|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| 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 | Interval |
| 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 | Ratio |

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

Ans:-

When three coins are tossed the sample space is given by,

S={HHH,HHT,HTH,HTT,,THH,THT,TTH,TTT }

P(S)=8

We have to find the probability having that two heads and one tail . this sample space we denoted as A & A sample space is given by,

A={HHT,HTH,THH}

P(A)= 3/8= 0.3

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:-

When two Dice are rolled

Total no. of outcomes=6^2=36

1. Probability=0 having the probability that sum is equal to 1
2. Less than equal to 4,

Possible events=(2,2),(3,1),(1,3),(1,1),(1,2),(2,1)=6

Probability=6/36=1/6

c)sum is divisible by 2 and 3,

Possible events = (3,3),(2,4),(4,2),(6,6),(1,5),(5,1)=6

Probability=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 number of events= ===21

Interested events===10

Probability that none of the balls is blue =10/21=0.47

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

Ans:-

Expected number = E(x)\*P(x)=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.

**U df\_7.describe()**

index , Points , Score, Weigh

**count 32.000000 32.000000 32.000000**

**mean 3.596563 3.217250 17.848750**

**std 0.534679 0.978457 1.786943**

**min 2.760000 1.513000 14.500000**

**25% 3.080000 2.581250 16.892500**

**50% 3.695000 3.325000 17.710000**

**75% 3.920000 3.610000 18.900000**

**max 4.930000 5.424000 22.900000**

**df\_7.mode()**

**index Points Score Weigh**

**0 3.07 3.44 17.02**

**1 3.92 NaN 18.90**

**df\_7['Points'].max()-df\_7['Points'].min()#range\_points**

**df\_7['Score'].max()-df\_7['Score'].min()#range\_score**

**df\_7['Weigh'].max()-df\_7['Weigh'].min() #range\_weigh**

**df\_7.var()**

Points 0.285881

Score 0.957379

Weigh 3.193166

dtype: float64

**df\_7['Points'].hist()**

<AxesSubplot:>

**se Q7.csv file**

**Ans:-**

**Ans.** df\_7.describe()

index, Points, Score, Weigh

count,32.0,32.0,32.0

mean,3.5965625,3.2172500000000004,17.848750000000003

std,0.5346787360709715,0.9784574429896966,1.7869432360968431

min,2.76,1.513,14.5

25%,3.08,2.58125,16.8925

50%,3.6950000000000003,3.325,17.71

75%,3.92,3.61,18.9

max,4.93,5.424,22.9

df\_7.mode()

index,Points,Score,Weigh

0,3.07,3.44,17.02

1,3.92,NaN,18.9

df\_7['Points'].max()-df\_7['Points'].min()= 2.17#range\_points

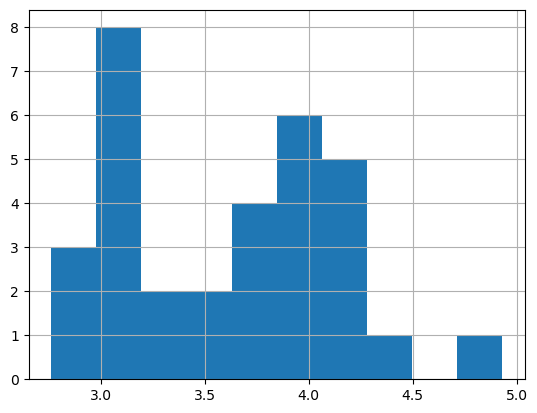
df\_7['Score'].max()-df\_7['Score'].min()=3.9110000000000005#range\_score

df\_7['Weigh'].max()-df\_7['Weigh'].min()=8.399999999999999 #range\_weigh

df\_7.var()

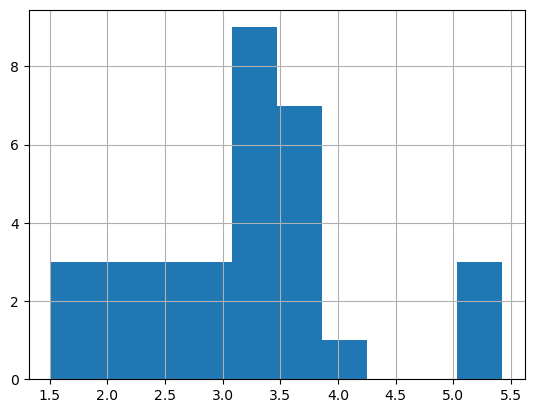
Points: 0.285881 ,Score: 0.957379 ,Weigh: 3.193166 ,dtype: float64

df\_7['Points'].hist()

****

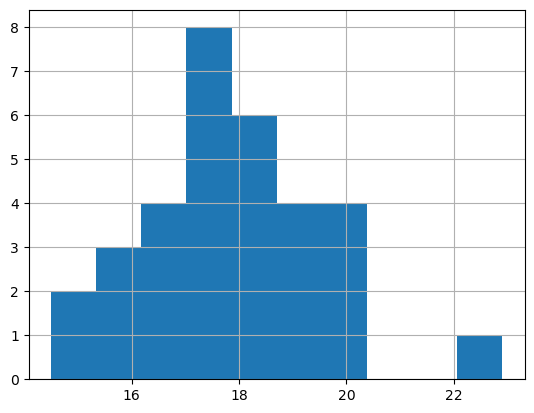
**df\_7['Score'].hist()**

AxesSubplot:>

****

**df\_7['Weigh'].hist()**

<AxesSubplot:>

****

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?

weight=[108,110,123,134,135,145,167,187,199]

weight\_mean=np.mean(weight)

print(weight\_mean)

