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

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Ordinal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

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

**A]** S = {HHH, HHT, HTH. HTT, TTT, TTH. THT, THH} = 8

P(2H & 1T) = 3/8 = 0.375

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 2and 3

**A]** n(s) = 36

a) P(Sum = 1) = 0

b) P(Sum ≤ 4) = 6/36 = 0.16667

c) P(Sum is divisible by 2 and 3) = 6/36 = 0.16667

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?

**A]** Total number of balls = (2+3+2) = 7

n(S) = no. of ways that 2 balls drawn out of 7 = 7C2 = 7\*6/2\*1 = 21

Let E = Event of 2 balls that none of which is blue

n(E) = no. of ways that 2 balls drawn out of 5 = 5C2 = 5\*4/2=10

Therefore, P(E) = n(E)/n(S) = 10/21

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

**A]** Expected value = (1\*0,015)+(4+0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120)

= 3.09

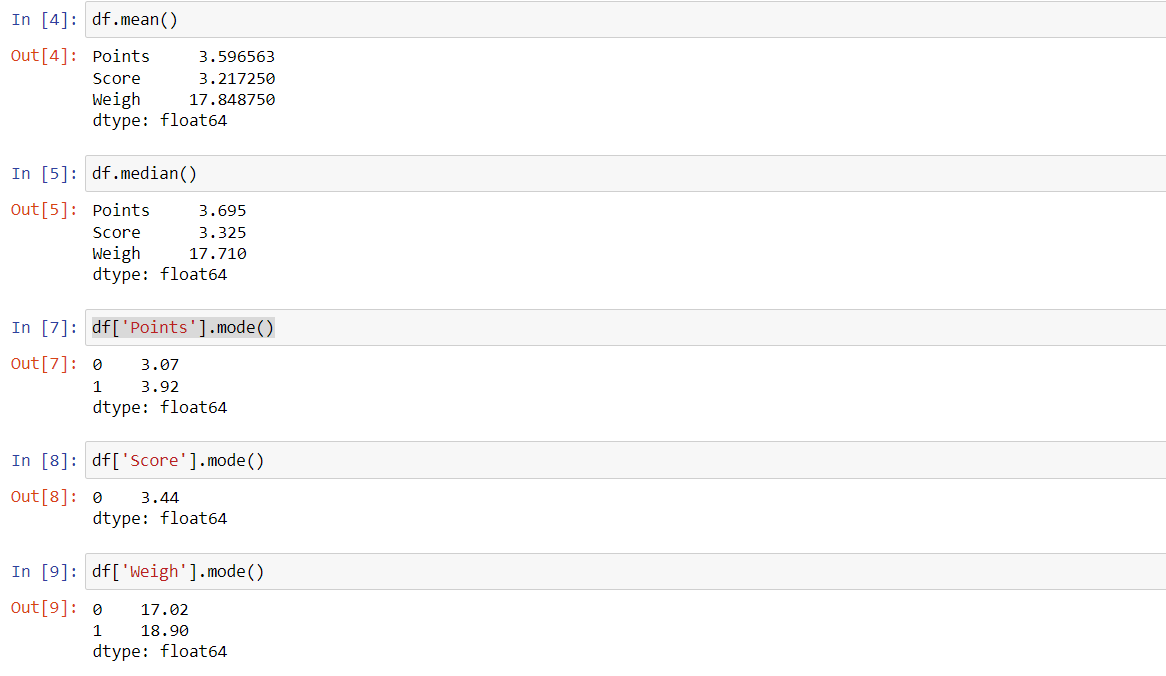
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

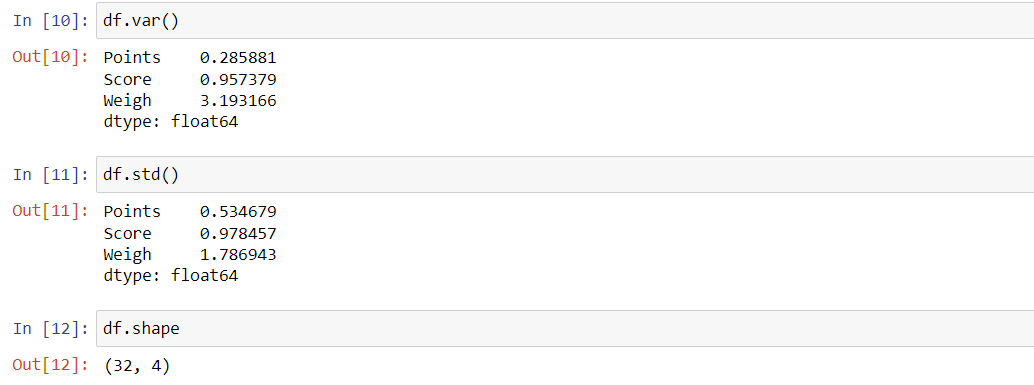
* For Points,Score,Weigh>

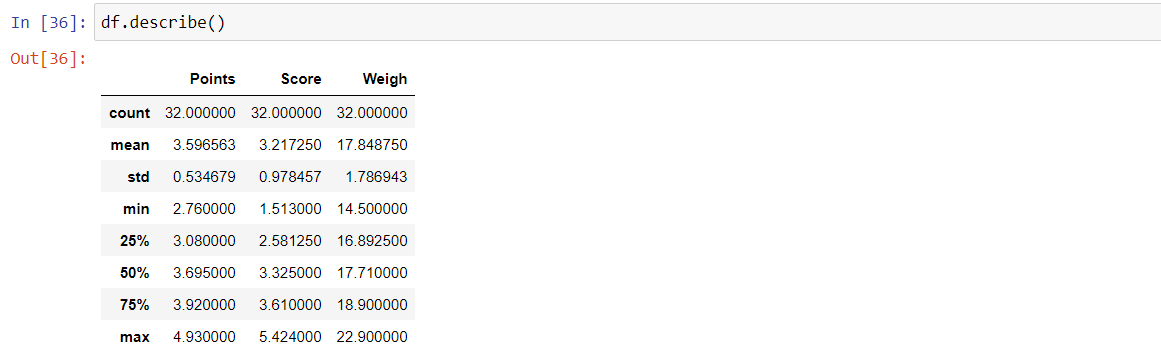
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

