Q1) Identify the Data type for the Following:

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

Name : Deepak Kumar Singh

Course : Data Science-Weekday-Hyderabad-Nilakshi-3 rd August-9:30AM-11:30AM

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	Interval
Type of living accommodation	Nominal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Ratio
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Internal
Number of Children	Ratio
Religious Preference	Nominal
Barometer Pressure	Ratio
SAT Scores	Ratio
Years of Education	Ratio

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

Ans: The probability of getting two heads and one tail is 0.375 or 37.5%
Explanation: The possible outcome of the three coins are HHH, HHT, HTT, HTH, TTT, THH, TTH, THT. The required outcomes are two heads and one tail which are HHT, HTH, THH Probability = (no.of favorable outcomes)/(no.of.outcomes) => 3/8 => 0.375

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

a) Equal to 1
   Ans: The probability when two dice are rolled and the sum of it must be 1 is 0. Explanation: Because even in the worst/least condition the value would be 2 (1,1)

b) Less than or equal to 4
   Ans: The Probability of getting the sum of two dice is less than or equal to 4 is 16.67%. Explanation: The favorable outcomes are [(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)] = 6 which is divide by the total number of outcome which is 36. Probability = (no.of favorable outcomes)/(no.of.outcomes) => 6/36 => 16.67%

c) Sum is divisible by 2 and 3
   Ans: The Probability of getting the sum divisible by 2 and 3 are 16.67% Explanation: The favorable outcomes are (1,2), (2,1), (2,4), (4,2), (3,3), (6,6) =6 which is divide by the total number of outcomes which is 36. Probability = 6/36 =>16.67%

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: 47.62% Explanation: Number of balls inside the bag is (2+3+2)=7, and the balls to be drawn are 2 so the combination of both are 7c2 = 7!/2!(7-2)! =21. The probability that none of the balls drawn are blue (2 red+ 3green) =5, 5c2=5! /2! (5-1)! =>10 so, Probability = (no.of favorable outcomes)/(no.of.outcomes)=> 10/21 =>0.4761

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of the count of candies for children (ignoring the nature of the child-generalised 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: The Expected Number of candies for a randomly selected Child is 3.09
Explanation: (candies count of for child * probability of the child)
(1*0.015)+(4*0.20)+(3*0.65)+ (5*0.005)+ (6*0.01)+ (2*0.120) =>
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

- 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

X	Mean	Median	Mode	Variance	SD	Range
Point	3.596563	3.695	3.9	0.285881	0.534679	2.17
Score	3.21725	3.325	2.875	0.957379	0.978457	3.911
Weigh	17.84875	17.71	18.61	3.193166	1.786943	8.4

Ans: **Comment about the values:**

Mean:

Summing up all values and dividing with the number of values m=(sum of all values)/ no. of values

Median:

Arrange the num in ascending order and find the middle value if the dataset has an even value (n/2+((n/2)+1))/2 if the dataset has an odd value (n+1)/2

Mode:

The value which appears more frequently Variance: average squared deviation from mean

Standard Deviation:

The standard deviation by taking the square root of the variance

Range:

Calculate the range by subtracting the minimum value from the maximum value

Inference:

Points:

The 'Points' column's statistical analysis reveals interesting insights about the driving characteristics of the vehicles. The mean points value of approximately 3.60 indicates that, on average, the vehicles have a balanced performance score. The median value of 3.695 being slightly higher than the mean suggests that there might be a few vehicles with relatively higher scores pulling the distribution slightly to the right. The presence of two modes, 3.07 and 3.92, indicates the existence of distinct subgroups of vehicles, possibly representing different performance profiles. The low variance and standard deviation values (0.29 and 0.53, respectively) imply that the scores are closely clustered around the mean,

reflecting consistent driving performance. The range of 2.17 highlights the spread of scores among the vehicles, encompassing a range of performance levels.

