|  |  |
| --- | --- |
| **Activity** | **Data Type** |
| Number of beatings from Wife | Numeric Data - Discrete |
| Results of rolling a dice | Numeric Data - Discrete |
| Weight of a person | Numeric Data - Continuous |
| Weight of Gold | Numeric Data - Continuous |
| Distance between two places | Numeric Data - Continuous |
| Length of a leaf | Numeric Data - Continuous |
| Dog's weight | Numeric Data - Continuous |
| Blue Color | Categorical Data |
| Number of kids | Numeric Data - Discrete |
| Number of tickets in Indian railways | Numeric Data - Discrete |
| Number of times married | Numeric Data - Discrete |
| Gender (Male or Female) | Categorical Data |

*Q1) Identify the Data type for the Following:*

*Q2) 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 Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

*Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?*

**Analysis:**

Let, H = Heads, T = Tails

Possible outcomes:

(H, H, H), **(H, H, T)**, **(H, T, H)**, (H, T, T), **(T, H, H)**, (T, H, T), (T, T, H), (T, T, T)

No. of **possible outcomes**: 8

Number of outcomes that gives two heads and one tail = 3

i.e, (H,H,T), (H,T,H), (T,H,H)

Thus, number of **favorable outcomes** = 3

Probability of an event (E) = No. of favorable outcomes / No. of possible outcomes

= **3/8**

Thus, the probability of getting two heads and one tail on tossing three coins at once is equal to **3/8**

*Q4) Two Dice are rolled, find the probability that sum is*

1. *Equal to 1*
2. *Less than or equal to 4*
3. *Sum is divisible by 2 and 3*

**Analysis:**

* With two dice, there are (6) \* (6) = **36** possible combinations of numbers.
* The 36 combinations which meet this requirement are given below:

(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)

(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)

(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5) (4, 6)

(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)

(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6).

* The minimum sum possible for the two dice thrown is (1, 1) = a sum of 2
* The maximum sum possible for the two dice thrown is (6, 6) = a sum of 12.

a) The probability that sum is **equal to 1**

The minimum possible sum is (1, 1) = 2

P (1) =0

i.e., not possible that sum always exceed to 1

Therefore P (1) = 0/ 36 = **0**

b) The probability that sum is **less than or equal to 4**

* A probability that sum is less than 4 can be achieved with number combinations (1, 1), (1, 2) and (2, 1), that is only with **3** combinations of numbers.
* A probability that sum is **equal to 4** can be achieved with number combinations (2, 2), (1, 3) and (3, 1), that is only with **3** combinations of numbers.

Therefore, P (**less than or equal to 4**) = 6 / 36 = 1/12 = 0.083 =**8.33%**

c) The probability that sum is **divisible by 2 and 3**

* A probability that sum is divisible by 2 and 3 can be achieved with number combination (1, 5), (3, 3), (4, 2), (5, 1), (6, 6). Therefore, number of favorable outcomes = 5

Therefore, P (**divisible by 2 and 3**) = 5 / 36 = 0.1389 = **13.89%**

*Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?*

Total number of balls = (2+3+2) = 7

n(S) = Number of ways of drawing 2 balls out of 7 = 7 \*2 = 21

Let E = Event of drawing 2 balls, none of which is blue.