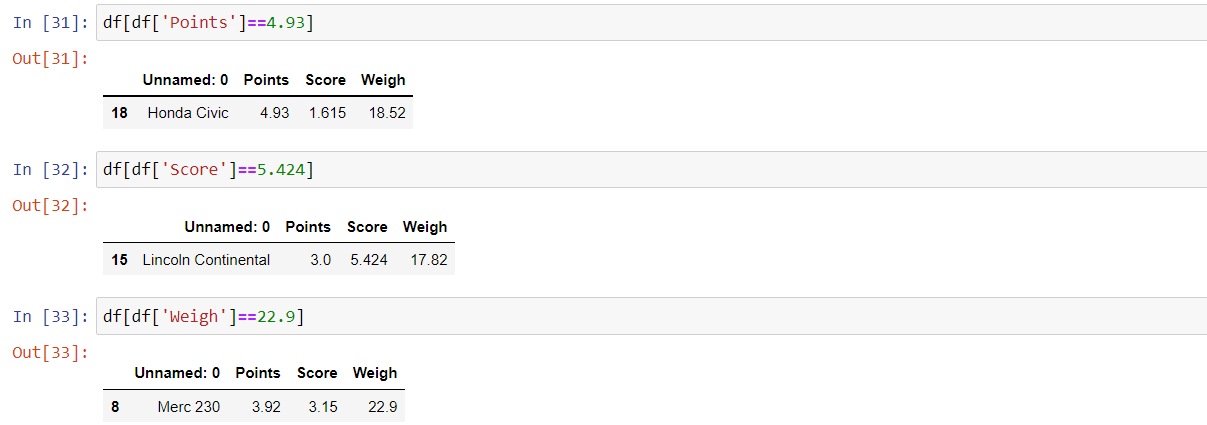
**Use Q7.csv file**

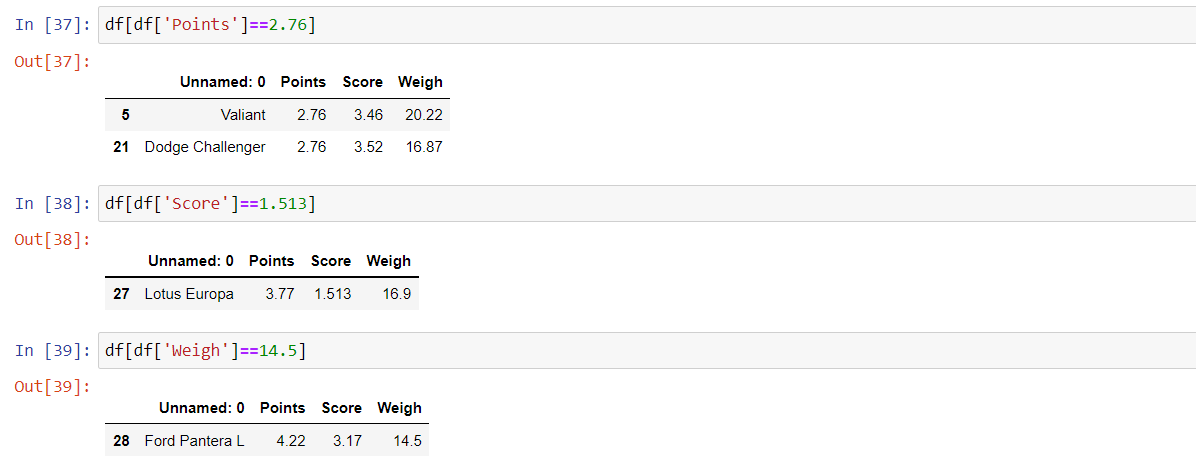
**A]**

****









**Inclusions:**

1. For the above inclusions, we can say that most types of cars have an average points of 3.596563, score of 3.21725 and weight of 17.84875. Since the mean and median values are close to each other, we can say that the most of the points lie near to the median.
2. The car having the most number of points is Honda Civic, whereas the car with highest score is Lincoln Continental and the car having the highest weight is Merc 230.
3. The car having the least number of points is Valiant and Dodge Challenger, whereas the car with lowest score is Lotus Europa and the car weighing the least is Ford Pantera L.

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?

