|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continues data |
| Weight of Gold | Continues data |
| Distance between two places | Continues data |
| Length of a leaf | Continues data |
| Dog's weight | Continues data |
| Blue Color | Nominal data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Nominal 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 | Ratio |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Ratio |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| 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 | Interval |
| Religious Preference | Ordinal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Interval |

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

**Answer-**

Sample size = 23 = 8

Sample space – [TTT, THT, TTH, THH, HTH, HHT, THH, TTH]

* Probability of 2 heads and 1 tail is 3/8 i.e., 0.375 or 37.5%

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

**Answer-**

sample space-

(1,1)(1,2)(1,3)(1,4)(1,5)(1,6)

(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)

(3,1)(3,2)(3,3)(3,4)(3,5)(3,6)

(4,1)(4,2)(4,3)(4,4)(4,5)(4,6)

(5,1)(5,2)(5,3)(5,4)(5,5)(5,6)

(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)

1. Probability of sum equal to 1 is zero. Because rolling a dice is a case of continuous distribution and in continuous distribution probability of single value will always be zero.
2. Probability of sum less than or equal to 4 is 1/6 or 16.66 %

Sample space=[(1,1)(1,2)(1,3) (2,1)(2,2) (3,1)]

1. Probability of sum divisible by 2 and 3 is 5/36 or 13.88 %.

Sample space= [(1,5) (2,4) (3,3) (4,2) (5,1)]

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?

**Answer**-

Total number of balls – 2+3+2=7

Sample space- 7C2 = 21

Drawing balls other than blue = 2 red +3 green =5

Number of ways drawing 2 balls from set of 5 balls = 5C2 =10

# Probability that none of the balls drawn is blue = 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

**Answer-**

Expected number of candies (1\*0.015)+(4\*0.2)+(3\*0.65)+(5\*0.005)+(6\*0.001)+(2\*0.12)

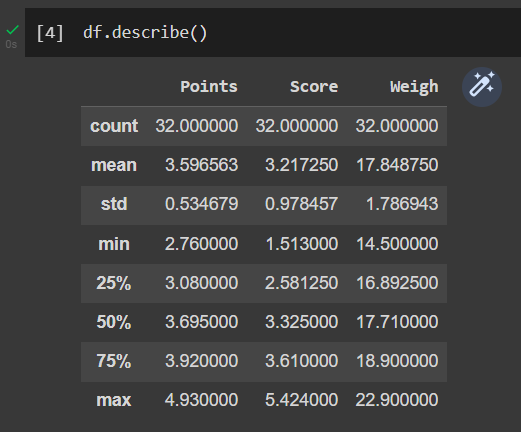
= 3.09

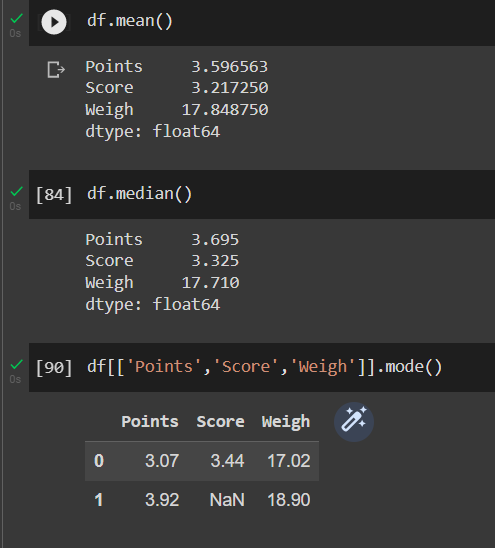
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

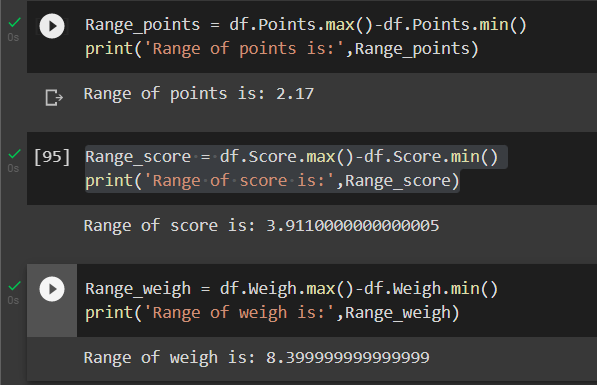
* For Points, Score, Weigh>

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

**Answer-**







Points:

* Mean =3.59
* Median=3.695
* Mode=3.92
* Variance=0.2809
* Standard Deviation=0.53
* Range=2.17

Score:

* Mean =3.217
* Median=3.325
* Mode=3.44
* Variance=0.956
* Standard Deviation=0.978
* Range=3.911

Weigh:

* Mean =17.84
* Median=17.71
* Mode=17.02
* Variance=3.189
* Standard Deviation=1.786
* Range=8.4

Interpretation-

* Mean<median<mode in case of Point and Score and hence these two are Negatively skewed.
* Mean>Median>mode in case of Weigh and hence it is positively skewed.

