|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | continues |
| Results of rolling a dice | discrete |
| Weight of a person | continues |
| Weight of Gold | continues |
| Distance between two places | continues |
| Length of a leaf | continues |
| Dog's weight | continues |
| Blue Color | categorical |
| Number of kids | discrete |
| Number of tickets in Indian railways | categorical |
| Number of times married | discrete |
| 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 | interval |
| Weight | interval |
| Hair Color | nominal |
| Socioeconomic Status | interval |
| Fahrenheit Temperature | interval |
| Height | interval |
| Type of living accommodation | interval |
| Level of Agreement | ratio |
| IQ(Intelligence Scale) | interval |
| Sales Figures | ratio |
| Blood Group | nominal |
| Time Of Day | interval |
| Time on a Clock with Hands | ratio |
| Number of Children | nominal |
| 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) three coins are tossed = 8 {hhh,ttt,hht,thh,hth,tth,htt,tht}

Probability of 2 heads and one tail = {,thh,hth,hht }= 3

probability that two heads and one tail are

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) total possible out come s = = 36

1. Sum =1 not possible sum of 2 dice rolled will exceed 1
2. Less than equal to 4 🡪 6/36 = 1/6
3. Sum is divisible by 2 and 3 – 16/ 36 🡺 0.444

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 out comes 7c2 == 7x6 / 2x1 🡺21

Possible outcomes 5c2 = 5x4/2x1 🡺 10

== 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) E(x) *=*

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

= 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 :

Mean : 3.596563 ; the average points gained

Median: 3.695 ; middle most value in the dataset

Mode:3.07 ; most repeated value

Variance: 0.28; distance from midpoint to ends

Standard Deviation: 0.534 ; sqareroot of variance which shows deviation of our data points range

Range:2.17 ; (max- 4.93, min -2.76) distance from max to min value in the data set.

For the data we can say that points data has no outliers and data matches with our normal distribution

**Score**

Mean :3.217250; Median: 3.325; Mode: 3.44; Variance: 0.957;Standard Deviation:0.978 ; Range:3.911 ;

From central tendency values we can say more number scored average but and

mode>median>mean there are outliers present so we can say the data in negatively skewed

**Weigh :**

Mean:17.848750; Median: 17.710; Mode:17.02; Variance:3.193 ;Standard Deviation:1.786 ; Range:8.399 ;

As from range value is larger the data scattered or deviated throughout the data points and mode<median<mean its is some what positively skewed or right skewness

Ite range is more so we can say the data scattered more wider

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) excepted value = ∑P(x)\*E(x)

= 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

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

= (1/9)  (  1308)

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans**) **Skewness** :speed🡪mean<median<mode:15.4<15 73 its is negative skewness And its -ve skewness value is -0.117

Distance🡪 mode<median<mean : 26<36<42 that means it’s a positive skewness

And its +ve skewness value is 0.80

**Kurtosis:**

:speed🡪 value -0.5 negative kurtosis cause kurtosis is< 0

Dist 🡪value 0.2 positive kurtosis cause kurtiosis >0

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS) Skewness:** sp🡪 mean > median >mode : 127>118.28>118.20 its positively skewed And its +ve skewness value is 1.61

Weight🡪mean<median: 32.41<32.73 its is negative skewness And its -ve skewness value is -0.61

**Kurtosis:**

**SP🡪** value 2.72 positive kurtosis cause kurtiosis >0

Weight--< value 0.81 positive kurtosis cause kurtiosis >0

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



ANS) 🡪mode = at 50-100 ,median and mean lies at 200 , range = 0-400 its 400 sowe can say the data has deviated more above the stranded deviation range

🡪peakon left and the tail ant the right-side it shows it’s a positively skewed data



Ans) 🡪the upper whisker in longer so we can say that more outliers present between the 75% to 100 of the data

**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= 2000,std s= 30, mean = 200 x ±z

94% 🡪 200

98% --> 200

96% -->200

**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 🡺 738/18 🡺41.0 ,

Median 🡺 40+40/2 🡺40.5

Variance🡪24.111

standard deviation🡪4.9103

1. What can we say about the student marks?

As per the data the std range shows more number of students scored between 36.1 to 45.9 there is 1 student below the std range , 2 students scored above std range more number of students preforming at average rate

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

Ans) it’s a no skew and when mean=median its normal distribution perfectly well shaped curve

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

Ans) if mean > median its positively skewed

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

Ans) if median > mean its negatively skewed

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

Ans) It has heavey tails that means there are more outliers on both ends ,thin bell at peak less data at peak

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

Ans) negative kurtosis shows thin tail and flat peak it shows data is lack of outliers and and data is uniformly distributed at extreme case

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



What can we say about the distribution of the data?

Ans) median is closer to upperquartile so the data is negatively skeweed more outliers present at lower quartile (0-25%)

What is nature of skewness of the data?

Ans) the data is negatively skeweed

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

Ans) the IQR of the data (approximately) will be 8.5 iqr = q3-q1=>18.5-10=>8.5

Q19) Comment on the below Boxplot visualizations?



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

Ans) 1 and 2 have similar center

1 boxplot represents data point consistently hover around center points

2 boxplot represent have a much larger variability than the 1

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) =>z=x-µ/ => 38 – 34.4/9.07🡺0.396🡺p(z>0.392)🡺1-0.651(from ztable)🡺0.349
  2. P(MPG<40)🡺40-34.4/9.13🡺5.6/9.13🡺0.613(from ztable)🡺0.7291

c. P (20<MPG<50)

🡺20-34.42/9.13🡺-14.4/9.13🡺-1.57

🡺50-34.42/9.13🡺15.6/9.13🡺1.70

P(-1.57 < x< 1.70)🡺(-0.582 <x<1.70) (from ztable)

🡺0.9554-0.582🡺0.3734

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans)The peak is slightly wider than ideal normal distributions peak but the overall graph shows similar to normal distributions

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

Dataset: wc-at.csv

Ans) AT🡪 median and mean are almost equal and the plot almost similar to Normal Distribution so it does follow Normal Distribution

Waist 🡪 median and mean are almost equal and the plot almost similar to Normal Distribution so it does follow Normal Distribution

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

Ans) z score of 90% 🡪1+90 /2🡪0.95-🡪z score of 0.95 is 1.65

z score of 94% 🡪1+94 /2🡪0.97-🡪z score of 0.97 is 1.89

z score of 60% 🡪1+60 /2🡪0.60-🡪z score of 0.8 is 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\_crit96 = np.abs(t.ppf((1-0.96)/2,24)) 🡪2.1715

t\_crit99 = np.abs(t.ppf((1-0.99)/2,24))🡪 2.7969

t\_crit95 = np.abs(t.ppf((1-0.95)/2,24))🡪2.0638

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 freedomc

ans) µ= 270

sµ = 260

sample std = 90

n= 18

α= x-µ/s🡪260-270/90🡪 -0.47

t = stats.t.sf((1-0.47)/2,17)🡪0.3970