Score:

Analyzing the 'Score' column reveals performance scores that are moderately distributed. The mean score of around 3.22 indicates a relatively balanced performance across the vehicles. The median score of 3.325 is slightly higher than the mean, hinting at a potential right skewness in the distribution, suggesting the presence of a few vehicles with higher scores. The mode at 3.44 indicates a prominent peak in the data, signifying a commonly occurring performance level. The moderate variance (0.96) and standard deviation (0.98) values suggest that the performance scores exhibit moderate variability around the mean. The range of 3.911 underscores the diversity in scores among the vehicles, possibly reflecting a variety of driving characteristics

Weigh:

The 'Weigh' column's statistical analysis provides insights into the weight distribution of the vehicles. The mean weight of around 17.85 reflects the average weight of the vehicles. The median weight of 17.71 being close to the mean suggests a relatively symmetric distribution of weights. The presence of two modes, 17.02 and 18.9, indicates potential subgroups of vehicles with different weight characteristics. The calculated variance (3.19) and standard deviation (1.79) values reveal moderate variability in weights, suggesting a range of vehicle sizes and types within the dataset. The range of 8.4 emphasizes the diversity in vehicle weights, indicating a significant difference between the lightest and heaviest vehicles.

Q8) Calculate the Expected Value for the problem below

a) 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: Weight(X) = 145.33

Explanation:

To find the Expected value which is also called as Mean value is by summing up all the given weights and dividing by the number of values in this case 108+110+123+134+135+145+167+187+199=1308 which is divided by 9 the 1308/9=145.33

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

Cars speed and distance

Use Q9_a.csv

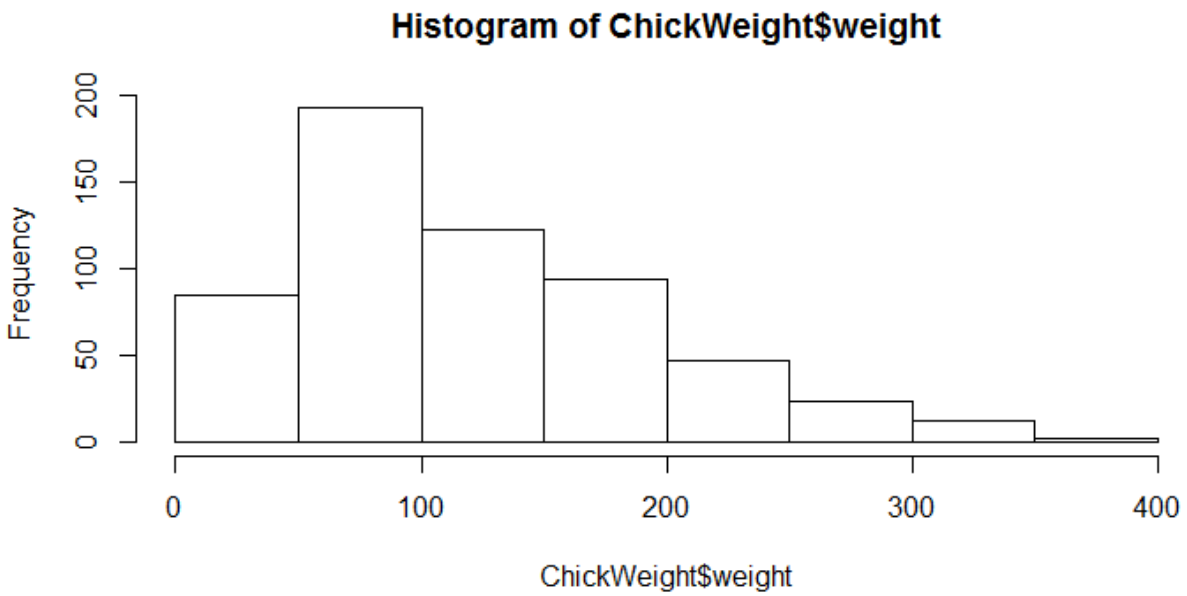
Inference:

The "Speed" values have a slightly negatively skewed distribution with a relatively normal level of kurtosis. The "Distance" values have a positively skewed distribution with heavier tails and a more peaked shape compared to a normal distribution.