n(E) = Number of ways of drawing 2 balls out of (2+3) balls = 5\*2 = 10

So, probability = 10/21 = 0.47 = **47%**

*Q6) Calculate the Expected number of candies for a randomly selected child*

*Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)*

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

*Child A – probability of having 1 candy = 0.015.*

*Child B – probability of having 4 candies = 0.20*

Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

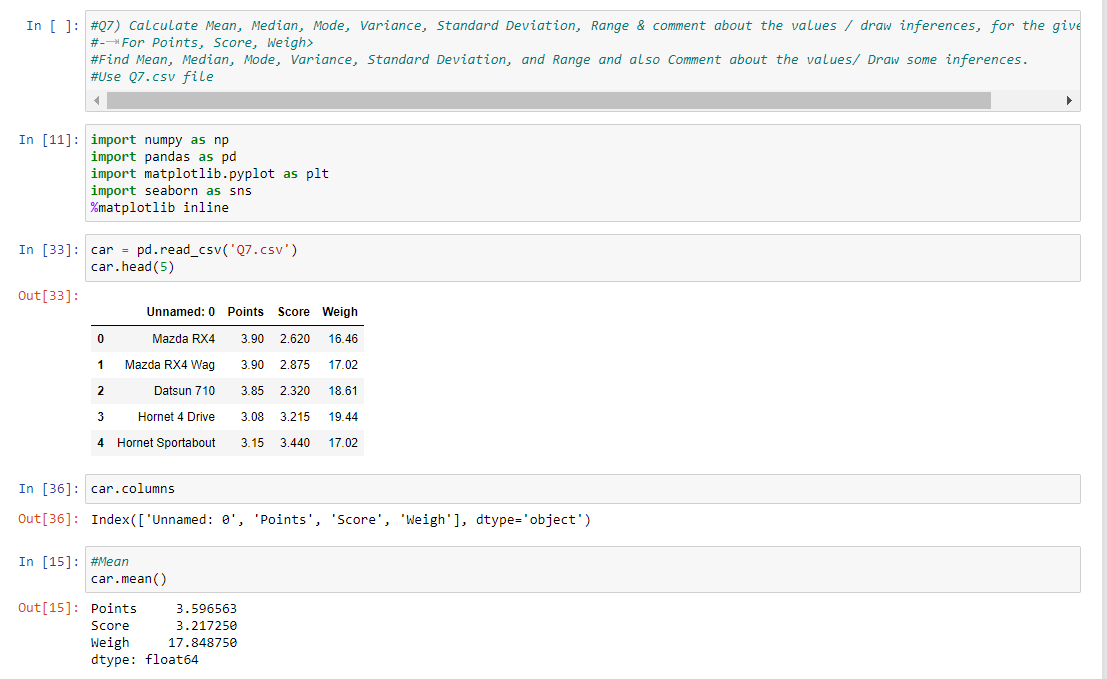
= **3.09**

*Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset*

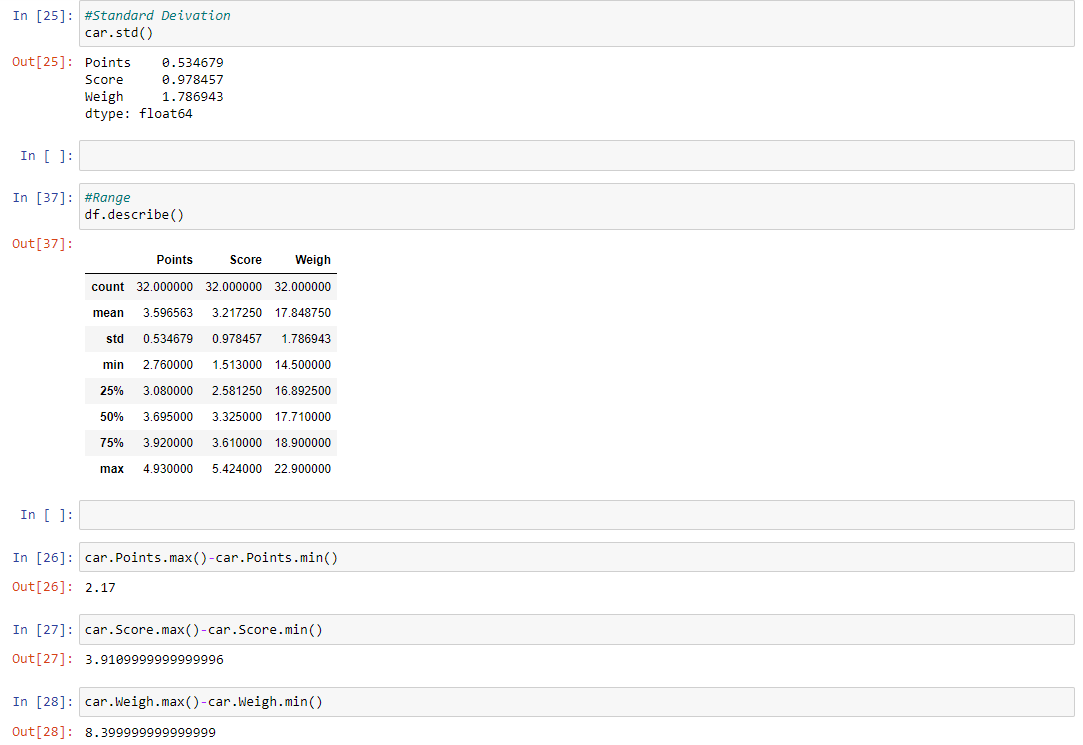
* *For Points, Score, Weigh>*

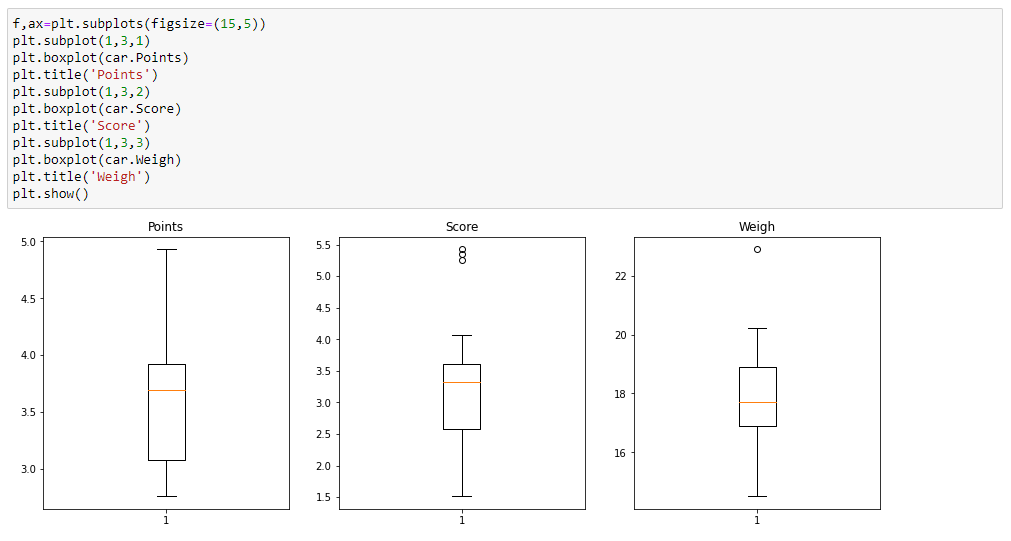
*Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.*

***Use Q7.csv file***









### Inferences:

#### a) For Points dataset:

* The data is concentrated around Median - 3.695
* There are no outliers
* The distribution is Left skewed

#### b) For Score dataset:

* The data is concentrated around Median - 3.325
* There are 3 outliers: 5.250, 5.424, 5.345
* The distribution is Left skewed

#### c) For Weigh dataset:

* The data is concentrated around Median - 17.710
* There are 1 outlier: 22.90
* The distribution is Right skewed

*Q8) Calculate Expected Value for the problem below*

1. *The weights (X) of patients at a clinic (in pounds), are*

*108, 110, 123, 134, 135, 145, 167, 187, 199*

*Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?*

Answer:

Expected Value = ∑ (Probability \* Value)

  =∑ P(x). E(x)

there are 9 patients

Probability of selecting each patient = 1/9

E(x) : 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) : 1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

Expected Value = (1/9) \*(108) + (1/9) \*(110) + (1/9) \*(123) + (1/9)\*(134) +

(1/9) \*(135) + (1/9) \*(145) + (1/9)\*(167) + (1/9)\*(187) + (1/9)199

= (1/9) \*( 108 + 110 + 123 + 134 + 135 + 145 + 167 +

187 + 199)

= (1/9) \* (1308)

