|  |  |
| --- | --- |
| 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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Nominal |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

Sol) Prob of getting a head = P(H) = 1/2

Prob of getting a tail = P(T) = 1/2

Let ‘E’ be the event for which 3 coins are tossed and prob of getting 2 heads and 1 tail be P(E).

P(E) = P(H) P(H) P(T)+ P(H) P(T) P(H)+ P(T) P(H) P(H)

= + +

= 3\*[(1/2)3]

=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

Sol) a) Sum is equal to 1

Two dice are rolled, Total possible outcomes = 62 = 36

The numbering on dice starts from 1 to 6, when two dice are rolled there are zero chances that sum of numbers is equal to 1. Let E be the event for sum is equal to 1.

P(E) = 0/36 = 0

b) Less than or equal to 4

Possible chances of getting the sum of less than or equal to 4 are {(1,1),(1,2),(2,1),(1,3),(3,1),(2,2)} i.e., = 6.

If E1 be the event for sum less than or equal to 4, then P(E1) is

P(E1) = 6/36 = 0.16667

c) Sum is divisible by 2 and 3

For divisibility by 2, the sum should be an even number. Possible chances that sum of 2 number when two dice are rolled are from 2 to 12, out of which only even are under our interest. Let E2 be the event

Sum is 2 = {(1,1),(1,1)} = 2

Sum is 4 = {(1,3),(3,1),(2,2)} = 3

Sum is 6 = {(1,5),(5,1),(2,4),(4,2),(3,3)} = 5

Sum is 8 = {(2,6),(6,2),(3,5),(5,3),(4,4)} = 5

Sum is 10 = {(4,6),(6,4),(5,5)} = 3

Sum is 12 = {(6,6)} = 1

Total possible chances for divisibility by 2 are 2+3+5+5+3+1 =19

P(E2) = 19/36 = 0.52778

For divisibility by 3, if sum of the digits is divisible by 3, then that number is divisible by 3. Possible chances that sum of 2 number when two dice are rolled are from 2 to 12, out of which only if sum of digits i.e., (3,6,9,12) are divisible by 3. Let E3 be the event

Sum is 3 = {(1,2),(2,1)} = 2

Sum is 6 = {(1,5),(5,1),(2,4),(4,2),(3,3)} = 5

Sum is 9 = {(3,6),(6,3),(4,5),(5,4)} = 4

Sum is 12 = {(6,6)} = 1

Total possible chances for divisibility by 3 are 2+5+4+1 = 12

P(E3) = 12/36 = 0.3333

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?

Sol) Total number of balls in the bag = 7

Out of 7 balls, 2 balls are drawn and prob that none of the ball is blue is

P(E) = Either 2 red are picked (or) Either 1 red and 1 green are picked (or) Either 2 green are picked out of all the 7 balls

P(E) =

P(E) =

P(E) = 10/21 = 0.4761

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

Sol) The expected number of candies is E(x) = Σ(x\*p(x))

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

E(x) = 237/100

E(x) = 2.37

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.

Sol) Writing the code in Spyder by using Q7.csv file

***Inferences drawn:***

-The given dataset lists out the different cars info about the points, score and weigh.

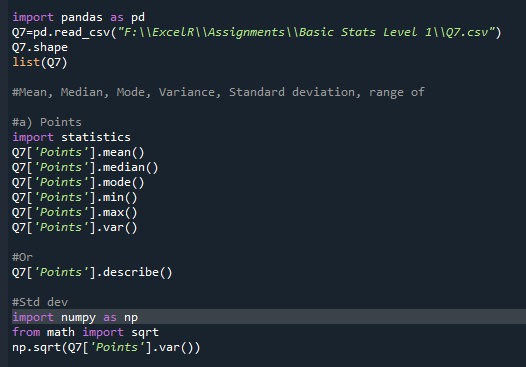
-Out of the 3 variables, Points show the least standard deviation from the mean.

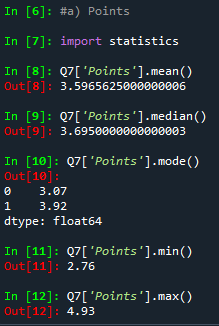
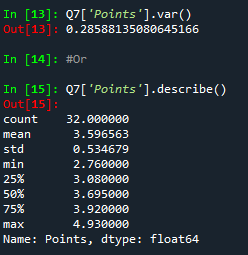
-The highest weigh among the given cars is about 22.9 units (Tons).

-The range of the scores varies from the min of 1.513 to the max of 5.424.

