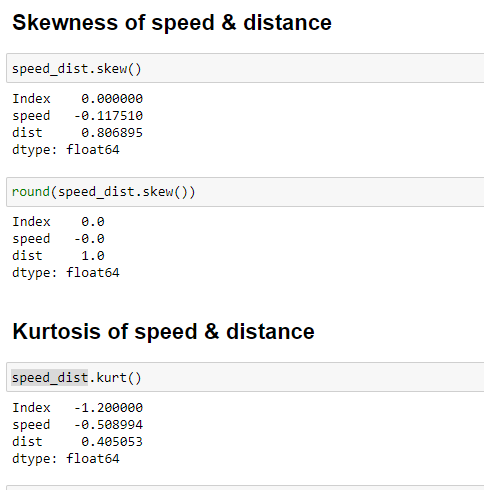
**And due to some sort of skewness we can say that it is not normally distributed**

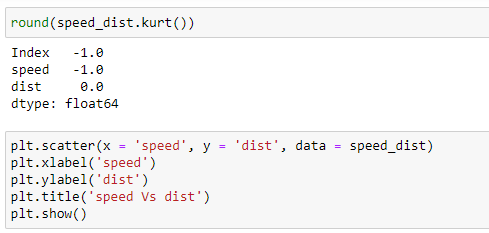


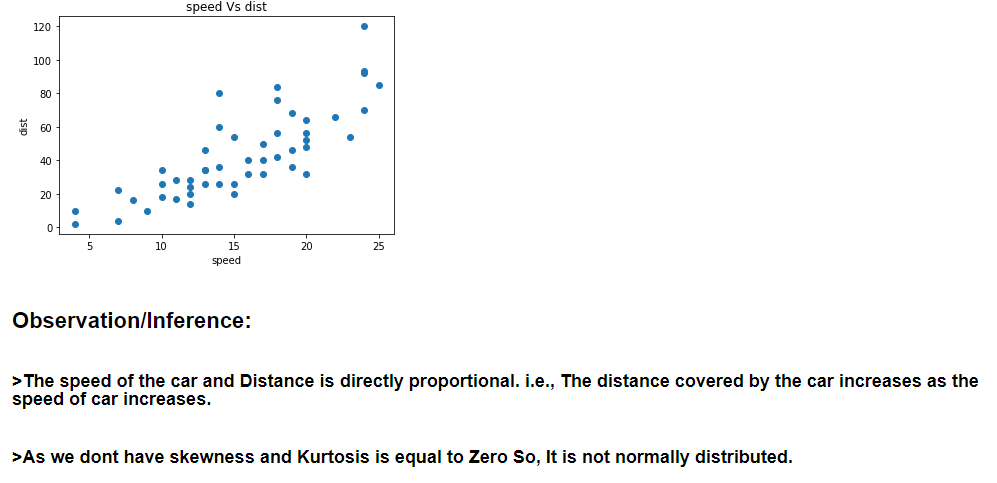


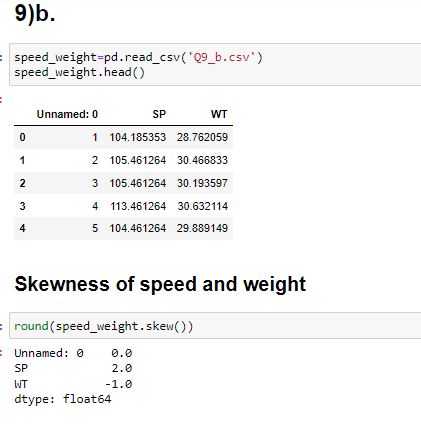


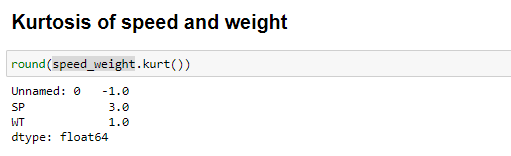


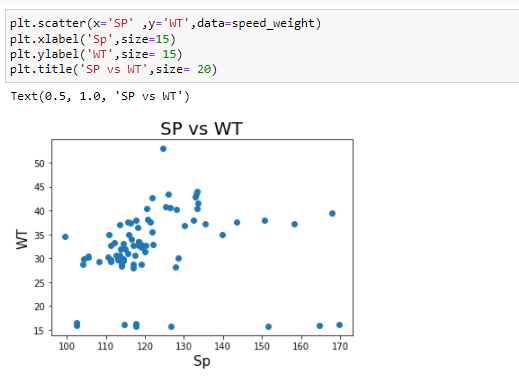


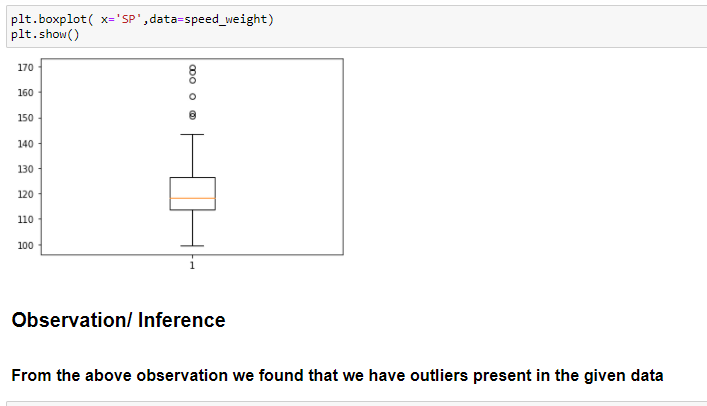




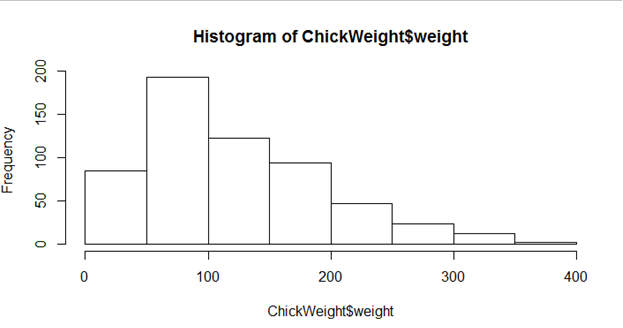