= **145.33**

Expected Value of the Weight of that patient = **145.33**

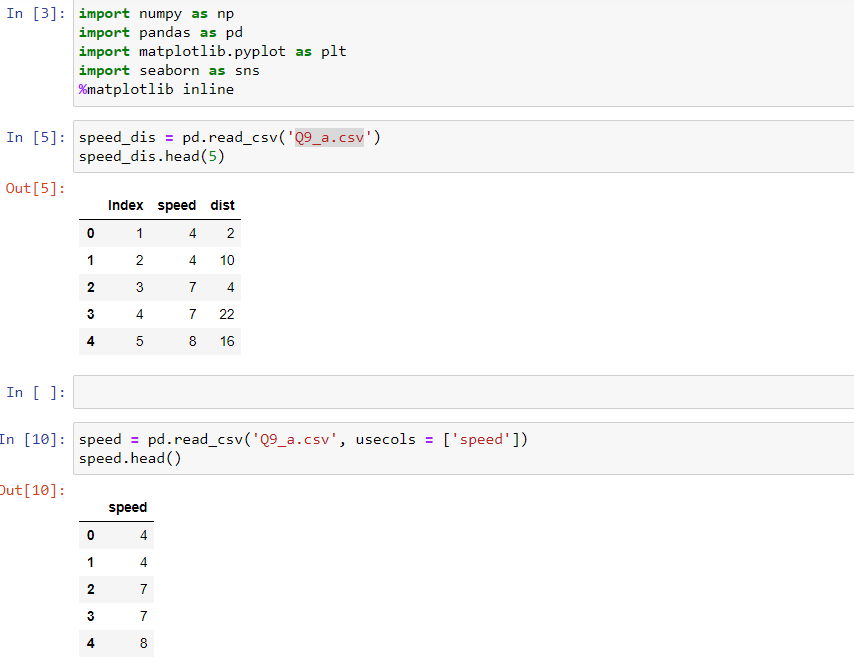
*Q9) Calculate Skewness, Kurtosis & draw inferences on the following data*

