|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Discrete data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete data |

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) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

Ans- P (Two heads and one tail) = N (Event (Two heads and one tail)) / N (Event (Three coins tossed)) = 3/8 = 0.375

the probability that two heads and one tail is 0.375

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 – a) P(sum is equal to 1)= 0

b) P(sum is less than or equal to 4)= N(event (sum is less than or equal to 4/

N(event(two dice rolled))

=6/36=1/6=0.166

c) P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3))/

N(Event (Two dice rolled)) = 6 / 36

= 1/6 = 0.16

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 number of balls =7 balls

Using nCr =n/r(n-r)

N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5!

= (7\*6\*5\*4\*3\*2\*1) /(2\*1) \* (5\*4\*3\*2\*1)

N (Event (2 balls are drawn randomly from bag) = (7\*6)/ (2\*1) = 21

If none of them drawn 2 balls are blue = 7 – 2 = 5

N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (5\*4) / (2\*1)

= 10

P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) /

N (Event (2 balls are drawn randomly from bag)

=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- 0.015+0.8+1.95+0.025+0.06+0.24 = 3.09

The 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**

Mean for Points = 3.59, Score = 3.21 and Weigh = 17.84

Median for Points = 3.69, Score = 3.32 and Weigh = 17.71

Mode for Points = 3.07, Score = 3.44 and Weigh = 17.02

Variance for Points = 0.28, Score = 0.95, Weigh = 3.19

Standard Deviation for Points = 0.53, Score = 0.97, Weigh = 1.78

Range [Min-Max] for Points [3.59 – 4.93], Score [3.21 – 5.42] and Weigh [17.84 – 22.9]

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 = Sum (X \* Probability of X)

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

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans-For distance,**

**Skewness=** 0.80689496 and Kurtosis= 0.40505258

For speed,

Skewness= -0.11750986 and Kurtosis = -0.50899442

Inference

- skewness

In the given dataset,the skewness of data set 'speed' is negatively skewed and normal curve tale is elongated towards left side with value -0.117 and the skewness of data set 'dist' is positively skewd and the normal curve is elongated towards right side with value 0.806.

- kurtosis

In the given datase, kurtosis of speed is -0.508 it means curve of 'speed' is flat i.e. platykurtic.

kurtosis of 'dist' is 0.405 which also a flat curve(platykurtic) but it has positive peak.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans- Skewness and Kurtosis of SP are** 1.6114502 and 2.9773289

Skewness and Kurtosis of WT are -0.61475333 and 0.95029149

#Inference

skewness- In the given dataset, the data of 'SP' is positively skewed with the value 1.6114 and the data of 'WT' is negatively skewed with the value -0.6147.As the the skewness refers to the measure of assymetry ,therefore both data 'SP' and 'WT' have assymetric curve with there 'SP' data set's curve elongated toward right side and 'WT' data set's curve elongated towards left side.

kurtosis- In the given dataset, the kurtosis of 'SP' = 2.977 and the kurtosis of 'WT' = 0.950. As the kurtosis is the degree of peakedness,therefore 'SP' has higher peak than 'WT'. we can conclude that 'SP' data set has mesokurtic and 'WT' data set has platykurtic.

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



Answer – Inference of Histogram

The given Histogram is plotted ChickWeight against Frequency. By observing the shape of histogram we can conclude, the given histogram depicts bell shaped normal curve with its tail elongated towards right side. We can also infer that histogram have outliers in the range 350-400.

About peakedness of histogram, it has peak can be mesokurtic and also by looking at peak we can also states that it’s a unimodal curve at chickWeight = 100.



Inference about Boxplot –

Box plot shows the five summary of a set of data includes minimum score, first quartile, median, third quartile and the maximum score. By observing above boxplot can conclude that, Minimum score has more variability in its 25% of the data as compare to the Maximum score which has slighter variability in its 25% of the data. Above box plot is negatively skwed as the distance from median to the minimum is greater than median to the maximum. This box plot also implies inconsistency. There are also some outliers present above the minimum scores.

