Name - Niranjan Nevase
Batch - 5th June Data Science

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	Nominal
Number of kids	Discrete
Number of tickets in Indian railways	Ordinal
Number of times married	Discrete
Gender (Male or Female)	Nominal

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	Nominal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Ordinal
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Interval
Time on a Clock with Hands	Interval
Number of Children	Ratio
Religious Preference	Nominal

Name - Niranjan Nevase
Batch - 5th June Data Science

Barometer Pressure	Interval
SAT Scores	Interval
Years of Education	Ratio

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

Total no. of Events (N) = 8

Total no. of favorable outcome (f) = 3

Probability (P) = **3/8 = 0.375 = 37.5%**

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

Sample space for 2 rolled dice - 36

a) Equal to 1 = 0/36 = **0**
b) Less than or equal to 4 = 6/36 = **1/6**
c) Sum is divisible by 2 and 3 = 6/36 = **1/6**

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: Case 1 – If drawing is done with replacement, then

Probability (P) = (5/7) x (5/7) = **25/49 = 0.51**

Case 2 – If drawing is done without replacement

Probability (P) = (5/7) x (4/6) = 20/42 = **0.48**

Name - Niranjan Nevase
Batch - 5th June Data Science

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: Expected Value, $E(x) = \sum x_i \cdot P(x_i)$

= (1 x 0.015) + (4 x 0.20) + (3 x 0.65) + (5 x 0.005) + (6 x 0.01) + (2 x 0.120)

, E (x)= **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	Score	Weigh
Mean	3.60	3.217	17.85
Median	3.695	3.325	17.71
Mode	3.92	3.440	17.02
Variance	0.29	0.957	3.19
Std Deviation	0.53	0.978	1.79
Range	2.17	3.911	8.40

Name - Niranjan Nevase
Batch - 5th June Data Science

Inferences:

1. Standard deviation values shows that data points for Points & Scores are clustered tightly around the respective means while Weigh data points are more spread out.
2. Points & Scores data is Negatively Skewed while Weigh data is Positively Skewed

Q8) Calculate 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: Let's assume every patient has equal probability to get chosen which is 1/9.

Expected value, $E(x) = \sum x_i \cdot P(x_i)$ = **145 Pounds**

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

Cars speed and distance

Use Q9_a.csv

	Speed	Distance
Skewness	-0.118	0.807
Kurtosis	-0.509	0.405

- For Speed, mean<median<mode, while for Distance mean>median>mode.
- Both Kurtosis values are less than 3, which means that data has lighter tails and a lower peak (less peaked) than a normal distribution.