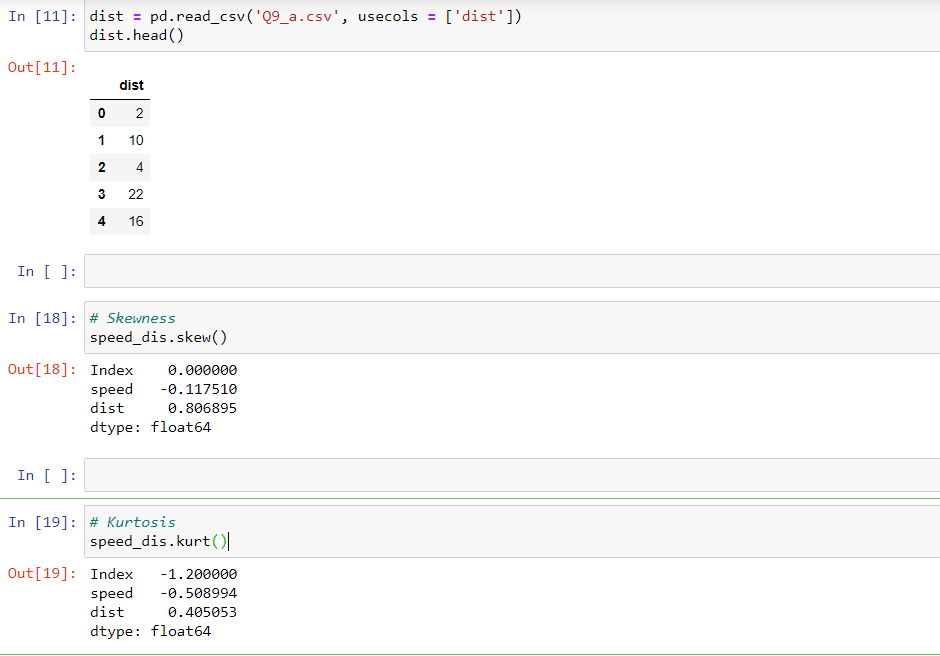
*Cars speed and distance*

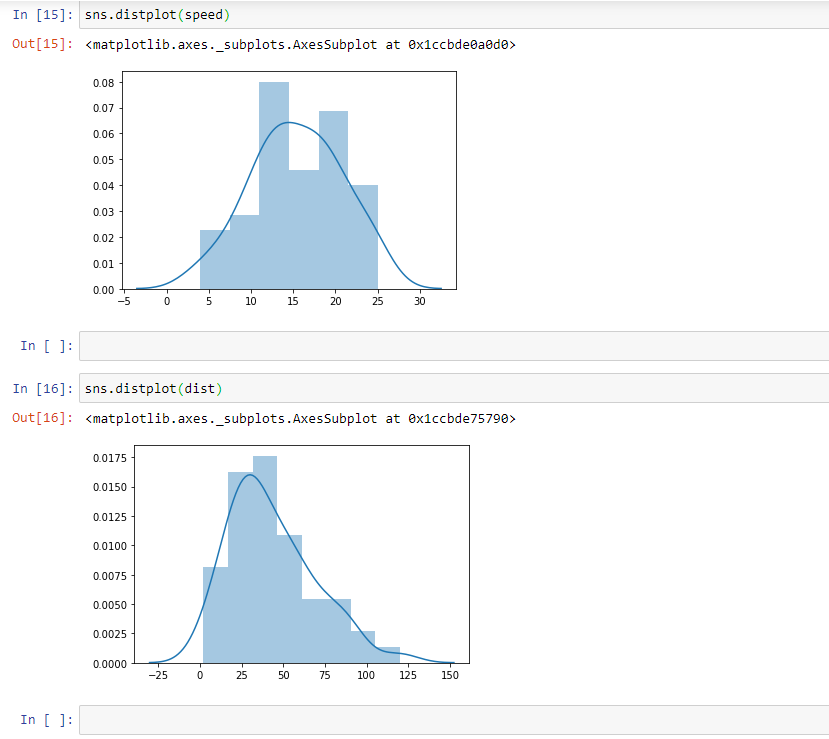
*Use Q9\_a.csv*

*SP and Weight (WT)*

*Use Q9\_b.csv*

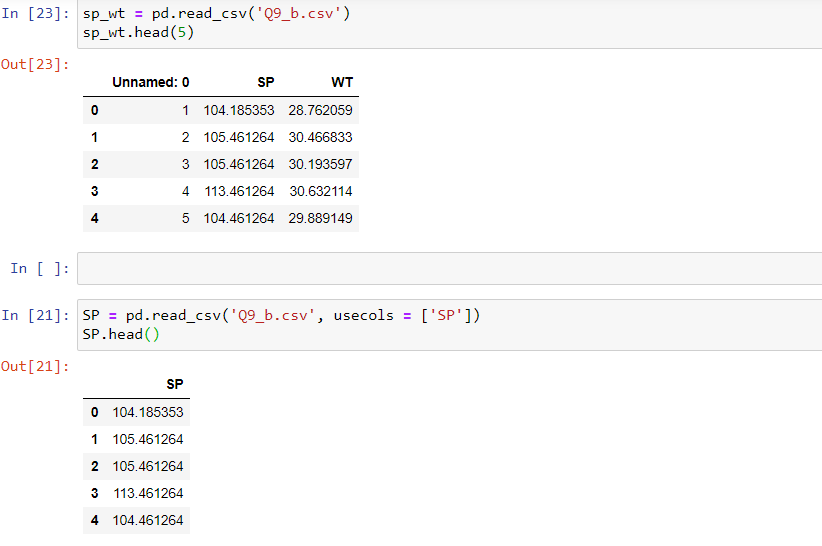


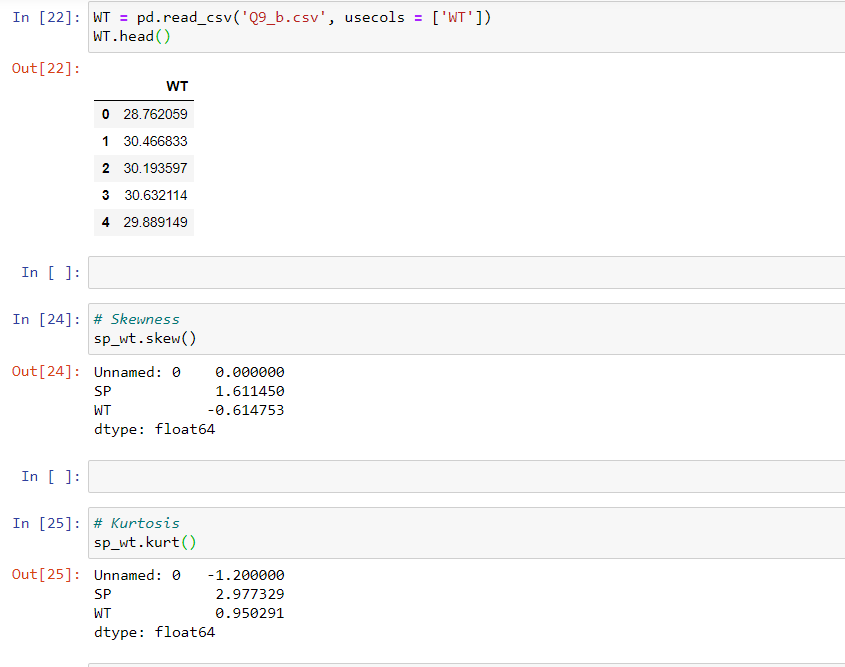


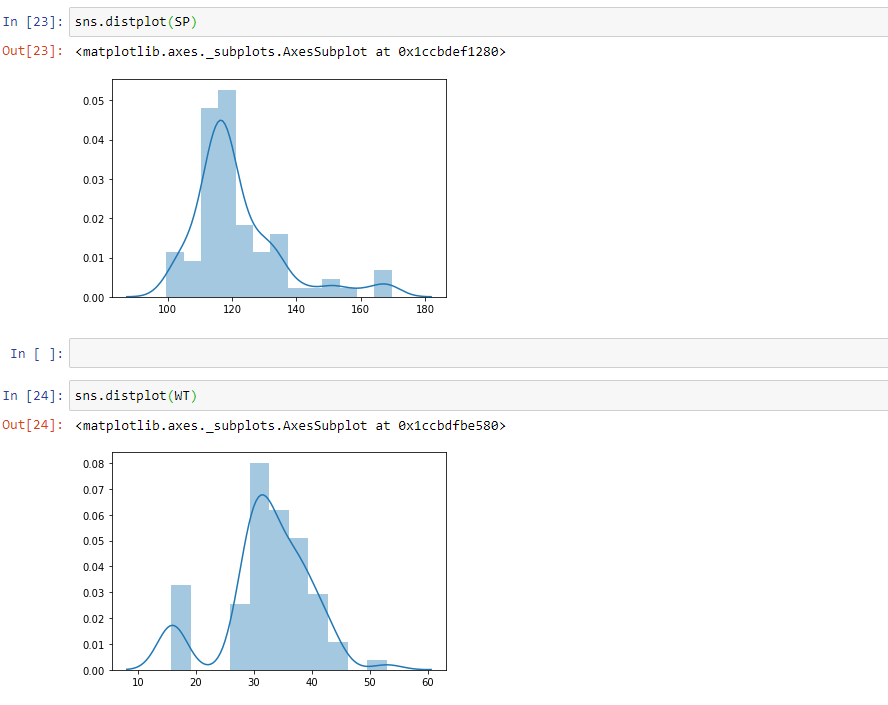


**Inferences for car speed and distance**

* Skewness value for car speed is -0.117510. Hence it is a **negative skewed** data (Left skewed) and also from the distribution the mass is towards right and tail is towards left.
* Skewness value for distance is 0.806895. Hence it is a **positive skewed** data (Right skewed) also from the distribution the mass is towards left and tail is towards right.
* Kurtosis value for the car speed is -0.508994. As it is having a negative value and flatter curve in the car speed distribution it is a **platykurtic**.
* Kurtosis value for the distance is 0.4050534. As it is having a positive value and sharp peeked data in the distance distribution it is a **leptokurtic**.