145.33333333333334

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

**Cars speed and distance**

**Use Q9\_a.csv**

**df\_9a=pd.read\_csv(r"C:\Users\POONAM PATIL\Desktop\Data Science Assignment\Basic stats1\Q9\_a (1).csv")**

**df\_9a**

**df\_9a=df\_9a.set\_index('Index')**

count 50.000000

mean 15.400000

std 5.287644

min 4.000000

25% 12.000000

50% 15.000000

75% 19.000000

max 25.000000

Name: speed, dtype: float64

**df\_9a['dist'].describe()**

count 50.000000

mean 42.980000

std 25.769377

min 2.000000

25% 26.000000

50% 36.000000

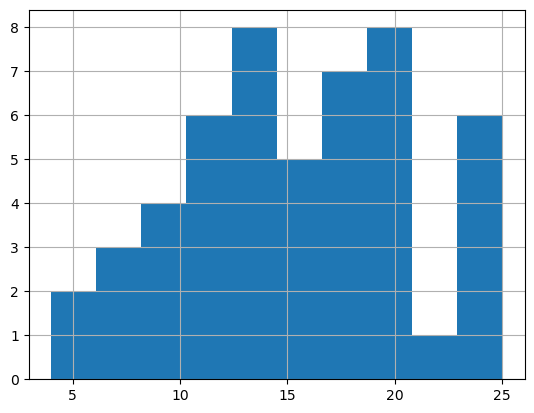
75% 56.000000

max 120.000000

Name: dist, dtype: float64

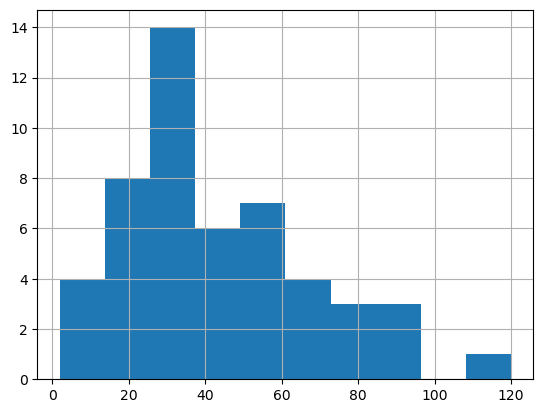
**df\_9a['speed'].hist()**

<AxesSubplot:>

****

**df\_9a['dist'].hist()**

<AxesSubplot:>

****

**df\_9a['dist'].skew()**0.8068949601674215

**df\_9a['speed'].skew()** -0.11750986144663393

**Observation**

"dist" is positively skewed where as “speed” is negatively skewed. Thus, dist has distribution of data concentrated on the left whereas speed has distribution on the right. As seen in the graph

**import scipy**

**from scipy import stats**

**df\_9a['speed'].kurtosis()** -0.5089944204057617

**df\_9a['dist'].kurtosis()** 0.4050525816795765

**Obsevation:**

"speed has negative kurtosis while "dist" has positive kurtosis

**SP and Weight(WT)**

**Use Q9\_b.csv**

**df\_9b=pd.read\_csv(r"C:\Users\POONAM PATIL\Desktop\Data Science Assignment\Basic stats1\Q9\_b (1).csv")**

**df\_9b**

**df\_9b.describe()**

**Unnamed: 0 SP WT**

**count 81.000000 81.000000 81.000000**

**mean 41.000000 121.540272 32.412577**

**std 23.526581 14.181432 7.492813**

**min 1.000000 99.564907 15.712859**

**25% 21.000000 113.829145 29.591768**

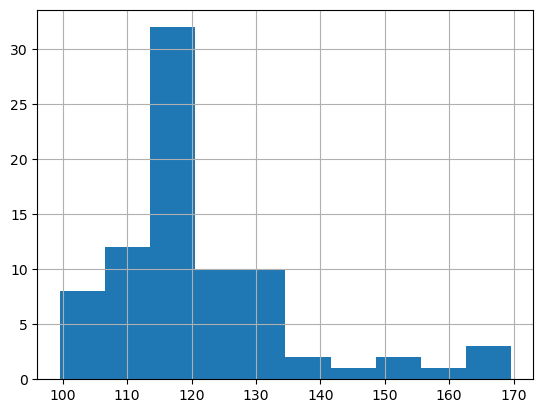
**50% 41.000000 118.208698 32.734518**

**75% 61.000000 126.404312 37.392524**

**max 81.000000 169.598513 52.997752**

**df\_9b['SP'].hist()**

<AxesSubplot:>

****

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

Ans:-

1. Majority of the chicks has weight in range 50-100, followed by 100-150 and 150-200
2. The data is positively skewed



Ans:- data has lots of outliers towards the 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:-

n=2000

= 200

s= 30

Confidence Interval Estimate= Z => 200 Z

**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?

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

Ans:- Skewness=0

Perfectly symmetric bell shaped curve

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

Ans:-

Skewness =positive .Data is distributed more on left.

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

Ans:- Skewness= . Data is distributed more on right .

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

Ans:- Peaked ness ( sharp peak) and less variation.

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

Ans: - less peaked Ness (broad peak) and more variation

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



What can we say about the distribution of the data?

Ans:-

In above data visualization maximum data points lie in lower extreme and there is one outliers present in lower extreme of the data visualization.

The data is not symmetric. Data is more concentrated towards right side

What is nature of skewness of the data?

Ans:- In above boxplot visualization left tell is longer than right tell then data is Negative skew

What will be the IQR of the data (approximately)?   
  
Ans:- IQR=Upper extreme- Lower Extreme

=18-10

IQR=8

Q19) Comment on the below Boxplot visualizations?



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

Ans:-

1. Data is normally distributed. No outliers. Center around262.5 . Comparitively, first boxplot has less range
2. Data is normally distributed. No outliers. Center around 262.5.

Comparitively , second boxplot has more range.

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

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