|  |  |
| --- | --- |
| 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 | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | 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 | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ordinal |
| Sales Figures | Ordinal |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ordinal |
| SAT Scores | Ordinal |
| Years of Education | Ratio |

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

**ANS:** 3 coins can be tossed in following ways-> (HHH, HHT, HTT, TTT, TTH, THH)

Therefore, Probability to get 2 heads and 1 tail is 1/3.

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

1. Equal to 1.

**ANS** = 0 Chances of getting sum equal to 1

1. Less than or equal to 4.

**ANS** = 6/36 = 1/6 Probability of getting sum less than of equal 4.

1. Sum is divisible by 2 and 3.

**ANS** = 6/36 = 1/6 Probability of getting sum divisible by 2 & 3.

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:** 1st ball not blue is = 5/7

2nd ball being not blue is = 4/6 = 2/3

Therefore, total probability of ball drawn being not being blue is = (5/7\*2/3) = 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:** Considering the above statement and give data, expected number of candies for a randomly selected child would be calculated as given below:

1 \* 0.015 + 4\*0.20 + 3 \*0.65 + 5\*0.005 + 6 \*0.01 + 2 \* 0.12

= 3.09

So, the expected number of candies for a randomly selected child would be 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:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| MEAN | 3.596563 | 3.21725 | 17.84875 |
| MEDIAN | 3.695 | 3.325 | 17.71 |
| MODE | 3.92 | 3.44 | 17.02 |
| VARIANCE | 0.286 | 0.957 | 3.193 |
| STANDARD DEVIATION | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.394 |

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:** The Expected value of Weight of random patient would be the mean of all the data values, i.e. 145.33.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS:**

|  |  |  |
| --- | --- | --- |
|  | Speed | Distance |
| Skewness | -0.1175 | 0.8068 |
| Kurtosis | -0.5089 | 0.405 |

* From the Data and the Graph plotted through it, we could understand that Speed is slightly left skewed indicating the negative value of skewness, whereas Distance having large positive value shows that the graph would be right skewed or positively skewed.
* Consider kurtosis value we can infer that the Distance is more peaked towards mean compared to Speed and so have a higher kurtosis value than Speed. It shows that Distance is more evenly plotted compared to Speed

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS:**

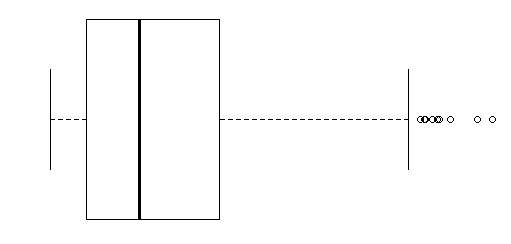
|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skewness | 1.611 | -0.615 |
| Kurtosis | 2.977 | 0.9502 |

* **From the values calculated we can understand that the SP is heavily Right skewed and the WT is moderately Left Skewed.**
* **Considering kurtosis, we understand that SP is more peaked at the mean compared to WT.**

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



* On X-axis the ChickWeight$weight data is being plotted and that on Y-axis the frequency of them is plotted.
* The range of the data points on Y-axis extends from 0-200(Where 200 means most frequent), and that of X-axis range extends from 0-400.
* Around 50-100 is the most frequent ChickWeight$weight. So, the mode of the ChickWeight$weight would lie around this range.
* The Histogram shows that the right tail is longer than the left tail. Therefore, we can conclude that it is a right skewed graph or positively skewed graph.
* There is an outlier present at the upper range.



Considering the above image is like shown.

From the above Box Plot we interpret that:

* The right whisker is longer than left whisker of the plot. Therefore, we can conclude that the graph is right skewed or positively skewed.
* Box Plot helps us in identifying if the outliers are present in the dataset. And from above graph we can conclude that there are a few outliers present at the upper extreme and NO outliers at the lower extreme.
* Assuming the dark vertical line shown is mean we interpret that the median is slightly left shifted.

**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:** Jupyter File

|  |  |
| --- | --- |
| 94% Confidence Interval | 198.738 – 201.261 |
| 98% Confidence Interval | 198.439 – 201.56 |
| 96% Confidence Interval | 198.62 – 201.377 |

**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:** Jupyter File

|  |  |
| --- | --- |
| MEAN | 41 |
| MEDIAN | 40.5 |
| VARIANCE | 24.11 |
| STANDARD DEVIATION | 4.91 |
| MODE | 41 |

The students mark ranges from 34 to 56, every student of the class have scored between this range.

The average marks score by students is 41, with more precision calculating median we get 40.5.

The variance and Standard Deviation both tells us about the spread of the data. Higher both the values more the data is spread.

In this case considering the Variance and STD values of 24.11 and 4.91 respectively, we can conclude that the datapoints area moderately spread.

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

**ANS:** If the mean and median are both equal then the distribution will have zero skewness and it will follow Normal Distribution

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

**ANS:** As the mean is greater than median therefore, the graph will be right skewed or we can say positively skewed.

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

**ANS:** If the median is greater than the mean then the distribution will be left skewed or we can say negatively skewed.

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

**ANS:** Positive skewness says that the distribution is peaked higher at the mean, with longer tails at both ends.

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

**ANS:** Negative kurtosis value indicates that the distribution doesn’t peak higher at the mean value and is more evenly distributed.

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



What can we say about the distribution of the data?

**ANS:** The data is left skewed with left whisker being longer than the right whisker. The data ranges between 10 – 18. The median of the data is slightly right shifted.

What is nature of skewness of the data?

**ANS:** The distribution is right skewed or we can say positively skewed.

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

**ANS:** IQR = (Q3-Q1) = (75th percentile – 25th percentile)

Therefore, IQR = 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:**

* We have 2 boxplot graphs plotted in the given distribution. One of them ranges from approximately, B1(250-280) and other B2(220-310).
* Median/50th percentile of both the Box Plots is around 265.
* There aren’t any outliers in both the boxplots. And both of them are normally distributed as both the upper and lower whiskers are of similar length.

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: Jupyter File

|  |  |
| --- | --- |
| P(MPG>38) | 0.407 |
| P(MPG<40) | 0.753 |
| P (20<MPG<50) | 0.851 |

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS:** Looking at the distplot of the MPG column and considering the mean and median values we say that it almost follows a normal distribution.

1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follow Normal Distribution

Dataset: wc-at.csv

**ANS:** Looking at the distplot and considering mean and median values of both AT and WT column we say that it doesn’t follow normal distribution.

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

**ANS:** To Calculate Z-scores we have a to first plot a normal distribution with success level of confidence level we desire to find Z-score for.

Then we will get the error possibility for the confidence level at the two extreme end of the normal distribution.

For 90% confidence error % would be 10% (i.e. 0.1). For 94% confidence level error % would be 6%(i.e. 0.6). For 60% confidence level the error % would be 40%(i.e. 0.4)

Now we divide the error region into 2 parts, So we get following values:

* For 90% - 0.05
* For 94% - 0.03
* For 60% - 0.2

Now, we look for these values in the negative Z-table to get the finalZ-score values.

Finally, we get:

For 90% - 1.64

For 94% - 1.88

For 60% - 0.84

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

**ANS:** Jupyter File

|  |  |
| --- | --- |
| 95% Confidence Level | 2.063 |
| 96% Confidence Level | 2.171 |
| 99% Confidence Level | 2.796 |

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:** The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.3216