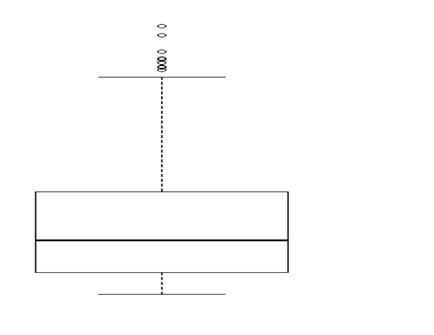




**Observation/ Inference:**

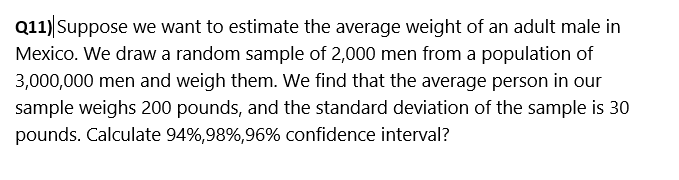
From the above Histogram we can inference that

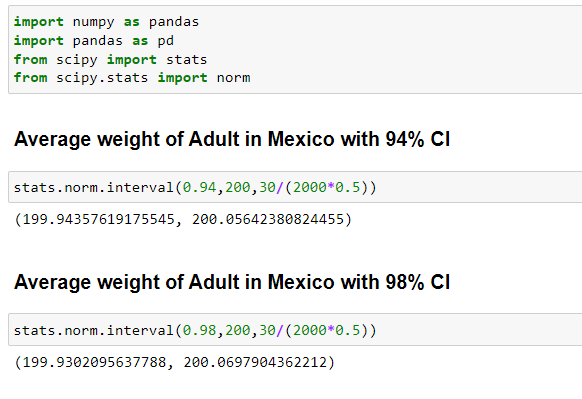
* It is a right-skewed normal distribution
* It is a univariate distribution of chickweight$weight
* As the weight of the chicken increased frequency has decreased
* We can conclude that there are outliers present in the distribution of the data

