|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete(Countable datatype) |
| Results of rolling a dice | Discrete(Countable datatype) |
| Weight of a person | Continuous datatype |
| Weight of Gold | Continuous datatype |
| Distance between two places | Continuous datatype |
| Length of a leaf | Continuous datatype |
| Dog's weight | Continuous datatype |
| Blue Color | Classificaion(Nominal)data type |
| Number of kids | Discrete(Countable datatype) |
| Number of tickets in Indian railways | Discrete(Countable datatype) |
| Number of times married | Discrete(Countable datatype) |
| Gender (Male or Female) | Classificaion(Nominal)data type |

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 | Nominal |
| Time on a Clock with Hands | Ordinal |
| Number of Children | Ratio |
| 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?

Ans: Total number of possible outcomes when three coins are tossed-2^3.Since we need 2 heads and 1 tail out of 3 coins. The outcomes comes as follows:

HHH,HHT,HTH,THH,TTT,TTH,THT,HTT=8

n(a)=event occurring two heads and one tail i.e., HHT,HTH,THH=3

n(s)=total number of possible outcomes i.e.,8

Probability=n(a)/n(s)=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

Ans: When two dice are rolled, the total number of possible outcomes are 6\*6=36 possibilities.They are:-

(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)

a)Equal to 1:

The probability of occurring the sum is equal to 1 is 0.As there is no possibility of occurring the sum equal to 1 when two dice are rolled. It is possible as the probability always lies between 0<=P(A)<=1.

b)sum is less than or equal to 4:

Ans: Total number of possibilities=n(A)={(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}=6

Total number of possible outcomes=n(S)=36

Total Probabilty = n(A)/n(S)=6/36=1/6.

c)Sum is divisible by 2 and 3:

Ans: The numbers which are divisible by 2 and 3 are 6 and 12

While **12** can only be made **1 way** (double 6) **6** can be made **5 ways** {(1,5) (2,4) (3,3) (4,2) (5,1) }.

N(a)=The possibilities of getting a sum of 12 or 6=6

N(s)=Total number of possible outcomes=36

Total Probability=N(a)/N(s)=6/36=1/6.

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?

Ans: Total number of balls=>2+3+2=7

Number of ways of drawing two balls out of 7 =7C2=(7\*6)/(2\*1)=42/2=21

We have to select the two days that none of the balls drawn is blue so we need to consider only red and green balls. There are 2 red balls and 3 green balls with the total of 5 balls.

Therefore, the possibilities are:-5C2=(5\*4)/(2\*1)=20/2=10

The required probability=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)

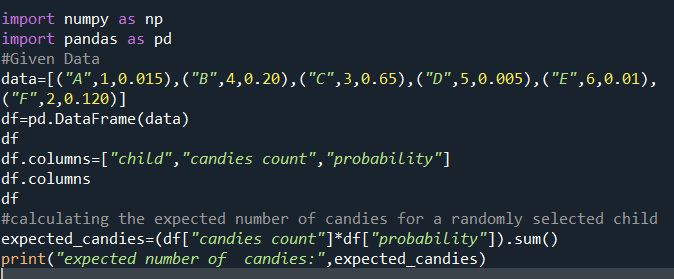
|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

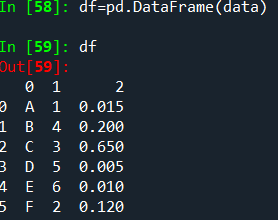
Child A – probability of having 1 candy = 0.015.

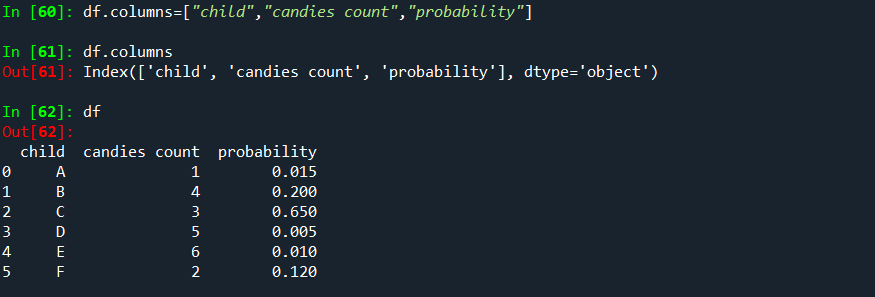
Child B – probability of having 4 candies = 0.20

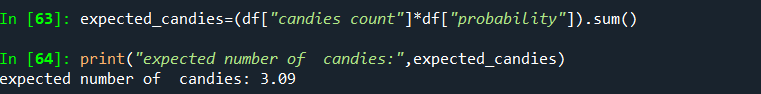
Ans: Expected Value= ∑(Value× Probability)

Here, values defines the number of candies and Probability defines the having number of candies by each individual. This can be achieved by pandas. First we need to numpy data using numpy library and change it to pandas data and assign the column names to each column and apply the formula to pandas data.The code for this is :-