Quartile 1 and Quartile 3 each represents 25% 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-

conf\_94 =stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))

print(np.round(conf\_94,0))

print(conf\_94)

For 94% confidence interval Range is [ 198.73 – 201.26]

For 98% confidence interval range is [198.43 – 201.56]

For 96% confidence interval range is [198.62 – 201.37]

**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 =41.0, Median =40.5 Variance =24.11111111111111

Standard deviation =4.910306620885412

1. What can we say about the student marks?

Ans- we don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

Ans- when mean and median are equal then skewness =0 it means distribution is symmetric i.e. normally distributed.

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

Ans- when mean>median,the nature of skewness is negatively skewed(skewed left).

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

Ans-when median>mean,the nature of the skewness is positively skewed (skewed right).

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

Ans- Positive kurtosis shows leptokurtic distribution.leptokurtic is peaked and possesses thick tail,it means more of the data located in the tail of the distribution intead of around the mean. Positive kurtosis also indicates that extreme values i.e. outliers present at the peak.

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

Ans- Negative kurtosis indicates that a distribution of the data is flat and has thin tails.

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



What can we say about the distribution of the data?

Ans- The median is not in the middle of the box and the whisker are not same on the both sides, therefore the distribution of the given boxplot is not symmetrical and the data is not consistent.

What is nature of skewness of the data?

Ans- The nature of the skewness of the data is natively skewed(skewed left).

What will be the IQR of the data (approximately)?   
Ans- IQR of the data is 8.

Q19) Comment on the below Boxplot visualizations?



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

Ans- The range of the Boxplot 1 is smaller than Boxplot 2 it implies that data of Boxplot 1 is less scatter than the Boxplot 2. Data of Boxplot 1 is more concentrated towards median than Boxplot 2. Median of both Boxplots are same.

By comparing IQR of both boxplots, Boxplot 1 is less dispersed as compare to

Boxplot 2.

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- Probability of MPG of cars are as follows

1. P(MPG>38) =0.3475
2. P(MPG<40) =0.7293

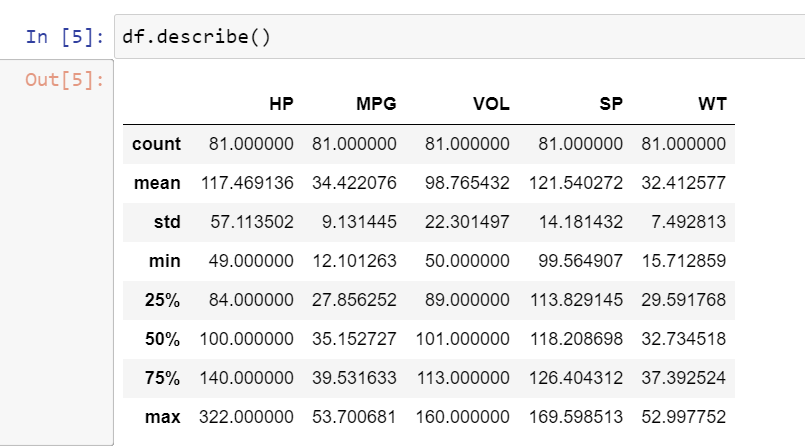
c. P (20<MPG<50) =0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

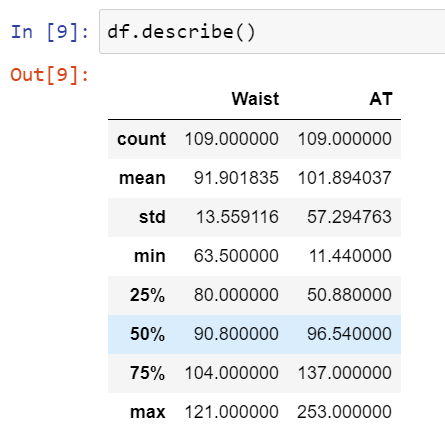
Ans-



By looking at the result of df.describe(), 50% of the data lies towards the mean. Hence the MPG of cars follows 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-

By observing the above result, 50% of the data lies towards the mean.

Therefore wc-at data set follws Normal Distribution.

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

Ans- Z-score of 90% CI is 1.28, for 94% and 60% Z-scores are 1.55 and 0.25.

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% CI is 1.71 and that for 96% and 99% is 1.82 and 2.49

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- t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

= (260-270)/90/np.sqrt(18))

t\_score = -0.471

= stats.t.cdf(t\_score, df = 17)

0.32 = 32%