**Inferences for SP and Weight (WT)**

* Skewness value for SP is 1.611450. Hence it is a **positive skewed** data (Right skewed) also from the distribution the mass is towards left and tail is towards right.
* Skewness value for Weight is -0.614753. Hence it is a **negative skewed** data (Left skewed) and also from the distribution the mass is towards right and tail is towards left.
* Kurtosis value for the SP is 2.977329. As it is having a positive value and sharp peeked data in the distance distribution it is a **leptokurtic**.
* Kurtosis value for the Weight is 0.950291. As it is having a positive value and sharp peeked data in the distance distribution it is a **leptokurtic.**

***Q10) Draw inferences about the following boxplot & histogram***



***Inferences from the histogram and boxplot***

* From the plots we can say, it is **Right side skewed or positively skewed.**
* **Maximum values** fall at the range 50 – 100 and the **minimum values** fall at the range 300 - 350
* There are few **outliers** at the range 350 – 400

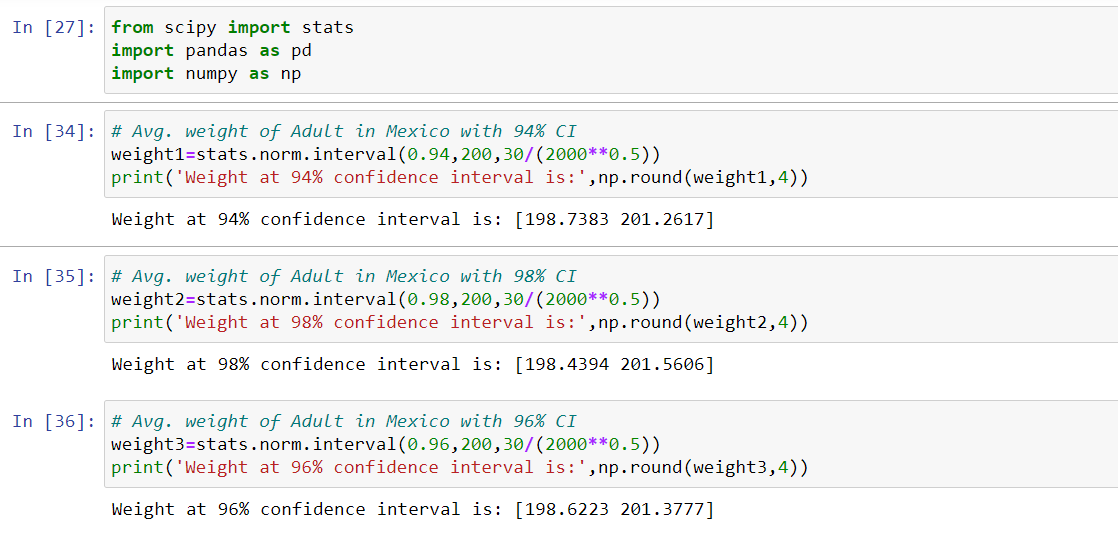
***Q11) Suppose*** *we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?*

Sample size, n = 2000

Population size, N =3000000

Sample mean, x̄ = 200

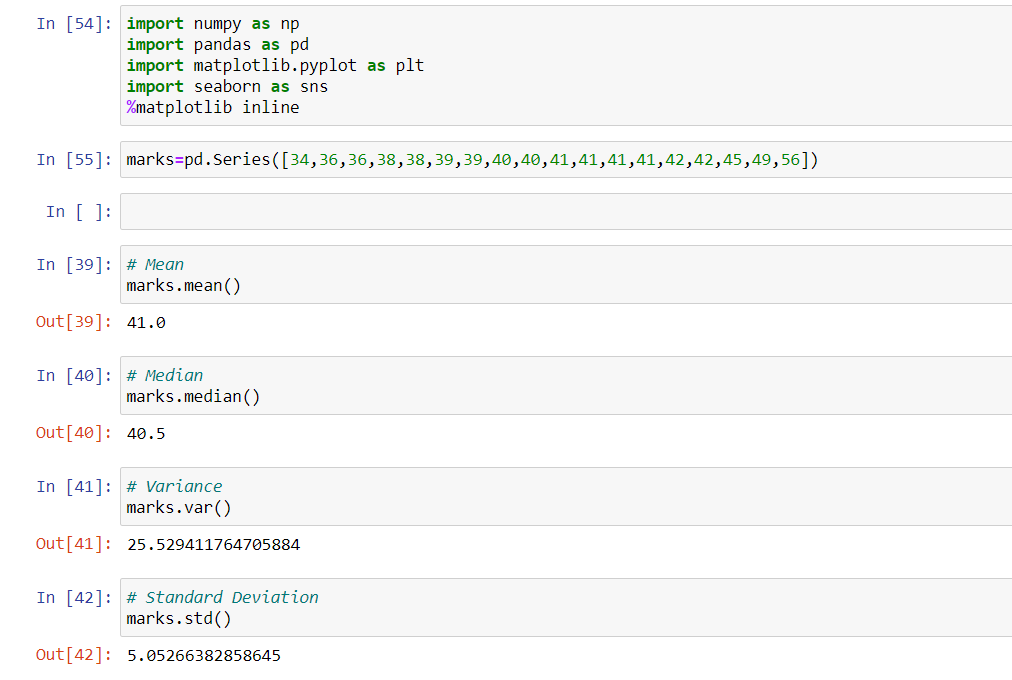
Standard deviation, σ = 30



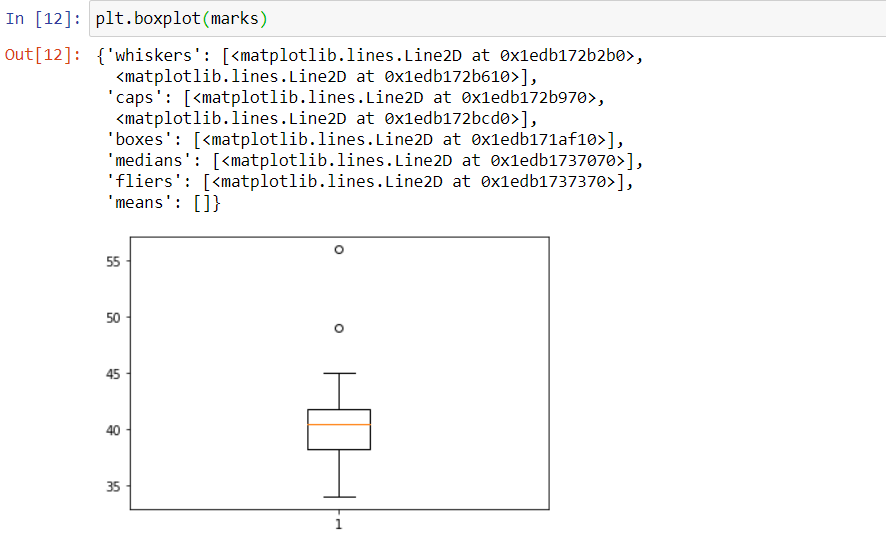
***Q12)*** *Below are the scores obtained by a student in tests*

