|  |  |
| --- | --- |
| 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 | Ordinal |
| 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?

Tossed Three Coins:

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

Probability for Two heads and One tail:

P(X) = Desired outcome/Total np of Events

P(X= 2H and 1T) = 3/8 = 0.375=37.5%

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

Rolling Dice -Total Possible outcomes = 62 = 36

1. P(X=1) = (0/36) =0
2. P(X<=4) = Favorable outcomes = (1 , 1) , (1 , 2) , (1 , 3) , (2, 1) , (2 , 2),(3,1)

(6/36) = 1/6=0.16 = 16.66%

1. Favorable outcomes = (1 ,5), (2,4), (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6)

P (Sum divisible by 2 and 3) = 6/36 =1/6=0.16 = 16.66%

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  
Let S be the sample space.  
Then, n(S) = Number of ways of drawing 2 balls out of 7  
= 7C2​  
= (7×6)​/ (2×1)  
=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.  
= 5C2​  
= (5×4)​/ (2×1)  
= 10  
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

Ans: 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\*0.120)

= 3.09

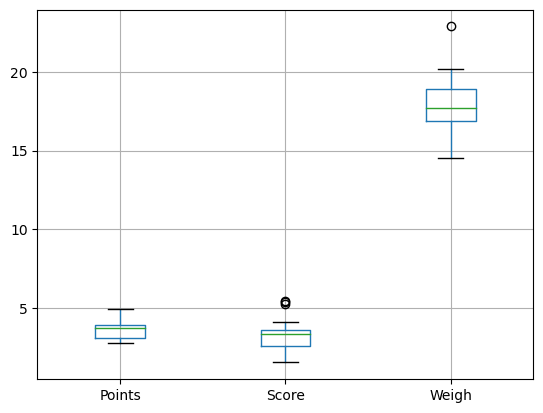
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* 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**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weight |
| Mean | 3.60 | 3.22 | 17.85 |
| Median | 3.70 | 3.33 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.29 | 0.96 | 3.19 |
| Standard Deviation | 0.53 | 0.98 | 1.79 |
| Range | 2.17 | 3.91 | 8.40 |



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?

The Weight of that patient is 118.

=RANDBETWEEN(C9,C17)

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans: import pandas as pd import numpy as np from scipy import stats

df=pd.read\_csv('Q9\_a.csv',index\_col = 'Index')

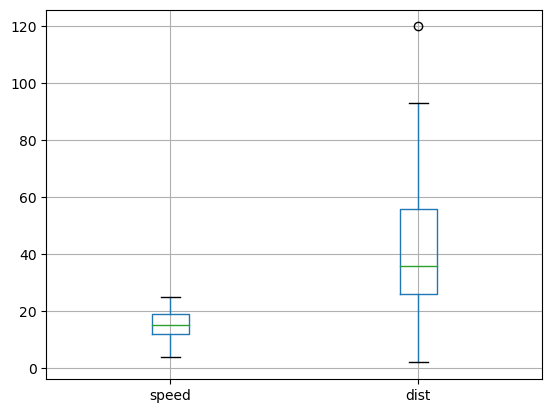
print('For Cars Speed', "Skewness value=", np.round(df.speed.skew(),2), 'and' , 'Kurtosis value=', np.round(df.speed.kurt(),2))

For Cars Speed Skewness value= -0.12 and Kurtosis value= -0.51

print('For Cars Distance','Skewness value =', np.round(df.dist.skew(),2),'and', 'Kurtosis value =', np.round(df.dist.kurt(),2))

For Cars Distance Skewness value = 0.81 and Kurtosis value = 0.41

df.boxplot()



**Use Q9\_b.csv**

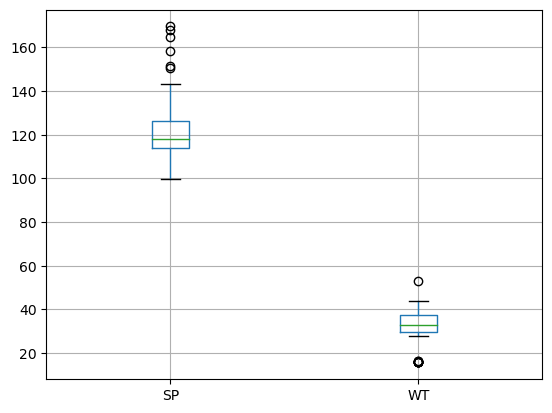
df1=pd.read\_csv('Q9\_b.csv',index\_col=0)

print('For Cars SP', "Skewness value=", np.round(df1.SP.skew(),2), 'and' , 'Kurtosis value=', np.round(df1.SP.kurt(),2))

For Cars SP Skewness value= 1.61 and Kurtosis value= 2.98

Print('For Cars WT', "Skewness value=", np.round(df1.WT.skew(),2), 'and' , 'Kurtosis value=', np.round(df1.WT.kurt(),2))

For Cars WT Skewness value= -0.61 and Kurtosis value= 0.95



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

Ans: The histograms peak has positive or right skew and tail is on right. Mean > Median. We have outliers on the higher side.