Name - Niranjan Nevase
Batch - 5th June Data Science

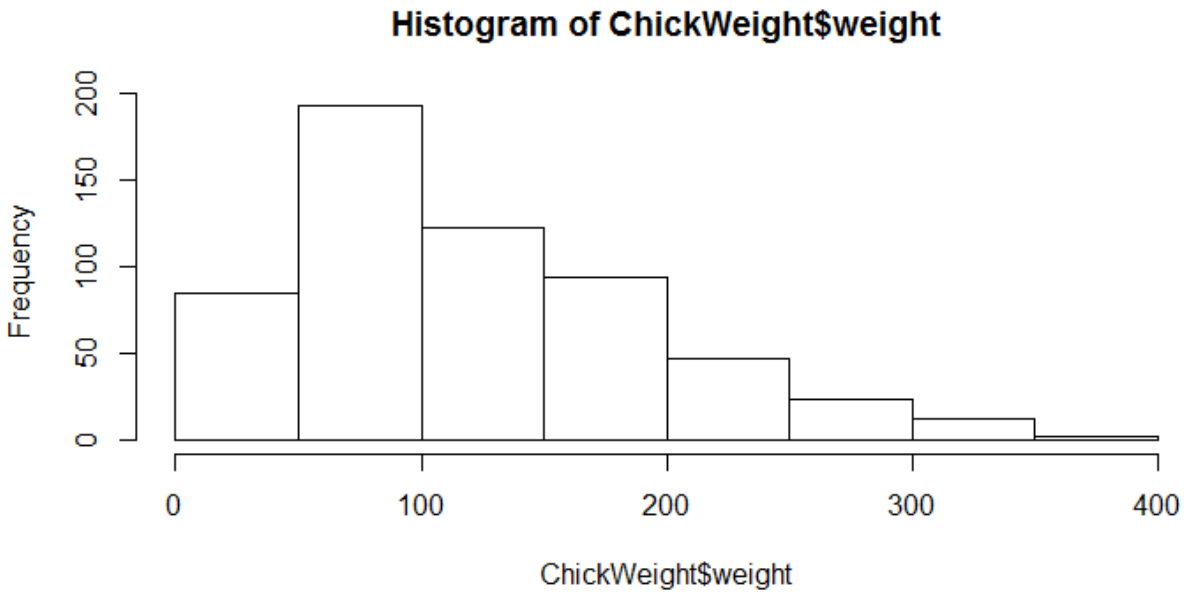
- **SP and Weight(WT)**

Use Q9_b.csv

	SP	WT
Skewness	1.611	-0.615
Kurtosis	2.977	0.950

- For SP, mean>median>mode, while for WT mean<median<mode.
- SP Kurtosis values is around 3, which means that data distribution is same as normal distribution, while for WT, that data has lighter tails and a lower peak (less peaked) than a normal distribution.

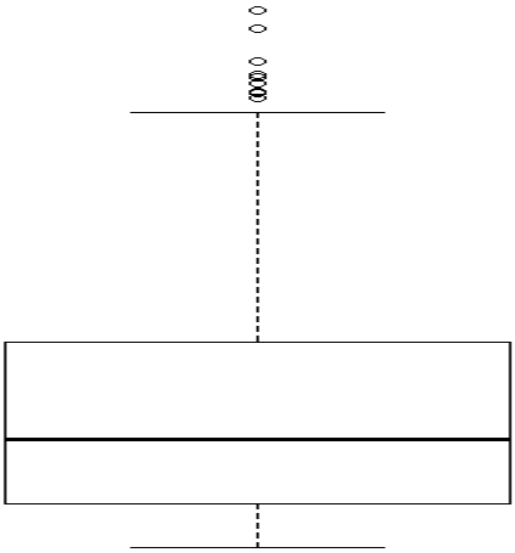
Q10) Draw inferences about the following boxplot & histogram



Ans:

- Data is positively skewed, which means most values are less than mean.
- For above distribution we can say that Mean>Median>Mode

Name - Niranjan Nevase
Batch - 5th June Data Science



Ans :

- Box plot has a higher upper whisker length than lower whisker length, it means that the distribution of the data is right-skewed or positively skewed. This means that most of the data values are below or near the median, and there are some high outliers on the right.

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?

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval
$\bar{x}$ = sample mean
z = confidence level value
s = sample standard deviation
n = sample size

1. 94% = 198.73 – 201.26
2. 98% = 198.43 – 201.56
3. 96% = 198.62 – 201.37

Name - Niranjan Nevase
Batch - 5th June Data Science

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.
2) What can we say about the student marks?

Mean	41.00
Median	40.50
Variance	25.53
Std Dev	5.05

- The mean and the median are close to each other, which suggests that the distribution of scores is symmetric and not skewed. This means that there are no extreme values or outliers that affect the center of the distribution.
- The variance and the standard deviation are relatively low, which indicates that the scores are not very spread out and have low variability. This means that most of the scores are close to the mean and there is less diversity or difference among the scores.
- The range of scores is from 34 to 56, which is 22 points wide.

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

- The nature of skewness is zero skew for mean = median.
- This means that the distribution is symmetrical, and its left and right sides are mirror images of each other.

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

- The nature of skewness when the mean is greater than the median is right skew or positive skew. This means that the distribution is asymmetrical and has a long tail on its right side.
- A right-skewed distribution indicates that there are more values below the mean than above it, and that there are some extreme values or outliers on the right side that pull the mean up.

Name - Niranjan Nevase
Batch - 5th June Data Science

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

- The nature of skewness when the median is greater than the mean is left skew or negative skew. This means that the distribution is asymmetrical and has a long tail on its left side.
- A left-skewed distribution indicates that there are more values above the mean than below it, and that there are some extreme values or outliers on the left side that pull the mean down.

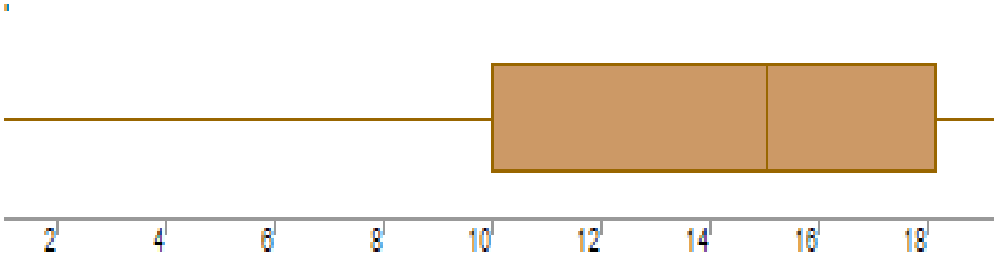
Q16) What does positive kurtosis value indicates for a data ?