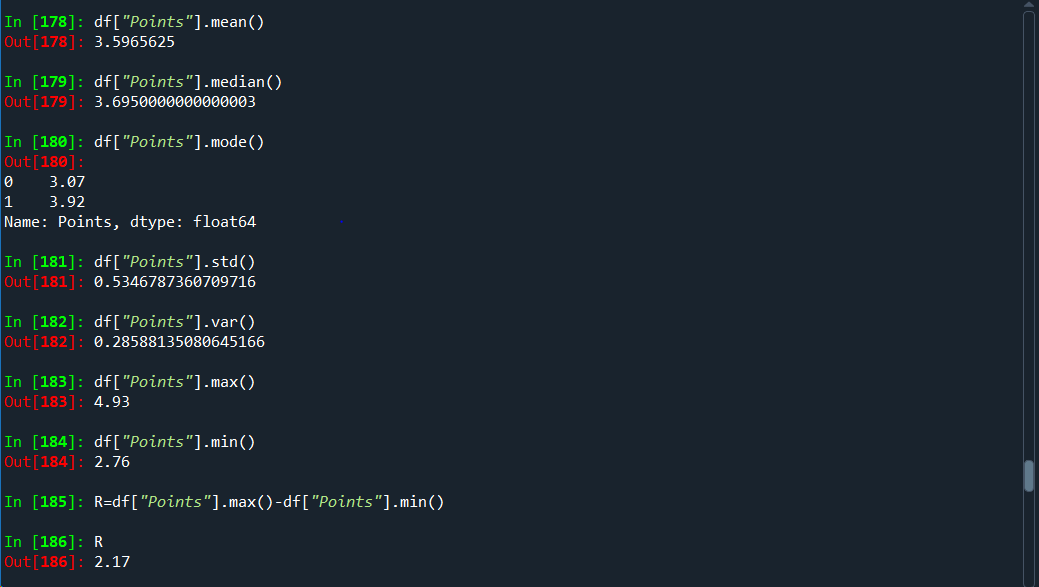
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

