|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | INTEGER (INT)(Discrete) |
| Results of rolling a dice | INTEGER (INT)( Discrete) |
| Weight of a person | FLOAT(Continuous) |
| Weight of Gold | FLOAT(Continuous) |
| Distance between two places | FLOAT(Continuous) |
| Length of a leaf | FLOAT(Continuous) |
| Dog's weight | FLOAT(Continuous) |
| Blue Color | STRING(Nominal) |
| Number of kids | INTEGER (INT)(Discrete) |
| Number of tickets in Indian railways | INTEGER (INT)(Discrete) |
| Number of times married | INTEGER (INT)(Discrete) |
| Gender (Male or Female) | STRING(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 | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Interval |
| 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 : Possible out come : {HHH , TTT,TTH , HHT,THT,HTH,THH,HTT}

Favourable out come : {HHT,HTH,THH}

Probability : 3 / 8 = 0.375

**Q4) Two Dice are rolled, find the probability that sum is**

**Ans : Possible outcomes :** *(1, 1) (1, 2) (1, 3) (1, 4) (1, 5) (1, 6)*

*(2, 1) (2, 2) (2, 3) (2, 4) (2, 5) (2, 6)*

*(3, 1) (3, 2) (3, 3) (3, 4) (3, 5) (3, 6)*

*(4, 1) (4, 2) (4, 3) (4, 4) (4, 5) (4, 6)*

*(5, 1) (5, 2) (5, 3) (5, 4) (5, 5) (5, 6)*

*(6, 1) (6, 2) (6, 3) (6, 4) (6, 5) (6, 6)*

1. **Equal to 1**

Ans : 0

1. **Less than or equal to 4**

Ans : Probability : 3/36 = 1/12 = 0.17

1. **Sum is divisible by 2 and 3**

Ans : Probability : 6/36 = 1/6 = 0.17

**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 outcome = 7C2 = 21 (total number of balls 2 + 3 + 2 of which 2 are drawn therefore using combination formula nCr = n! / r! \* (n-r)! we get total outcome)

Favourable outcome = 5C2 = 10 (total number of balls that we now want is 5 since no blue ball of which 2 are drawn therefore using combination formula nCr = n! / r! \* (n-r)! we get Favourable outcome)

Probability = 10/21 = 0.467

**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 Candies = 1 x 0.015 + 4 x 0.20 + 3 x 0.65 + 5 x 0.005 + 6 x 0.01 + 2 x 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.**

Ans : Points : {'Mean': 3.6, 'Median': 3.7, 'Mode': 3.07, 'variance': 0.29, 'standard deviation': 0.53, 'range': [2.76, 4.93]}

Score : {'Mean': 3.22, 'Median': 3.32, 'Mode': 3.44, 'variance': 0.96, 'standard deviation': 0.98, 'range': [1.513, 5.424]}

Weigh : {'Mean': 17.85, 'Median': 17.71, 'Mode': 17.02, 'variance': 3.19, 'standard deviation': 1.79, 'range': [14.5, 22.9]}

Inference : Score and weigh have outliers in upper range, all the data set are fairly symmetrical since skewness value is between -0.5 to 0.5 . The mean value for Score and weigh is affected by the outlier so they are not reliable.

**Use Q7.csv file**

**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 = (1/9)(108) + (1/9)110 + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

= 145.33

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

**Cars speed and distance**

Ans : speed:= Skewness : -0.12

kutrosis : -0.51

dist:= Skewness : 0.81

kutrosis : 0.41

speed is fairly symmetrical since skewness is between -0.5 to 0.5 means there is likely no outliers and the dist column is positively skewed that means there is outlier in upper region of distribution.

**Use Q9\_a.csv**

**SP and Weight(WT)**

Ans : SP:= Skewness : 1.61

kutrosis : 2.98

WT:= Skewness : -0.61

kutrosis : 0.95

SP is highly skewed and has outliers in upper limits, WT is moderately skewed and has outliers in both upper and lower limits

**Use Q9\_b.csv**

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



Ans : The distribution is positively skewed as seen in histogram and by observing box plot we see the sample has outliers in upper limit.

**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 : 1)for 94% : t-score = 1.88 ,

interval range = [200 – 1.88 \* 30/ √2000 , 200 + 1.88 \* 30/ √2000]

interval range = [198.74, 201.26]

2)for 98% : t-score = 2.33 ,

interval range = [200 – 2.33 \* 30/ √2000 , 200 + 2.33 \* 30/ √2000]

interval range = [198.44, 201.56]

3)for 96% : t-score = 2.055 ,

interval range = [200 – 2.33 \* 30/ √2000 , 200 + 2.33 \* 30/ √2000]

interval range = [198.62, 201.38]

**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 = 41.0, median=40.5, variance=24.11 , std = 4.910306620885412

1. **What can we say about the student marks?**

Ans : Since mean is slightly greater than media we observe positive skewness.

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

Ans : The skewness is 0 and distribution is symmetrical

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

Ans : The skewness is positive

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

Ans: The Skewness is Negative

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

Ans : Distribution is peaked and number of values present in the data is near mean.

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

Ans : Distribution is wide spread and number of values present in the data is more towards the tail i.e away from mean.

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



**What can we say about the distribution of the data?**

Ans : It is observed that the mean of the distribution is less than the median.

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

Ans : Since the mean is less than median the distribution is negatively skewed.

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

Ans : IQR = 75 percentile – 25 percentile = 18 – 10 = 8 (approx.)

**Q19) Comment on the below Boxplot visualizations?**



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

Ans : Both the distribution have same median. The range of values for distribution 2 is higher than range of values of distribution 1. There is no skewness in either of the distribution and none of them contain outliers.

**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 : 1 - stats.norm.cdf(38,34.42,9.07) = 0.3465292624921241

* 1. **P(MPG<40)**

Ans : stats.norm.cdf(40,34.42,9.07) = 0.7307936266788738

**c. P (20<MPG<50)**

Ans: stats.norm.cdf(50,34.42,9.07) – (1- stats.norm.cdf(20,34.42,9.07)) = 0.013012521130224997

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

Ans: The MPG distribution does not 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 : The distribution of the Adipose Tissue (AT) and Waist Circumference(Waist) does not follow normal distribution.

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

Ans : for 90% z-score : 1.6448536269514722

For 94% z-score : 1.8807936081512509

For 60% z-score : 0.8416212335729143

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

Ans : for 95% t-score : 2.0638985616280205

For 96% t-score : 2.1715446760080677

For 99% t-score : 2.796939504772804

**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: Null Hypothesis : Population mean is equal to 270

Alternate Hypothesis : population mean < 270

tstat = (260-270)/(90/18\*\*0.5) = -0.4714045207910317

pvalue = stats.t.cdf(-0.4714045207910317,17)

pvalue(0.32167253567098353) > alpha value(0.05) therefore the assumptions made are correct and the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.32167253567098353