- A positive kurtosis value indicates that the distribution has heavier tails than the normal distribution. This means that there are more outliers or extreme values in the data than in a normal distribution.
- A positive kurtosis value also indicates that the distribution is peaked and has a narrower center than a normal distribution.

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

- A negative kurtosis value indicates that the distribution has lighter tails than the normal distribution. This means that there are fewer outliers or extreme values in the data than in a normal distribution.
- A negative kurtosis value also indicates that the distribution is flatter and has a wider center than a normal distribution.

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



Name - Niranjan Nevase
Batch - 5th June Data Science

What can we say about the distribution of the data?

- The lower whisker length is longer than the upper whisker length, which indicates that the data is skewed to the left or negatively skewed. This means that there are more values on the left side of the median than on the right side, and that there are some extreme values or outliers on the left side that pull the mean down.

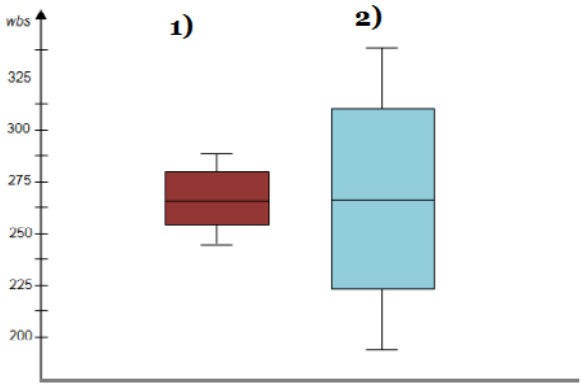
What is nature of skewness of the data?

- Data is skewed to the left or negatively skewed.

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

- IQR = Q3 – Q1 = 18 – 10 = **8 (approx..)**

Q19) Comment on the below Boxplot visualizations?



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

- Both boxplots have the same median value, which means that the middle point of the data is the same for both distributions. This also implies that the mean of the data is likely to be similar, unless there are extreme outliers that skew the mean.

- IQR Boxplot 2 is greater than boxplot 1. Boxplot 2 has a wider range of values and more dispersion in its data than boxplot 1, even though they have the same median value.
- Boxplot 2 also has longer whiskers than boxplot 1, which means that the minimum and maximum values of the data are farther away from the median for boxplot 2 than for boxplot 1. This also suggests that boxplot 2 has more variability or dispersion in its data than boxplot 1.
- Longer whiskers may also indicate the presence of outliers, which are values that are unusually high or low compared to the rest of the data.

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

Population mean, $\mu = 34.42208$

Population Std. dev, $\sigma = 9.074903$

a. P(MPG>38)
   z-score = $(x - \mu) / \sigma = (38 - 34.42208) / 9.074903 = 0.39427$
   From z able, prob for MPG < 38 = 0.6517
   P (MPG > 38) = 1 – 0.6517 = **0.3483 = 34.83%**

b. P(MPG<40)
   z-score = $(x - \mu) / \sigma = (40 - 34.42208) / 9.074903 = 0.61465$
   From z able, P(MPG<40) = **0.7324 = 73.24%**

c. P (20<MPG<50)
   Ans: prob_MPG_greater_than_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)
   print('p(MPG>20)=',(prob_MPG_greater_than_20))

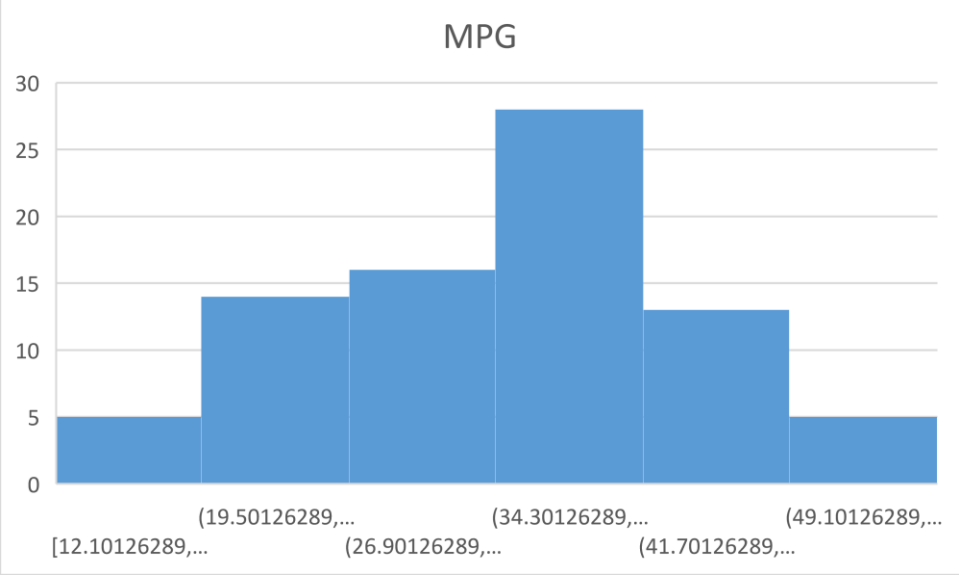
p(MPG>20)= 0.943

prob_MPG_less_than_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)
print('P(MPG<50)=',(prob_MPG_less_than_50))
P(MPG<50)= 0.956