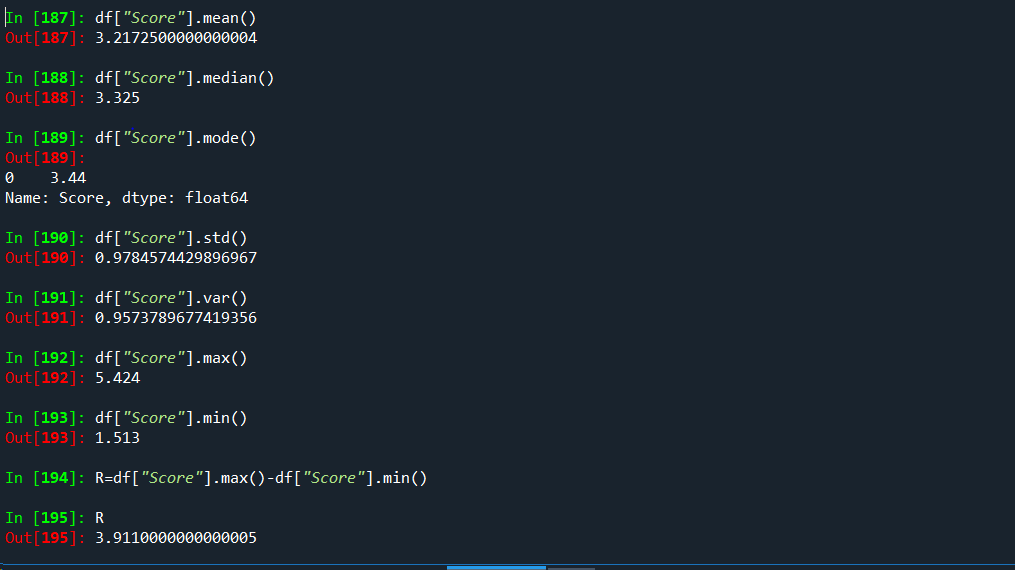
* For Points,Score,Weigh>

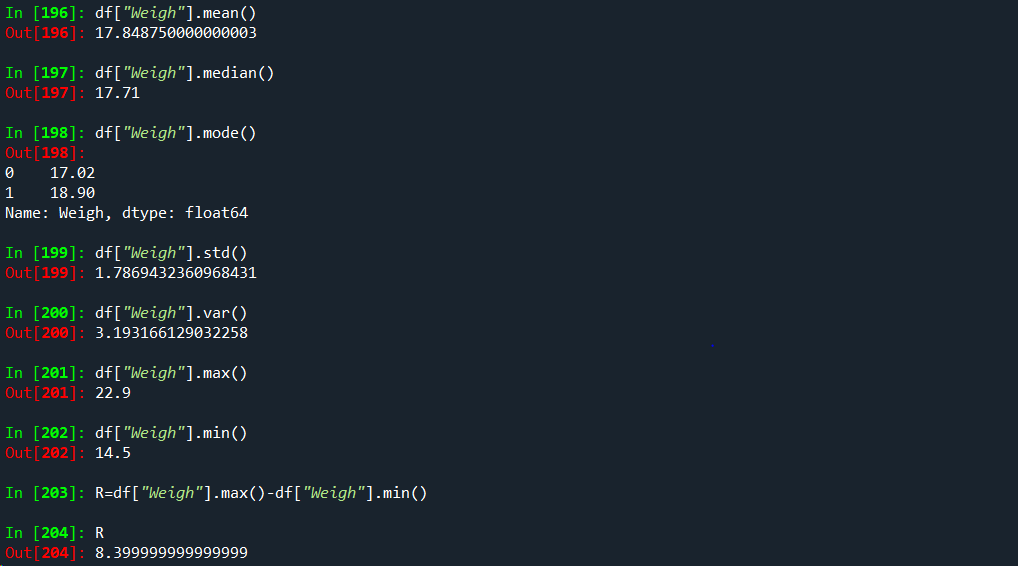
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Points** | **Score** | **Weigh** |
| Mazda RX4 | 3.9 | 2.62 | 16.46 |
| Mazda RX4 Wag | 3.9 | 2.875 | 17.02 |
| Datsun 710 | 3.85 | 2.32 | 18.61 |
| Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| Hornet Sportabout | 3.15 | 3.44 | 17.02 |
| Valiant | 2.76 | 3.46 | 20.22 |
| Duster 360 | 3.21 | 3.57 | 15.84 |
| Merc 240D | 3.69 | 3.19 | 20 |
| Merc 230 | 3.92 | 3.15 | 22.9 |
| Merc 280 | 3.92 | 3.44 | 18.3 |
| Merc 280C | 3.92 | 3.44 | 18.9 |
| Merc 450SE | 3.07 | 4.07 | 17.4 |
| Merc 450SL | 3.07 | 3.73 | 17.6 |
| Merc 450SLC | 3.07 | 3.78 | 18 |
| Cadillac Fleetwood | 2.93 | 5.25 | 17.98 |
| Lincoln Continental | 3 | 5.424 | 17.82 |
| Chrysler Imperial | 3.23 | 5.345 | 17.42 |
| Fiat 128 | 4.08 | 2.2 | 19.47 |
| Honda Civic | 4.93 | 1.615 | 18.52 |
| Toyota Corolla | 4.22 | 1.835 | 19.9 |
| Toyota Corona | 3.7 | 2.465 | 20.01 |
| Dodge Challenger | 2.76 | 3.52 | 16.87 |
| AMC Javelin | 3.15 | 3.435 | 17.3 |
| Camaro Z28 | 3.73 | 3.84 | 15.41 |
| Pontiac Firebird | 3.08 | 3.845 | 17.05 |
| Fiat X1-9 | 4.08 | 1.935 | 18.9 |
| Porsche 914-2 | 4.43 | 2.14 | 16.7 |
| Lotus Europa | 3.77 | 1.513 | 16.9 |
| Ford Pantera L | 4.22 | 3.17 | 14.5 |
| Ferrari Dino | 3.62 | 2.77 | 15.5 |
| Maserati Bora | 3.54 | 3.57 | 14.6 |
| Volvo 142E | 4.11 | 2.78 | 18.6 |
| **Mean** | 3.596 | 3.217 | 17.84875 |
| **Median** | 3.695 | 3.325 | 17.71 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Standard Deviation** | 0.534679 | 0.978457 | 1.786943 |
| **Variance** | 0.285881 | 0.957379 | 3.193166 |
| **Range** | 2.17 | 3.911 | 8.4 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |







**Inference:**

**1)**The average value of the “Points” column is approximately 3.596.

“Score” column is approximately 3.217.

“Weigh” column is approximately 17.848.

It says that the average point value falls around this number.

**2)** The standard deviation value of the “Points” column is approximately 0.534.

“Score” column & “Weigh” column is approximately 0.978 & 1.786

**Standard Deviation** is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a “typical” deviation from the mean.

3. The Variance value of the “Points” column is approximately 0.2858.

“Score” column is approximately 0.9573.

“Weigh” column is approximately 3.1931.

The term variance refers to a statistical measurement of the spread between numbers in a data set. It quantifies how much the values in a dataset differ from the mean value.

4)The Range value of the “Points” column is approximately 2.17.

“Score” column is approximately 3.911.

“Weigh” column is approximately 8.4.

The range refers to the difference between the maximum and minimum values in a dataset.

Mathematically, the range of a dataset is calculated as:

Range=Maximum value−Minimum value

5.The Median value of the “Points” column is approximately 3.695.

“Score” column is approximately 3.325.

“Weigh” column is approximately 17.71.

The median is a measure of central tendency that represents the middle value of a dataset when it is arranged in ascending or descending order. It divides the dataset into two equal halves, where half of the values are below the median, and the other half are above it.

6. The Mode value of the “Points” column is approximately 3.92.

“Score” column is approximately 3.44.

“Weigh” column is approximately 17.02.

The mode refers to the value or values that appear most frequently in a dataset. It represents the data point(s) with the highest frequency or occurrence within the dataset.

Q8) Calculate Expected Value for the problem below

a)The weights (X) of patients at a clinic (in pounds), are

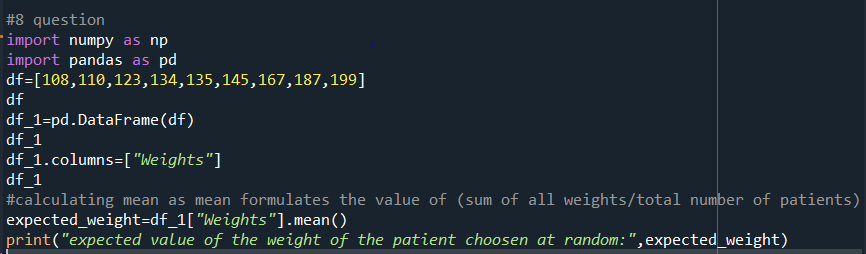
108, 110, 123, 134, 135, 145, 167, 187, 199

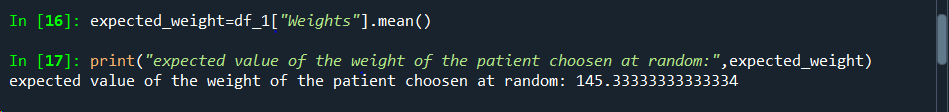
Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Ans: To calculate the expected value of the patients we use the formula:-