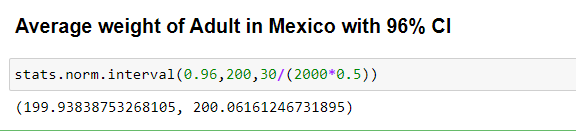


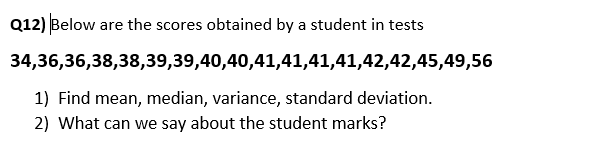
**Observation/Inference:**

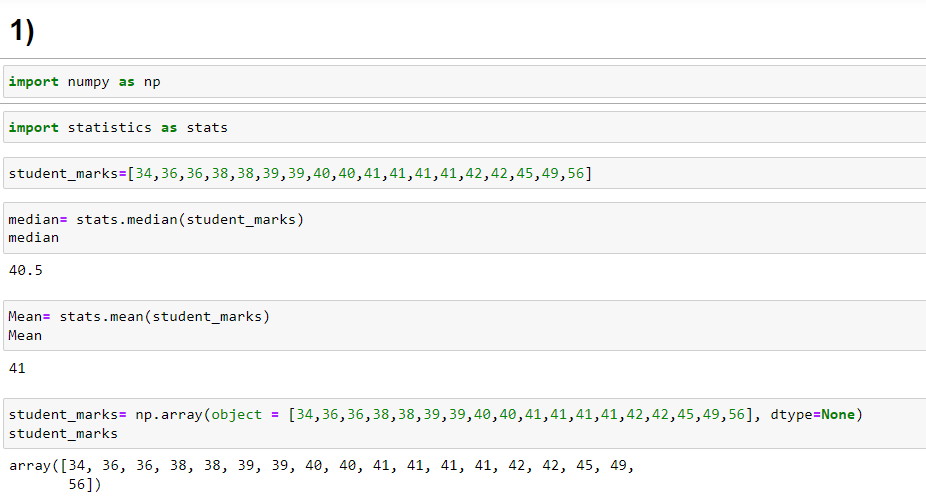
* When the same data is distributed in the form of boxplot we can infer that there are many outliers present in the upper quartile of the boxplot.
* It is mainly used to identify the outliers clearly

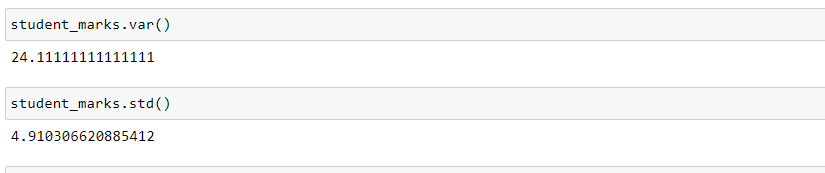










 **2)** The average marks obtained by the student is ‘41’, and the highest marks obtained is 56

**Q13) What is the nature of skewness when mean, median of data are equal?**

If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the

Mean= median= mode

**Q14) What is the nature of skewness when mean > median?**

If the mean is greater than the median, the distribution is positively skewed. In this the tail will be on the right side of the distribution.

**Q15) What is the nature of skewness when median > mean?**

If the median is greater than the mean or mean is less than median the distribution is negatively skewed. In this the tail will be on the left side of the distribution.

**Q16) What does positive kurtosis value indicates for a data?**

Positive excess values of kurtosis (>3) indicate that distribution is peaked and possesses thick tails. A leptokurtic distribution has a higher peak (thin bell) and a taller tail than a normal distribution.

**Q17) What does negative kurtosis value indicates for a data?**

Negative excess values of kurtosis (<3) indicate that distribution is flat and has thin tails. A platykurtic distribution is a flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter tails.

**Q18) Answer the below questions using the below boxplot visualization.**

****

**What can we say about the distribution of the data?**

**What is nature of skewness of the data?**

**What will be the IQR of the data (approximately)?**

>> It is a boxplot distribution, that helps to visualize the variability of a distribution. It displays data based on five number

“Minimum”

“First quartile”

“Median”

“Third quartile”

“Maximum”

>> It is a negatively skewed distribution.

>>IQR is interquartile range,

IQR= q3-q1=18-10=8

the value is lower than the 1.5\*IQR below the lower quartile(Q1), the value will be considered as an outlier.

**Q19) Comment on the below Boxplot visualizations?**

****

**Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.**

>>The data in both the boxplot are normally distributed.

