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

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

Ans:

|  |
| --- |
| **By Probability Method**  **Total Number of events**  = Number of Possibilities per Experiments Number of Experiments = 23= 8  **Total Number of Interested Events**  = nCr = 3C2 = = 3  **Probability of Interested Events**  P(X=2H) = = =0.375 = 3.75 % |

|  |
| --- |
| **By Probability Mass Function formula**  p.m.f.= nCx \* P x \* (1-P) n-x  where, n = Number of trails  x = Number of success required  p = probability of getting success in one trail  P(X=2H) = 3C2 \* (0.5)2 \* (1-0.5)3-2  P (X= 2H) = \* 0.25 \* 0.5  P(X=2H) = 3\* 1.25  P(X=2H) = 3.75% |
| * **By Python** |

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

1. Equal to 1

|  |
| --- |
| When we roll two dies,  **Total Number of events**  = Number of Possibilities per Experiments Number of Experiments = 62= 36  We will get minimum sum of 2, Therefore  **Total Number of Interested Events = 0**  **Probability of Interested Events**  P(sum=1) = = =0 = 0 % |

1. Less than or equal to 4

|  |
| --- |
| When we roll two dies,  **Total Number of events**  = Number of Possibilities per Experiments Number of Experiments = 62= 36  **Total Number of Interested Events**  **= Number below that gives sum of 4 or less**  = [(3,1), (2,2), (1,3)] = 3  **Probability of Interested Events**  P(sum4) = = =0.0833 = 8.33 % |

1. Sum is divisible by 2 and 3

|  |
| --- |
| When we roll two dies,  **Total Number of events**  =Number of Possibilities per Experiments Number of Experiments = 62= 36  Number below that divisible by 2&3 both = [6,12]  Number of Combinations that gives sum 6  = [(5,1), (4,2), (3,3), (2,4), (1,5)] = 5  Number of Combinations that gives sum 12  = (6,6) = 1  **Total Number of Interested Events =** 5 + 1 = 6  **Probability of Interested Events**  = = =0.166= 1.66 % |

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 events**  = 7C2 =7C2 = = 21  **Total Number of Interested Events**  = nCr = 5C2 = = 10  **Probability of Interested Events**  P(none of the ball is blue) = = =0.476 =47.6% |

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:

|  |
| --- |
| Excepted Random value =  = (1\*0.015) + (4\*0.20) + (3\*0.65) + (5 \* 0.005) + (6\*0.01) + (2\*0.120)  =3.09  Expected number of candies for a randomly selected child = 3.09 |

