|  |  |
| --- | --- |
| Activity | Data Type (Continuous/district) |
| Number of beatings from Wife | District |
| Results of rolling a dice | District |
| 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 | District |
| Number of kids | District |
| Number of tickets in Indian railways | District |
| Number of times married | District |
| Gender (Male or Female) | District |

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) | Ordinal |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Ordinal |
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Ratio |

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

ANSR .-P={HHH,HHT,HTH,THH,HTT,THT,TTH,TTT}

P(X=2H AND 1T)=3/8=0.375.

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

ANSR .- no of outcomes=6^2=36 nos.

(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. Sum is equal to 1= no. of outcomes=0, i.e=P=0/36=0
2. Sum is less than or equal to 4:

No of outcomes={(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}=6

i.e. P=6/36 = 0.17.

c) Sum is divisible by 2 and 3:

No of outcomes={(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}=6

i.e. P=6/36 = 0.17.

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?

ANSR. Total no. of ball=7

Number of ways of drawing 2 balls out of 7 is:

=7C2

=(7X6)/(2X1)

=21

No. of ways of drawing 2 balls, none of ball is blue:

i.e- no. of ways of drawing 2 balls out of (2+3) balls.

=5C2

=(5X4)/(2X1)

=20/2

=10

**i.e- probability is 10/21=0.476**

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

ANSR. Expected number of candies for a randomly selected child:

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

=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**

**ANSR Please go through the Jupiter notebook attached.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | **3.60** | **3.22** | **17.85** |
| **Median** | **3.70** | **3.32** | **17.71** |
| **Mode** | **3.07 and 3.92** | **3.44** | **17.02 and 18.90** |
| **Variance** | **0.285** | **0.957** | **3.19** |
| **Standard Deviation** | **0.53** | **0.98** | **1.79** |
| **Range** | **4.93-2.76** | **5.424-1.513** | **22.9-14.5** |
|  |  |  |  |

Inferences: Mean Value is almost equal to Median Value, so no skewness. Also no ouliers.

Q8) Calculate Expected Value(Mean 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?

ANSR = Mean Value=(108+110+123+134+135+145+167+187+199)/9

=145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANSR: Please go through the Jupiter notebook attached.**

**ANSR.Skewness:**

speed -0.117510; negatively moderately skewed.

dist 0.806895; Positively highly skewed. As value is greater than 0.5. Outlier Present.

**Kurtosis:**

speed -0.508994, Negative Kurtosis,indicates a shape flatter than normal

dist 0.405053,

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANSR.Skewness:**

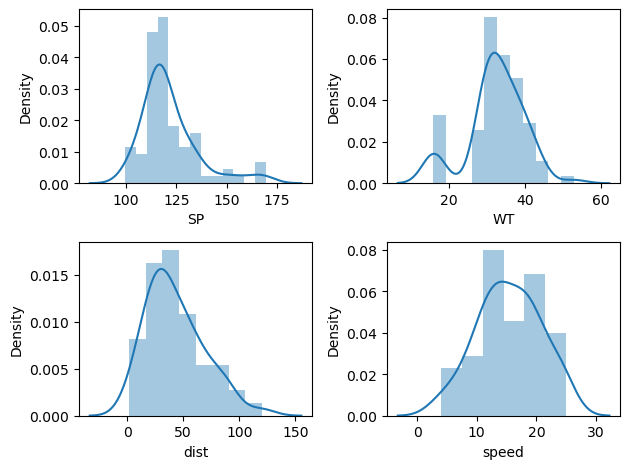
SP= 1.611450 ; Highly positive skewed.

WT= -0.614753 ;Value between -1 to -0.5. i.e data is moderately symentric

**Kurtosis:**

SP =2.977, Value is positive this indicates that the distribution is more peaked than normal

dist 0.950, Negative Kurtosis,indicates a shape flatter than normal

****

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



Q.10 Ans. Histogram is not symmetric. And is Positively Skewed. If we draw the tips of Histogram, then we get the density plot. Which shows that long tail at right side, so consequently it’s a positively skewed.