***34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56***

1. *Find mean, median, variance, standard deviation.*
2. *What can we say about the student marks?*







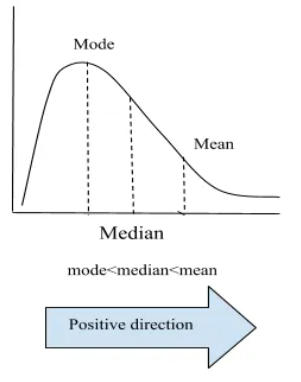
**Inference**:

* There are 2 Outliers in Student's marks: 49 and 56
* The data is concentrated around median 40.5
* The distribution is right skewed
* The average score achieved by the students is 41.0

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

Skewness is a measure of the symmetry of a distribution. Skewness can be positive, negative or zero. In a **perfectly symmetric distribution** the mean, median and mode everything is at one point which is equal to each other. In such cases there is no skewness

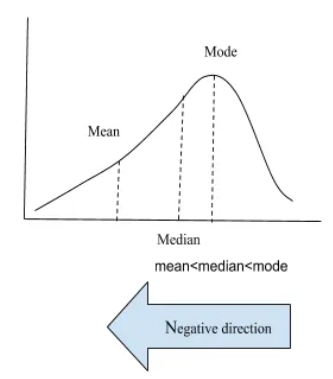
*Q14) What is the nature of skewness when mean > median?*



The [mean, mode and median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/) can be used to figure out if you have a positively or negatively skewed distribution.

* If the [mean](https://www.statisticshowto.com/mean/)is greater than the mode, the distribution **is positively skewed. (Right Skewed)**

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



The [mean, mode and median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/) can be used to figure out if you have a positively or negatively skewed distribution.

* If the mean is less than the median, the distribution is **negatively skewed (Left Skewed)**

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

* If the kurtosis is positive value, then we get a very peeked data, which is called Leptokurtic.
* The leptokurtic distribution shows heavy tails on either side, indicating large outliers.

E.g.: In finance, a leptokurtic distribution shows that the investment returns may be prone to extreme values on either side. Therefore, an investment whose returns follow a leptokurtic distribution is considered to be risky.

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

* If the kurtosis is negative value, then we get a flatter curve compare to normal distribution, which is called platykurtic.
* The platykurtic kurtosis reveals a distribution with flat tails. The flat tails indicate the small outliers in a distribution.

E.g.: In the finance, the platykurtic distribution of the investment returns is desirable for investors because there is a small probability that the investment would experience extreme returns.

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



What can we say about the distribution of the data?

* The following boxplot shows that the median is approximately at 15.2.
* Most of the data is distributed between 10 and 18, but there are distributions that are as low as 1.5 and as high as 18.5
* 1st Quartile is at 10 and 3rd Quartile is at 18.
* The above boxplot shows a left skewness
* There are no outliers seen

What is nature of skewness of the data?

Ans) The distribution is left skewed

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

Ans) Median of LW(Q1) =10

Median of UW(Q3) = 18

IQR=Q3-Q1

= 18-10 = **8**

Q19) Comment on the below Boxplot visualizations?



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

* In the above plot, sample 1 and 2 appear to have similar centers or median at the datapoint 262.5
* sample 1 and 2 are symmetric
* There are no obvious outliers in any of the samples.
* The two plots show the difference graphically for distributions with the same [mean](https://statisticsbyjim.com/glossary/mean/) but more and less dispersion. The panel on the left (Sample 1) shows a distribution that is tightly clustered around the [average](https://statisticsbyjim.com/glossary/mean/), while the distribution in the right panel (Sample 2) is more spread out.

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

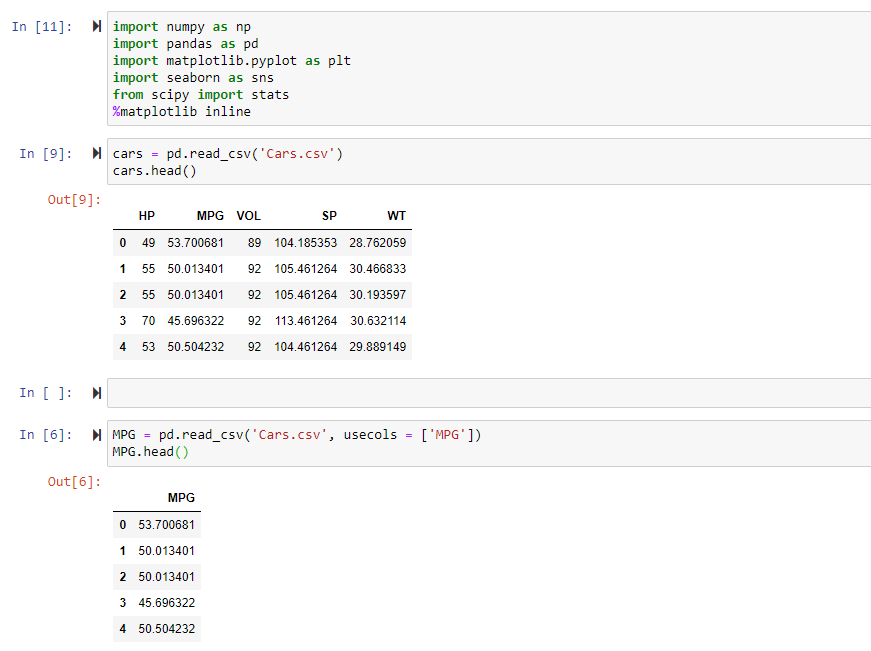
Data \_set: Cars.csv

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)