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

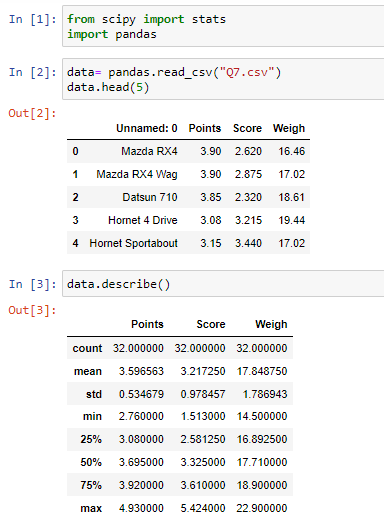
* For Points, Score, Weight

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

**Use Q7.csv file**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **FORMULA(STEPS)** | **POINTS** | **SCORE** | **WEIGHT** | **INFERANCE** |
| MEAN |  | 3.6 | 3.22 | 17.85 | \* More number of observations near mean \*Probability of Getting Mean is High \*Mean will be influenced by Outliers. |
| MEDIAN | \*ARRANGR IN ASSENDIG OR DESSENDING ORDER  \*Finding Middle Value | 3.7 | 3.33 | 17.71 | \*Media will Not be Influenced by the outliers, there for the we are seeing variations in Mean & Median  \*IN points +ve Outliers are there as Mean < Median \*IN Score +ve Outliers are there as Mean < Median \*IN Weight -ve Outliers are there as Mean > Median |
| MODE | \*MOST FREQUENTLY OCCURRED NUMBER | 3.92 | 3.44 | 17.02 | \*Median will be useful in the case of Categorical data |
| VARIENCE | For Population  2=  For Sample  S2= | 0.29 | 0.96 | 3.19 | \*A large number of Variance indicates that the data in data set are far from mean and far from each other and smaller Variance is opposite to larger number's indication. \*Here Weight is having more variance and it indicates that data points in weight are far from mean & far from each other as compared to Points & Score |
| STANDARD DIVIATION |  | 0.53 | 0.98 | 1.79 | \*Standard Deviation tells us how the spread of data around the Mean of data  \*Here Point's data are closely clustered around the Mean as compared to Others. |
| RANGE | Range = Max(xi) - Min (Xi) | 2.17 | 3.91 | 8.4 | \*Range tells us that how much spread is there from the lowest value to the Highest values. |

**By Using Python**

****

|  |  |  |
| --- | --- | --- |
|  |  |  |

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?

|  |  |  |
| --- | --- | --- |
| Xi | P(Xi) | P(Xi)\*Xi |
| 108 | 0.111111 | 12 |
| 110 | 0.111111 | 12.22222 |
| 123 | 0.111111 | 13.66667 |
| 134 | 0.111111 | 14.88889 |
| 135 | 0.111111 | 15 |
| 145 | 0.111111 | 16.11111 |
| 167 | 0.111111 | 18.55556 |
| 187 | 0.111111 | 20.77778 |
| 199 | 0.111111 | 22.11111 |
|  |  | 145.3333 |

Expected Value of the Weight of Randomly chosen patient = 145.33 Pounds

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

|  |  |
| --- | --- |
| **Cars speed and distance**  **Use Q9\_a.csv** | **SP and Weight (WT)**  **Use Q9\_b.csv** |
|  |  |

Inference:

**Speed:**

1. Skewness = -0.117

* data is slightly Negatively Skewed or Left Skewed data (Mass of data is on right side of median),
* means data spared is More on left side of the Median

1. Kurtosis = -0.508

* Data has platykurtic distribution& has thin tails compared to Normal dist.,
* The distribution is flat as compared to Normal distribution.

**Distance:**

1. Skewness = 0.806

* data is skewed Positively or Right skewed data (Mass of data is on left side of median),
* Means data spared is more on right side of the Median

1. Kurtosis = 0.405

* Data has Leptokurtic distribution & has thick tails as compared to normal dist.,
* The distribution is peak as compared to Normal Distribution.

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

|  |  |
| --- | --- |
| histogram | Boxplot1 |
| **Inferance :**   * Positivelly Skewed data(Right Skewed data) * Spared of the data on right side of the distribution is More & Mass of data is on left side of Median * Frequency of the data between 50 – 100 is more | **Interance:**   * Positivelly Skewed data or Right skewed data (Whisker is More on right side on median) * Spared of the data on right side of the distribution is More & Mass of data is on left side of Median * Positive Outliers are there on Right side of the distribution |

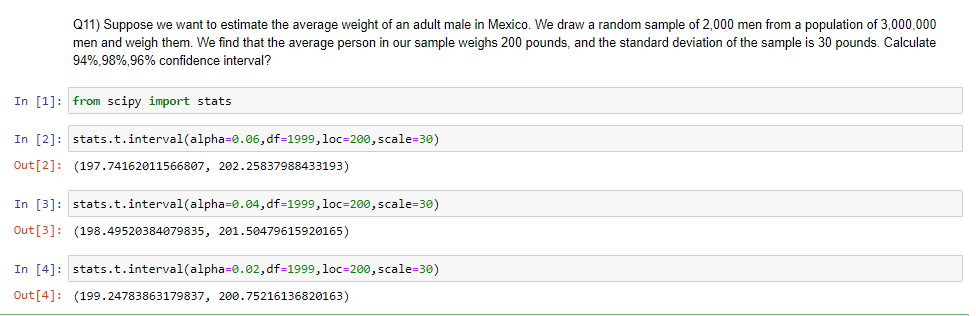
**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