>>Both Boxplots when plotted as a normal distribution we can infer that Boxplot1 might be leptokurtic, & Boxplot2 might be Mesokurtic.

>>The interquartile range of boxplot1 is lower when compared to boxplot2, i.e., (Q3-Q1)

>>50% of the data lies above the median and 50% of the data lies below the median

**Q 20) Calculate probability from the given dataset for the below cases**

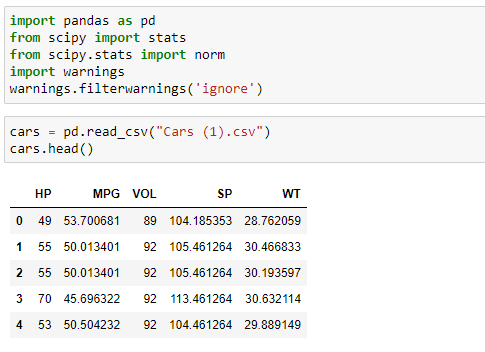
**Data \_set: Cars.csv**

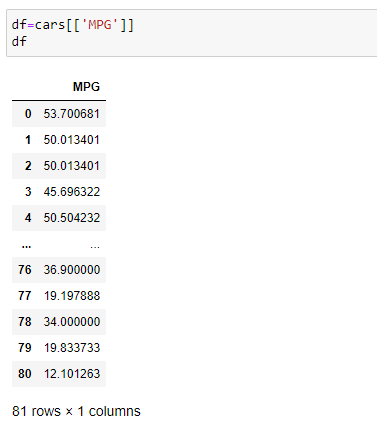
**Calculate the probability of MPG of Cars for the below cases.**