SP and Weight(WT)

Use Q9_b.csv

Q10) Draw inferences about the following boxplot & histogram

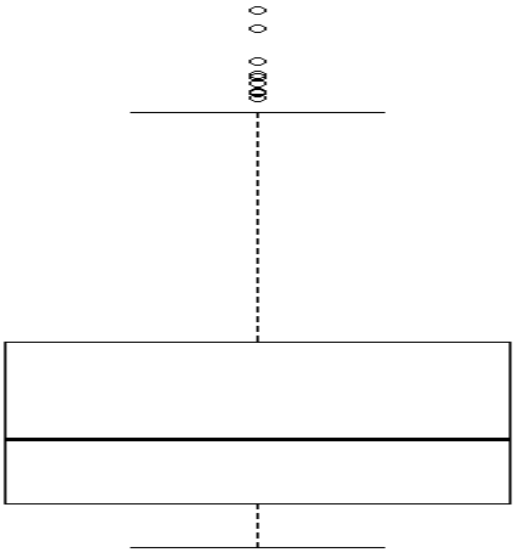


Histogram:

1.Chick weight is a Right skewed or positively skewed

2. Fifty percent of the chick weight falls between 50 to 150

3. Highest Chick Weight is between 50-100



Box Plot:

1. It is Right skewed or positively skewed

2. It has outliers on the top 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:

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,
Median=9.5,
Mode=41,
Standard deviation=5.05,
Variance=25.5

- Mean is the sum of the value of each observation in dataset divided by the number of observation. Median divides the distribution into half
- Standard Deviation finds how far the score lies from the mean (x-x^2)/n-1 Variance is the average squared of the standard deviation

2) What can we say about the student marks?

Ans:

- There might be a group of students with consistently high marks around 41. There could be a tail of higher marks that are pulling the mean upwards, causing a larger difference between the mean and median.
- The relatively high standard deviation and variance indicate that there is a wide range of marks, indicating diverse performance levels among the students

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

if the mean, median, and mode are equal, the distribution is symmetric, and the skewness is zero. This symmetry suggests a balanced spread of data points around the central value.

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

when the mean is greater than the median, it suggests a right-skewed distribution with a tail extending towards higher values.

This indicates that there are some data points with values significantly higher than the majority of the data, contributing to the skewness in the positive direction.

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

when the median is greater than the mean, it suggests a left-skewed distribution with a tail extending towards lower values.

This indicates that there are some data points with values significantly lower than the majority of the data, contributing to the skewness in the negative direction.

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

A positive kurtosis value indicates that the distribution has a more concentrated center with heavier tails.

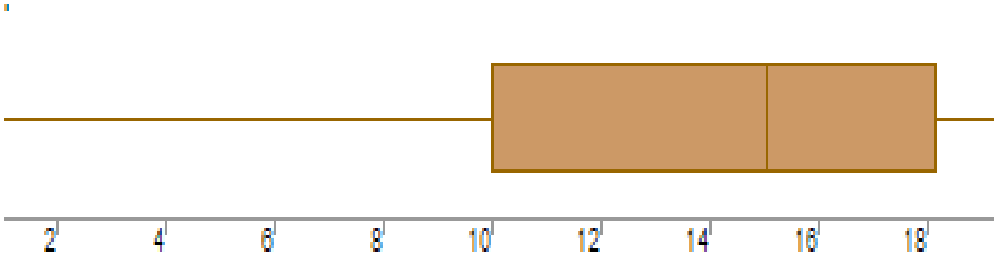
It's essential to interpret kurtosis in conjunction with other statistical measures and consider the context of the data to understand the distribution's shape and characteristics.

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

a negative kurtosis value indicates that the distribution has lighter tails and a flatter peak.

It's important to interpret kurtosis in the context of other statistical measures and consider the specific characteristics of the dataset.

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



What can we say about the distribution of the data?

Ans:

The Boxplot is normally distributed and the median is towards higher value.

What is nature of skewness of the data?

Ans:

The Data is skewed towards left and the whisker range of minimum value is greater than maximum

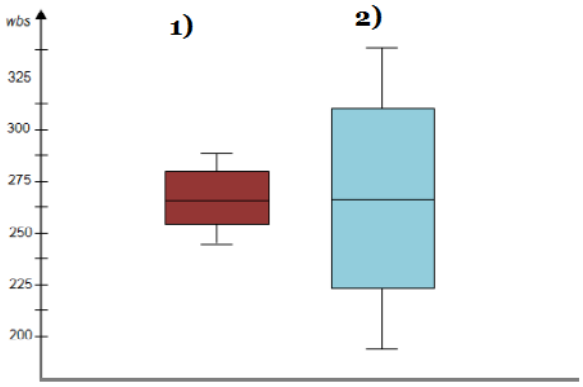
What will be the IQR of the data (approximately)?