**Use Q7.csv file**

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?

**Answer-**

Total number of patients = 9

Probability of choosing one patient = 1/9

Expected value of weight = 108/9+110/9+123/9+134/9+135/9+145/9+167/9+187/9+199/9

145.3

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

**Car’s speed and distance**

**Use Q9\_a.csv**

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Answer-**

**Note-** Python Notebook for this question is attached with the mail.

1. SKEWNESS:

If the skewness is between -0.5 to +0.5 then the data is nearly symmetrical.

If the data is between 0.5 to 1 (positively skewed) or between -1 to -0.5 (Negatively skewed) then data is slightly skewed.

Hence from above statements, it is clear that speed is nearly symmetrical and distance is positively skewed

KURTOSIS:

Excess Kurtosis = Kurtosis-3

Kurtosis for speed=-0.5-3=-3.5

Kurtosis for distance=0.4-3=-2.6

Kurtosis value for speed is more than 3, it is called as leptokurtic (short tailed) with low standard deviation and more data concentrated near the mean.

Kurtosis value for distance is less than 3, it is called as Platykurtic (long tailed) with high standard deviation and shows presence of outliers.

1. SKEWNESS:

Skewness value of SP is 1.611. If skewness is more than 1 it is called as highly skewed. SP distribution is positively skewed.

Skewness value of weight is -0.61, hence it is Negatively skewed.

KURTOSIS:

Kurtosis for SP = 2.97-3 = 0.03

Kurtosis for weight = 0.95-3=2.05

Kurtosis of SP is near to zero. It is called as Mesokurtic. Curve is similar to normal distribution.

Weight data has High standard Deviation and outliers are present.

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



**Answer-**

HISTOGRAM:

* Histogram depicts maximum weight is occurring between 50 to 100 and tail is on right side so it shows plot is positively skewed.
* In above histogram mean >median>mode.

BOXPLOT:

* Outliers are present at upper extreme side.
* Very less data points are present below 25%.
* Median is slightly shifted below 50% which indicates data is not symmetrical.

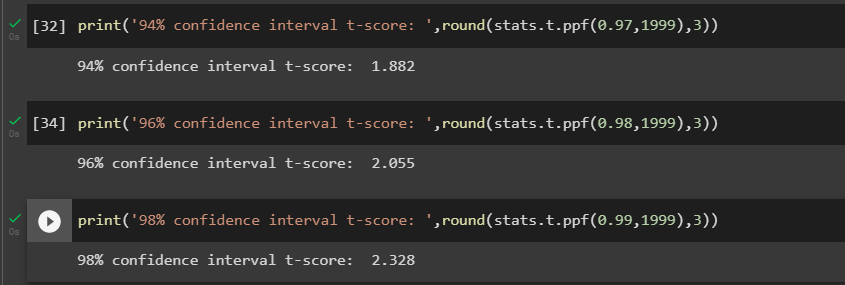
**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate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

**Answer-**

Given data:

* Sample size - 2000
* Population size – 3000000
* Average weight of sample – 200 ponds
* Standard Deviation of sample – 30 pounds

Calculating t-score,



Calculating confidence interval using formula,



After putting values in above formula,

Confidence interval of 94% is [201.262,198.738]

Confidence interval of 98% is [201.562,198.438]

Confidence interval of 96% is [201.378,198.622]

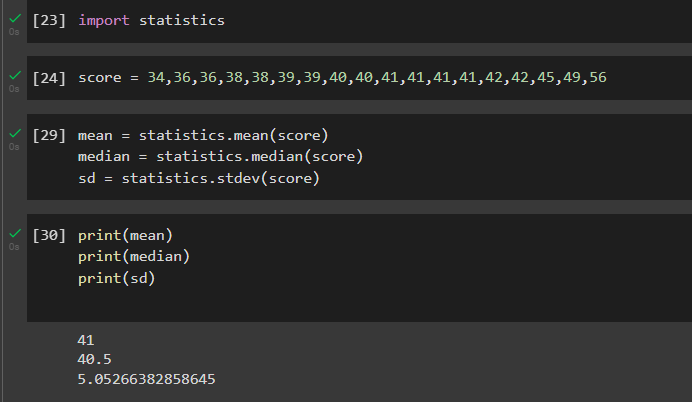
**Q12)** Below are the scores obtained by a student in tests

**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?

**Answer-**

1)



2)Interpretation –

* Distribution of marks is positively skewed.
* The marks are spread between 34 and 56 with a standard deviation of 5.05 which indicates that variability is more.

