|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Descrete |
| Results of rolling a dice | Descrete |
| 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 | Descrete |
| Number of tickets in Indian railways | Descrete |
| Number of times married | Descrete |
| Gender (Male or Female) | Categorical/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 | Interval |
| Celsius Temperature | Ratio |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Interval |

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

**Ans:** events 3 out comes /possibilities 8. Probability is 3/8=0.375 or 37.50%

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:**

1. Zero
2. Events 3 for 4, events 2 for 3, event 1 for 2, Possibilities/outcomes is 36. Probability is 6/36=0.1667 or 16.67%
3. Events 5 for 6, events 1 for 12. Possibilities/outcomes is 36. Probability is 6/36=0.1667 or 16.67%

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:** events 6, Possibilities/outcomes is 16. Probability is 6/16=0.375 or 37.50%

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:** Highest probability is 0.65 & candies count is 3. So expected number of candies for a randomly selected child is 3

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 | MEDIAN | Variance | Std. Dev | Range |
| Points | 3.5966 | 3.695 | 0.2859 | 0.5347 | 2.1700 |
| Score | 3.2173 | 3.325 | 0.9574 | 0.9785 | 3.9110 |
| Weigh | 17.8487 | 17.710 | 3.1932 | 1.7869 | 8.4000 |

Mode Points: Multimodel mode (values 2.76,3.07,3.08,3.15,3.9,3.92,4.08,4.22 are frequently observed.

Mode Score: Bimodel mode ( 3.44 observed three times, 3.57 observed twice)

Mode Weigh: Bimodel mode ( 17.07 observed twice times, 18.9 observed twice)

For all parameters (points, score and weigh) : mean, median and mode not equal.

All three datasets, points, score and weigh are continuous data type.

Based on the graph “score” and “Weigh” has outliers

Note : Jupyter Notebook file attached

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 is 145.3333. Expected value of weight of patient is 145.33 pounds

Note : Jupyter Notebook file attached

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans(Q9\_a):**

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| speed | -0.1175 | -0.5089 |
| dist | 0.8068 | 0.4050 |

“dist” is positively skewed and “speed” is negatively skewed

With reference to graph, dist has distribution of data concentrated on the left whereas speed has distribution of data on the right.

dist has positive Kurtosis with more outliers and speed has negative Kurtosis with less outliers

Note : Jupyter Notebook file attached

**Ans.(Q9\_b) =**

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.61 | 2.97 |
| WT | -0.61 | 0.95 |

“SP” is positively skewed and “WT” is negatively skewed

With reference to graph, SP has distribution of data concentrated on the left and WT has distribution on the right.

Both WT and SP has positive Kurtosis with more outliers

Note : Jupyter Notebook file attached

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



**Ans:**

Histogram of chickweight$weight:

Histogram represents frequency distribution.

Range of Chickweight$weight is 400. Range of frequency is 200

Majority of the Chicks has weight in range 50 – 100, followed by 100 -150 and 150 – 200

Chickweight$weight bin with sub range 350-400 has lowest frequency of around 5.

Most of the data points of this histogram are towards left and data points are decreasing as we move towards right. Hence it is positively skewed and Mean will be greater than median.

Three categories can be defined (Under weight<50; Avg weight 51 – 150; Over weight >150)

Box Plot:

Shape, rectangle

Description automatically generated

Outliers ( around seven observed))

Inter Quartile (50% data points)

Median

Whisker (25% data points) Extreme

Whisker (25% data points) Extreme

Lower Extreme

Lower Quartile

Upper Quartile

Upper Extreme

Rage: it is represented on box plot by distance between smallest and largest value including outliers. If outliers are ignored, then it is distance between opposite ends of whiskers.