Ans: The boxplot has outliers on the maximum 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?

Sample n =2000

Population =3,000,00

Sample mean=loc=200

Standard deviation of sample=30

Scale=30/np.sqrt(2000)

Confidence interval for 94%

stats.t.interval (0.94,df=1999,loc=200,scale=30/np.sqrt(2000))

Ans: (198.74, 201.26)

In [15]:

Confidence interval for 98%

stats.t.interval(0.98,df=1999,loc=200,scale=30/np.sqrt(2000))

(198.43, 201.56)

Confidence interval for 96%

stats.t.interval(0.96,df=1999,loc=200,scale=30/np.sqrt(2000))

(198.62, 201.37)

**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.

Ans: Mean =41, Median =40.5, Variance =25.52 and

Standard Deviation =5.05

1. What can we say about the student marks?

Ans: we don’t have outliers and the data is slightly skewed towards right because mean is greater than median(mean>median).

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

Ans: No skewness is present we have a perfect symmetrical distribution(normal Distribution)

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

Ans: Skewness and tail is towards Right (Positive or Right Skewness)

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

Ans: Skewness and tail is towards left (Negative or left Skewness)

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

Ans: Positive kurtosis means the curve is more peaked and it is Leptokurtic

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

Ans: Negative Kurtosis means the curve will be flatter and broader

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



What can we say about the distribution of the data?

Ans: The above Boxplot is not normally distributed the median is towards the higher value

What is nature of skewness of the data?

Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum

What will be the IQR of the data (approximately)?   
Ans: The Inter Quantile Range IQR = Q3 Upper quartile – Q1 Lower Quartile =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: First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 250 to 275 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker 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. A .P(MPG>38)

Ans: Prob\_gr\_38 = np.round(1 - stats.norm.cdf(38, loc= df.MPG.mean(), scale= df.MPG.std()),3) print('P(MPG>38)=',Prob\_gr\_38)

P(MPG>38)= 0.348

1. P(MPG<40)

Ans: Prob\_ls\_40 = np.round(stats.norm.cdf(40, loc= df.MPG.mean(), scale= df.MPG.std()),3) print('P(MPG<38)=',Prob\_ls\_40)

P(MPG<38)= 0.729

1. P (20<MPG<50)

Ans: Prob\_gr\_20 = np.round(1 - stats.norm.cdf(20, loc= df.MPG.mean(), scale= df.MPG.std()),3) print('P(MPG>20)=',Prob\_gr\_20) P(MPG>20)= 0.943

Prob\_ls\_50 = np.round(stats.norm.cdf(50, loc= df.MPG.mean(), scale= df.MPG.std()),3) print('P(MPG<50)=',Prob\_ls\_50)

P(MPG<50) = 0.956

prob\_20\_50 = Prob\_ls\_50-Prob\_gr\_20 print('P(20<MPG<50) =', prob\_20\_50)

P(20<MPG<50) = 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: a.) MPG of cars follows 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

 Ans: Adipose Tissue (AT) and Waist does not follow Normal Distribution

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

print('Z score for 90% Confidence Interval =',np.round(stats.norm.ppf(.05),4)) Z score for 90% Confidence Interval = -1.6449

print('Z score for 94% Confidence Interval=',np.round(stats.norm.ppf(.03),4)) Z score for 94% Confidence Interval = -1.8808

print('Z score for 60% Confidence Interval=',np.round(stats.norm.ppf(.2),4)) Z score for 60% Confidence Interval = -0.8416

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

print('t score for 95% Confidence Interval =', np.round(stats.t.ppf(.025,df=24),4))

t score for 95% Confidence Interval = -2.0639

print('t score for 96% Confidence Interval =', np.round(stats.t.ppf(.02,df=24),4))

t score for 96% Confidence Interval = -1.8281

print('t score for 99% Confidence Interval =', np.round(stats.t.ppf(.005,df=24),4))

t score for 99% Confidence Interval = -2.7969

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

Ans: t \_score = (x - pop mean) / (sample standard daviation / square root of sample size) =(260-270)/(90/np.sqrt(18))

stats.t.cdf(t\_score, df=17) =0.32=32%

The probability of 18 randomly selected bulbs would have an average life of no more than 260 days is 32%