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

**Answer –** Skewness will be zero when mean and median are equal. Because if mean=median=mode then dataset follows normal distribution and skewness for normal distribution is zero.

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

**Answer-** If mean is greater than median it indicates Tail is on the right side of the distribution and it is positively skewed.

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

**Answer –** If Median is greater than mean it indicates tail is on left side of the distribution and it is negatively skewed.

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

**Answer –**Positive kurtosis indicates distribution contains peak value and has thick tails.

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

**Answer –** Negative kurtosis indicates distribution is flat and has thin tails.

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



What can we say about the distribution of the data?

**Answer –** data is heavily distributed towards right side and median is between 14 and 16.

What is nature of skewness of the data?

**Answer –** Median is greater than mean. It indicates given boxplot is negatively skewed.

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

**Answer –** 18-10= 8

Q19) Comment on the below Boxplot visualizations?



**Answer –**

* Interquartile range is between 250 and 280 (approximately) .
* Distribution is symmetric and dataset is following normal distribution.
* Extreme quartile is not so far from median.



* Interquartile range is between 225 and 315 (approximately).
* Distribution is symmetric and dataset is following normal distribution.
* Whiskers are having longer Length; it indicates extreme values are far from median.

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

Interpretation:

* Tail of first dataset is shorter than second.
* More data is stored in second dataset than first.

Q 20) Calculate probability from the given dataset for the below cases

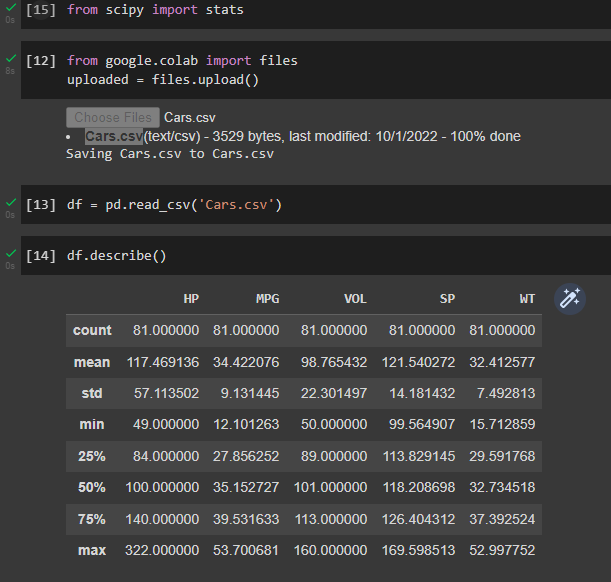
Data \_set: Cars.csv

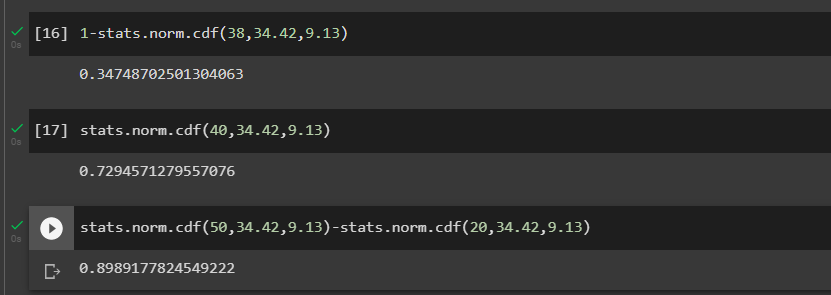
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

