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

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 |  | nominal |
| Level of Agreement |  | ordinal |
| IQ(Intelligence Scale) |  | interval |
| Sales Figures |  | ratio |
| Blood Group |  | nominal |
| Time Of Day |  | nominal |
| Time on a Clock with Hands |  | ordinal |
| Number of Children |  | ratio |
| Religious Preference |  | nominal |
| 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?  
= 3/8

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

1. Equal to 1 = 1/36
2. Less than or equal to 4 = 1/6
3. Sum is divisible by 2 and 3 =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?

= 25/42

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

= E(x) = ∑(Xi \* Pi)

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

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

= 3.125

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: |  |
| --- | --- |
| Points: | 3.5965625 |
| Score: | 3.21725 |
| Weigh: | 17.84875 |

| Median: |  |
| --- | --- |
| Points: | 3.695 |
| Score: | 3.325 |
| Weigh: | 17.71 |

| Mode: |  |
| --- | --- |
| Points: | 3.92 |
| Score: | 3.44 |
| Weigh: | 17.02 |

| Variance: |  |
| --- | --- |
| Points: | 0.2858813508 |
| Score: | 0.9573789677 |
| Weigh: | 3.193166129 |

| Standard deviation: |  |
| --- | --- |
| Points: | 0.5346787361 |
| Score: | 0.978457443 |
| Weigh: | 1.786943236 |

| Range: |  |
| --- | --- |
| Points: | 2.17 |
| Score: | 3.911 |
| Weigh: | 8.4 |

**Inference:**

**The mean, median, and the mode of the given data tell us about the average data as we can see the values are approximately equal. The variance and the standard deviation of the given data tell us how widely spread the data is from the mean of the data. And from the range of the given data we come to know the difference between the maximum value of the data and the minimum value of the data.**

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?

= The expected weight of a randomly chosen patient is 145.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**=**

| Skewness: |  |
| --- | --- |
| Speed= | -0.1175098614 |
| Distance= | 0.8068949602 |

| Kurtosis: |  |
| --- | --- |
| Speed= | -0.5089944204 |
| Distance= | 0.4050525817 |

**Inference:**

**From the above table, we can see the skewness for the speed is negative which indicates left-skewed towards the lower values, suggesting that the most speed is concentrated in the lower end with lower values. And, for the distance, it is the opposite as it is positive, indicating it is right-skewed and has most of its concentration towards the higher values.**

**For the kurtosis of the given data, the above table indicates the kurtosis to be Platykurtic, which means there are few extreme data and the values are concentrated near the mean of the data.**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**=**

| Skewness: |  |
| --- | --- |
| SP= | 1.611450196 |
| WT= | -0.6147533255 |

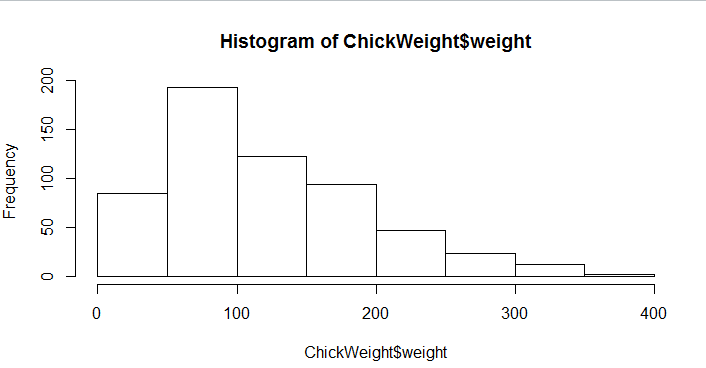
| Kurtosis: |  |
| --- | --- |
| SP= | 2.977328944 |
| WT= | 0.950291491 |

**Inference:**

**From the skewness, we can understand that the values of the speed(SP) are right-skewed and the majority of the values are concentrated towards the higher values. But for weight(WT) we see it is left-skewed with a negative value which indicates that most of the values are concentrated towards the lower values.**

**The kurtosis of the given data indicates that the data has values which are not extreme and are concentrated near the mean of the data.**

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



= The histogram shows the frequency of the ChickWeight$weight, through this visualisation we can understand that frequency is higher towards the start with the lower range of values but extremely less towards the higher values. The highest that we can see is 50 to 100 with a frequency of around 200 and the lowest frequency is in the range of 350 to 400.



= The boxplot tells us about the many outliers lying extremely far from the average value of the data. And, we can see the data is asymmetrical since the distance from the Q1 and the Q3 quartiles are different.