Boxplot:-Outliers are found at the up side of the upper Extreme.

**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?

ANSR

Number of sample=n=2000

Mean of sample, x\_bar=200

Std deviation of sample, s=30

Formula=sample mean value+-(z value (1-alpha)\*(std deviation of sample/root of no of sample)

94% confidence interval:

Lower value=200 –(1.89\*(30/root of 2000))

=200-(1.89\*(30/44.7214))

=198.73

Upper Value=200 +(1.89\*(30/root of 2000))

201.27

94% confidence interval: 198.73-201.27

98% confidence interval:

Lower value=200 –(2.33\*(30/root of 2000))

=200-(2.33\*(30/44.7214))

=198.43

Upper Value=200 +(2.33\*(30/root of 2000))

201.56

98% confidence interval: 198.43-201.56

96% confidence interval:

Lower value=200 –(2.06\*(30/root of 2000))

=200-(2.06\*(30/44.7214))

=198.62

Upper Value=200 +(2.33\*(30/root of 2000))

201.38

96% confidence interval: 198.62-201.38

**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?

**ANSR : Please go through the Jupiter notebook attached.**

Ans.-count=18.000000

mean=41.000000

std=5.052664

min=34.000000

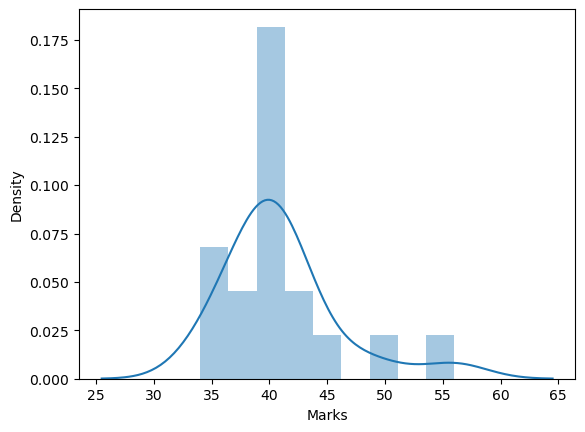
25%=38.250000

50%(Median)=40.500000

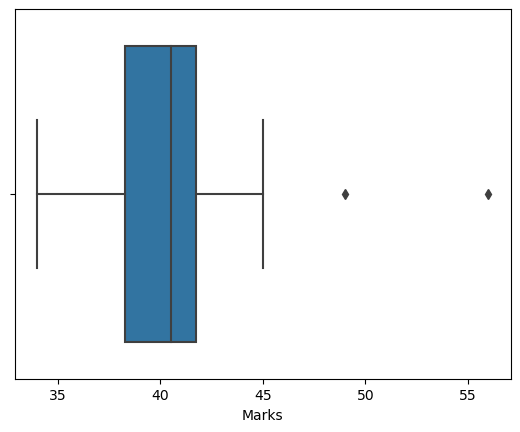
75%=41.750000

Max=56.000000

Var=25.5294



Data is positively skewed. Outliers are presents.



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

ANSR. No Skewness.

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

ANSR. Positive Skewness.

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

ANSR. Negative Skewness.

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

ANSR. Positive kurtosis indicated the distribution is peaked. And possesses the thick tail.

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

ANSR. Positive kurtosis indicated the distribution has lighter tails than the normal distribution

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



What can we say about the distribution of the data?

ANSR. Box plot is not symmetric, As median line is not equidistance from min and max line.

What is nature of skewness of the data?

ANSR. Distance of median to min is greater than the distance between median to max. so Nature of skewness is negative.

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

A. ISRQR=Q3-Q1

=18-10

IQR=8.

Q19) Comment on the below Boxplot visualizations?



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

Q19 ANSR. Median Value is same for boxplot 1 and boxplot 2. Both box are symentric.

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)

**ANSR : Please go through the Jupiter notebook attached.**

Ans. A) 0.34759015198895815

B) 0.7293536121889511

c) 0.8988704032941203

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANSR : Please go through the Jupiter notebook attached.**

Ans Q21)-a). As per density plot, it is not symentic.

