|  |  |
| --- | --- |
| 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 | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans: Sample Space={HHH,HTT,TTT,HHT,TTH,THH,HTH,THT}

A= {HHT, THH, HTH}

P(A)= No. of outcomes/No. of outcomes in Sample space

=3/8

Hence **P(A)=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 2 and 3

Ans: Sample Space ={(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) }

1. Equal to 1

**P(a)= 0/36=0**

b)Less than or equal to 4

**P(b)= 6/36=1/6**

c)Sum is divisible by 2 and 3

**P(c)=6/36=1/6**

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?

Ans: Total no. of balls = 2+3+2=7

Let S be the Sample Space, n(S)= 7C2 i.e. No. of ways of drawing 2 balls out of 7 balls

n(S)=(7\*6)/(2\*1)=21

Let A be an event of drawing no blue ball, i.e. 2 balls will be drawn out of 2 red and 3 green balls. Then n(A)=5C2

n(A)=(5\*4)/(2\*1)=10

P(A)=n(A)/n(S)=10/21

Hence **P(A)=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

Ans: Expected number of candies for a randomly selected child=∑*P*(*Xi*​)×*Xi*​​  
=1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

=0.015+0.80+1.95+0.025+0.06+0.24

**Expected number of candies for a randomly selected child =3.09**

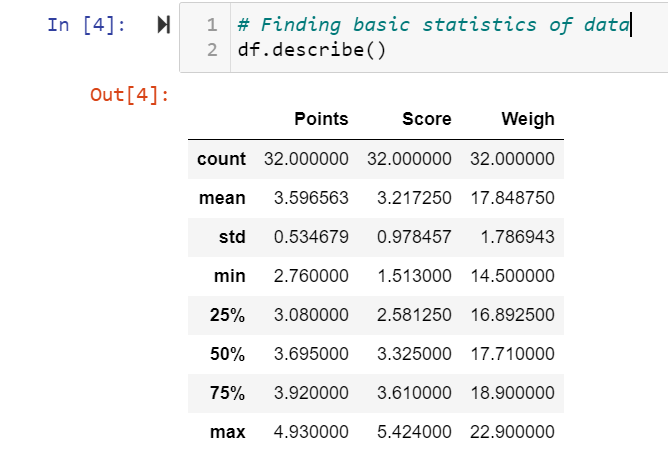
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

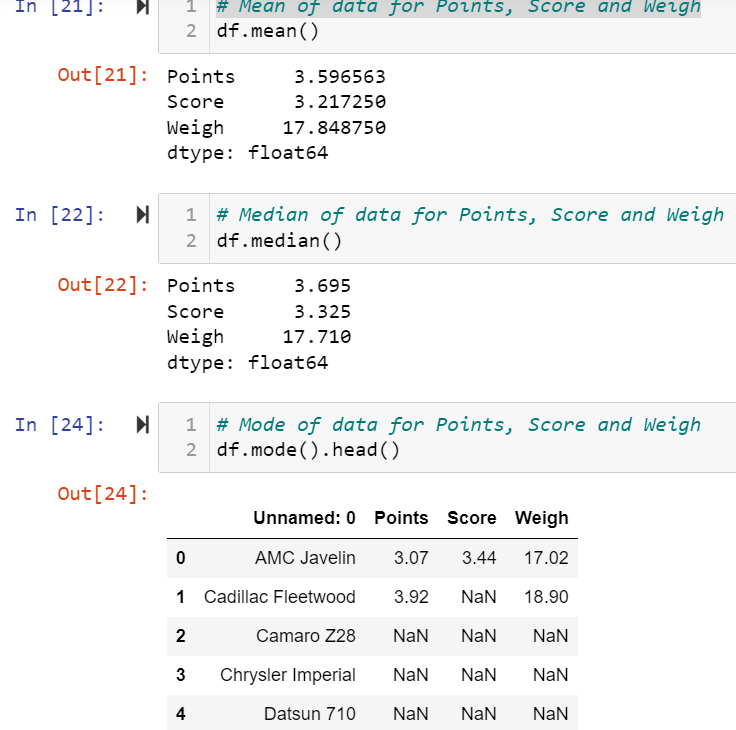
* For Points,Score,Weigh>

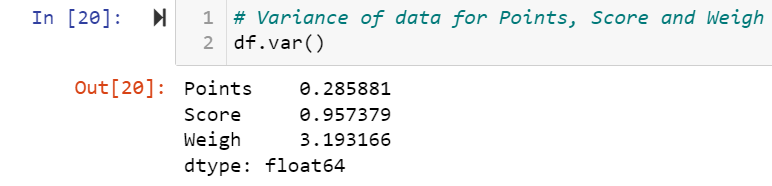
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

