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

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

**= Possible Combinations : (HHH, HHT, HTH, THH, TTH, THT, HTT, TTT)  
The number of combinations which have two heads and one tail are:  
(HHT, HTH, TTH)  
Three coins are tossed, the probability of two heads and one tail is 3/8.**

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
4. **Probability equal to 1** = 0 because minimum number we should get is 2 when we roll 2 Dice.

**b)   Probability to get less than or equal to 4** = Below are the pair of outcomes which give sum of 4.

die 1   die 2  sum

 1        1       2

 1        2       3

 1        3       4

 2        1       3

 2        2       4

 3        1       4

So we have in total 6 ways of outcomes give sum of 4.

Answers will be 6/36 = 1/6.

**c)   Sum is divisible by 2 and 3=**

Numbers that can be divided by 2&3 between 2 to 12 are 6,12.

Favorable outcomes are (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6)

Probability = 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?

**= Total Number of balls = (2 + 3 + 2) = 7  
n(S) = Number of ways of drawing 2 balls out of 7**

**= 7 C2**

**=**

**= 21**

**N(E) = Number of ways of drawing 2 balls out of (2+3) balls.**

**= 5C2**

**=**

**= 10**

**P(E) =**

**=**

**The probability of none of the drawn ball blue is 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

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

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

**= 3.09**

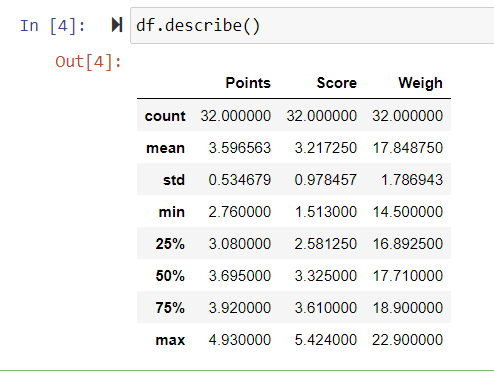
**The expected number of candies for randomly selected child is 3.09**

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

* For Points, Score, Weight>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

= **Use Q7.csv file**



**Mode :**

| **Points** | **Score** | **Weigh** |
| --- | --- | --- |
|  |  | **3.07** | **3.44** | **17.02** |

**Variance : Points 0.285881**

**Score 0.957379**

**Weigh 3.193166**

**Here mean and median values are quite similar to each other hence the distribution of data is symmetric.**

**Range of Weight column is larger than all other two columns.**

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?