Expected value=Sum of weights/number of patients

Total number of patients=9

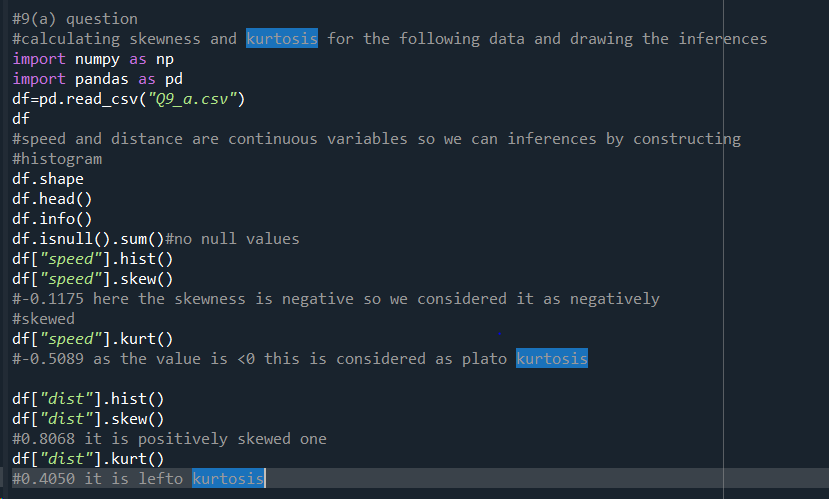


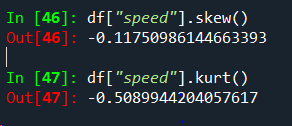
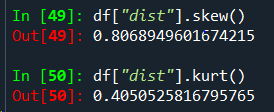