n = 2000 Pounds, N = 3000000, Ẋ = 200, S = 30 Pounds

t =

CI = Ẋ ± t\*

|  |  |  |
| --- | --- | --- |
|  |  |  |

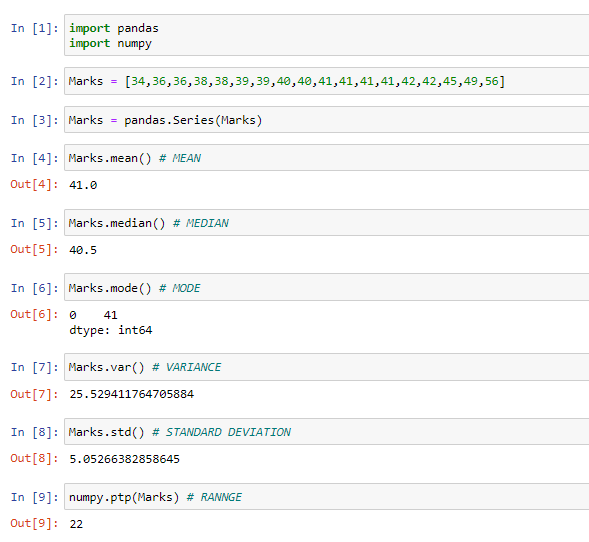


**Q12)** Below are the scores obtained by a student in tests

**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.

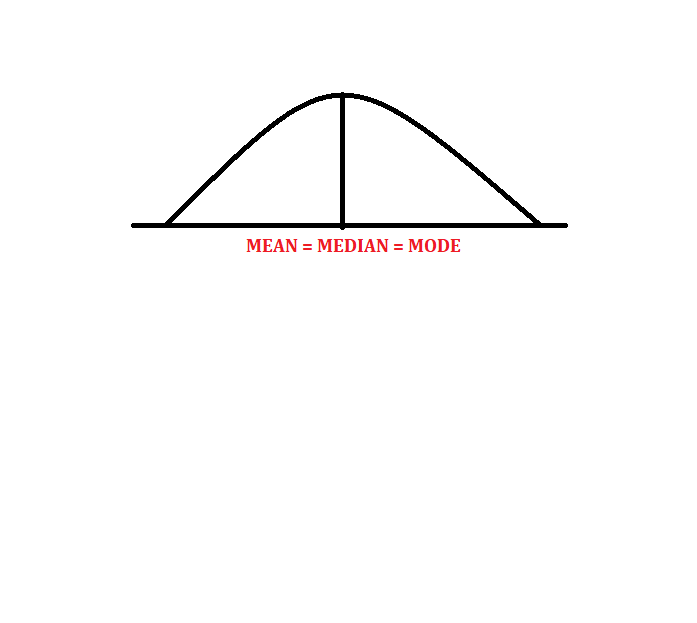
* Mean =   = = 41
* Median = = 40.5
* Variance = S2=   = 434 =25.52941
* Standard Deviation =   = = 5.052664