Data has outliers and 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:** N:200000

n:2000

Sample mean: 200pounds

Standard Deviation of sample:30 pound

|  |  |  |
| --- | --- | --- |
| t for 94 % confidancee ( 0.97) | 1.96 |  |
|  |  |  |
| 94% confidance interval |  |  |
|  | 201.3148 | pound |
|  | 200.0000 | pound |

|  |  |  |
| --- | --- | --- |
| t for 96 % confidance ( 0.98) | 1.96 |  |
|  |  |  |
| 96% confidance interval |  |  |
|  | 201.3148 | pound |
|  | 198.6852 | pound |

|  |  |  |
| --- | --- | --- |
| t for 98 % confidance ( 0.99) | 2.326 |  |
|  |  |  |
| 98% confidance interval |  |  |
|  | 201.5603 | pound |
|  | 198.4397 | pound |

**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?

**Ans**

1)

|  |  |
| --- | --- |
| Mean | 41 |
| Median | 40.5 |
| Variance (population formula) | 24.11 |
| standard deviation (population formula) | 4.91 |

2) Lowest score obtained by student is 34 .

Highest score obtained by student is 56

Range of score obtained by student is 22.

score obtained by students is multimodel

Majority of the students scored between 35 – 45 Marks

Data has outlier

Not normally distributed

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

**Ans:** If mean and median of data are equal then skewness of distribution is zero.

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

**Ans:** When mean > median, then distribution skewness is positive. Data is distributed more on Left

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

**Ans:** When median > mean then distribution skewness is negative. Data is distributed more on Right .

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

**Ans:** Positive Kurtosis indicate narrow peak and thicker tails.

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

**Ans:** Negative Kurtosis indicate wider peak and thinner tails.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

**Ans:**

Distribution is not symmetric

Skewness is negarive.

IQR (approximate) will be 8.

Q19) Comment on the below Boxplot visualizations?



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

**Ans:**

Median of both plots is same 262.5 (approximately)

No outliers in both plots.

Range of box plot 1 is 47.5 (approximately). Range of box plot 1 is 160. (approximately). Range of Box plot 2 is higher than box plot 1.

IQR of box plot 1 is 25. IQR of box plot 2 is 93.5 . IQR of box plot 2 is higher than box plot 1.

Box plot 2 is bigger than box plot 1.

Data is Normally Distributed for both

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)
  3. P (20<MPG<50)

**Ans:**

|  |  |
| --- | --- |
| mean | 34.42 |
| Standard deviation | 9.075 |

Note : Standard deviation calculated considering population data.

|  |  |
| --- | --- |
| Z-38 | 0.394 |
| Prob | 0.652 |

|  |  |
| --- | --- |
| Z-40 | 0.615 |
| Prob | 0.729 |

|  |  |
| --- | --- |
| Z-20 | -1.589 |
| Prob | 0.056 |

|  |  |
| --- | --- |
| Z-50 | 1.717 |
| Prob | 0.957 |

|  |  |
| --- | --- |
| P(MPG>38) | 0.3483 |
| P(MPG<40) | 0.7291 |
| P (20<MPG<50) | 0.9014 |

Alternate method (**Note:** Jupyter Notebook file attached)

1. P(MPG>38)

Ans. = 0.3475

1. P(MPG<40)

Ans. = 0.7293

1. P (20<MPG<50)

Ans. = 0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** Mean of MPG is 34.4220. Median is 35.1527. both are very close. MPG of Cars follows Normal Distribution

**Note:** Jupyter Notebook file attached

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 Circumference Mean : 91.9018,

Waist Circumference Median :90.80

Adipose Tissue Mean : 101.8940

Adipose Tissue Median : 96.54

Both AT and Waist doesn’t follow Normal Distribution

**Note:** Jupyter Notebook file attached

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

**Ans:**

|  |  |
| --- | --- |
| Z- Score for 90% confidence interval (0.95) | 1.65 |

|  |  |
| --- | --- |
| Z- Score for 94% confidence interval (0.97) | 1.89 |

|  |  |
| --- | --- |
| Z- Score for 60% confidence interval (0.8) | 0.85 |

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 ( Sample Size 25) | 2.060 |

|  |  |
| --- | --- |
| T- Score for 96% confidence interval ( Sample Size 25) | 2.060 |

|  |  |
| --- | --- |
| T- Score for 99% confidence interval ( Sample Size 25) | 2.787 |

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:**

Popullation mean:270

Sample mean=260

n=Sample size =18

s = standard deviation of sample=90 days

t= -0.4714

Degree of freedom = n-1=18-1=17

r code used as below:

pt(-0.4714,17)= 0.3216741