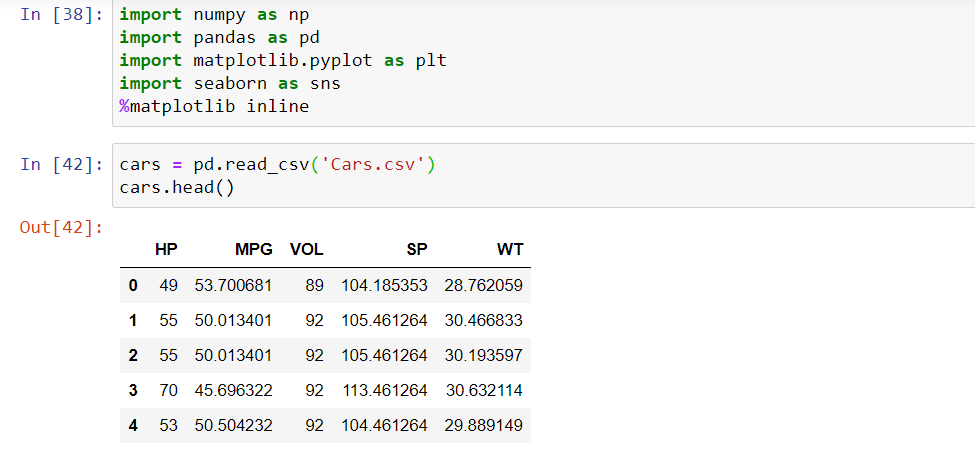


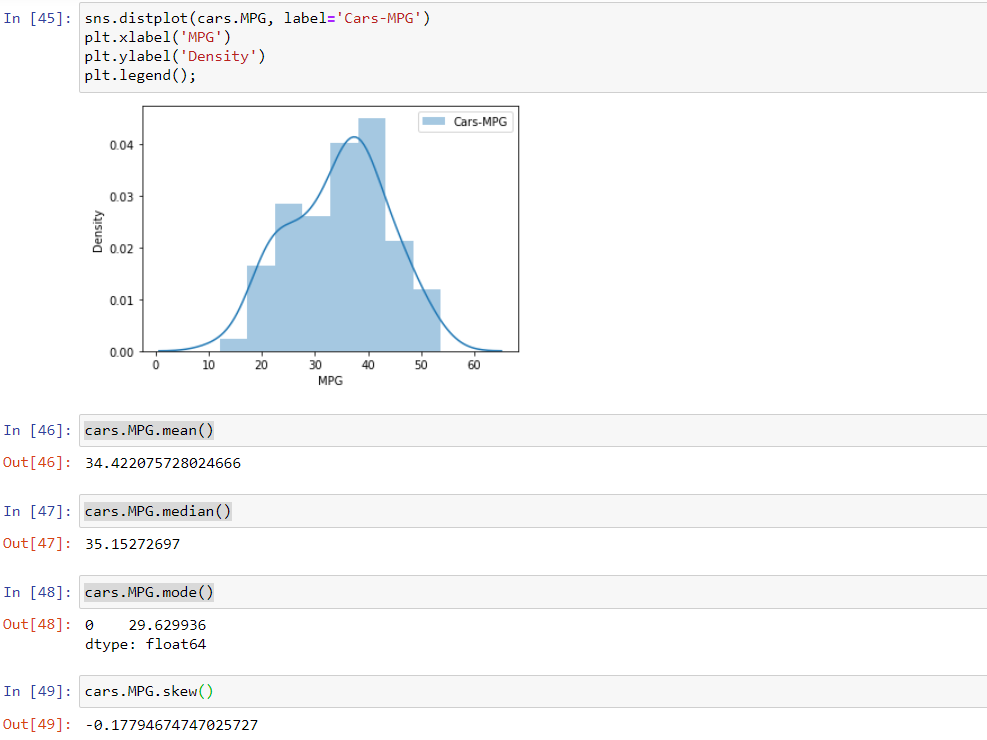


*Q 21) Check whether the data follows normal distribution*

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

*Dataset: Cars.csv*



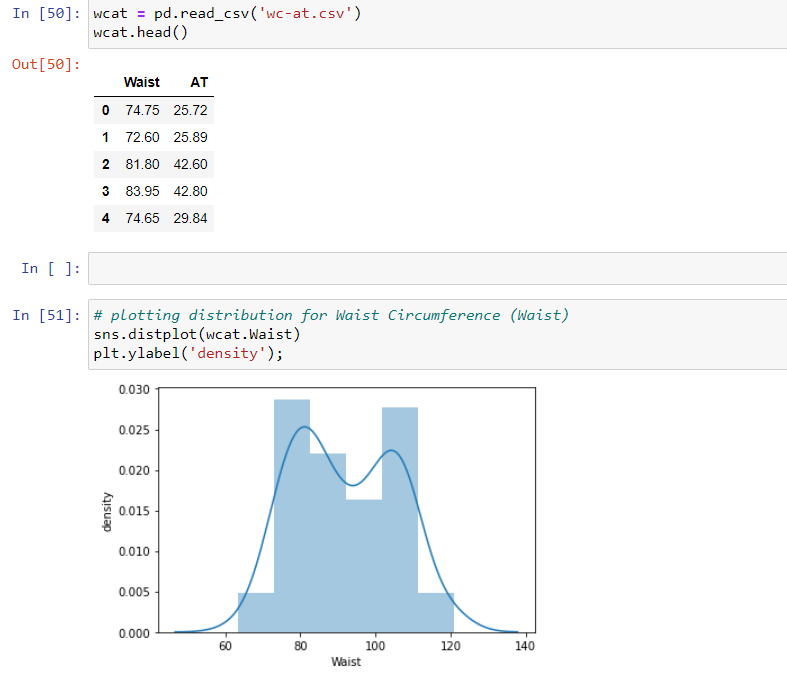


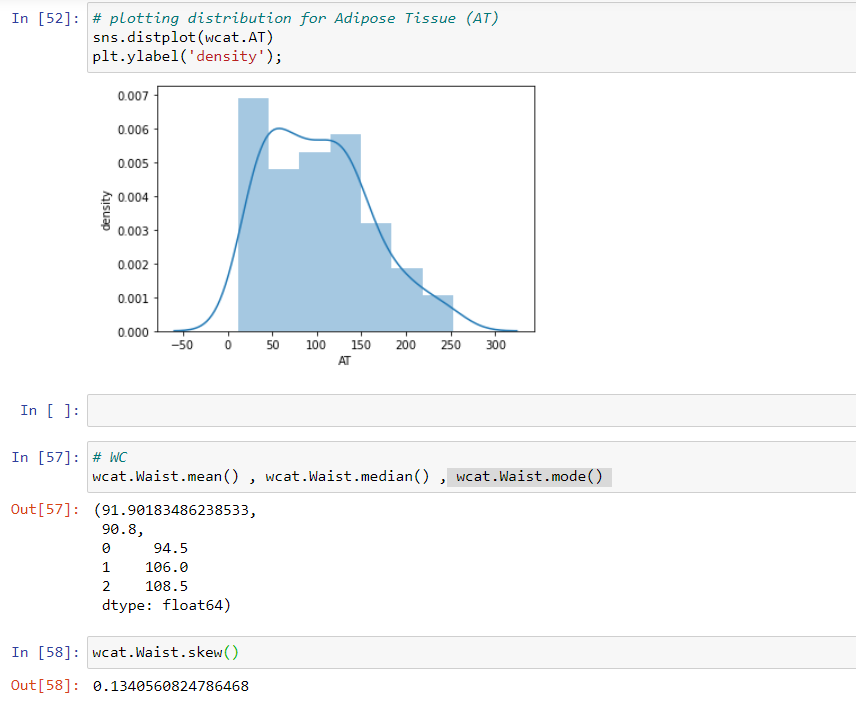
#### **Inference**

* MPG of Cars does not follow Normal Distribution because as per the property of normal distribution mean, median and mode will be equal to each other.
* Skewness will be zero if it’s a normal distributio

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

Dataset: wc-at.csv

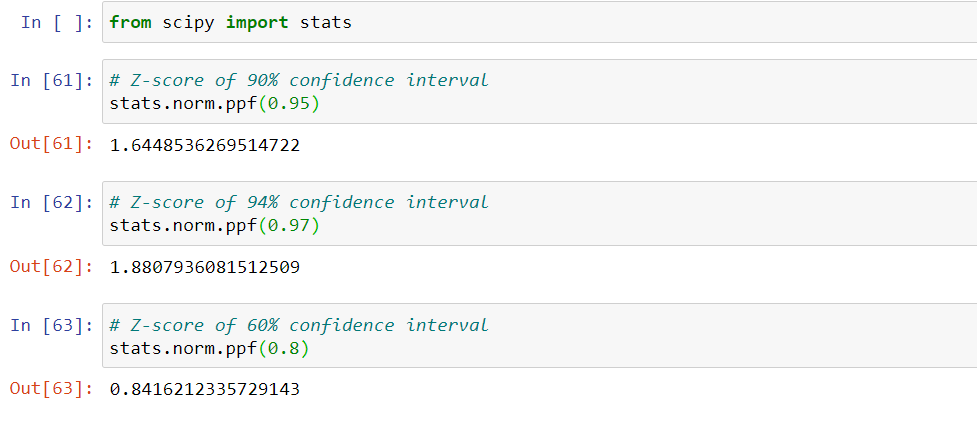




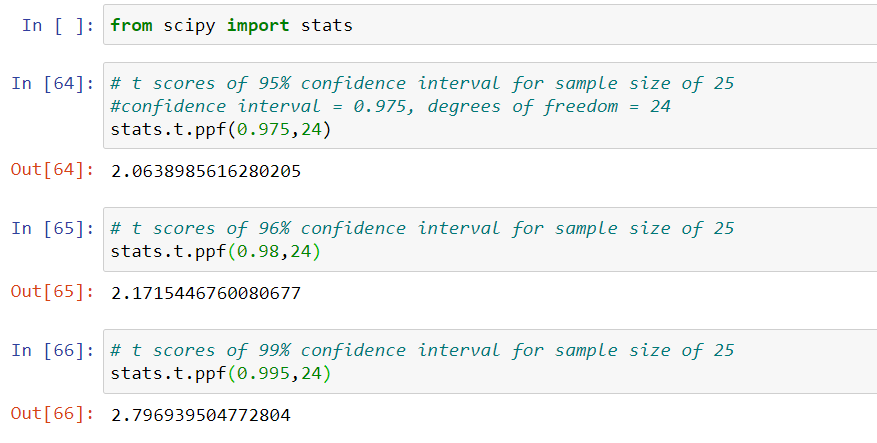