**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

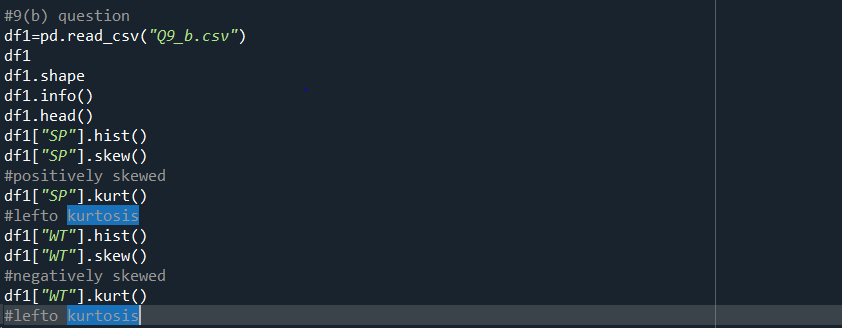
**Use Q9\_a.csv**

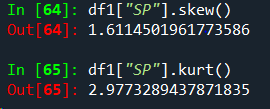


**SP and Weight(WT)**

**Use Q9\_b.csv**





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

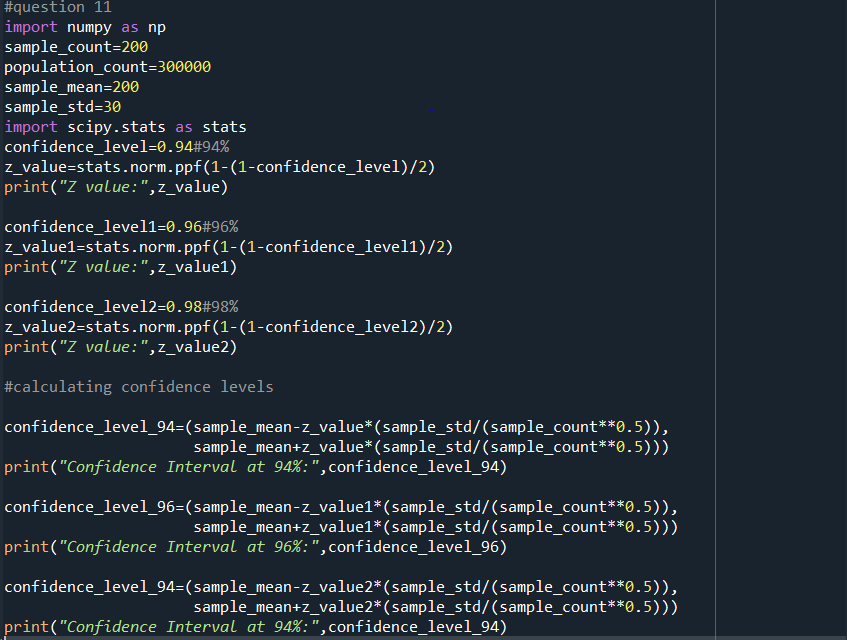


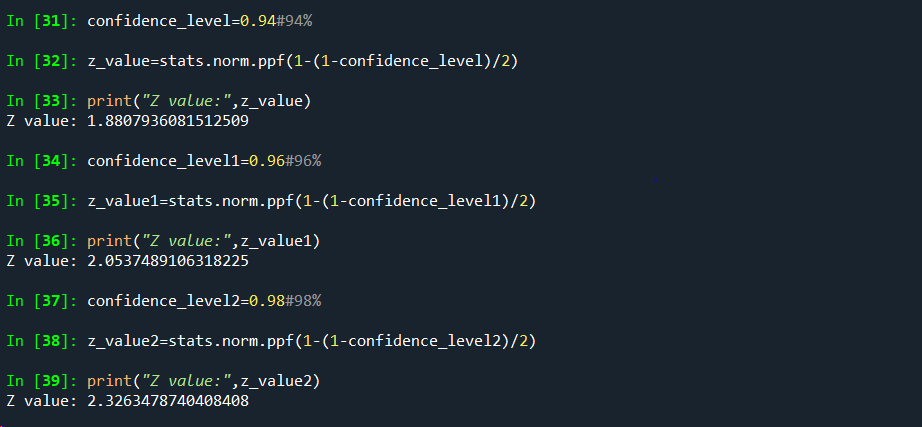
Ans: In Histogram, the majority of the data points are clustered towards the left side of the histogram with a fewer values streching towards the right side.So it is considered as positively skewed.

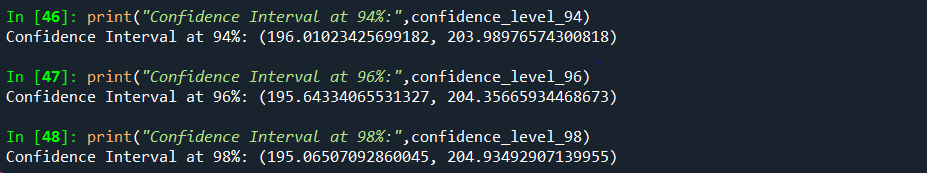


Box plot is used to identify the outliers. The lines (whiskers) extend from the edges of the box to indicate the range of the data. The points which are far away from the whiskers are considered as outliers and are plotted as individual points. Here,the whiskers are unequal in length, or there are many outliers present, it suggests potential asymmetry or variability in the dataset.

**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?





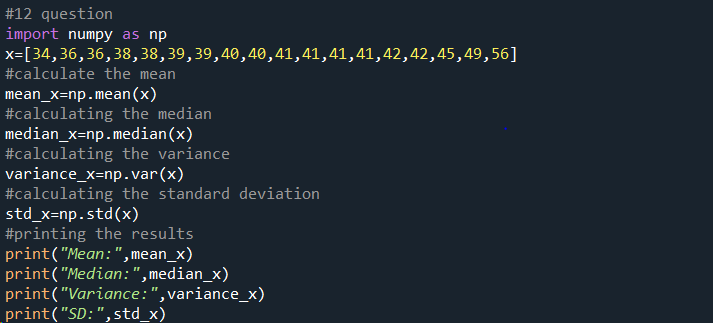


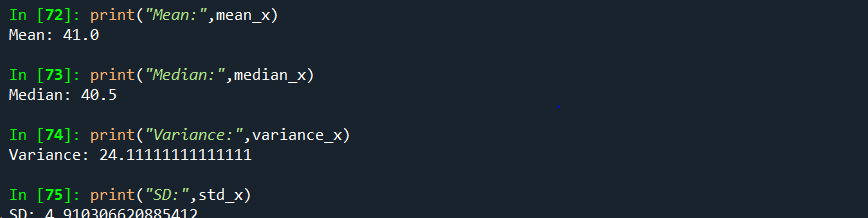
**Q12)** Below are the scores obtained by a student in tests

**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?

1)





2) **Mean:**The mean is the average of all the students.In this case it is approximately 41.0.This indicates that,on average, the student marks are close to 41.0.

**Median:** The median is the middle value in a dataset when the values are arranged in the ascending order or descending order. If the dataset has an odd number of values, the median is the middle number. If the dataset has an even number of values, the median is the average of the two middle numbers. In this case, the median value is 40.5.This implies half of the student marks are below 40.5 and other half are above 40.5.

**Variance:** Variance is a statistical measure that quantifies the dispersion or spread of a set of data points around their mean (average).In this case, the variance is approximately 24.11. The higher variance indicates that the student marks are somewhat spread out from the mean, suggesting some variability in scores.

**Standard Deviation:** Standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a dataset. It measures how spread out the values in a dataset are around the mean. It is the square root of the variance. In this case, the standard deviation is approximately 4.91. The larger standard deviation indicates that there is a noticeable amount of variability in the student marks from the mean.

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

Ans: When the mean and median are equal, the data distribution is symmetric, and the skewness of the distribution is close to zero. There is a balanced spread of data on both sides of the center, resembling a symmetric or normal distribution. Skewness refers to the measure of the asymmetry of the probability distribution of a real-valued random variable. When the mean and median are equal, the skewness of the distribution is close to zero or effectively zero. This indicates that there is no skewness or minimal skewness in the distribution, signifying a symmetric distribution.Top of Form

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

Ans: When the mean of a dataset is greater than the median, it indicates that the distribution is positively skewed, also known as right-skewed. Positively skewed distributions often occur when there are a few extremely high values (outliers) that significantly affect the mean, pulling it towards the right side.

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

Ans: When the median of a dataset is greater than the mean, it suggests that the distribution is negatively skewed, also known as left-skewed. Negatively skewed distributions often occur when there are a few extremely low values (outliers) that significantly affect the mean, pulling it towards the left side.

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

Ans: A positive kurtosis value indicates that a dataset has heavier tails and a more peaked central peak (leptokurtic distribution) compared to a normal distribution.

Positive Kurtosis (or excess kurtosis > 0).