Ans:

The Inner Quartile Range of the given data is 8

IQR = Q3(upper quartile)-Q1(lower quartile)

Q19) Comment on the below Boxplot visualizations?



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

Ans:

We can see that both the Boxplot doesn't have an Outlier. Both the data have the same point as a Median of around 262.

They are normally distributed with no skewness either 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

Ans:

First Find the mean and standard Deviation of MPG through using mean () function and std () function where mean is 34.42 and std is 9.13

a. P(MPG>38)
   Ans:
   To find the Probability of greater than 38 we can use (1-(stats.norm.cdf (38,34.42,9.13))) hence the answer is 35%

b. P(MPG<40)
   Ans:
   To find the Probability of lesser than 40 we can use (stats.norm.cdf (38,34.42,9.13) hence the answer is 73%

c. P (20<MPG<50)

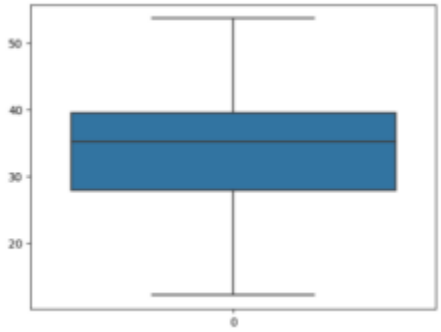
   Ans:

   To find the Probability of greater than 38 we can use (1-(stats.norm.cdf(50,34.42,9.13))-((stats.norm.cdf(20,34.42,9.13)) hence the answer is 1.31%

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
   Dataset: Cars.csv

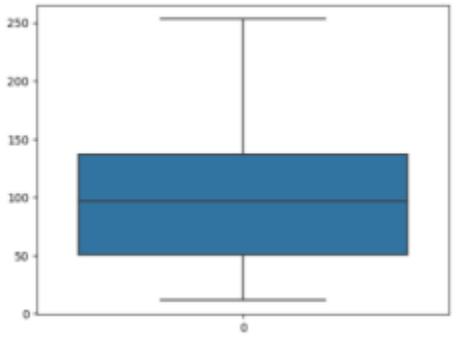
Ans:



The MPG of Cars follows Normal Distribution with Median is 35.

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

Ans:



Adipose Tissue Does not follow Normal Distribution.

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

Ans:

90% Confidence:

To find the Z-score value we need to find the Area towards Left of the curve the formulae is AL= (1+CL)/2, our CL is 90% hence AL= 0.95 with the help of AL search the 0.95 value in Z-score table and add the two values to find our Z-score. The Z-Score of 90% is 1.65

94% Confidence:

To find the Z-score value we need to find the Area towards Left of the curve the formulae is AL= (1+CL)/2, our CL is 90% hence AL= 0.97 with the help of AL search the 0.97 value in Z-score table and add the two values to find our Z-score. The Z-Score of 94% is 1.89

60% Confidence:

To find the Z-score value we need to find the Area towards Left of the curve the formulae is AL= (1+CL)/2, our CL is 90% hence AL= 0.8 with the help of AL search the 0.8 value in Z-score table and add the two values to find our Z-score. The Z-Score of 60% 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:

95% Confidence Interval

First find out the Df value by using the formulae Df=n-1, where n is the sample size Df=24 find the Df value in the T-score table and find the critical Value the TScore value of 95% is 1.711

96% Confidence Interval

First find out the Df value by using the formulae Df=n-1, where n is the sample size Df=24 find the Df value in the T-score table and find the critical Value the TScore value of 96% is 2.064

99% Confidence Interval

First find out the Df value by using the formulae Df=n-1, where n is the sample size Df=24 find the Df value in the T-score table and find the critical Value the TScore value of 99% is 3.467

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:

import numpy as np Import scipy as stats t_score = (x - pop mean) / (sample standard deviation / 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%

Name : Deepak Kumar Singh

Course : Data Science-Weekday-Hyderabad-Nilakshi-3 rd August-9:30AM-11:30AM