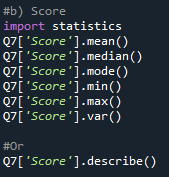
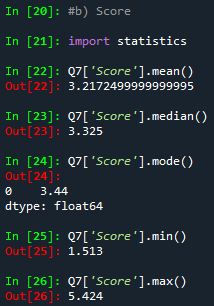
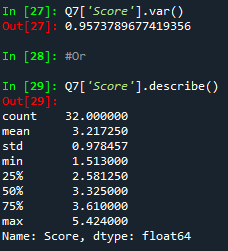
-All the 3 variables are of floating datatype

**a)Points:**

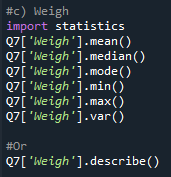
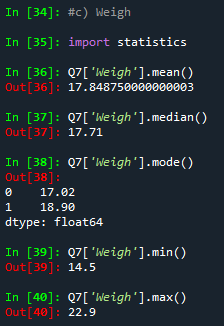
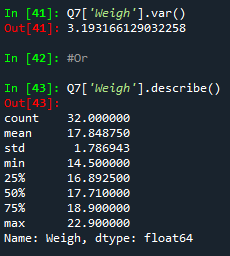


**b) Score:**

**c)Weigh:**

Q8) Calculate Expected Value for the problem below

The weights (X) of patients at a clinic (in pounds), are 108, 110, 123, 134, 135, 145, 167, 187, and 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

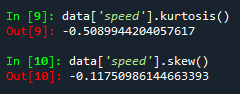
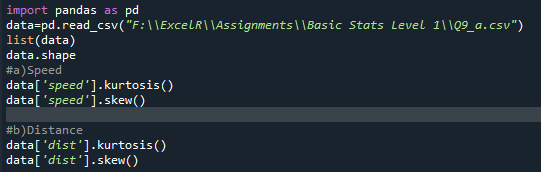
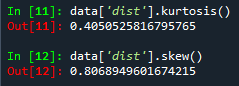
Sol) The expected value of the weight of the patient is nothing but mean(X).

Mean(X) = = 145.33 pounds

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

Cars speed and distance (Use Q9\_a.csv)

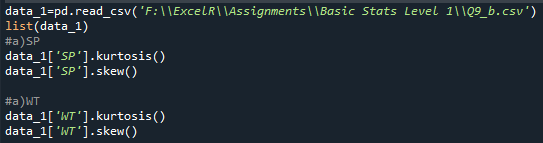
Sol) By writing the code in python and checking the skewness and kurtosis

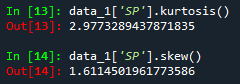
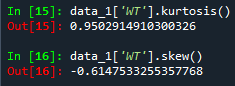
 

Inference: From the result we can see that variable ‘Speed’ it is negatively skewed data i.e.,(Mean<Median<Mode) and also have Negative kurtosis i.e.,(Platy Kurtosis)

From the result we can see that variable ‘Dist’ it is positively skewed data i.e.,(Mean>Median>Mode) and also have Positive kurtosis i.e.,(Lefto Kurtosis)

SP and Weight(WT),Using Q9\_b.csv and checking its skewness and kurtosis



** **

Inference: From the result we can see that variable ‘SP’ it is positively skewed data i.e.,( Mean>Median>Mode) and also have Positive kurtosis i.e.,(Lefto Kurtosis)

From the result we can see that variable ‘WT’ it is negatively skewed data i.e.,(Mean<Median<Mode) and also have Positive kurtosis i.e.,(Lefto Kurtosis)

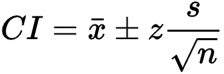
**Q10) Draw inferences about the following boxplot & histogram**



Inference for the histogram, the data speaks about the chick weights which looks to be positively skewed in which Mean>Median>Mode. Most of the chick weights fall under the category of 50-100.

Inference for the boxplot, there are outliers present in the graph. The lower line of the box plot represents Q1(Quartile 1-25 percentile), the middle line represents Median(50 percentile) and the last above line represents Q3(Quartile 3-75 percentile).

**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?



Here the parameters are as follows:

Population = N = 3,000,000

Sample size = n = 2000

Confidence level value = Z

Mean = x =200

Standard deviation = s = 30

We can calculate confidence intervals with the formula = x̅ ± Z(σ / √n)

1. For 94% confidence interval: 200 ± 1.881(30/√2000) = 200 ± 1.2681
2. For 98% confidence interval: 200 ± 2.326(30/√2000) = 200 ± 1.5603
3. For 96% confidence interval: 200 ± 2.05(30/√2000) = 200 ± 1.3751

**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?

Sol) Mean(X)=

Mean(X)= 41 marks

Median 🡺 Arrange in ascending order

34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

As no. of observations are even, median is avg. of middle most two values

Median = = 40.5 marks

Mode 🡺 Most repeated observation is 41 marks, so mode = 41 marks

**Inference:** we can observe that the Mean = Mode >Median

The marks obtained by students can be positively skewed in nature

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

Sol: When Mean=Median=Mode, It is symmetrical in nature.

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

Sol: When Mean>Median, It is positively skewed in nature.

