Q1) Identify the Data type for the Following:

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

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 | **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 | **Nominal** |
| Religious Preference | **Nominal** |
| Barometer Pressure | **Ratio** |
| SAT Scores | **Interval** |
| Years of Education | **Ratio** |

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

**Ans) Possible outcomes: HHH , HHT, HTH , THH, TTH, THT , HTT , TTT**

**Probability of 2 heads and 1 tail = no of times 2 heads & 1 head / total no of**   **outcomes**

**Probability =3/8 = 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) (a) zero . because it is not possible to get the sum equals 1. (lowest possible**  **sum is 2)**

**(b) (1,1),(1,2),(1,3),(2,2),(3,1),(2,1) = 6**

**Probability = 6/36 =1/6**

**(C) sum divisible by 2 and 3 are 6 and 12**

**Sum is 6 = (1,5),(2,4),(3,3),(4,2),(5,1) =5**

**Sum is 12 =(6,6) =1**

**Probability =5+1 /36 = 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 possible outcome =7\*6 = 42**

**Total possible outcomes without blue balls = 5\*4 =20**

**Probability = 20/42 =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 \* 0.12**

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

**= 3.090**

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)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | **3.5965625** | **3.21725** | **17.84875** |
| **Median** | **3.695** | **3.325** | **17.71** |
| **Mode** | **3.92** | **3.44** | **17.02** |
| **Variance** | **0.276947559** | **0.927460875** | **3.093379688** |
| **Standard Dev** | **0.526258072** | **0.963047701** | **1.758800639** |
| **Range** | **2.17** | **3.911** | **8.4** |

**Inferences:**

* **Points and score have less standard deviation and variance as they lies within a small range**
* **Weigh datas have high spread , as their std dev , variance are more.**

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) mean (X) = 1308/9**

**= 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans)**

|  |  |  |
| --- | --- | --- |
|  | **speed** | **distance** |
| **skewness** | -0.11395477 | 0.782483517 |
| **kurtosis** | -0.50899442 | 0.405052582 |

**Inference:**

* **The data distribution of speed is left skewed (negative skewness) and wide peak and short tail (negative kurtosis)**
* **The data distribution of distance is right skewed (positive skewness) and wide peak and small tail (less than 3)**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans)**

|  |  |  |
| --- | --- | --- |
|  | **speed** | **weight** |
| **skewness** | 1.581453679 | -0.603309932 |
| **kurtosis** | 2.977328944 | 0.950291491 |

**Inference:**

* **Data distribution of speed is right skewed ( positive skewness) and have approximate normal distribution( value near 3)**
* **Data distribution of weight is left skewed (negative skewness) and have wide peak and short tails( less than 3)**

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



**Ans) HISTOGRAM:**

* **Right skewed data distribution.**
* **Range is 0 to 400.**
* **Majority of data is in between 0 to 200.**
* **Most frequent data in 50 to 100.**

**BOX PLOT :**

* **50 percent of data lies in small range.**
* **Last 25 percent data are spread in a large range.**
* **There are many outliers .**



**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 n=2,000**

**Population size N=3,000,000**

**Sample mean x` =200**

**Sample std dev s=30**

**94 % confidence ---> CI =200 +- 1.261**

**96 % confidence ---> CI = 200+- 1.375**

**98 % confidence ---> CI = 200+- 1.563**

**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.

**Mean= 41.56**

**Median=40.5**

**Mode=41**

**Variance=95.06**

**Standard Deviation=9.75**

1. What can we say about the student marks?

**Student’s mark is increasing gradually. Most of the marks lies in between 38 and 42. The mean , median and mode is very near lies between 40 and 42.**

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

**Ans) Skewness will be zero. Because it is symmetric distribution when mean and mode is equal.**

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

**Ans)positive skewness . It is right tailed distribution when mean is greater than mode.**

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

**Ans) negative skewness . It is left tailed distribution when mean is less than mode.**

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

**Ans) positive kutosis values indicates the sharper peak and long tails of the data distribution. It says that most of the data lies on the tails than the centre.**

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

**Ans) it indicates the flatter peak and short tails of data distribution.**

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



What can we say about the distribution of the data?

**Ans) left tailed distribution. First 25% data lies within a wide range from 1 to 10.**

**What is nature of skewness of the data?**

**Ans)negative skewness**

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

**Ans) IQR = upper quartile – 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) Boxplot 1 data is densely packed and lies within a small range .it has small IQR and small whiskers. Boxplot 2 data is widely packed and lies within a large range .it has larger IQR and large whiskers. The median of both boxplot is equal.**

**Boxplot 1 has less range and boxplot 2 has larger 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)

**Ans) a)1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())**

**0.3475939251582705**

**B)# P(MPG<40)**

**stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())**

**0.7293498762151616**

**C) # P (20<MPG<50)**

**stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())- stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.std())**

**1.243096879732761**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans) Mean = 34.42207573**

**Median = 35.15272697**

**Skewness = -0.174634338**

**Mean ~ median and also skewness is near to zero**

**So, the data approximately 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

**Ans)**

|  |  |  |
| --- | --- | --- |
|  | **Waist** | **AT** |
| **mean** | **91.90183486** | **101.8940367** |
| **median** | **90.8** | **96.54** |
| **skewness** | **0.132204176** | **0.576789698** |

**Waist is approximately normally distributed as mean ~ median ~mode and also skewness nearer to 0**

**AT is not normally distributed as mean and median is differ also skewness is larger than 0.**

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

**Ans) CI 90% ----> Z score =1.642**

**CI 94% ----> Z score =1.880**

**CI 60% ----> Z score =0.842**

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

**Ans)**

**CI 95% ---> t score =2.064**

**CI 96% ---> t score =2.171**

**CI 99% ---> t score =2.797**

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 à pt(tscore,df)

df à degrees of freedom

**Ans) sample size n=18 population mean µ =270**

**sample mean x`=260**

**Sample std dev = 90**

**t= ---> 260-270/90/**

**t = -0.4714**

**Df=n-1 = 17**

**P(t=-0.417, df=17) =0.32**

**----> 32%**