#### **Inference**

* Adipose Tissue (AT) and Waist Circumference (Waist) does not follow Normal Distribution because as per the properties of normal distribution mean, median and mode will be equal to each other.
* Skewness will be zero if it’s a normal distribution

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



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



*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*

t-statistics for the data is given as follows:

t = (x-μ) /( s/√n)

where,

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t = (x-μ) /( s/√n)

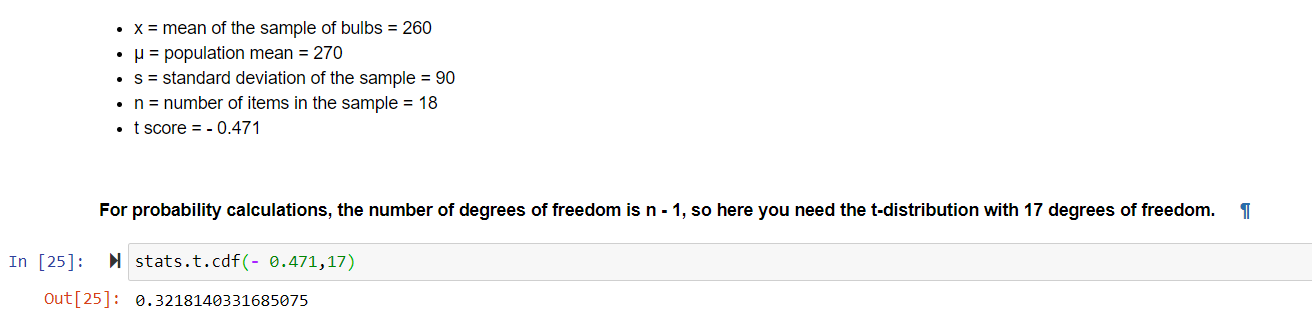
t = (260 – 270) / (90/√18)

= -10 / (90/3√2)

= -10 / 30/√2

= (-1 \* √2 ) / 3

t = - **0.471**



The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean 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.