**MPG <- Cars $MPG**

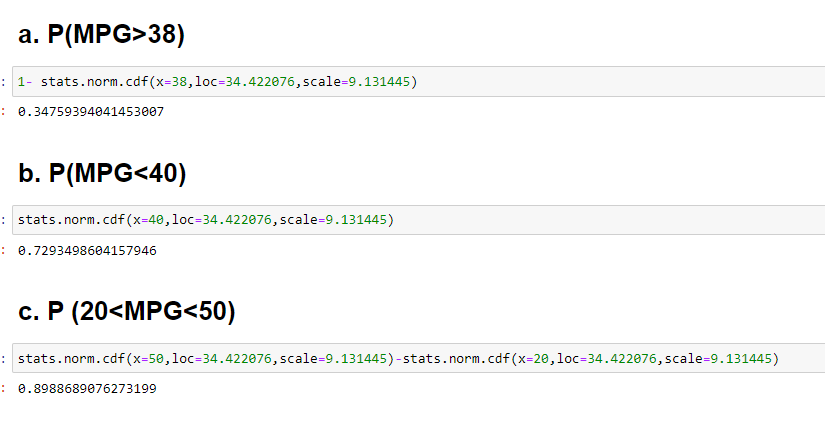
* 1. **P(MPG>38)**
  2. **P(MPG<40)**

**c. P (20<MPG<50)**

****

****

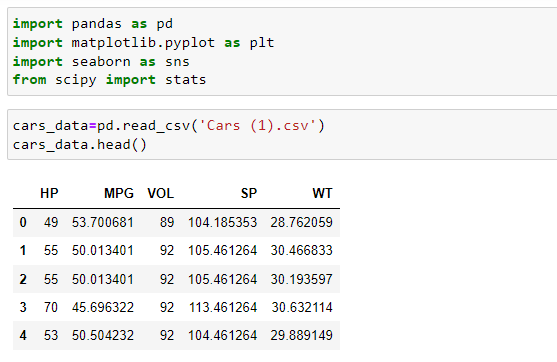
****

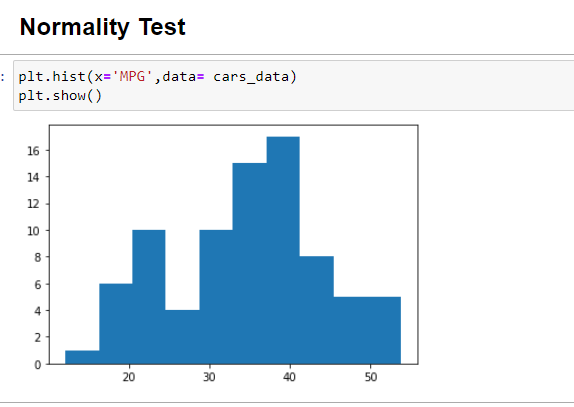
****

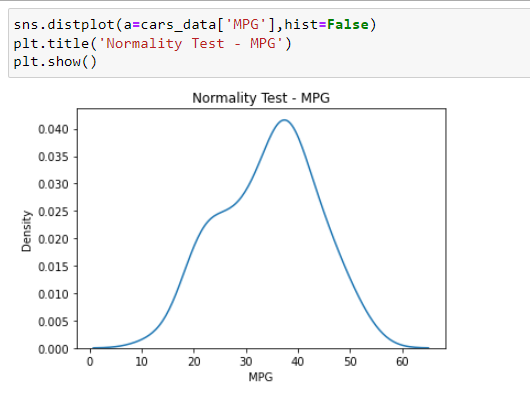
**Q 21) Check whether the data follows normal distribution**

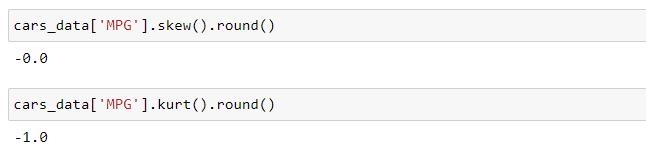
1. **Check whether the MPG of Cars follows Normal Distribution**