**A]** There are 9 patients so the probability of single patient is 1/9.

Expected value = (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)

= 145.33

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

**Cars speed and distance**

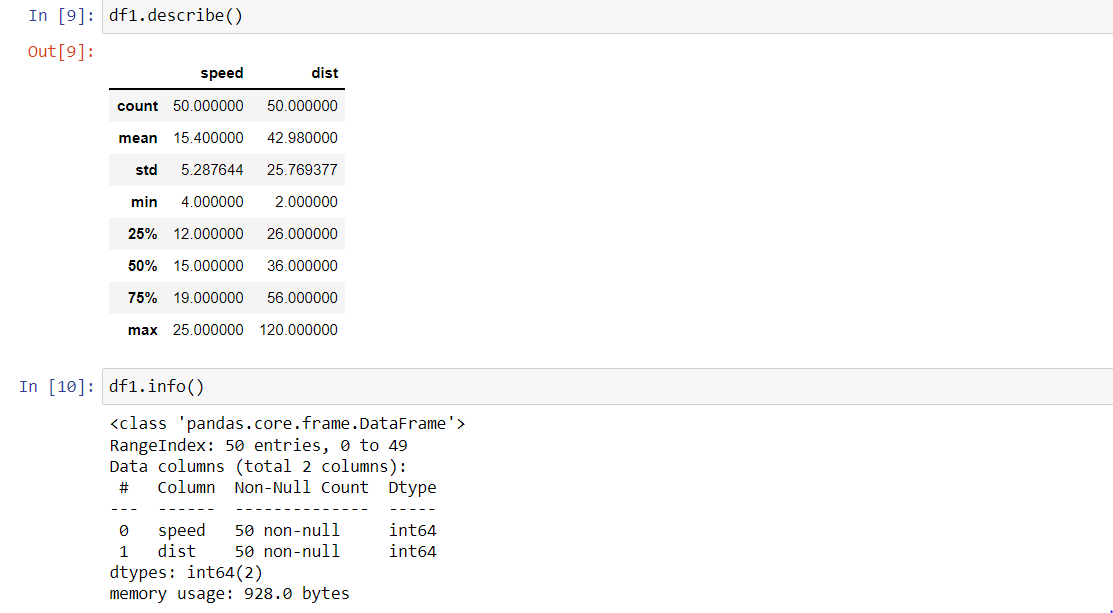
**Use Q9\_a.csv**

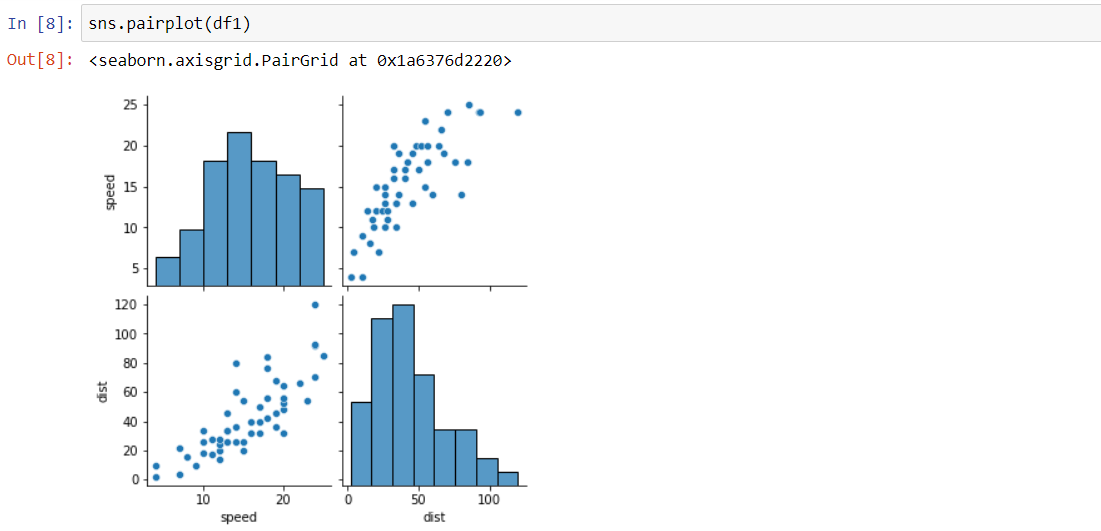
**SP and Weight(WT)**

**Use Q9\_b.csv**

**A] Q9\_a.csv**

****

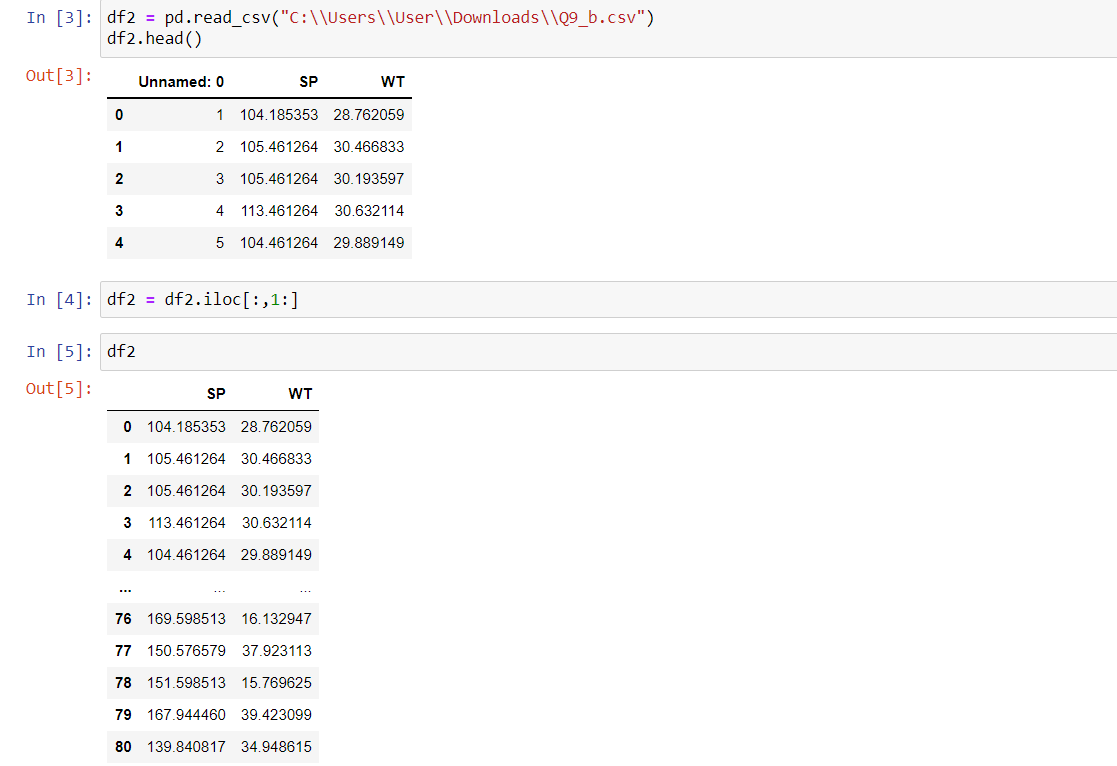
****

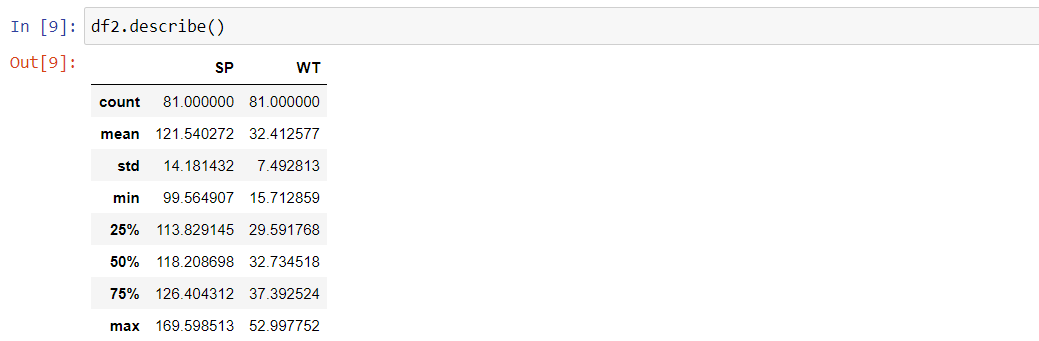
