|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Numerical |
| Results of rolling a dice | Numerical |
| Weight of a person | Numerical |
| Weight of Gold | numerical |
| Distance between two places | numerical |
| Length of a leaf | numerical |
| Dog's weight | numerical |
| Blue Color | categorical |
| Number of kids | numerical |
| Number of tickets in Indian railways | numerical |
| Number of times married | numerical |
| 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 | ratio |
| 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) | ratio |
| Sales Figures | interval |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | ordinal |
| Number of Children | Interval |
| 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:#since the possible combinations for 2 heads and one tail is 3

HHT,HTH,THH.

Two heads = 0.375 or 37.5%

One tail = 0.375 or 37.5%

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

1. Equal to 1

Probability = 0

Ans:Since P value for each dice is 0.4,so P2-P1 that equals to 0

1. Less than or equal to 4

Probability= 0.1

Ans: since there are only 6 possible outcomes and total no of outcomes is 36, so 6/36.

1. Sum is divisible by 2 and 3

Probability = 0.1

Ans:since there are only 6 possible outcomes and total no of outcomes is 36,so 6/36

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 :#since there are 7 balls in total so the probability of first ball not being blue is 5 out of 7 and leaves us with 6 balls out which two are blue

Probability of none of the balls drawn = 0.4or 40 %

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

As the probability of each candy count is given so multiplication of probability to the candies count and summation gives us the result.

1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

Solution: Expected number of candies for a randomly selected child = 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 score weigh

Mean 3.59 3.21 17.84

Median 3.69 3.32 17.7

Std 0.53 0.97 1.78

Range 0.84 1.03 2.01

Mode 3.07 3.44 17.02

Variance 0.28 0.95 3.19

The average points for all cars is 3.59,score is 3.21 and weight is 17.84 respectively .the variance for points is 0.28, for score is 0.95 and for weigh it is 3.19.the standard deviation for points is 0.53, for score 0.97 and weigh it is 1.78.since all the values are on the positive side it is a right tailed data distribution,therefore the peak is positive .this data has no outliers .

**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:the expected value of weight for the patient randomly selected is

145.33 pounds.as we calculate the average of summation of weights of all patients divided by number of patients .

Code: #sum of x divided by total number of patients

sum([108,110,123,134,135,145,167,187,199])/9

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans: skewness for speed is -0.11 and for distance it is 0.80,kurtosis for speed is**

**-0.50 and dist is 0.40. since the skewness is negative for speed it is left tailed distribution and for distance it is positive so it is right tailed data,the kurtosis peak for speed is flat since the value is negative and the peak for distance is pointy and steep with thick tails since the value is positive.**

**SP and Weight(WT)**

**Ans: skewness for sp is 1.61 and wt is -0.61 and kurtosis for sp is 2.97 and wt is 0.95. since the skewness for sp is positive the distribution is right tailed and for wt is in negative the distribution is left tailed,the peak for sp is pointy and steep with thick tails while for wt the peak is flat and has an extreme value.**

**Use Q9\_b.csv**

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



According to the Histogram shown here, the peak has a right skew and tail is on the right side of the mean .we have outliers on the higher side of the histogram.



According to the Box plot shown above,the distribution is not symmetrical,the median divides into 75% and 25% of the data,the whiskers are asymmetrically distributed and we have outliers on the maximum side.

**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: Confidence Interval are as follows:

For 94%-198.73-201.26 pounds

For 98%-198.43-201.56 pounds

For 96%-198.62-201.37pounds

Therefore the average weight of an adult male in Mexico at different confidence intervals are given as shown above.since we know the mean,standard deviation and alpha we calculate using the formula.

stats.norm.interval(0.96,200,30/np.sqrt(2000))

**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,Variance=25.52,Standard deviation=5.05.

1. What can we say about the student marks?

Ans: the maximum score by any student is 56 and minimum score by any student is 34, while the average score by students is 41 .we don’t have any outliers, since the mean is greater than median we don’t have skewed data

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

Ans: no skewness is present we have a perfect symmetrical distribution

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

Ans: the nature of skewness when mean is greater than median is towards right

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

The nature of the skewness when median is greater than mean is towards left.

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

The top curve is more peaked or leptokurtic

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

The peak is flatter and broader also the tails are heavily distributed

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



What can we say about the distribution of the data?

Ans:The distribution is positively distributed by data

What is nature of skewness of the data?

Ans:the data is skewed towards left, the whisker range of minimum value is greater than maximum value

What will be the IQR of the data (approximately)?   
IQR= Q3 upper quartile - Q1 lower quartile

=18-10  
IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

According to the plots shown, there are no outliers,the median of both the plot are similar as shown ,that is the range from 250-275 ,both the data are normally distributed with zero to no skewness at the minimum or maximum whisker 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)

Ans: since the prob is greater than 38 therefore 1- stats.norm.cdf(38,34.42,9.13)

0.34 or 34%

* 1. P(MPG<40)

Since the prob is lesser than 40 ,stats.norm.cdf(40,34.42,9.13)

Ans: 0.72 or 72%

* 1. P (20<MPG<50)

Since probability is greater than 20 but lesser than 50 therefore (1-stats.norm.cdf(20,34.42,9.13))-(stats.norm.cdf(50,34.42,9.13))

Ans:0.01 or 1%

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 cars does follow normal distribution .since the mean and median both are located at center of the 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 adipose tissue and waist circumference does not follow normal distribution since the mean and median values are not located at center of the mean.

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

Ans: Z score for 90% = 1.64

Z score for 60%= 0.84

Z score for 94% = 1.8

Python code :#Z score for 90% confidence interval

stats.norm.interval(0.90)

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% is 2.06

t score for 96% is 2.17

T score for 99% is 2.79

Python code :#t score for 95% confidence interval

stats.t.interval(0.99,24)

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 for 18 randomly selected bulbs that would have an average life not more than 260 days is 0.32or 32%,since the sample size is less than 30 we use t test and find the probability.

Python code:((260-270))/(90/np.sqrt(18))

1-stats.t.cdf(0.471,17)