Also mean value is not equal to median.

mean 34.422076

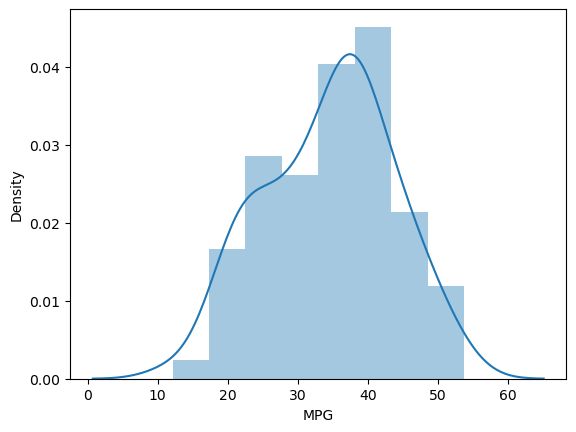
std 9.131445

median 35.152727

Skewness -0.17794674747025727

Kurtosis -0.6116786559430913

The MPG of Cars does not follows 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

ANSR Q21)-a).

As per density plot, it is not symentic.

Also mean value is not equal to median.

Kde plot is not bell shape.

|  |  |  |
| --- | --- | --- |
|  | Waist | AT |
| **count** | 109.000000 | 109.000000 |
| **mean** | 91.901835 | 101.894037 |
| **std** | 13.559116 | 57.294763 |
| **min** | 63.500000 | 11.440000 |
| **25%** | 80.000000 | 50.880000 |
| **50%** | 90.800000 | 96.540000 |
| **75%** | 104.000000 | 137.000000 |
| **max** | 121.000000 | 253.000000 |

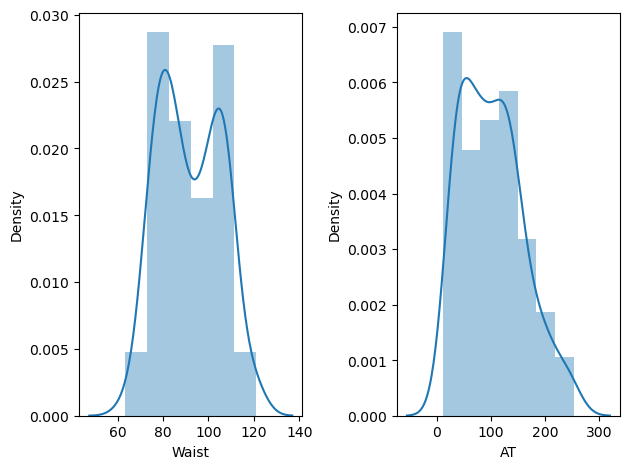
Skewness Waist 0.134056

AT 0.584869

Kurtosis Waist -1.102667

AT -0.285576

The Waist and AT in datasets does not follows Normal Distribution.



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

Z scores of 90% confidence interval:

Alfa=0.05

i.e=0.90+0.05=0.95

stats.norm.ppf(0.95)= 1.6448536269514722

OR From Z table-1.65

Z scores of 94% confidence interval:

Alfa=0.03

i.e=0.94+0.03=0.97

stats.norm.ppf(0.97)= 1.8807936081512509

OR From Z table-1.89

Z scores of 60% confidence interval:

Alfa=0.20

i.e=0.60+0.20=0.80

stats.norm.ppf(0.80)= 0.8416212335729143

OR From Z table-0.85

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

ANSR.:

#t scores of 95% confidence interval

stats.t.ppf(0.975,24)= 2.0638985616280205

#t scores of 96% confidence interval

stats.t.ppf(0.98,24)= 2.1715446760080677

#t scores of 99% confidence interval

stats.t.ppf(0.995,24)= 2.796939504772804

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

**ANSR : Please go through the Jupiter notebook attached.**

#P\_value(0.32167411684460556) is greater than the 0.05 so we fail to reject the null hypothesis.

# i.e Average life of bulb is greater than 260 days.