****

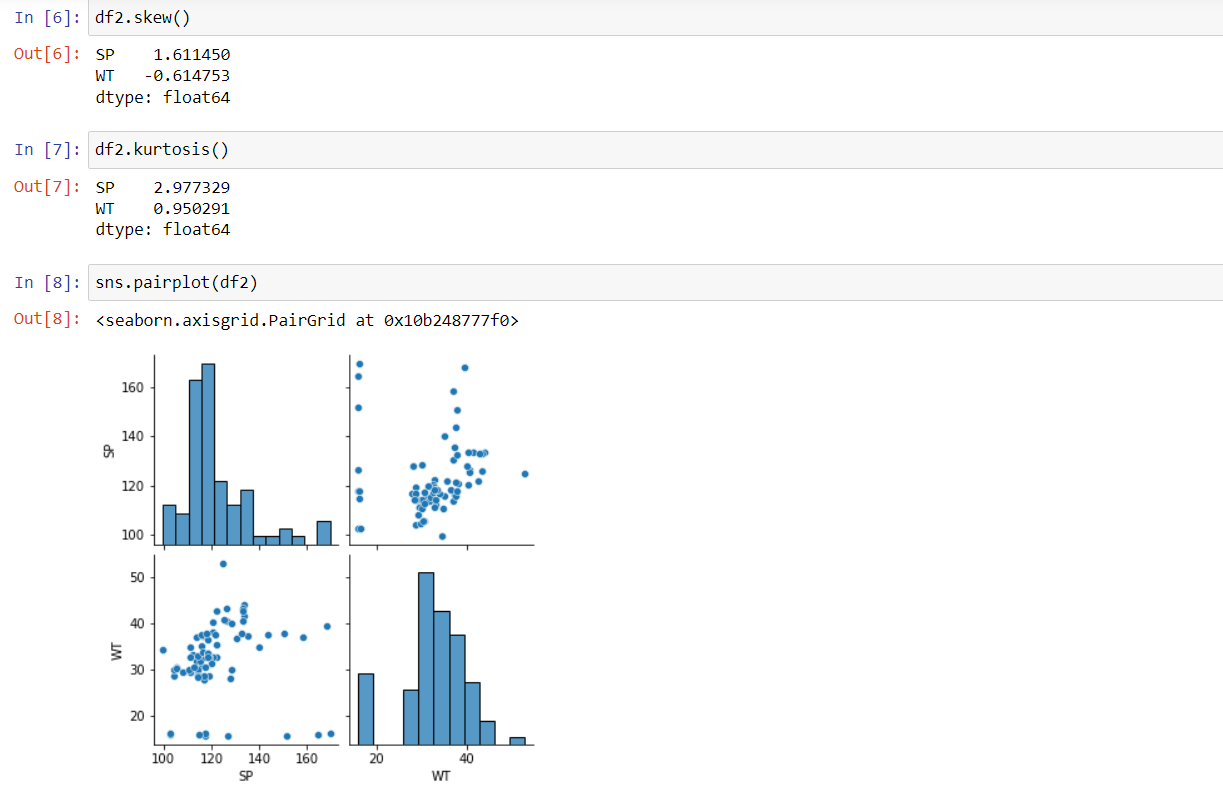
**Inferences:**

1. For speed both skewness and kurtosis are negative which means that the data is more towards right side of the mean value. It means that the distribution is negative or left skewed. It means that the mean is lesser than the median.
2. For distance both skewness and kurtosis are positive which means that the data is more towards left side of the mean value. It means that the distribution is positive or right skewed. It means that the mean is greater than the median.