**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?

=

| Confidence Interval | Lower Bound (pounds) | Upper Bound (pounds) |
| --- | --- | --- |
| 94% | 194.47 | 205.53 |
| 98% | 193.02 | 206.98 |
| 96% | 193.87 | 206.13 |

**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?

=

1)

Mean = 41

Median = 40.5

Variance = 25.53

Standard deviation = 5.05

2) We can see that the student's mean score is 41, which is slightly higher than the median score of 40.5. This indicates that the distribution of scores is slightly skewed to the right, meaning there are more scores above the median than below. The variance is 25.53 and the standard deviation is 5.05, which suggests that the scores are moderately spread out.

There are two outliers in the data set: 49 and 56. These scores are more than 1.5 standard deviations away from the mean.

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

= When the mean and median of a data set are equal, it indicates that the distribution is symmetrical. There is no skewness towards the left or right side of the distribution.

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

= When the mean of a data set is greater than the median, it typically indicates a positively skewed distribution. This means the distribution has a longer "tail" towards higher values compared to the left side.

Therefore, a mean exceeding the median suggests that the majority of the data points are clustered towards the lower end, while a few data points skew the mean towards the higher end.

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

= When the median of a data set is greater than the mean, it typically indicates a negatively skewed distribution. This means the distribution has a longer "tail" towards lower values compared to the right side.

Therefore, a median exceeding the mean suggests that the majority of the data points are clustered towards the higher end, while a few data points skew the mean towards the lower end.

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

= Positive kurtosis:

* Indicates a distribution with heavier tails and a sharper peak compared to a normal distribution.
* More data points fall in the tails, meaning there are more extreme values (both high and low) compared to the centre.
* Interpretation:

-More potential outliers

-Less concentration of data around the mean

-More variability in the data

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

= Negative kurtosis:

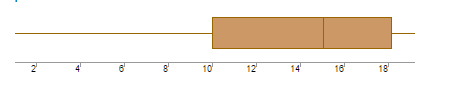
* Indicates a distribution with lighter tails and a flatter peak compared to a normal distribution.
* Fewer data points fall in the tails, meaning there are fewer extreme values (both high and low) compared to the centre.
* Interpretation:

-Less variability in the data

-Data tends to be concentrated around the mean

-Less likely to encounter outliers

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



What can we say about the distribution of the data?

* The median is greater than the mean of the data.
* The interquartile range is 8
* The 1st quartile is 10
* The median is between 14 and 16
* And, the 3rd quartile is 18.

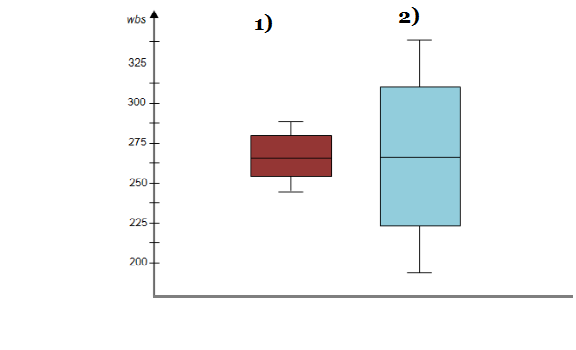
What is nature of skewness of the data?

= The distribution of the data is left-skewed and the median is greater than the mean of the data.

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

= The interquartile range (IQR) is the difference between the Q1 and the Q3 of the data. Here, the Q1 is 10 and the Q3 is 18 so the IQR = Q3-Q1 = 18-10 = **8**.

Q19) Comment on the below Boxplot visualizations?



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

= We can see that the is a lot more variation going on in Boxplot 1 than the Boxplot 2. Since the mean and the median of both datasets are equal, there is no skewness in both of the datasets.

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)

= Probability of MPG of Cars is:

a. P(MPG>38) = 0.4074074074

b. P(MPG<40) = 0.7530864198

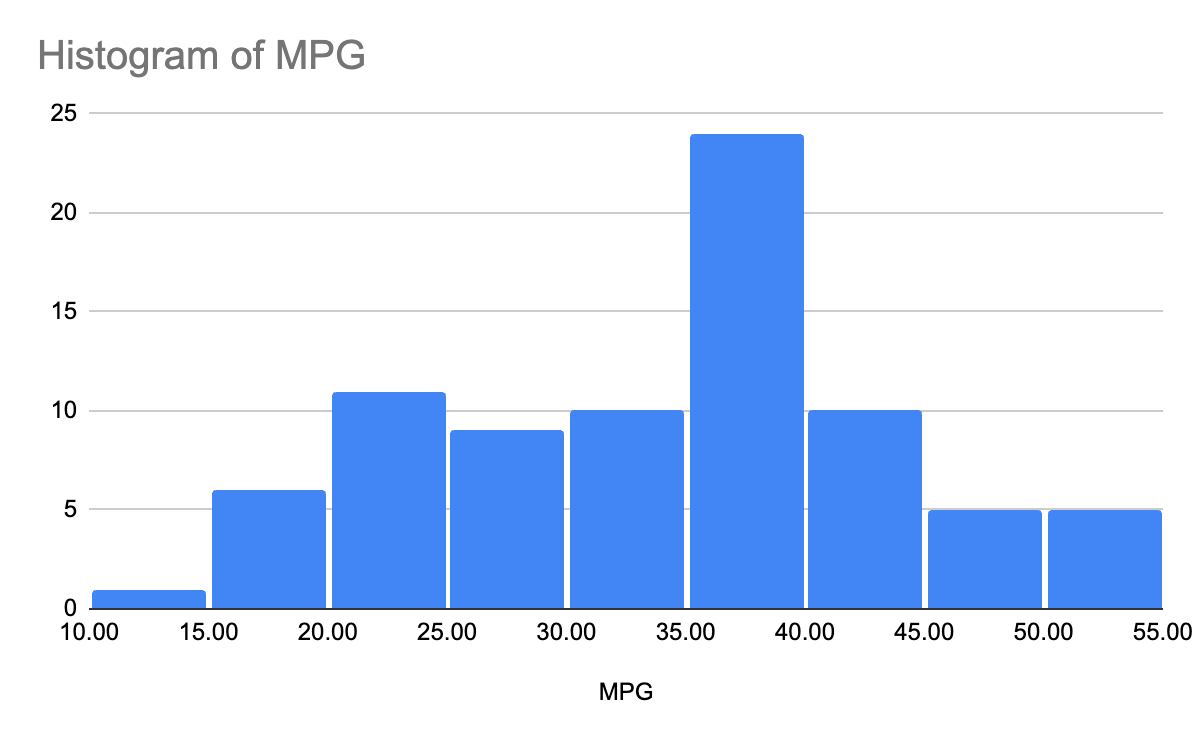
c, P (20<MPG<50) = 0.8518518519

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

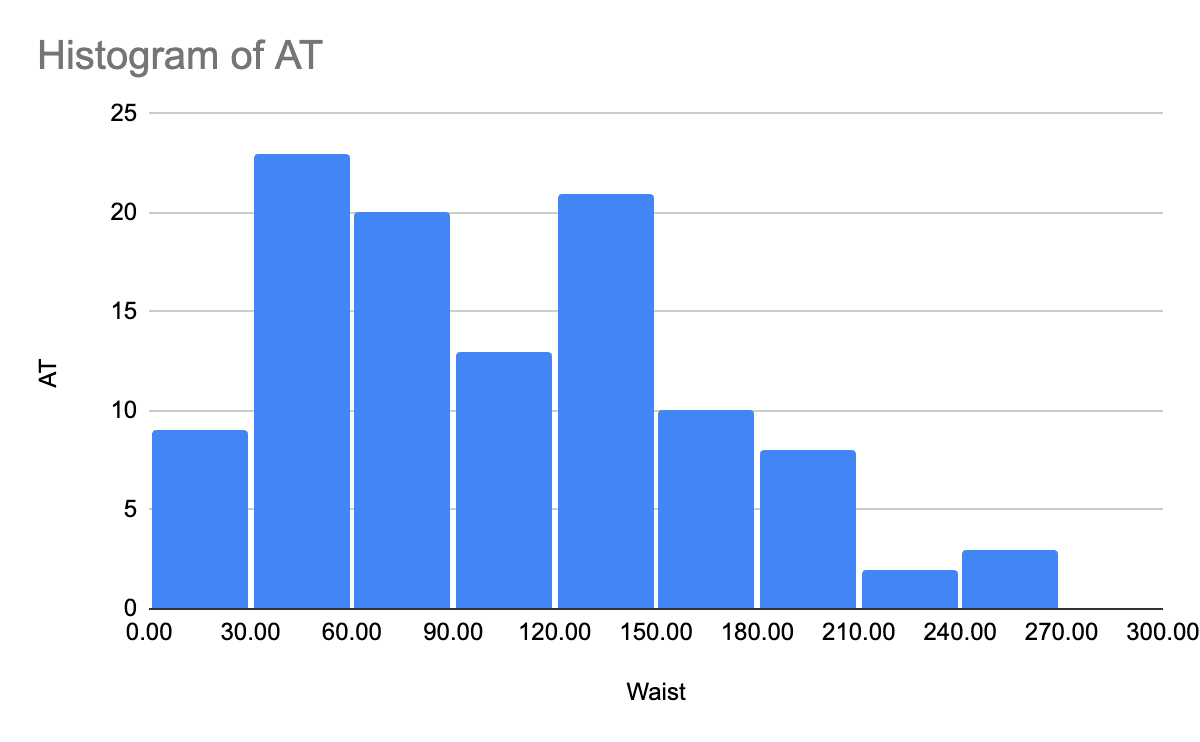
= The MPG of Cars does follow Normal distribution. As per the histogram, the data is symmetrical which indicates that it does not follow normal distribution. The graph forms an inverted U-shaped curve to indicate it is a normal distribution.