**Expected Value =**

**= 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199 / 9**

**= 1308 / 9**

**= 145.33**

**= Expected value of the weight of patient is 145.33.**

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

Cars speed and distance

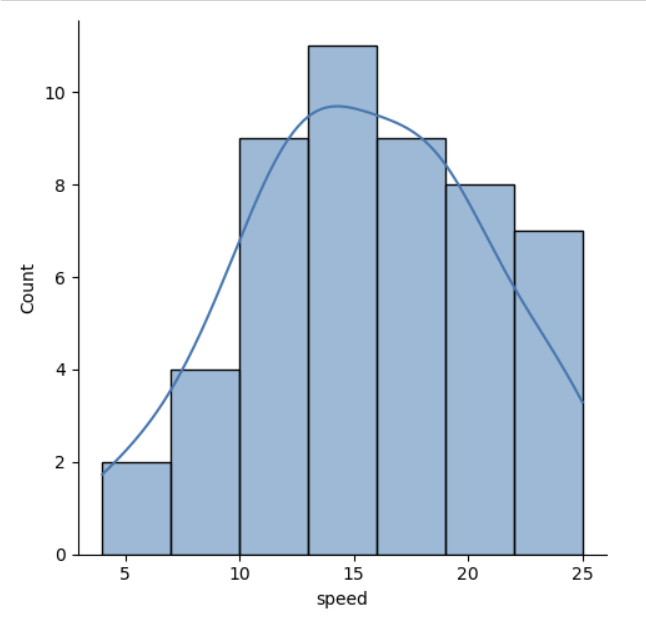
Use Q9\_a.csv

**= Skewness of Speed = -0.11750986144663393**

**# Tail of the distribution is skewed to the left.**

**Kurtosis of Speed = -0.5089944204057617**

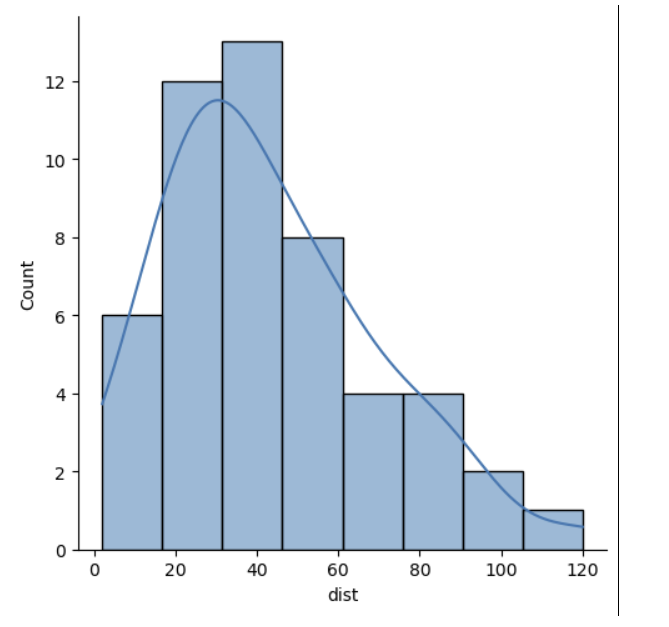
**# The peakness of the distribution is light.**

****

**= Skewness of Distance = 0.8068949601674215**

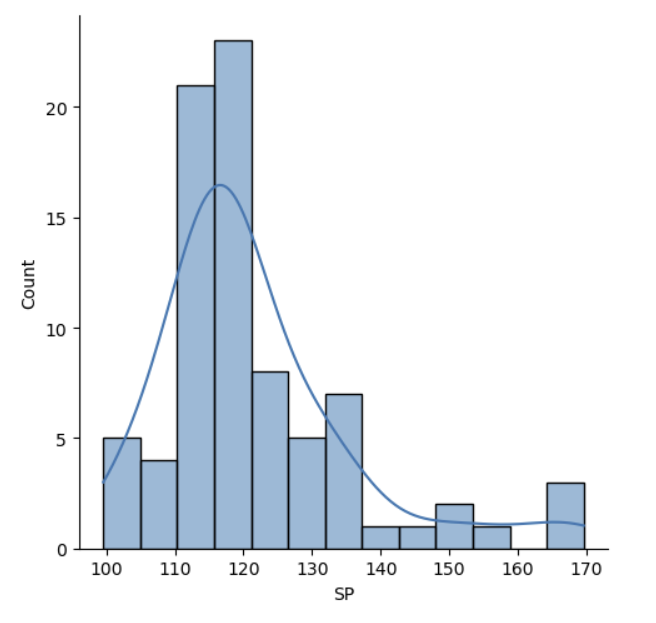
**# Tail of the distribution is skewed to the Right.**

**Kurtosis of Distance = 0.4050525816795765**

**# The peakness of the distribution is heavy peak.**

SP and Weight(WT)

Use Q9\_b.csv

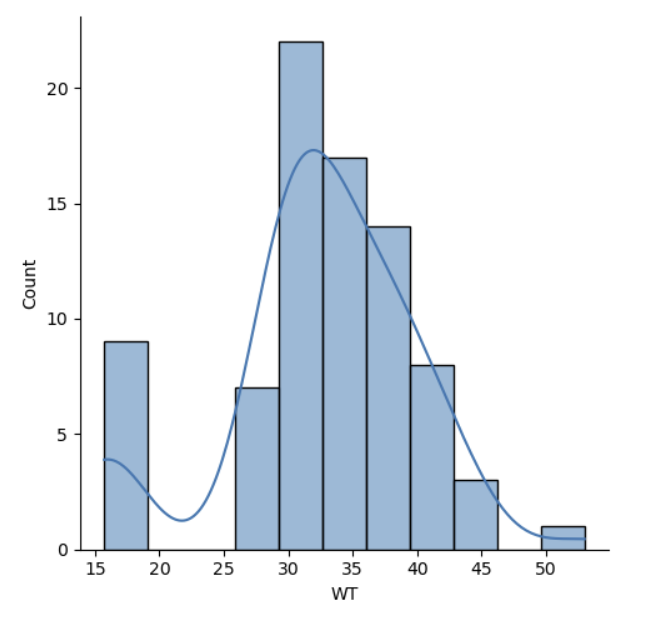


**= Skewness of Distance =** 1.6114501961773586

**# Tail of the distribution is skewed to the Right.**

**Kurtosis of Distance =** 2.9773289437871835

**# The peakness of the distribution is heavy peak.**

****

**= Skewness of Distance =** -0.6147533255357768

**# Tail of the distribution is skewed to the Left.**

**Kurtosis of Distance =** 0.9502914910300326

**# The peakness of the distribution is heavy peak.**

Q10) Draw inferences about the following boxplot & histogram



**The above histogram shows that it is not a normal distribution as it is right skewed.  
There might be chance of present of outliers in ChickWeight as we can see there is very a smaller number of datapoint in right bin of 350 to 400.  
Central tendency of data is between 50-100.  
Most datapoints fall under range of 50-100 where between 300-400 very a smaller number of datapoints are present.**