1. What can we say about the student marks?

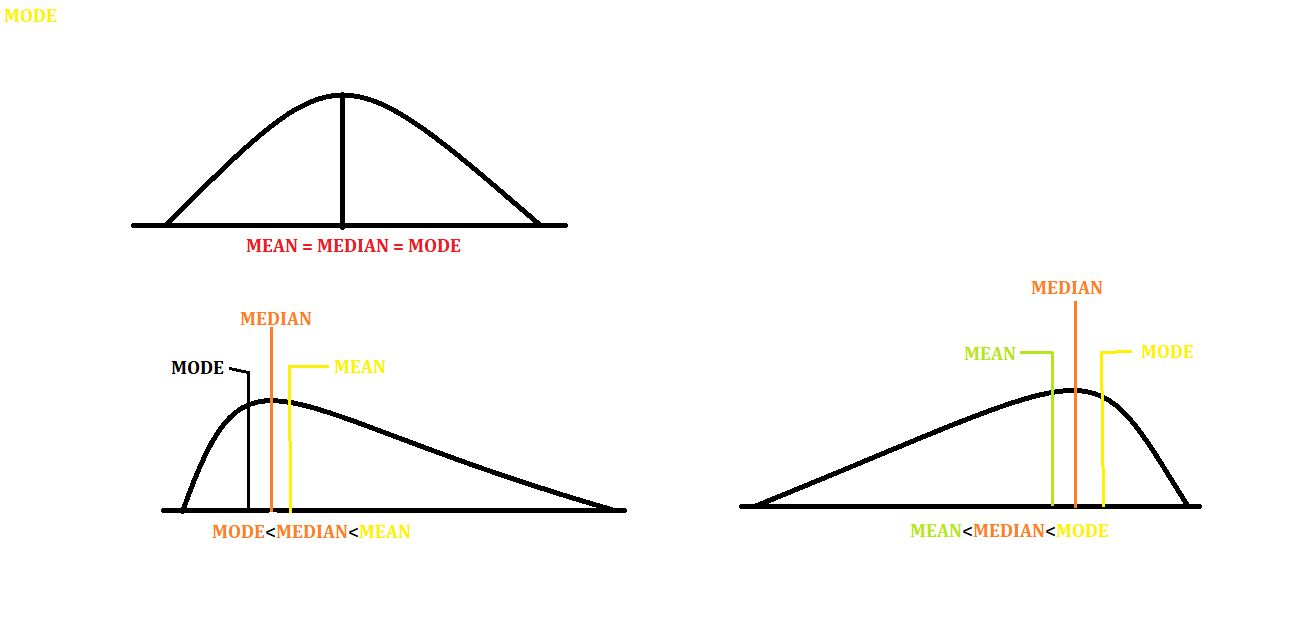
|  |
| --- |
| **Mean = 41**   * Most of students’ marks are nearer to 41 |
| **Median = 40.5 Mean**   * There is no too high (like 98,76) & too low marks (like 0,2) (Outliers) present |
| **Standard deviation = 5.05**  As mean is approximately equal to median follows Normal distribution,   * 1 = (41-5 =36, 41+5 = 47) * 68% of students are scored between 36 to 47 * 2 = (41-10=31,41+10=51) * 95% of students are scored between 41 to 51 * 3SD= (41-15=26,41+16=57) * All most all (99.7%) students are scored between 26 to 57 |

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



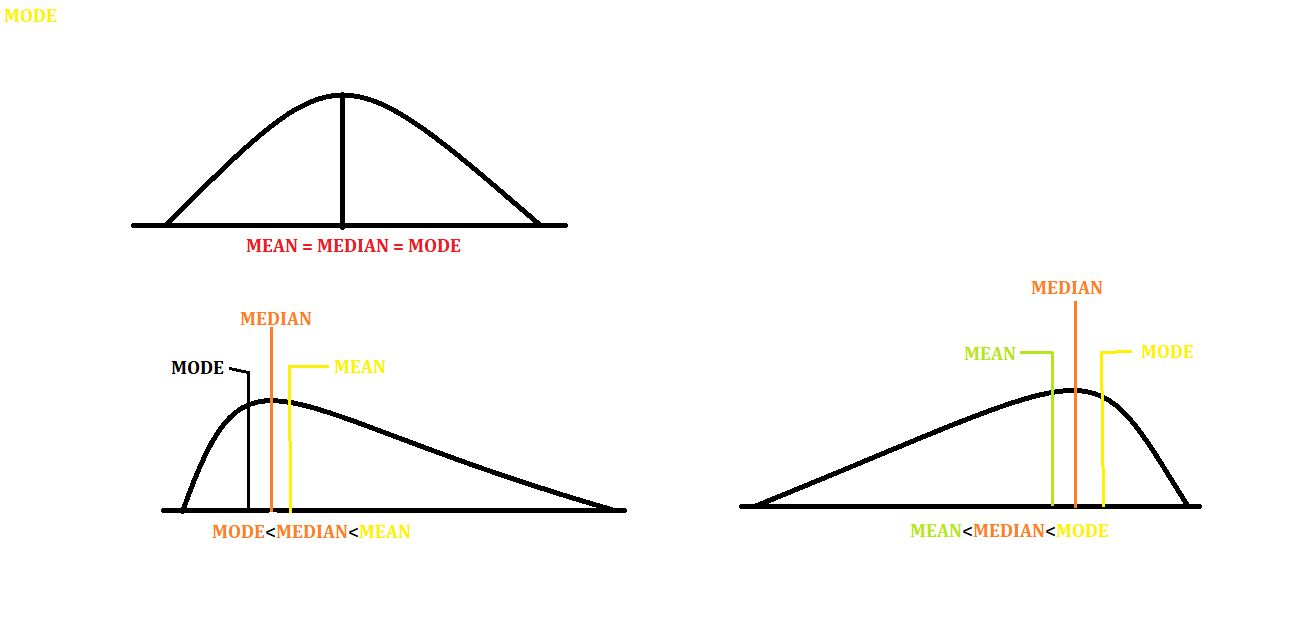
**When Mean = Median,** we can say data is Normally Distributed.