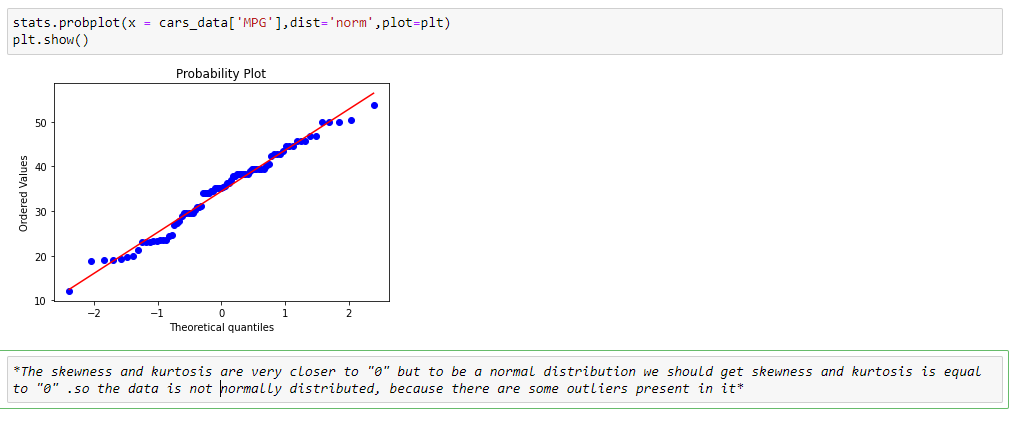
**Dataset: Cars.csv**

****

****

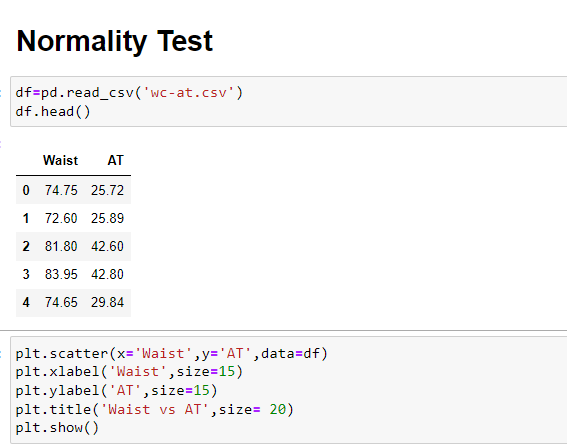
****

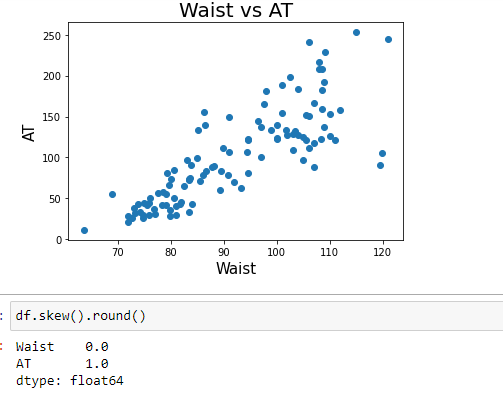
****

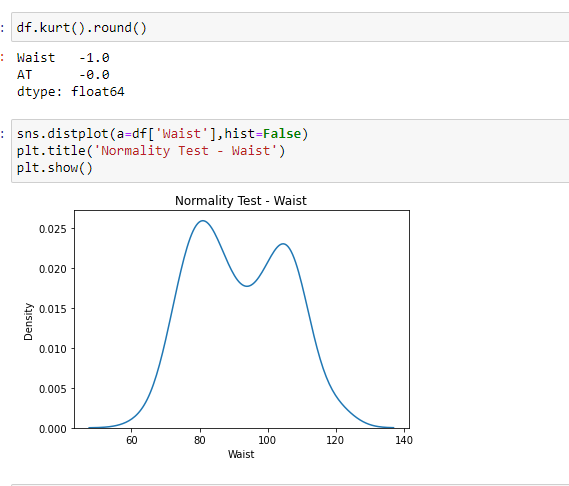
****

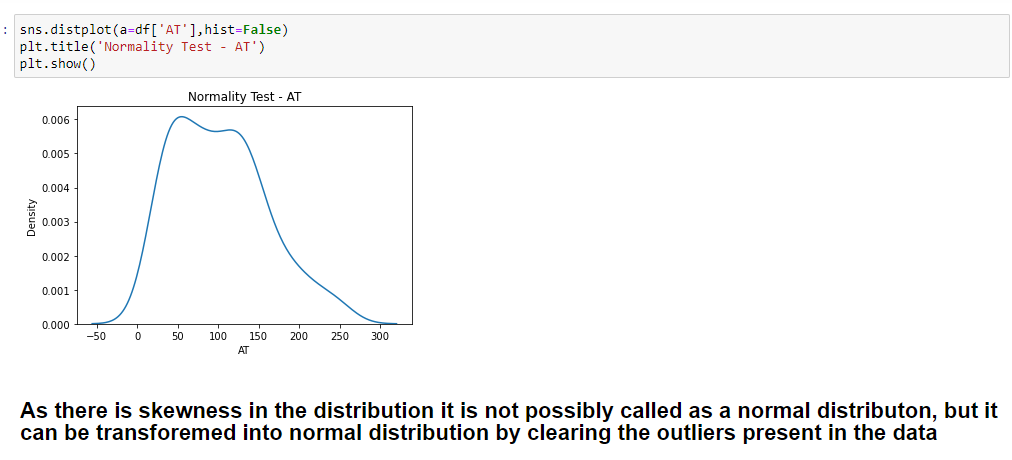
1. **Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution**

**Dataset: wc-at.csv**

****

****

****

****

**Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval**

***>> Z score of 90% confidence interval is 1.645***

***>>Z score of 94% confidence interval is 1.880***

***>>Z score of 60% confidence interval is 0.253***

**Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25**

Given sample size(n) = 25

Formula = t(confidence interval , n-1)

**95% confidence interval:** t(95%, (25-1))

t(95%, 24)

= 2.0639

**96% confidence interval:** t(96%, (25-1))

t(96%, 24)

=2.172

**99% confidence interval:** t(99%, (25-1))

t(99%, 24)

= 2.797

**Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days**

**Hint:**

**rcode 🡪 pt(tscore,df)**

**df 🡪 degrees of freedom**

Given

A Government company claims that an average light bulb lasts 270 days.

A researcher randomly selects 18 bulbs for testing

The sampled bulbs last an average of 260 days, with a standard deviation of 90 days.

t- statistics for the data is given as follows:

x=mean of the sample of bubs = 260

u= population mean = 270

s= standard deviation of the sample =90

n= number of items in the sample = 18

t= -0.471

For probability calculations, the number of degrees of freedom is n-1, so here we need the t-distribution with 17 degrees of freedom.

The probability that t<-0.471 with 17 degrees of freedom assuming the population means is true, the t- value is less than the t-value obtained with 17 degrees of freedom and a t score of -0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulbs is 300 days.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*THE END\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*