**Q9\_b.csv**

****

****

****

**Inferences:**

1. For SP, both skewness and kurtosis are positive which means that the data is more towards left side of the mean value. It means that the distribution is positive or right skewed. It means that the mean is greater than the median.
2. For WT, the skewness is negative whereas the kurtosis is positive which means that the data is more towards the right side of the mean value. It means that the distribution is negative or left skewed. In negative skewed, mean is less than median.

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



**A]** From the above histogram and box plot, we can say that distribution has outliers present at the end (i.e. histogram tail side and box plot upper extreme). Most of the frequency for the ChickWeight$weight lies between 50 to 100. Hence we can say that the distribution is right skewed or positive skewed.

**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?

**A]** Sample mean, = 200

Sample size, n = 2000

Standard deviation, σ= 30

Confidence interval = ± z\*(σ/√n)

For 94% confidence interval, z=1.89

Confidence interval for 94% = 200 ± 1.89\*(30/√2000)

= (198.73, 201.26)

For 98% confidence interval, z=2.33

Confidence interval for 98% = 200 ± 2.33\*(30/√2000)

= (198.43, 201.56)

For 96% confidence interval, z=2.06

Confidence interval for 96% = 200 ± 2.06\*(30/√2000)

= (198.61, 201.38)

**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.

A] Mean: 41

Median: 40.5

Variance: 25.529412

Std: 5.052664



1. What can we say about the student marks?

**A]** From the above values we can say that the student scores an average of 41 marks most of the times, the lowest marks scored by the student is 34 and the highest marks scored is 56.

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

**A]** The nature of skewness is zero.

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

**A]** The nature of skewness is right skewed or positive skewed.

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

**A]** The nature of skewness is left skewed or negative skewed.

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

**A]** Positive kurtosis value means that the data is high peaked and is light-tailed. There are no outliers.

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

**A]** Negative kurtosis value indicates that the data is less peaked and is heavy-tailed. Here you can expect outliers on the higher and lower side.

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



What can we say about the distribution of the data?

**A]** The data distribution lies between the range 10 to 18.2.

What is nature of skewness of the data?

**A]** The data is negative-skewed since it is more towards the left side.

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

**A]** The IQR of the data will be 8.2 approximately.

Q19) Comment on the below Boxplot visualizations?



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

**A]** From the above boxplot 1 and 2, we can say that the data looks symmetric. We can also say that both are normally distributed. The mean of the data is 262 for both the box plots.

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

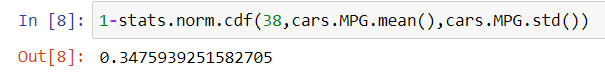
* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

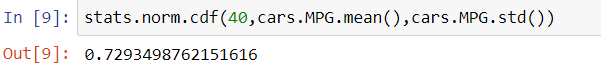
**A]** mean: 34.422076

std: 9.131445

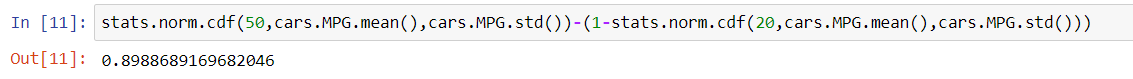
1. P(MPG>38)



