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

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 | Interval |
| 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 | Nominal |
| 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?

**Ans:**

No. of favorable outcomes is 3 & Total outcomes is 8

Then P(HHT)= No. of favorable outcomes/ No. of Total outcomes=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

**Ans:** a) 0

b) 1/6

c) 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:** No. of favorable outcomes=5C2=5\*4/2=10

No. of total outcomes=7C2=7\*6/2=21

P(!B) =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= (Σ Candies count \*Probability)

E(C)=1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

E(C)=3.09

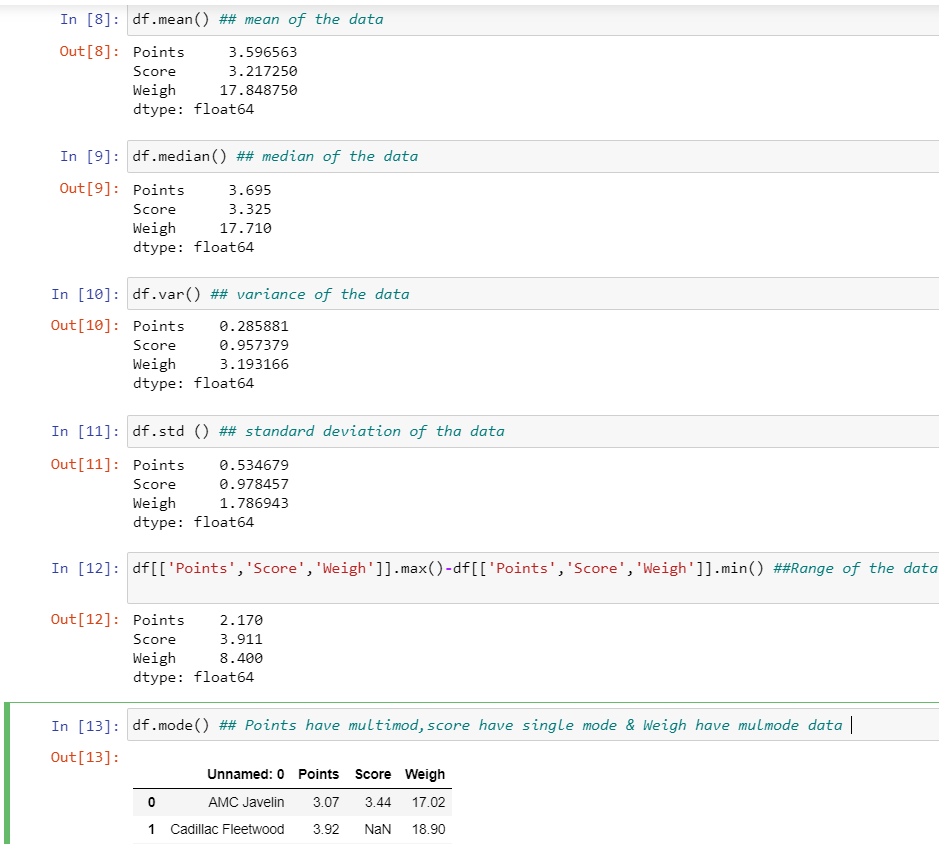
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* 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:**



**Inference Drawn:**

* The mean is useful for spotting trends in the data because we can compare means over a time period to spot trends. The mean is the most common measure of central tendency.
* The median divides a sample of data in half; it is the middle score. The median is a useful statistic if we think our data have some extreme cases. The median is not impacted by extreme cases, but the mean is.
* In the dataset Points variable don’t have outlier but it is approximately normally distributed, Score variable have outlier but it is positive skewed & Weigh also have outlier but it is positive 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?

**Ans:**

Expected Value of the Weight of that patient=Sum of all patients weight/Total no. of patient

Expected Value of the Weight of that patient= (108+110+123+134+135+145+167+187+199)/9

Expected Value of the Weight of that patient=145.3333333

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

**Cars speed and distance**

**Use Q9\_a.csv**



**Ans:**

**A) SPEED: -** Skewness is negative, that tells us that the distribution is skewed towards left. Mean of distribution is less than the Median. Kurtosis Value is less than 0, that tells us that the distribution has broad peak and thin tails as evident from the histogram.

**B) DISTANCE:** - Skewness is positive, that tells us that the distribution is skewed towards right. Mean of distribution is more than the Median. Kurtosis Value is more than 0, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.