prob_MPG_greaterthan20_and_lessthan50= (prob_MPG_less_than_50) - (prob_MPG_greater_than_20)
print('P(20<MPG<50)=',(prob_MPG_greaterthan20_and_lessthan50))
P(20<MPG<50)= 0.0130

Q 21) Check whether the data follows normal distribution

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

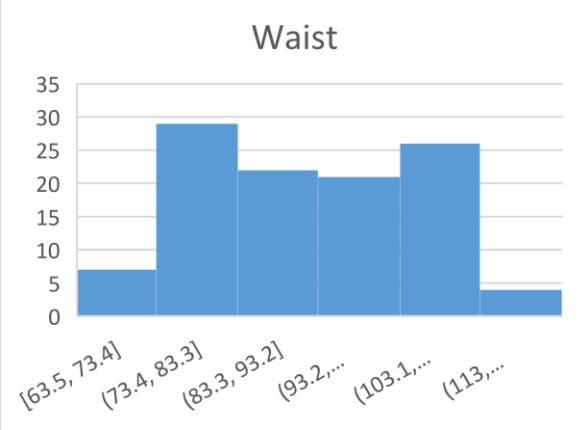


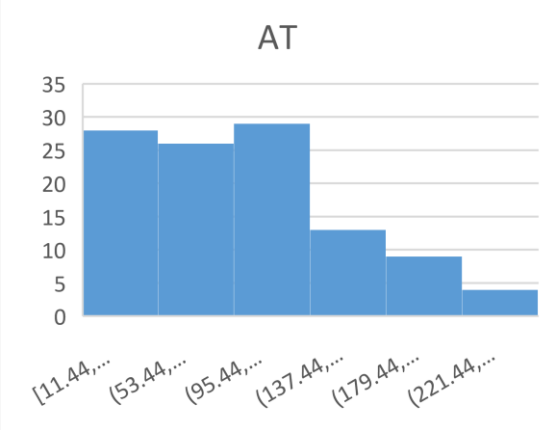
- Histogram shows MPG is data is Normally distributed (bell curve) with slightly negative skewness.

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

Name - Niranjan Nevase
Batch - 5th June Data Science

Dataset: wc-at.csv





- We can conclude from histograms that AT shows bell curve, while there is no bell curve like shape in Waist histogram.

So AT is normally distributed with positive skewness while Waist is not normally distributed.

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

For 90% confidence interval , Z score = 1.64485

For 94% confidence interval , Z score = 1.88079

For 60% confidence interval , Z score = 0.84162

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

Sample Size = 25, Degrees of freedom, df = 25 -1 = 24

- t score for 95% confidence interval = 2.064
- t score for 96% confidence interval = 2.171
- t score for 99% confidence interval = 2.797

Name - Niranjan Nevase
Batch - 5th June Data Science

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

- To determine the probability that 18 randomly selected bulbs would have an average life of no more than 260 days if the CEO's claim were true, you can use the z-test for the population mean.

- The null hypothesis (H0) is that the average light bulb lasts 270 days, and the alternative hypothesis (H1) is that the average light bulb does not last 270 days. In this case, we're interested in the probability that the sample mean is less than or equal to 260 days.

Population mean (μ): 270 days

Sample mean ($\bar{x}$): 260 days

Standard deviation (σ): 90 days

Sample size (n): 18 bulbs

First, calculate the standard error (SE) of the sample mean using the formula:

$SE = \sigma / \sqrt{n}$

Name - Niranjan Nevase
Batch - 5th June Data Science

SE = 90 / √18 ≈ 21.21

Next, calculate the z-score using the formula:

$z = (\bar{x} - \mu) / SE$

z = (260 - 270) / 21.21 ≈ -0.4713

Now, we can find the probability (p-value) associated with this z-score. Since we are interested in the probability that the sample mean is less than or equal to 260 days, you'll find the cumulative probability to the left of the z-score (-0.4713) using a standard normal distribution table which is approximately 0.3192.

So, the probability that 18 randomly selected bulbs would have an average life of no more than 260 days, assuming the CEO's claim is true, is approximately **0.3192** or **31.92%.**