**The above boxplot shows that the Data is more concentrated in the lower half.  
The IQR is smaller so we can say that there is less variability in datapoints.  
Given distribution of data is asymmetric as it has longer whisker than other.  
There are more thane one outliers are present.**

**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 =**

**For 94% Confidence Interval :**

**200 1.8808 (30 / )**

**200 1.8808 (30/44.72)**

**200 1.8808 (0.67)**

**200 1.260136**

**200 + 1.260136 = 201.260136 & 200 - 1.260136 = 198.739864**

**For 98% Confidence Interval :**

**200 2.3263 (30 / )**

**200 2.3263 (30/44.72)**

**200 2.3263 (0.67)**

**200 1.558621**

**200 + 1.558621 = 201.558621 & 200 - 1.558621 = 198.441379**

**The Confidence interval for 98% is 201.558621, 198.441379.**

**For 96% Confidence Interval :**

**200 1.7507 (30 / )**

**200 1.7507 (30/44.72)**

**200 1.7507 (0.67)**

**200 1.172969**

**200 + = 201.172969 & 200 - = 198.827031**

**The Confidence interval for 94% is 201.260136, 198.739864.**

**The Confidence interval for 98% is 201.558621, 198.441379.**

**The Confidence interval for 96% is 201.172969, 198.827031.**

**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.

**import numpy as np**

**scores = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]**

**mean = np.mean(scores)**

**median = np.median(scores)**

**variance = np.var(scores)**

**std\_dev = np.std(scores)**

**Mean: 41.0**

**Median: 40.5**

**Variance: 24.11111111111111**

**Standard Deviation: 4.910306620885412**

1. What can we say about the student marks?

**= Mean and Median is quite close to each other which means given distribution of data is not much skewed. Variance and Standard Deviation is high which means the spread of data is more.**

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

**= When Mean and Median of data is equal that means there is no skewness present in data and and the given data is normal distribution.**

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

**=When Mean is greater than Median, it means given data is positively skewed.**

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

**=When Median is greater than Mean, it means given data is negatively skewed.**

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

**= Positive kurtosis indicated that in given data outliers are present and high peakness.**

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

**= Positive kurtosis indicated that in given data outliers present is less and low peakness.**

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



What can we say about the distribution of the data?

**= Given data is left skewed and had median value of approx. 15.  
Most values lie between 10 to 18 data point.  
More data is concentrated one side.**

What is nature of skewness of the data?

**= Given data is left skewed.**

What will be the IQR of the data (approximately)?   
**= IQR of given boxplot is 10-18.**

Q19) Comment on the below Boxplot visualizations?



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

**= Boxplot 1 has less variance compare to boxplot 2 since we don’t have any idea about how many datapoints are present in both boxplot.  
Median point of both boxplot is equal.  
Data is normally distributed and there is no skewness present.**

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)

**= Total Number of MPG greater than 38 = 33**

**= 33/81**

**=0.40**

* 1. P(MPG<40)

**= Total Number of MPG less than 40 = 61**

**= 61/81**

**= 0.75**

c. P (20<MPG<50)

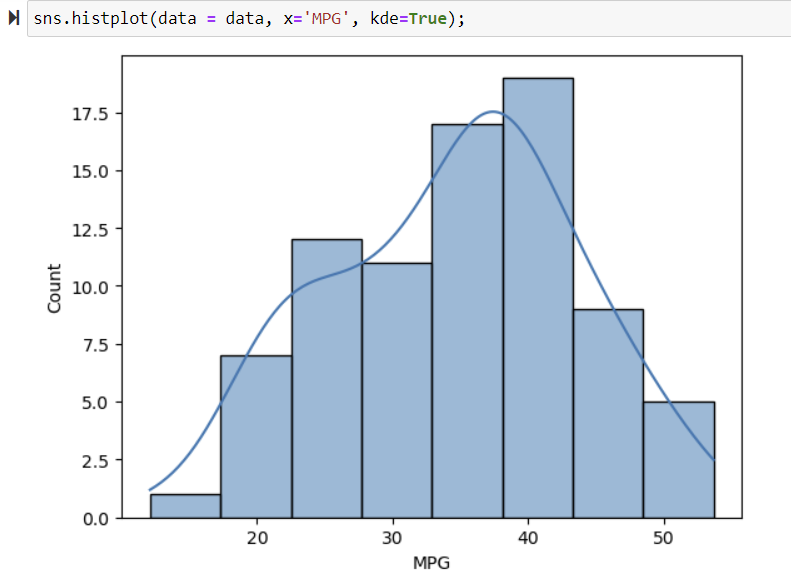
**= Total number of MPG between 20 and 50 = 69**