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

Ans: A negative kurtosis value indicates that a dataset has lighter tails and a flatter central peak (platykurtic distribution) compared to a normal distribution. Negative Kurtosis (or excess kurtosis < 0)

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



What can we say about the distribution of the data?

* The data is distributed across a range from around 2 to above 18.
* The upper whisker length extending beyond 18 suggests that the possibility of outliers or data points that are more spread out.
* The position of the median line between 14 and 16 suggests that the central value of the data is shifted slightly towards the higher end of the range (the line inside the boxplot indicates the median line)
* The point that the quartile range 1 being starts exactly at 10 shows that the minimum value of the data set starts from 10 and the lower 25% of data is clustered around a value around the 10.
* In the same way, the slight exceeding of Q3 beyond the 18 shows that the upper 25% of the data extends beyond 18.

What is nature of skewness of the data?

Ans: Q2(Median):

* The position Q2(median) can be calculated as the average of Q1 and Q3.
* Q1 = 10
* As Q3 exceeds 18, let’s consider Q3 as 18.5
* Q2 = (Q1+Q3)/2 = (10+18.5)/2 = 28.5/2 = 14.25

In the above case, it is difficult to determine the skewness without actual distribution. However, with the information we have for now there could be a mild positive skewness. This is because the tail (right side) might be longer ,with a few higher values pulling beyond the 18.

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

* The IQR is the difference between Q3 and Q1.
* Q1 = 10
* Let’s assume Q3 = 18.5

**.** IQR = Q3- Q1 = 18.5 – 10 = 8.5(approximately).  
  
  
  
Q19) Comment on the below Boxplot visualizations?

Ans: Here there is a representation of 2 box plots in which box plot 2) is highly distributed across the plane and 1) is slightly less distributed.(variance)

Whiskers in these diagrams also show this.100% of the data is spread across values from 350 in 2 whereas its spread in range 250-290 app x in 1.



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

Here when we compare box plot 1 with box plot 2 we can say that the data in box plot 1 is widely spread .Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2.The median in the 2 box plots are equal .And the data spread in both of them are symmetrical.

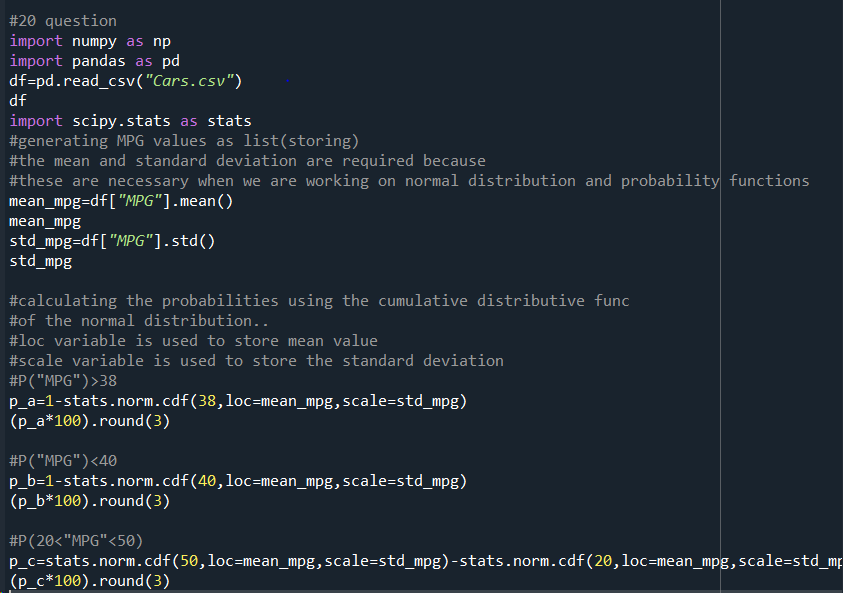
Q 20) Calculate probability from the given dataset for the below cases

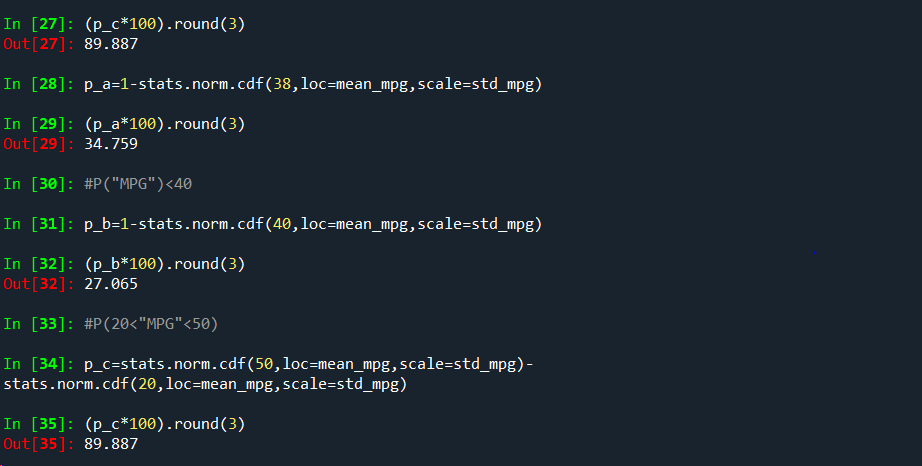
Data \_set: Cars.csv

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)