1. P(MPG<40)



1. P (20<MPG<50)

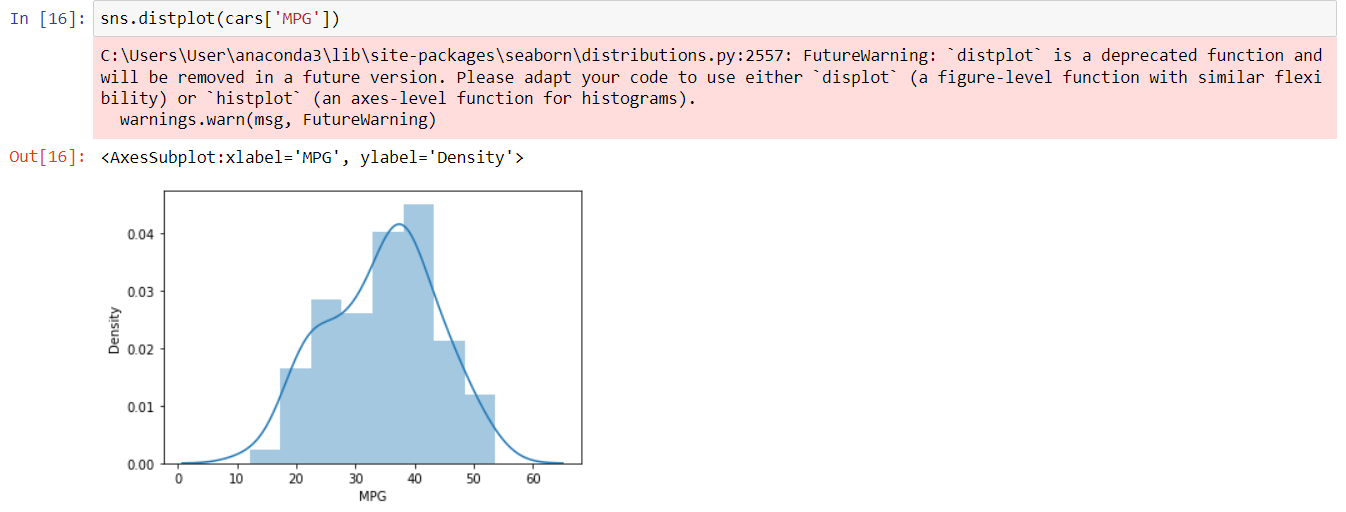
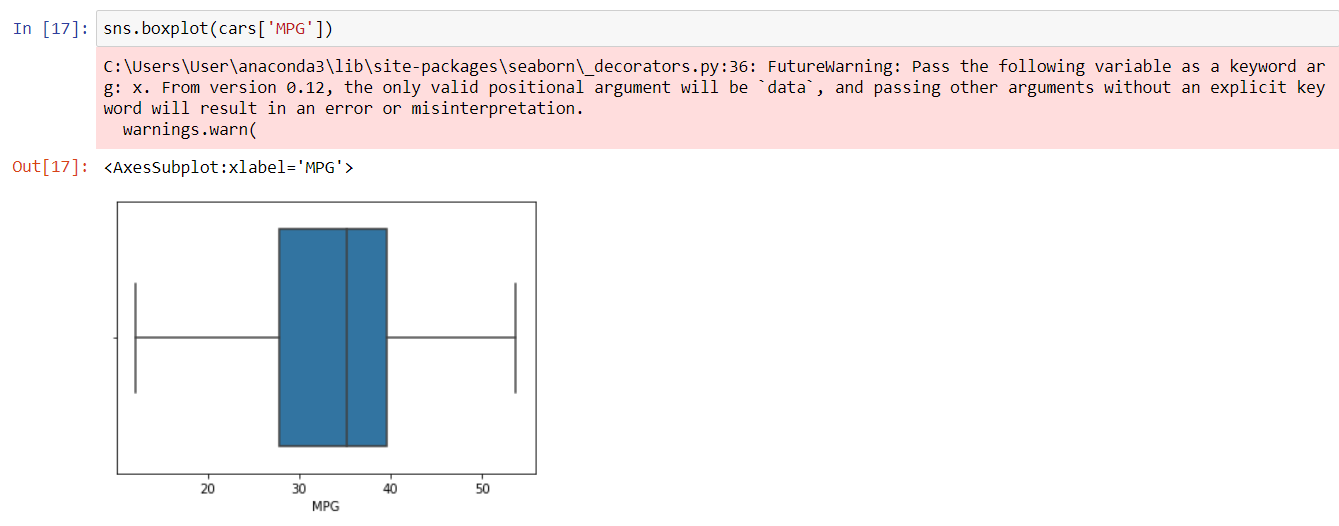