**= 69/81**

**= 0.85**

Q 21) Check whether the data follows normal distribution

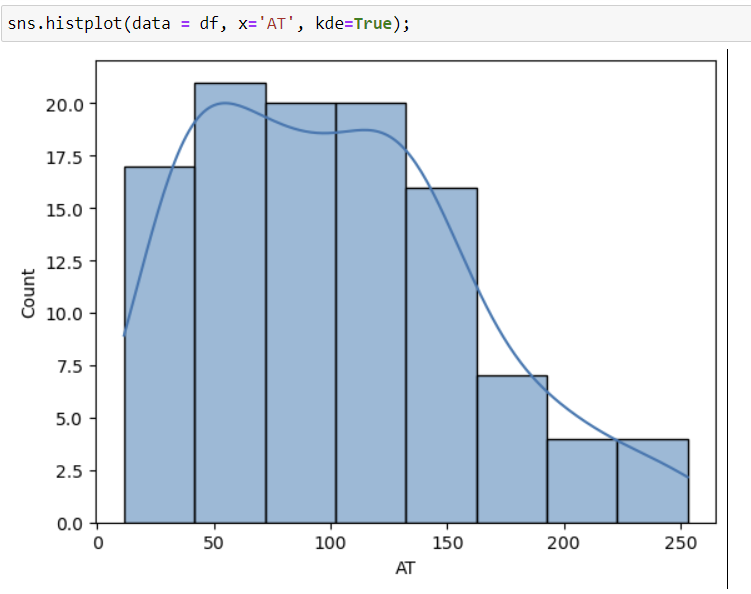
1. Check whether the MPG of Cars follows Normal Distribution

 Dataset: Cars.csv

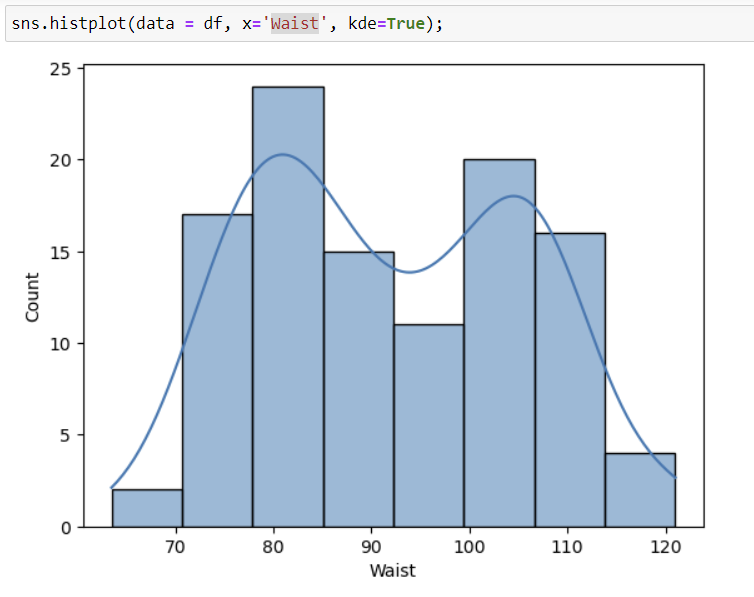
**MPG of Cars is slightly left skewed.**

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

Dataset: wc-at.csv



**Adipose Tissue is right skewed distribution.**



**Waist Circumference is not a normal distribution.**

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

**90% Confidence Interval:**

**C = 0.90**

**α = 1 – C = 0.10**

**Zα/2 = Z0.05​**

**Z = 1.645**

**94% Confidence Interval:**

**C = 0.94**

**α = 1 – C = 0.06**

**Zα/2 = Z0.03​**

**Z = 1.555**

**60% Confidence Interval:**

**C = 0.60**

**α = 1 – C = 0.40**

**Zα/2 = Z0.20**

**Z = 0.253**

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

**df = n – 1 = 25 - 1 = 24**

**1. 95% Confidence Interval:**

**C = 0.95**

**df = 24**

**tα/2 = t0.025**

**tα/2 = 2.06390**

**2. 96% Confidence Interval:**

**C = 0.96**

**df = 24**

**tα/2 = t0.02**

**tα/2 = 2.172**

**3. 99% Confidence Interval:**

**C = 0.99**

**df = 24**

**tα/2 = t0.005**

**tα/2 = 2.79694**

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

**T-Score =**

**=**

**=**

**=**

**=**

**= -0.471404521**

**Degree of freedom = n – 1 = 18 – 1 = 17**

**The probability t < -0.471 with 17 degree of freedom, the t-value is less than the t-value obtained.**

**The probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulb is 300 days.**