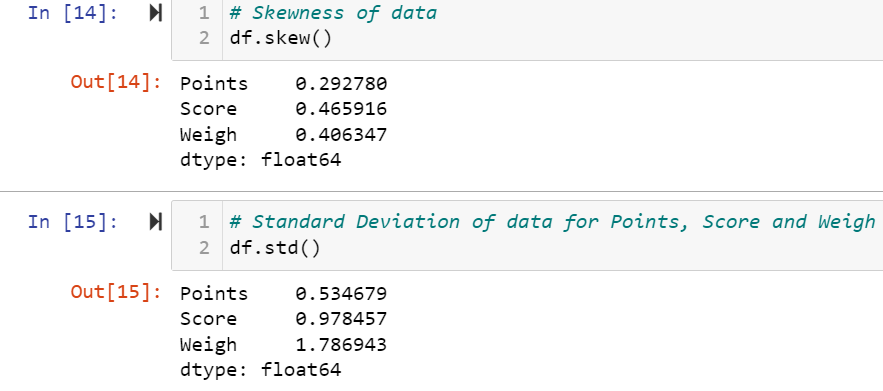
**Use Q7.csv file**

**Solution:**

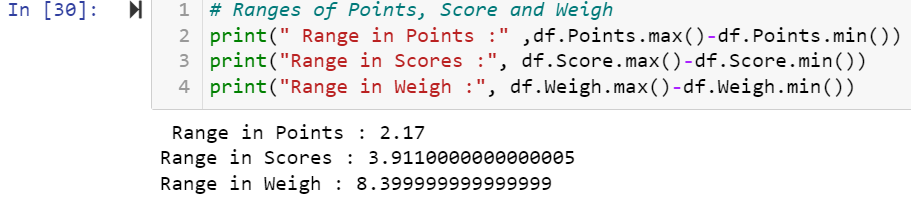








**Inferences: 1.** Here, data is having some sort skewness but as per the general thumb rule of skewness, if the skew values range from -0.5 to 0.5 , then the distribution is said to be approximately symmetric.



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?

**Solution**:

Expected Value = ∑ (probability \* Value)

Total no. of patients=9

Probability of selecting one patient =1/9

Thus, 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

=145.33

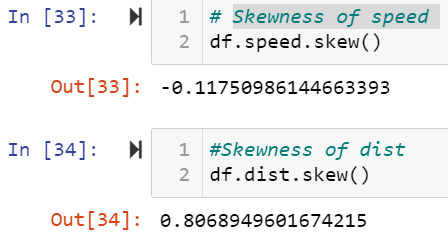
Hence, **Expected Value= 145.33**

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

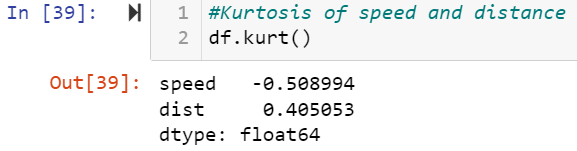
**Cars speed and distance**

**Use Q9\_a.csv**

**Solution:**



**Inference:** As skewness value for speed is less than -1 , we can say it is highly left skewed. Whereas, the skew value for distance is between 0.5 to 1 , we can say it is moderately skewed.

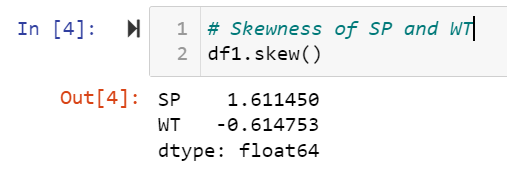


**Inference:** As the kurtosis values are less than 3 we can inference that data is thin-tailed with low outlier frequency.

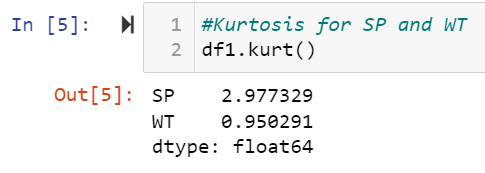
**SP and Weight(WT)**

**Use Q9\_b.csv**

**Solution:**



**Inference:** As the skew value for SP is greater than 1, it is highly skewed (positive skewed) on right side. The WT is also highly skewed (Negative Skewed) as the value is less than -1.



**Inference:** Kurtosis value for SP is close to 3, indicating the distribution is too peaked or tend to have heavy tail or outliers. Whereas, for WT the distribution is having shorter tail with fewer outliers.