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

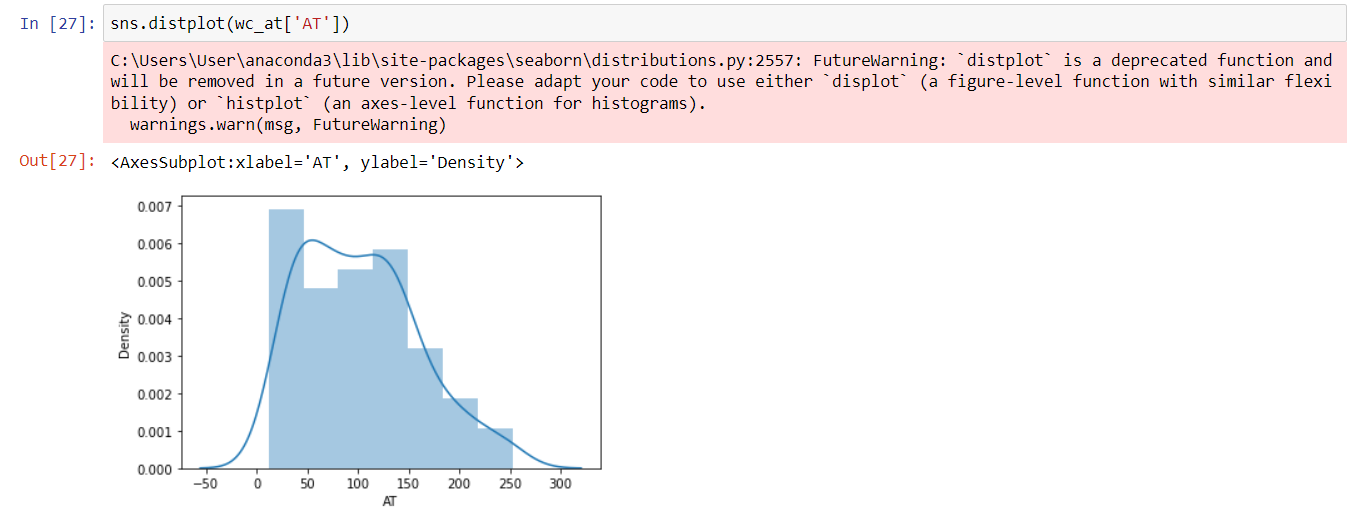
**A]**

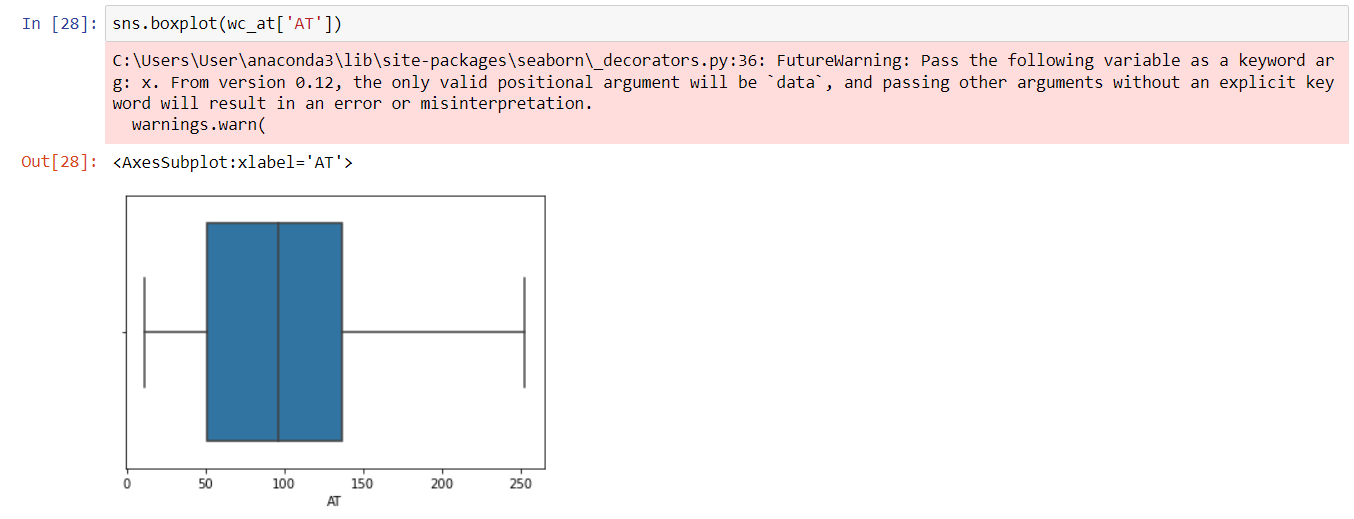
 

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

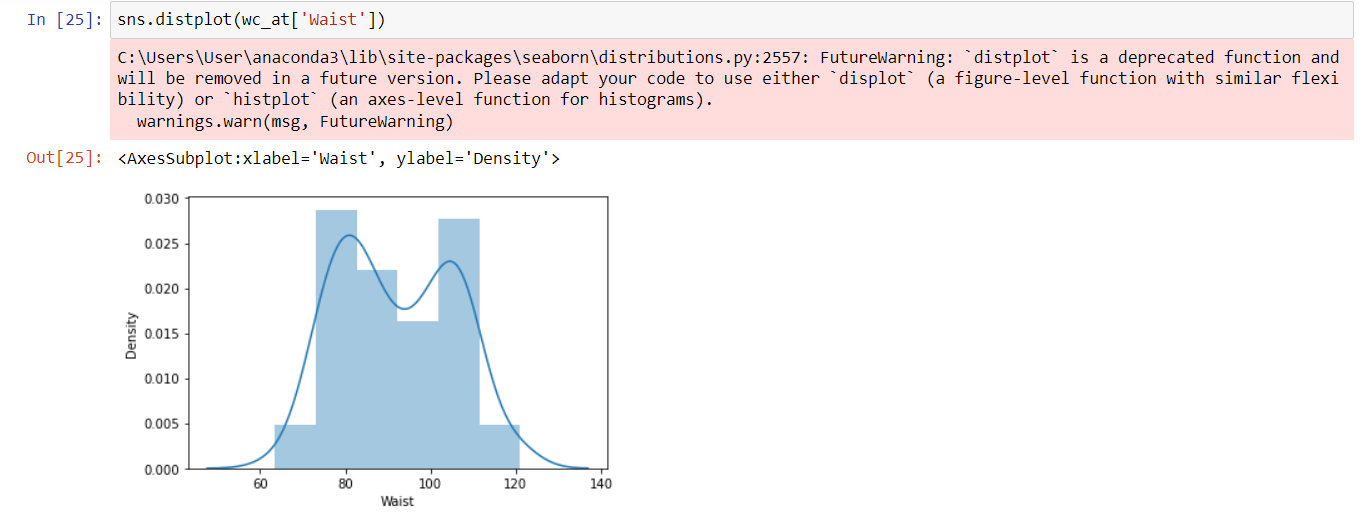
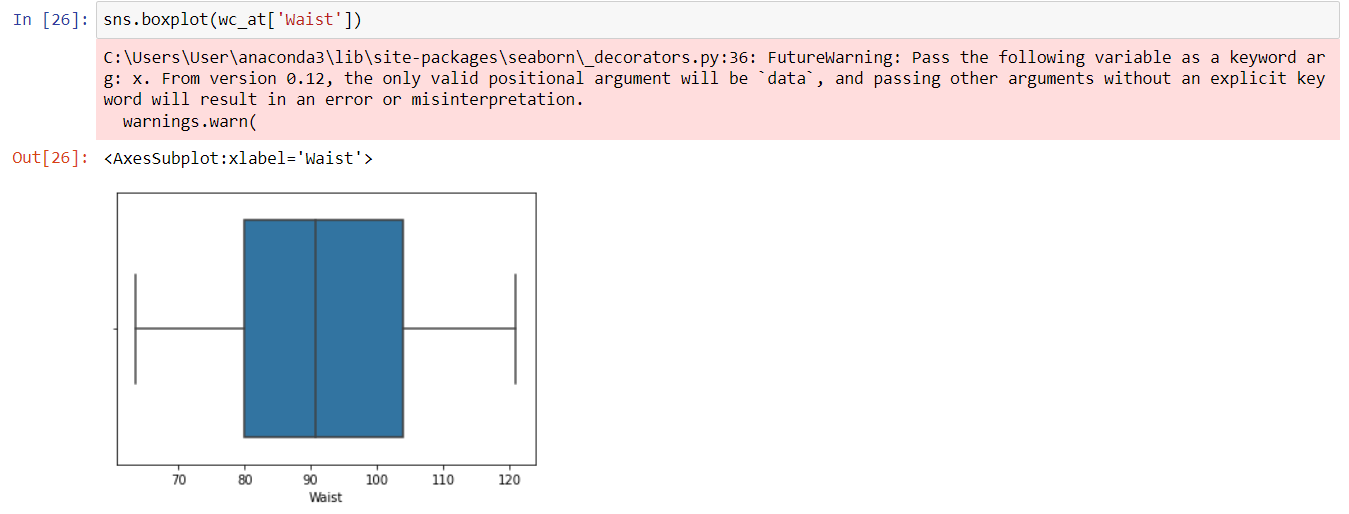
a) Adipose Tissue(AT)