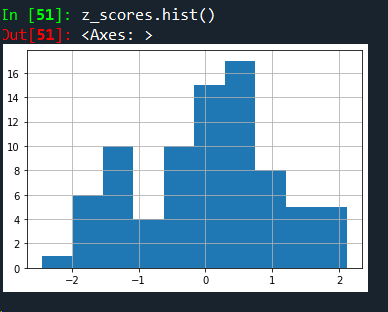


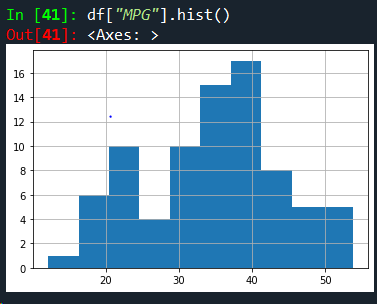


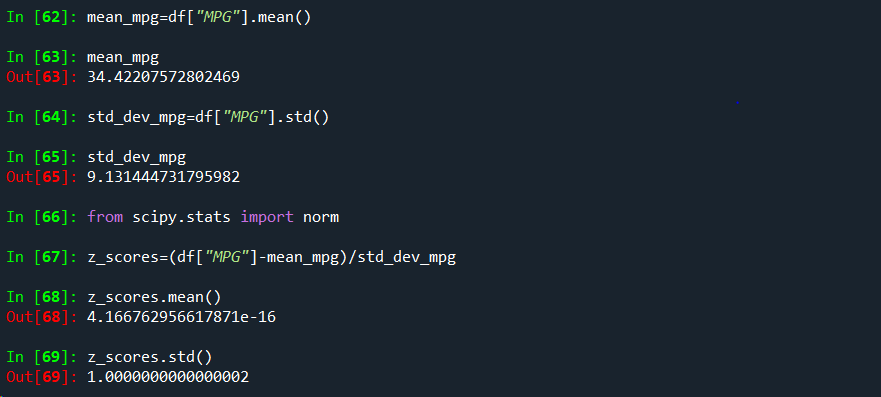
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv







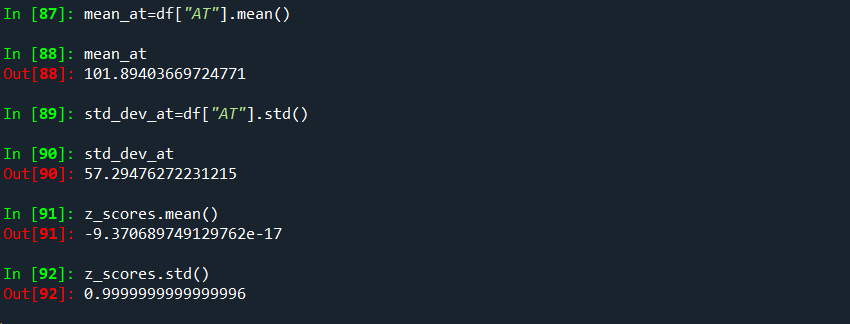
So after the Z-transformation we got the mean nearly equal to zero and standard deviation as equal to 1 so the “MPG” follows normal distribution…

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

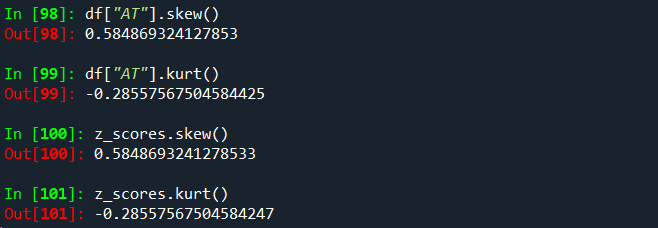
Dataset: wc-at.csv

For Adipose Tissue:





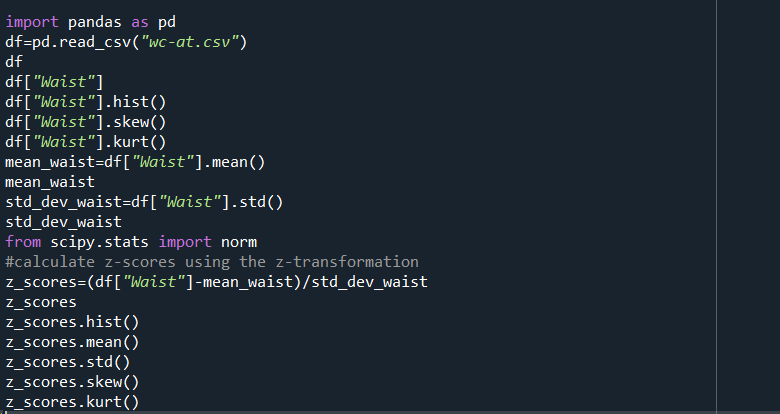
So after the z transformation we have got the values of mean and standard deviation nearly equal to 1 but not exactly equal to 1 so then we also observe skewness and kurtosis value for that variable so “For a standard normal distribution (a normal distribution with mean 0 and std 1 )the skewness is 0 and also excess kurtosis (kurtosis minus 3)is also 0.but here the values for the above measures is not 0 so the above variable is said to be not following “normal distribution”.

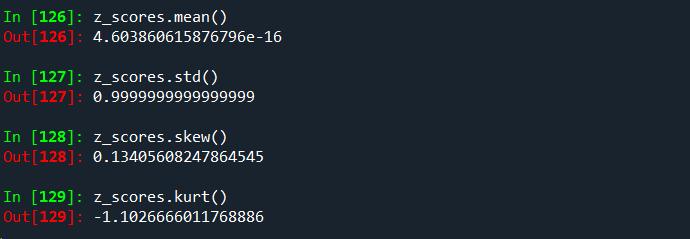




The skewness and kurtosis values are only valuable insights.