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



**Inferences:**

1. The data is right skewed/Positive skewed.
2. Most of the chickweight is between 50-100
3. More than 50% of chickweight falls in range of 50-150



**Inferences:**

1. The data is positively skewed.
2. It has outliers on the upper side

**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?

**Solution:**

Confidence interval (CI) = ‾X ± Z(S ÷ √n)

Given:

‾X= Sample Mean=200

S=Sample Standard Deviation=30

n= Sample Size=2000

As per calculated formula, below are the CI

1. 94% CI is (198.73,201.62) (z=1.8808)
2. 96% CI is (198.622,201.378) (z=2.0537)
3. 98% CI is (198.439,201.561) (z=2.3263)

**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?

**Solution:**

**1.**Mean= Sum of Elements/Number of Elements

=(34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

**Mean=41**

Median is the middle value for the dataset. Here Median will be as below.

Median=(9th Term+10th term)/2=(40+41)/2

**Median=40.5**

**Variance=25.52**

**Standard Deviation= 5.05**

**2.** Most of the students got marks ranging from 35 to 45

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

**Answer**: When Mean , Median and Mode of data are equal then the skewness is zero and the date is said to be normally distributed.

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

**Answer:** When Mean>Median the data is Right/Positive Skewed.

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

**Answer:** When Median> Mean the data is Left/Negatively Skewed.

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

**Answer:** Positive values of kurtosis indicates that the data is peaked and has thick tails.

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

**Answer:** Negative kurtosis indicates that the data is uniformly distributed and has no peak at all with flat distribution. Also, it has thinner tails, more data is located near the mean.

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



What can we say about the distribution of the data?

**Answer:** More data is distributed on the right side of the boxplot.

What is nature of skewness of the data?

**Answer:** It is negatively skewed data

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

**Answer: IQR= Q3-Q1**

Q3=18.1, Q1=10

**IQR=18.1-10=8.1**

Q19) Comment on the below Boxplot visualizations?



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

**Answer:** Both Boxplot 1 and Boxplot 2 are normally distributed.

Medians for both boxplots are same.50% of data lies above the median and 50% lies below the median.

IQR for boxplot1 is lower as compared to IQR of boxplot2.

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.

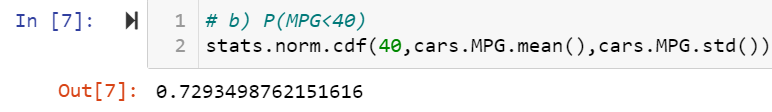
MPG <- Cars$MPG

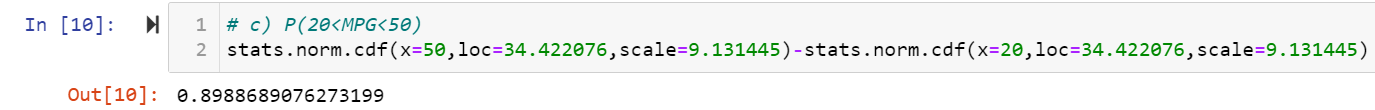
* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

**Solution:**





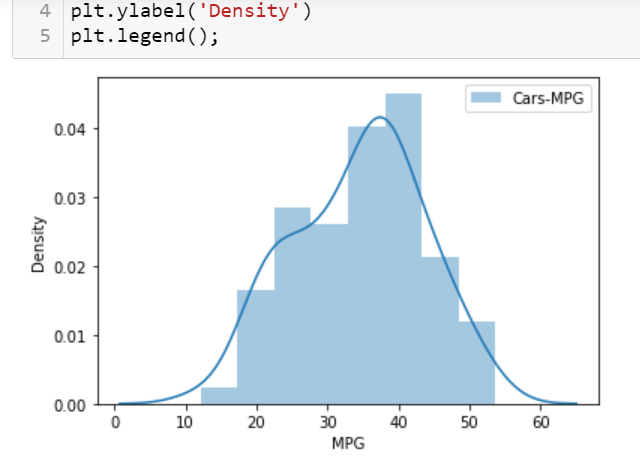


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Solution:**



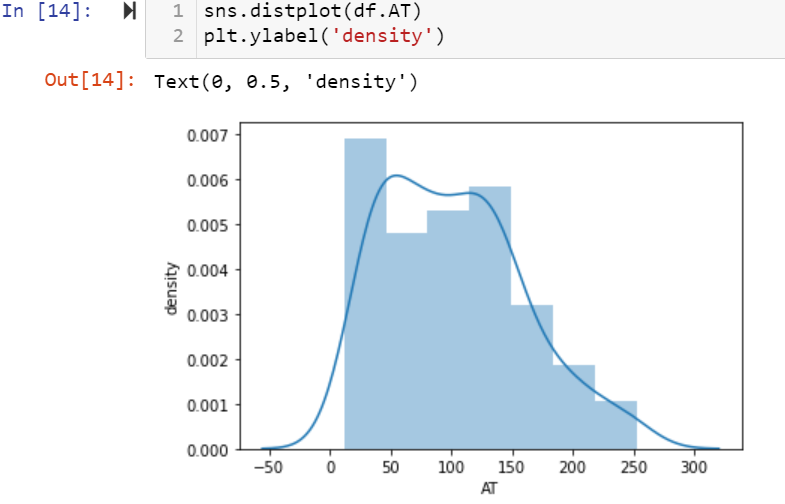
As per the graph we can say that MPG of cars does not follow normal distribution

