|  |  |
| --- | --- |
| 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 |

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

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

**Ans** : The total possible outcome 2^3=8

HHH,HHT,HTT,THT,TTH,HTH,THH,TTT

No of favorable outcomes =3

P=3/8=0.374

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**:

Number of possible outcomes for the above event is

N (Event (Two dice rolled)) = 6^2 = 36

1. P (sum is Equal to 1) = ‘0’ zero null nada none.
2. P (Sum is less than or equal to 4) = N (Event (Sum is less than or equal to

4)) / N (Event (Two dice rolled))

= 6 / 36 = 1/6 = 0.166 = 16.66%

1. P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3)) / N

(Event (Two dice rolled))

=1/ 36 = 0.139

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 balls =7 balls

N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5!

= (7\*6\*5\*4\*3\*2\*1) /

(2\*1) \* (5\*4\*3\*2\*1)

N (Event (2 balls are drawn randomly from bag) = (7\*6)/ (2\*1) = 21

If none of them drawn 2 balls are blue = 7 – 2 = 5

N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (5\*4) / (2\*1)

= 10

P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) /

N (Event (2 balls are drawn randomly from

bag)

= 10 / 21 = 0.476

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**:

0.015+0.8+1.95+0.025+0.06+0.24 = 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**

**Ans**:

Mean for Points = 3.59, Score = 3.22 and Weigh = 17.85

Median for Points = 3.70, Score = 3.33 and Weigh = 17.71

Mode for Points = 3.92, Score = 3.44 and Weigh = 17.02

Variance for Points = 0.29, Score = 0.96, Weigh = 3.19

Standard Deviation for Points = 0.53, Score = 0.98, Weigh = 1.79

Range Points= 2.17,Score=2.91 ,Weight=8.40

Draw Inferences



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**:

145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans**:

**df=read.csv('Q9\_a.csv')**

> # skewness and kurtosis for speed

> skewness(df$speed)

[1] -0.1139548

> kurtosis(df$speed)

[1] 2.42

>

> #skewness and kurtosis for dist

> skewness(df$dist)

[1] 0.76

> kurtosis(df$dist)

[1] 3.25

**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans :-

> df=read.csv('Q9\_b.csv')

> #skewness and kurtosis for SP

> skewness(df$SP)

[1] 1.55

> kurtosis(df$SP)

[1] 5.72

>

> #skewness and kurtosis for WT

> skewness(df$WT)

[1] -0.59

> kurtosis(df$WT)

[1] 3.82

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



**Ans**:

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



**Ans**:

Data has outliers

-Data is positively 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?

**Ans**:

conf\_94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))

print(np.round(conf\_94,0))

print(conf\_94)

For 94% confidence interval Range is [ 198.96 – 201.04]

For 98% confidence interval range is [198.62 – 201.38]

For 96% confidence interval range is [198.83 – 201.17]

|  |  |  |  |
| --- | --- | --- | --- |
|  | 94% | 98% | 96% |
| Upper | 201.04 | 201.38 | 201.17 |
| Lower | 198.96 | 198.62 | 198.83 |

**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 =25.53 and Standard Deviation =5.05

|  |  |
| --- | --- |
| Mean | 41 |
| Median | 40.50 |
| Variance | 25.53 |
| Standard Deviation | 5.05 |

1. What can we say about the student marks?

**Ans**:

1. Data has outlier
2. Not normally distributed
3. Number of the students scored between 35 – 45 Marks

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

**Ans**:

Skewness = 0.

- Symmetric bell shaped curve Perfectly

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 and tail is towards left

Skewness = Negative. Data is distributed more on right

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

**Ans**:

High and narrow peak on central part of the data

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

**Ans**:

wider peak on central part of the data

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



What can we say about the distribution of the data?

**Ans**:

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

What is nature of skewness of the data?

**Ans**:

Skewness = Negative

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

**Ans**:

The Inter Quantile Range = 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**:

1. Data is Normally Distributed.

No Outliers.

1st graph has less range.

Center around 262.5.

1. Data is Normally Distributed.

No Outliers.

2nd graph has more range.

Center around 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)

**Ans**:

Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3)

print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.4074074

* 1. P(MPG<40)

**Ans**:

prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.7530864

* 1. P (20<MPG<50)

**Ans**:

prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20))

p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))

P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)

print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

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**:

1. 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

**Ans**:

# z value for 90% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.05),4))

Z score for 60% Conifidence Intervla = -1.711

# z value for 94% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.03),4))

Z score for 60% Conifidence Intervla = -1.828

# z value for 60% confidence interval

print('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.2),4))

Z score for 60% Conifidence Intervla = -2.492

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

**Ans**:

# t score for 95% confidence interval

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))

T score for 95% Confidence Interval = -2.0639

# t value for 94% confidence interval

print('T score for 94% Confidence Inteval =',np.round(stats.t.ppf(0.03,df=24),4))

T score for 94% Confidence Inteval =-2.167

# t value for 99% Confidence Interval

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

T score for 95% 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**:

Import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

(260-270)/90/np.sqrt(18))

t\_score = -0.471

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

0.32 = 32%