**For Waist Circumference(Waist):**







As we can see after z transformation the column named “Waist” mean and standard deviation is nearly equal to zero and 1 but not exactly one so we can say this is not following the normal distribution . For our better understanding we can also take skewness and kurtosis values as powerfull and usefull insights and we are also not getting 0,1 values respectively as you can see it in the above picture.so clearly the column is not following normal distribution.

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

Ans: The Z-score for a confidence interval represents the number of standard deviations a data point is from the mean. To calculate the Z-scores for various confidence intervals, you can use the standard normal distribution table or a statistical calculator.

Z = Zα/2 ---🡪where Zα/2 is the corresponding Z-score to the desired significance level alpha.

90% confidence interval:

To calculate the Z-score for a 90% confidence interval, you'll use the standard normal distribution table or a statistical calculator. The Z-score associated with a 90% confidence level is approximately 1.645.

This value is used in statistics to determine the range around the mean within which the true population parameter is estimated to lie with 90% confidence, assuming a normal distribution or a sufficiently large sample size.

94% confidence interval:

To calculate the Z-score for a 94% confidence interval, you can use a standard normal distribution table or a statistical calculator. The Z-score associated with a 94% confidence level is approximately 1.88.

This value is used in statistics to determine the range around the mean within which the true population parameter is estimated to lie with 94% confidence, assuming a normal distribution or a sufficiently large sample size.

60% confidence interval:

To calculate the Z-score for a 60% confidence interval, you can use a standard normal distribution table or a statistical calculator. The Z-score associated with a 60% confidence level is approximately 0.84.

This value is used in statistics to determine the range around the mean within which the true population parameter is estimated to lie with 60% confidence, assuming a normal distribution or a sufficiently large sample size.

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

Ans: We will calculate the t scores when we have the sample size (<30). we would use the t-distribution instead of standard normal distribution.

t = tα/2, df

where tα/2  is the t-score corresponding to the desired significance level alpha.

And the degree of freedom**df** = n-1(sample size – 1).

**96% Confidence Level:**

The confidence level is 96% which means alpha = 1-0.96 = 0.04.so half of the value of alpha is 0.04/2 = 0.02.the degrees of freedom (df) for a sample size is 25-1 = 24.so now we have to look up the t-score corresponding to the cumulative probability of 0.980(1-0.02) and df = 24 in the t-distribution table which equals

2.398.

**95% Confidence Level:**

The confidence level is 95% which means alpha = 1-0.95 = 0.05.so half of the value of alpha is 0.05/2 = 0.025.the degrees of freedom (df) for a sample size is 25-1 = 24.so now we have to look up the t-score corresponding to the cumulative probability of 0.975(1-0.025) and df = 24 in the t-distribution table which equals 2.064.

**99% Confidence Level:**

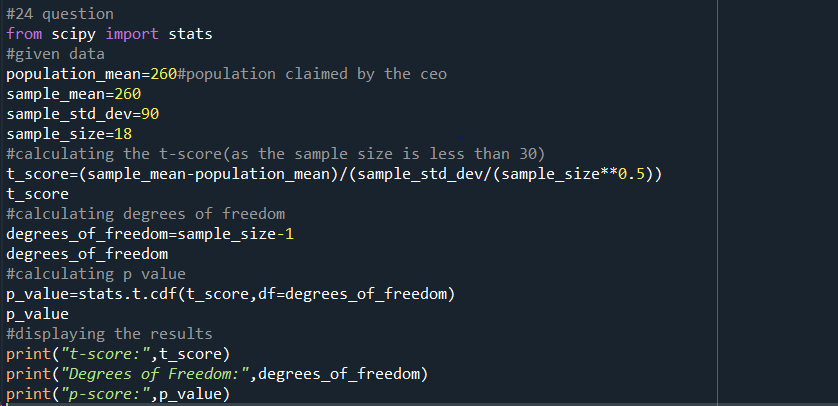
The confidence level is 99% which means alpha = 1-0.99= 0.01.so half of the value of alpha is 0.01/2 = 0.005.the degrees of freedom (df) for a sample size is 25-1 = 24.so now we have to look up the t-score corresponding to the cumulative probability of 0.995(1-0.005) and df = 24 in the t-distribution table which equals

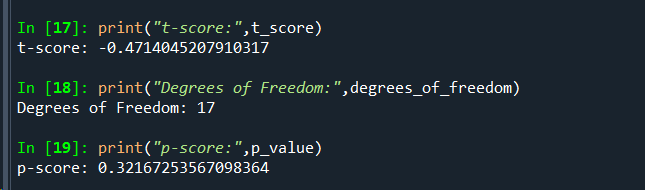
2.797.

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





1.Null Hypothesis(Ho):The average bulb life is 270 days(CEO’S claim).

Alternate Hypothesis(H1):The average bulb life is less than 270 days.

2.Alpha=95% confidence interval i.e.,0.05(Significance level).

3.Calculate t-scores:These measures how many standard error the sample mean is away from the population mean under the assumption of null hypothesis.

T=sample mean-population mean/sample standard deviation/(n)(1/2)

4.Find the p value and compare the p value with significance level.

Here p value is equal to 0.321 and our significance level is 0.05 so practically speaking the p value is greater than significance value then,

H0 is accepted and H1 is rejected.