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

Sol: When Median > Mean, It is negatively skewed in nature.

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

Sol: Positive kurtosis means peak of the normal distribution above it. The spread of the normal distribution is very less in positive kurtosis.

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

Sol: Negative kurtosis means peak of the normal distribution below it. The spread of the normal distribution is more in negative kurtosis.

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



What can we say about the distribution of the data?

Sol: The box plot looks to be concentrated more on right of the scale.

There are no outliers present in the plot.

What is nature of skewness of the data?

Sol: As per the box plot, the Mean<Median, it looks to be negatively skewed.

What will be the IQR of the data (approximately)?   
Sol: Quartile1=Q1=25 percentile=10

Quartile2=Q2=50 percentile (Median)=15

Quartile3=Q3=75 percentile=18

Inter Quartile Range(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.

Sol) (1) box-plot is more concentrated and while **(2)** box-plot has a more distributed data. When we talk about range of both box-plots **(1)** has the range of 237.5 - 287.5 = 50 and **(2)**  has the range of 200 - 350 = 150 and hence **(2)** has more range. The kurtosis value of **(1)** will be greater than **(2)** since **(1)** is more concentrated.

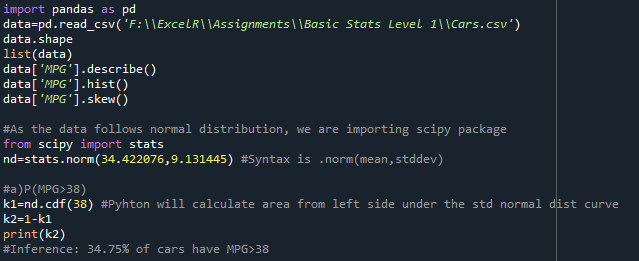
Q 20) Calculate probability from the given dataset for the below cases

(Data \_set: Cars.csv)

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

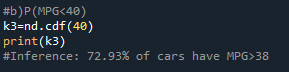
MPG <- Cars$MPG

* 1. P(MPG>38)

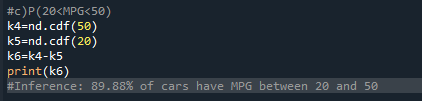
Inference: 34.75% of cars have MPG>38

* 1. P(MPG<40)

Inference: 34.75% of cars have MPG<40

* 1. P (20<MPG<50)

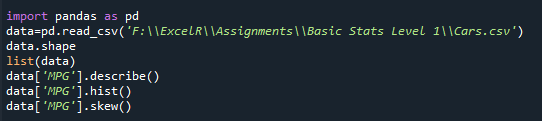
 

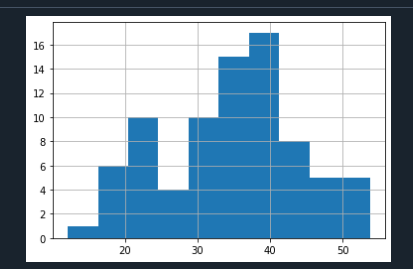
Inference: 34.75% of cars have MPG between 20 and 50.

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

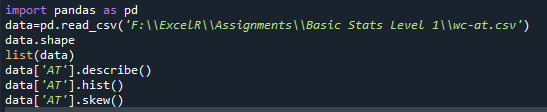


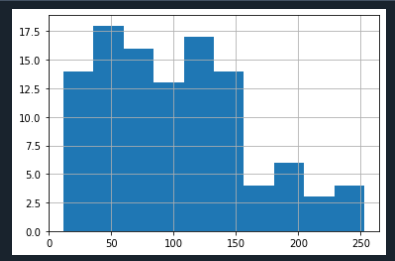
 

Inference: The data looks to be normally distributed and skewness is also under the acceptable range of -0.5 to +0.5

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

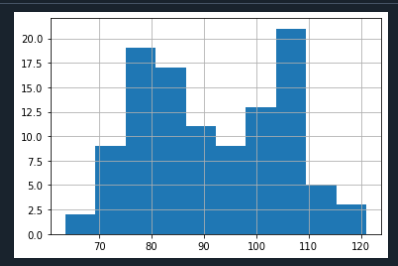
Dataset: wc-at.csv



Inference: The data looks to be positively skewed and skewness is slightly deviating from the acceptable range of -0.5 to +0.5



Inference: The data looks to be normally distributed and skewness is also under the acceptable range of -0.5 to +0.5

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

|  |  |
| --- | --- |
| Confidence Interval | Z-Scores |
| 90% | 1.645 |
| 94% | 1.881 |
| 60% |  |

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

|  |  |
| --- | --- |
| Confidence Interval | t-Scores |
| 95% | 2.262 |
| 96% | 2.054 |
| 99% |  |

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

Sol) P(X<=260) 🡺 converting into standard normal variate

= 260,

P(X<=260) : P() = P(

P(Z) = P(0) = 0.5