From the above histogram and boxplot, we can say that the Adipose Tissue(AT) follows Normal Distribution.

b) Waist Circumference(WT)

** **

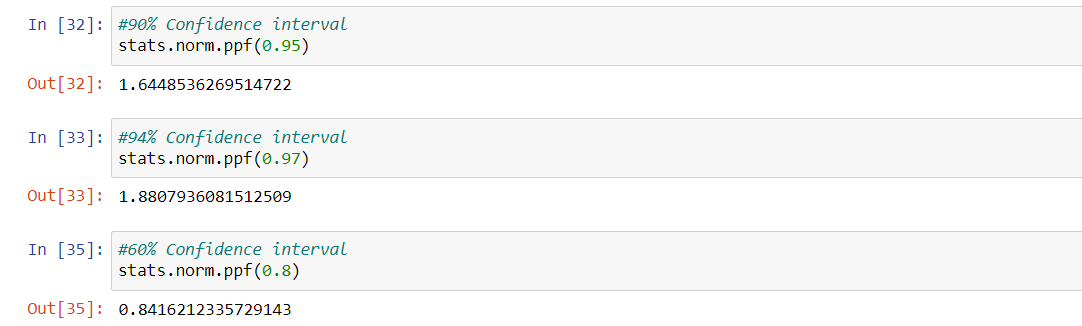
From the above histogram and boxplot, we can say that the Waist Circumference(Waist) follows Normal Distribution.

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

**A]** Z score for 90% CI = 1.6448

Z score for 94% CI = 1.88079

Z score for 60% CI = 0.8416



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

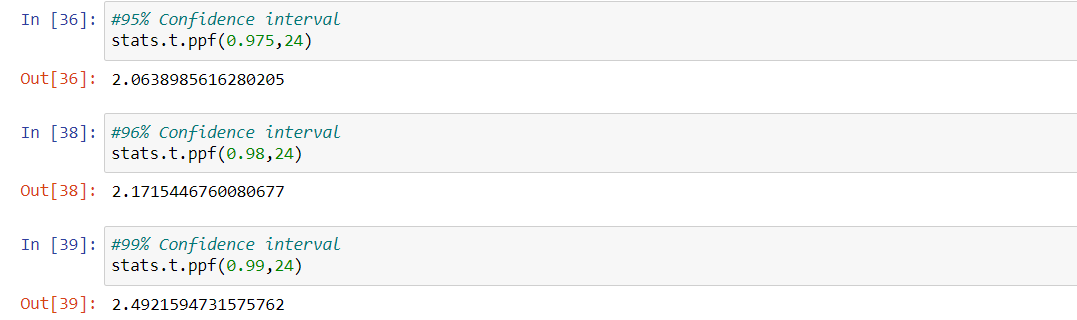
**A]** n = 25

df = n-1 = 25-1 = 24

t score of 95% CI = 2.0638

t score of 96% CI = 2.1715

t score of 99% CI = 2.4921



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

**A]** Sample mean, = 260

Sample size, n = 18

Population mean, μ = 270

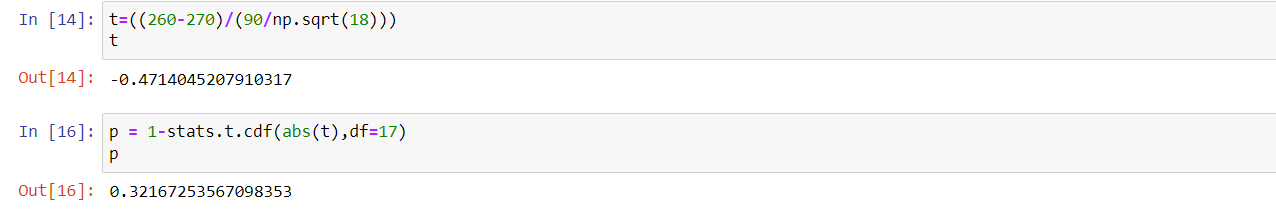
Standard deviation, σ = 90

t-score = ( – μ)/(σ/√n)

= (260-270)/(90/√18)

= -0.47

Degrees of freedom(df) = n-1 = 18-1 = 17



Probability = 0.32167