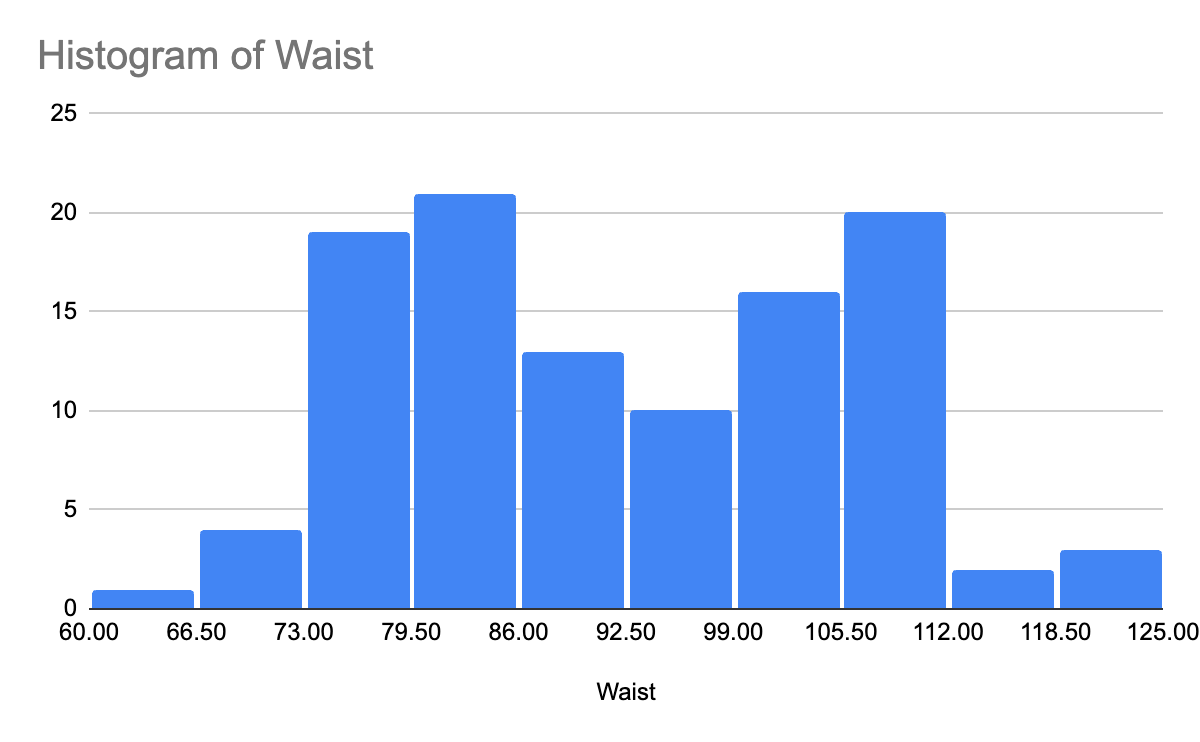
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

= According to the graph shown below the Adipose tissue (AT) is showing a normal distribution dataset. The curve will form an inverted U-shaped structure to ultimately suggest it is following a normal distribution.



= The Waist from the dataset is also following a normal distribution. Since the histogram will produce an inverted U-shaped curve it will indicate the dataset follows a normal distribution.



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

= The z-scores for the confidence intervals you requested are:

* 90% confidence interval: 1.645
* 94% confidence interval: 1.555
* 60% confidence interval: 0.253

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

= The t-scores for the 95%, 96%, and 99% confidence intervals with a sample size of 25:

* 95% confidence interval: 2.064
* 96% confidence interval: 2.171
* 99% confidence interval: 2.797

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

= The calculated probability is approximately 0.1428. Therefore, if the CEO's claim were true, the probability of randomly selecting 18 bulbs with an average lifespan of no more than 260 days is around 14.28%.

Working:

t = (Xbar - mu) / (s / sqrt(n))

t = (260 - 270) / (90 / sqrt(18)) = -1.074

df = 18 - 1 = 17

pt(-1.074, 17) = Ans.