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



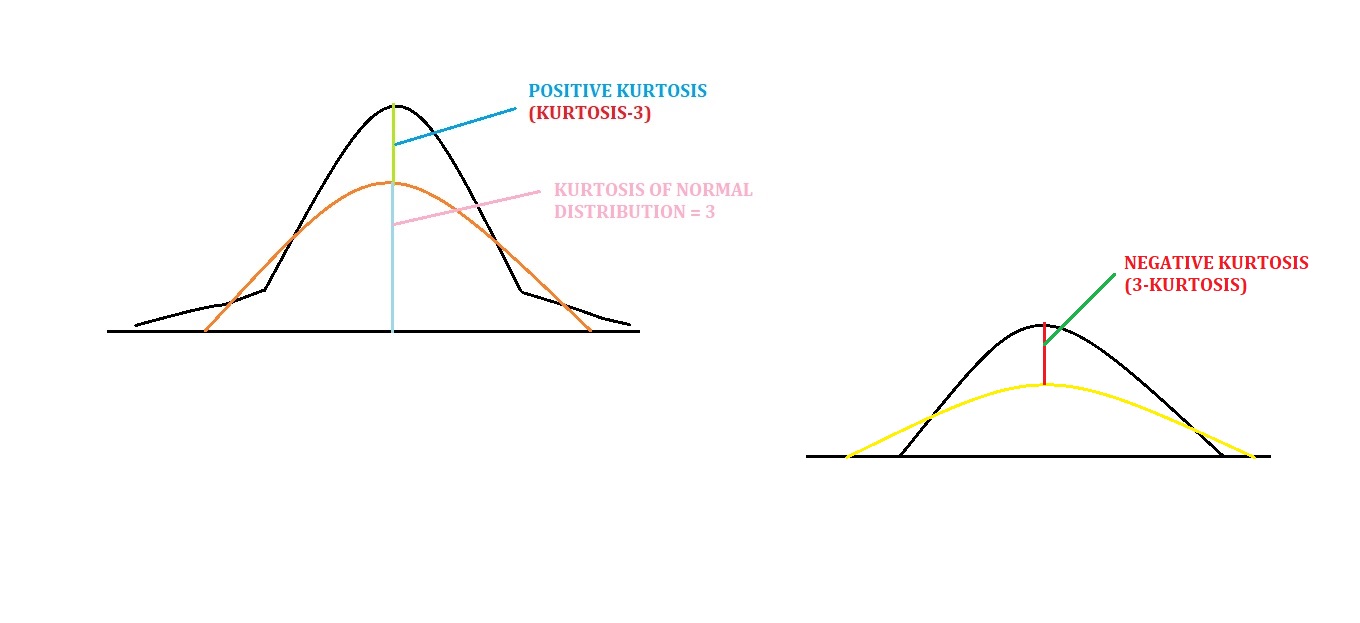
**When Mean > Median,** we can say Positively Skewed data (Right Skewed data).