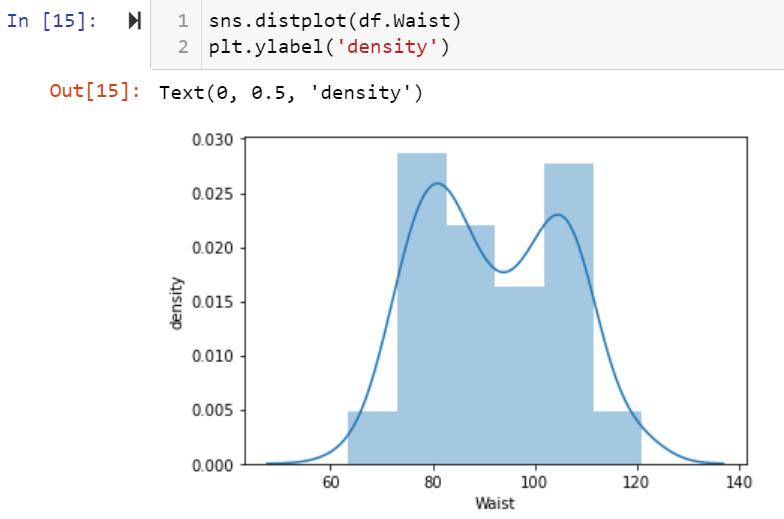
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**Solution:**

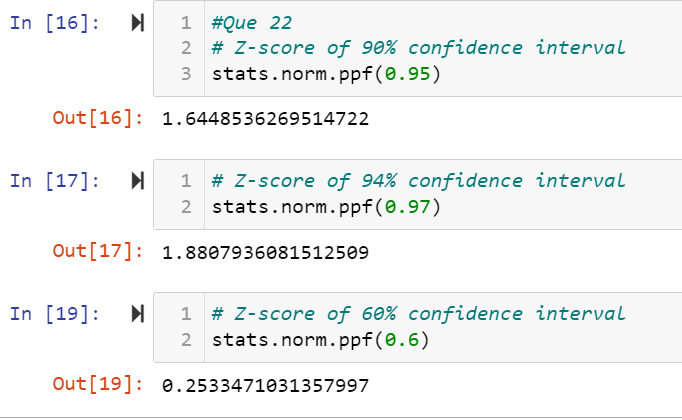
As per the below plots AT and Waist do not follow Normal Distribution





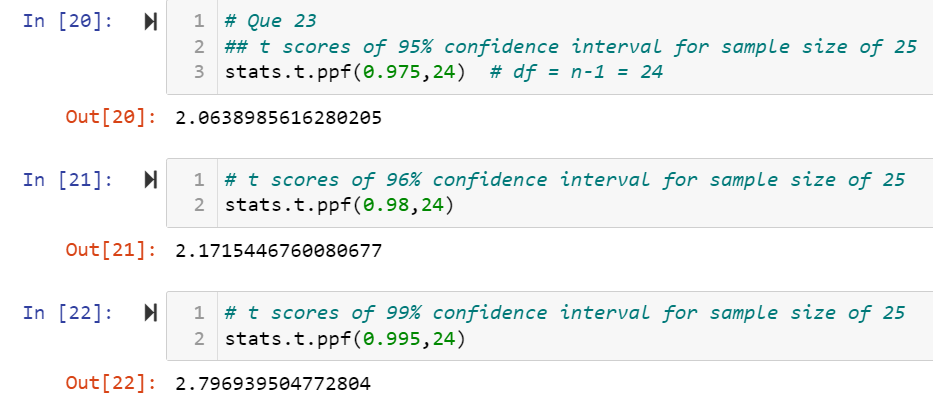
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Solution:**



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

**Solution:**



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

Solution:

The degree of freedom is n-1, so here we need the t-distribution with 17 degrees of freedom. T-score is -0.471.

The probability of the bulbs lasting less than 260 days on an avg of 0.321 assuming the mean life of the bulb is 300 days