**SP and Weight (WT)**

**Use Q9\_b.csv**



**Ans:**

**(A) SP:** Skewness is positive, that tells us that the distribution is skewed towards right. Mean of distribution is more than the Median. Kurtosis Value is more than 0, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.

**(B) WT:** Skewness is negative, that tells us that the distribution is skewed towards left. Mean of distribution is less than the Median. Kurtosis Value is more than 0, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.

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



**Ans:**

1. In histogram of chickweight we observed that the data is positive or right skewed & Mean>Median.
2. The boxplot is defined that the distribution has lot of outliers towards upper extreme

**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?

**Ans:** Sample size is 200

Sample weighs is (Mean of sample) is 200 pounds

Standard deviation is 30 of population Standard deviation of sample will be=30/sqrt (2000) =0.67

94% confidence interval=(200+1.88\*0.67)&(200-1.88\*0.67)=198.74 to 201.26

96% confidence interval=(200+2.05\*0.67)&(200-2.05\*0.67)=198.63 to 201.37

98% confidence interval=(200+2.33\*0.67)&(200-2.33\*0.67)=198.44 to 201.56

**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?

**Ans:**

1) Mean of student scores = 41

Median of student scores = 40.5

Standard deviation of student scores = 5.05

Variance of student scores = 25.53

2)There are two outliers present in student marks 49,56 who is affected for the mean/avg value, we can say mass of student marks concentered on left hand side & distribution is positive skewed and Mean>Median.

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

**Ans:** - The data is symmetrical distributed.

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

**Ans:** - The data is positive or right-hand side skewed.

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

**Ans:** - The data is negative or left-hand side skewed.

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

**Ans:** - Positive kurtosis implies that narrow peak & heavier tails.

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

**Ans:** - Negative kurtosis implies that wider peak & thinner 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)?

**Ans:** a) The data are not symmetrical or not normally distributed & outliers not present in the data

1. Data is negative or left skewed because mass of distribution is concentrated in right hand side
2. Q1= 10, Q2=15.5, Q3=18 then 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.

**Ans:** a) Both are normally distributed & we can say skewness is similar to zero

1. Mean & median for each dataset is similar (262.5)

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)
  3. 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

**Ans:** - As observed in histogram we can say MPG follow the normal distribution as well as Skewness of data is -0.17 is greater than -0.5 & mean (34.42) is similar to median (35.15)

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

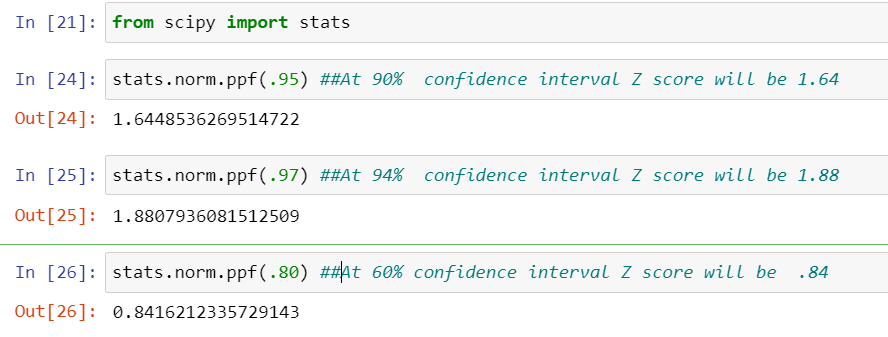
Dataset: wc-at.csv

**Ans: -**

1. Waist follow the normal distribution as well as Skewness of data is 0.13 is less than 0.5 & mean (91.9) is similar to median (90.8)
2. AT don’t follow the normal distribution as well as Skewness of data is 0.58 is greater than 0.5 & mean (101.89) is far from median (96.54)

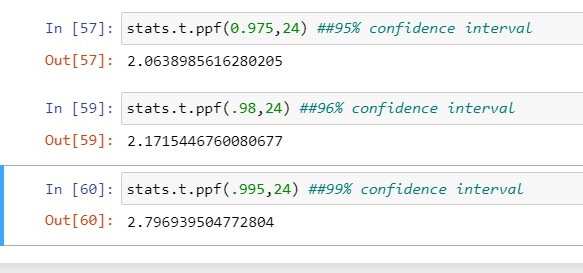
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**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:** 