**Answer-**





Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

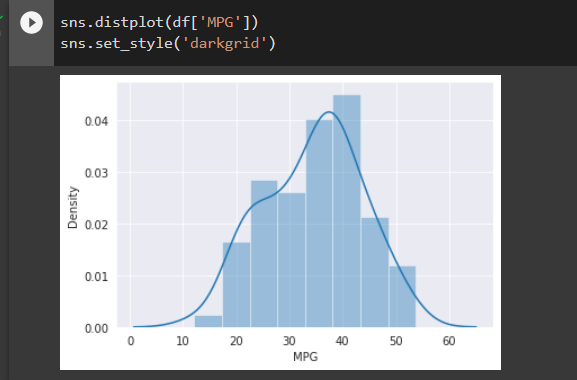
Dataset: Cars.csv

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

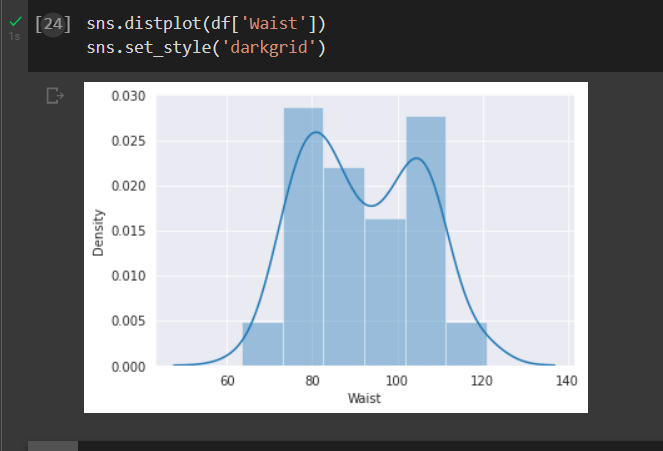
Dataset: wc-at.csv

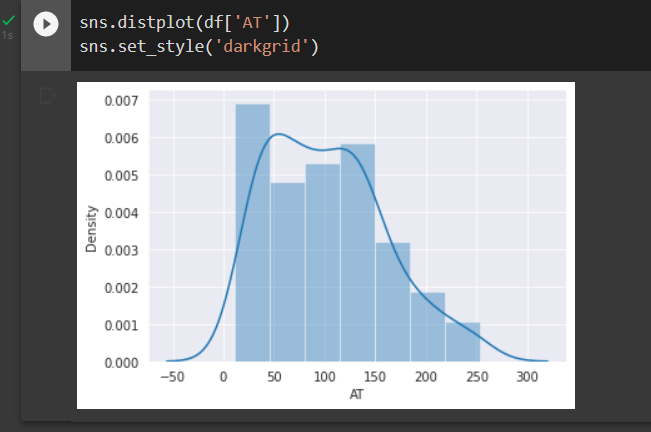
**Answer-**

a)



b)





A curve is said to be normally distributed, when data is symmetrically distributed with no skews.

From distribution plots, we can conclude that MPG, WAIST and AT plots are not normally distributed.

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

**Answer -**

90% confidence interval in decimal is 0.90

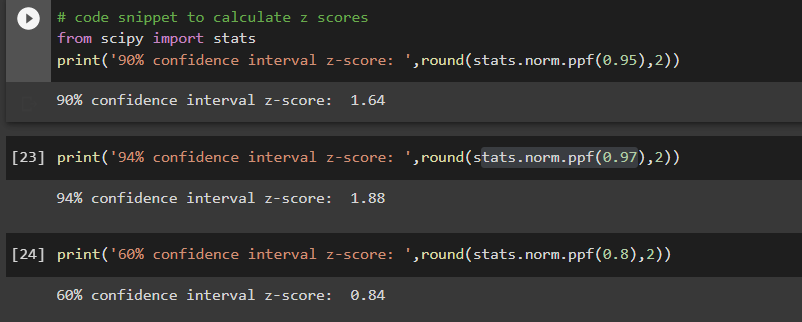
Calculating alpha = (1+0.90)/2= 0.95

94% confidence interval in decimal is 0.94

Alpha =(1+0.94)/2=0.97

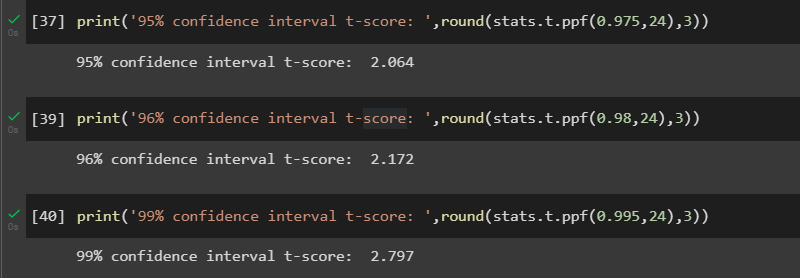
60% confidence interval in decimal is 0.60

Alpha = (1+0.60)/2 = 0.8



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

**Answer-**



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

**Answer-**

Population mean=270

Sample Mean = 260

Sample size = 18

Standard Deviation of Sample = 90

t-score = =- 0.471

