|  |  |
| --- | --- |
| 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 | discrete |
| Number of kids | discrete |
| Number of tickets in Indian railways | discrete |
| Number of times married | discrete |
| Gender (Male or Female) | discrete |

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 | ratio |
| Socioeconomic Status | interval |
| Fahrenheit Temperature | interval |
| Height | ratio |
| Type of living accommodation | ordinal |
| Level of Agreement | interval |
| IQ(Intelligence Scale) | interval |
| Sales Figures | interval |
| Blood Group | ratio |
| Time Of Day | interval |
| Time on a Clock with Hands | interval |
| Number of Children | ratio |
| Religious Preference | ratio |
|  |  |
| Barometer Pressure | interval |
| SAT Scores | ratio |
| Years of Education | nominal |

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

Ans : Three coins = 2 ^ (3) = 8

(H, H, T), (H, T, H), (T, H, H) = 3

P = 3/8

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

1. Equal to 1

Ans: P(X=1) = 0 / 36 = 0

(Minimum addition starts from (1+1) = 2)

1. Less than or equal to 4

Ans: (X <=4) = (1,1)(1,2)(1,3)(2,1)(2,2)(3,1) = 6

Two Dice = 6 ^ (2) = 36

P = 6/36 = 1/6

1. Sum is divisible by 2 and 3

Ans: (X/2 & X/3) = (1,5)(2,4)(3,3)(4,2)(5,1) = 5

P = 5/36

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: Total no of balls = 7  
 two balls drawn = 2  
 so 7C2 = 21 sample space  
 No blue ball should drawn = 2  
 except blue total no of ball = 5  
 so 5C2=10  
 **P =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 |

Ans = Expected number of candies for a randomly selected child

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

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

**Use Q7.csv file**

|  |  |  |  |
| --- | --- | --- | --- |
| Ans = **used aggregate methods mean(), median(), mode(), std(), var() and range = max() – min()** |  |  |  |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Points** | **Score** | **Weigh** | | **Mean** | 3.596563 | 3.21725 | 17.84875 | | **Median** | 3.695 | 3.325 | 17.71 | | **Mode** | 3.07 | 3.44 | 17.02 | | **Std dev** | 0.534679 | 0.978457 | 1.786943 | | **Variance** | 0.285881 | 0.957379 | 3.193166 | | **Range** | 2.17 | 3.9109999 | 8.3999999 | |  |  |  |
|  |  |  |  |
|  |  |  |  |

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 of weight for a randomly selected Patient

= 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)

= 12 + 12.222 + 13.666 + 14.888 + 15 + 16.111 + 18.555 + 20.777 + 22.110

= **145.33** pounds

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

**Cars speed and distance Use Q9\_a.csv**

Ans = **used method skew() and kurt()**

**Car Speed Distance**

**Skewness** -0.117510 0.806895  
  
 **kurtosis** -0.508994 0.405053

**SP and Weight(WT) Use Q9\_b.csv**

Ans = **used method skew() and kurt()**

**SP Weight**

**Skewness** 1.611450 -0.614753  
  
 **kurtosis** 2.977329 0.950291

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

Ans = **Right side skewed or positively skewed**



Ans = **The box plot is positively skewed**

**There are Outliers at upper 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 = Average weight of Adult in Mexico with:

94% C.I = **stats.norm.interval(0.94,200,30/np.sqrt(2000))**

**(198.738325292158, 201.261674707842)**

98% C.I = **stats.norm.interval(0.98,200,30/np.sqrt(2000))**

**(198.43943840429978, 201.56056159570022)**

96% C.I = **stats.norm.interval(0.96,200,30/np.sqrt(2000))**

**(198.62230334813333, 201.37769665186667)**

**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 = used aggregate methods mean(), median(), std(), var()**

|  |  |
| --- | --- |
| **Mean** | **41** |
| **Median** | **40.5** |
| **Std dev** | **5.05** |
| **Variance** | **25.52** |

1. What can we say about the student marks?

|  |
| --- |
| **Ans =** **used boxplot() method for the given list of scores,**  **There are 2 Outliers in Student's marks at 49 and 56**  **Mass of students Marks is between 38-42** |
|  |

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

Ans = Normalized Skewness or Normal Distribution or Zero Skewness because

the mean, median and mode are at the same point which means the

plotted cuve will be perfect bell shape.

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

Ans = It can be inferred that mean that is average of the data distribution is more than

and to the Right of the median that is the middle number of the data which

means it is Right Skewed or Positively Skewed Distribution

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

Ans = It can be inferred that mean that is average of the data distribution is less than

and to the Left of the median that is the middle number of the data which means

it is Left Skewed or Negatively Skewed Distribution

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

Ans = Data distribution has Sharp peak in the plot and heavy tail wich means

possibilities of heavy outliers.

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

Ans = Data distribution has Flat peak in the plot and Light tail wich means

possibilities of very less outliers.

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



What can we say about the distribution of the data?

Ans = The distribution of data of the given boxplot implies that most of the data

points lies between 15 and 18, and very less data points are there below 15

What is nature of skewness of the data?

Ans = Negative or Left side skewed because median or the middle point of the given

data is much closer to the third quartile and the whiskers are also long on the left

side.

What will be the IQR of the data (approximately)?   
  
Ans = Q3 - Q1

= 18 – 1

=  **8** is IQR

Q19) Comment on the below Boxplot visualizations?



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

Boxplot 2.

Ans = If we compare boxplot 1 with boxplot 2, it can be said that both the box

plots are symmetrical which means both are normally distributed with

zero skewness as both the whiskers of the two boxplots are equal.

Boxplot 1 has closed distribution of data compared to boxplot 2 where

the data is dispersed fair enough. The quartiles and extremes are far from

median in boxplot 2 as to that of boxplot 1 which means the range of

boxplot 2 is greater. The main difference between the two boxplots lies in

the concentration of data around the median.

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) = **0.3475939251** used syntax

1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

* 1. P(MPG<40) = **0.7293498762** used syntax

stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

* 1. P(20<MPG<50) = **0.8988689169** used syntax

stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std()) - stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans = The distplot of the given dataset shows its almost symmetrical which

Which means it’s a normal distribution.

The mean, median and mode values are also very close or almost equal

Which states its again an 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 Boxplot of the given dataset shows that waist circumference is

fairly symmetrical which means it’s a normal distribution.

The Boxplot of the given dataset shows that Adipose Tissue is

Positively skewed which means it deos not follows normal distribution.

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

Ans = Z – scores of :

90% is **1.6448536269** using stats.norm.ppf(0.95)

94% is **1.8807936081** using stats.norm.ppf(0.97)

60% is **0.8416212335** using stats.norm.ppf(0.8)

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

Ans = t – scores of 25 sample size at confidence interval of:

95% is **2.0638985616** using stats.t.ppf(0.975,24)

96% is **2.1715446760** using stats.t.ppf(0.98,24)

99% is **2.7969395047** using stats.t.ppf(0.995,24)

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 = here x=260, mu=270, n=18, df=17, sigma=90

So tscore at x = (260-270)/(90/np.sqrt(18))

= **-0.47140452079**

Now P(x<260) = 1 - P(x>=260)

= 1- stats.t.cdf(0.4714045,17) = **0.321672542**

probability that 18 randomly selected bulbs would have an average life no more than 260 days is 32.17%.

Assuming Significance value alpha(α)=0.05(standard value)

if p value < α ----reject Ho(null) and accept Ha(Alternative)

if p value > α ----accept Ho and Reject Ho

therefore: p value is > α , then accept Ho, The CEO claims are false and the average life of bulb > 260 days.