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



**When Mean < Median**, we can say Negatively Skewed data (left Skewed data).

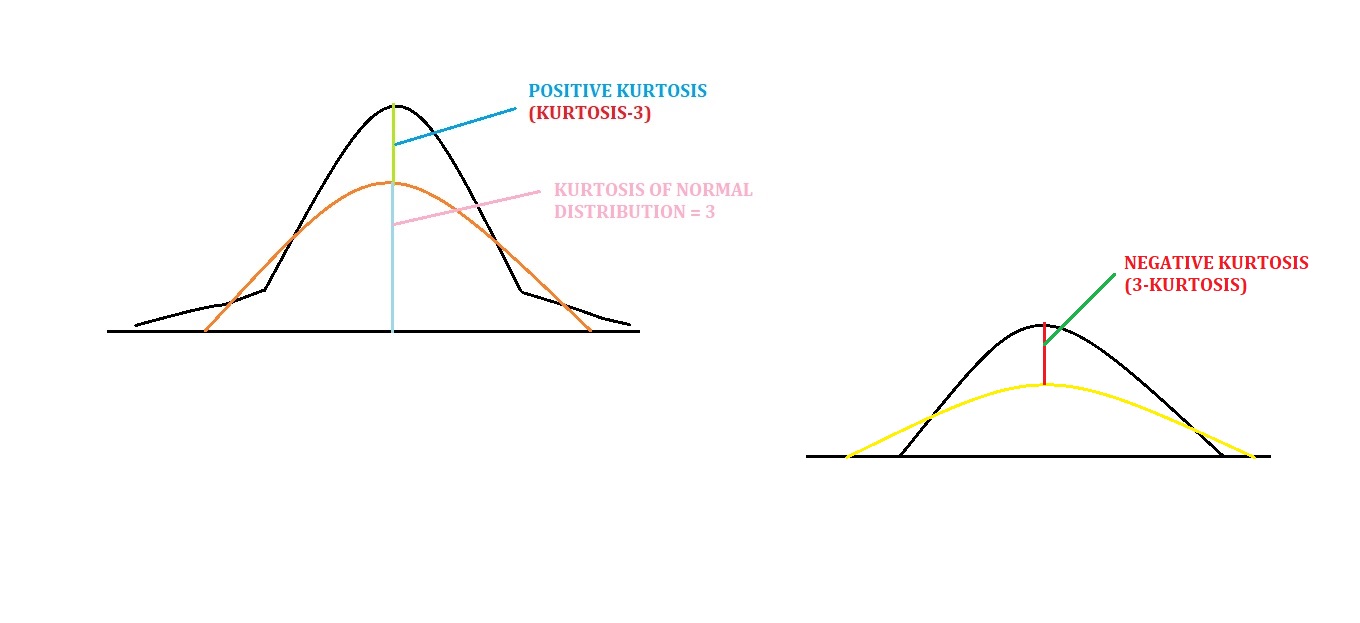
Q16) What does positive kurtosis value indicates for a data?



Positive Kurtosis (Excess Kurtosis) indicates that,

* Distribution is Leptokurtic (peak of bell curve is more as compared to Normal distribution)
* Spared There are more values around mean.

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



Negative Kurtosis indicates that,

* Distribution is Platykurtic (peak of bell curve is less as compared to Normal distribution)
* Spread of the data is More (There are more far values from mean).

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



What can we say about the distribution of the data?

* Most of the data lies between 10 t0 18.
* Q1 = Quartile 1 = 10
* Q2 = Quartile 2 = 15 = MEDIAN = 50th Percentile
* Q3 = Quartile 3 = 18

What is nature of skewness of the data?

**Negatively skewed data**: There are negative outliers present in the data

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

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.

|  |  |
| --- | --- |
| **Boxplot 1** | **Boxplot 2** |
| Data ranges between 240 to 280 | Data ranges between 190 to 340 |
| Mean = Median = Mode = Qurtile2(Q2) =260 | Mean = Median = Mode= Qurtile2(Q2) =260 |
| Normally Distributed | Normally Distributed |
| Quartile1 = 255 | Quartile1 = 220 |
| Quartile3 = 280 | Quartile3 = 310 |
| IQR (INTER QURTAIL RANGE) is less = 280-255 =25 | IQR (INTER QURTAIL RANGE) is more = 310-220 =90 |

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)
  3. P (20<MPG<50)

Mean = 34.42

STANDARD DEVIATION = S = 9.131445

P(XXi) = P(ZXi)

|  |  |  |  |
| --- | --- | --- | --- |
| **P(MPG>38)**  = 1 - P(MPG38)  = 1 -P  = 1 – P  = 1-P(Z0.392)  =1-0.6517  =0.3483  =34.83% | **P(MPG)**  = P(MPG)  = P  = P  = P(Z0.611)  = 0.7291  =72.91% | | **P(20<MPG<50)**  =P(MPG50) – P(MPG0)  = P-P  =P(Z1.707) – P(Z-1.579)  =0.9564-0.0571  =0.8993  =89.93% |
|  | |  | |
|  | |
|  | |

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**

|  |  |
| --- | --- |
|  |  |
|  |  |

Since,

1)MEAN ≠ MEDIAN,

2)Skewness = 0.177

3)Kurtosis = 0.6116

4)IN Box plot Q2 is not at center, whisker is more negative side , Midian(Q2) is nearer to Q3 and in bell curve skewed towards negative numbers

**We can Say That the “MPG” data is Sightlly Right skewed or Negatively Skewed data.**

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

Dataset: wc-at.csv

**Waist:**

|  |  |
| --- | --- |
|  |  |
|  |  |

Since,

1)MEAN = MEDIAN =91.9018,

2)Skewness = 0.134 0

3)Kurtosis = -1.01

4)IN Box plot Q2 is approximately at center

**We can Say That the “Waist” data is Normally Distributed**

**AT**

|  |  |
| --- | --- |
|  |  |
|  |  |

Since,

1)MEAN MEDIAN

2)Skewness, is not nearer zero

3)Kurtosis = -2.855 is not nearer to zero

4)IN Box plot Q2 is not at center and whisker is More in Positive side

**We can Say That the “AT” data is Moderately Positively Skewed Data.**

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

|  |  |  |
| --- | --- | --- |
| **Confidence Interval** | **Alpha(α)**  **=(1-CL)/2** | **Z score**  **(Z table)** |
| 90% | 0.10/2=0.05 | ±1.64 |
| 94% | 0.06/2=0.03 | ±1.88 |
| 60% | 0.40/2=0.20 | ±0.84 |

**Using Python:**

|  |  |  |
| --- | --- | --- |
|  |  |  |

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

|  |  |  |
| --- | --- | --- |
| **Confidence Interval** | **Df** | **T score**  **(t table)** |
| 95% | 25 | 2.060 |
| 96% | 2.060 |
| 99% | 2.787 |

|  |  |  |
| --- | --- | --- |
|  |  |  |

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:

Claim: an average light bulb lasts 270 days

= = 2720

Number of Sample bulbs= n = 18

Average days of sample = Ẋ = 260 days

Standard Deviation of Sample = S =90 days

To find probability that 18 randomly selected bulbs would have an average life of no more than 260 days,

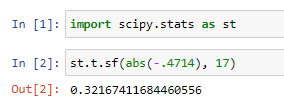
We need to calculate t statistics for given data,

t =

t =

t = -0.4714

Pt=-0.471, df=17

=

18 randomly selected bulbs would have an average life of no more than 260 days = 0.3216 = 32.16 %