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

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?

The possible outcomes when three coins are tossed is

HHH, HHT, HTH, THH, TTT, TTH, THT, HTTàn(s)=8

Therefore, The probability of getting two heads and one tail are obtained is 3/8 or 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

When two dices are rolled, the total number of outcomes 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)

So, there are 36 possibilities

a)There is no possible way to get a sum is equal to 1 when two dices are rolled,as the minimum value on each die is 1.Hence,the probability is 0

It is possible as the probability always lies between **0 <=P(A)<=1.**

b)The total number of possibilities occurring when the probability that sum is less than or equal to 4 is {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)} i.e.

The total number of possibilities = 36

Therefore, the probability that sum is Less than or equal to 4=6/36=1/36

c)The sums that are divisible by both by 2 and 3 are 6 and 12.

So,the possible outcomes are (1,5),(2,4),(3,3),(4,2),(5,1),(6,6)) i.e.,6 outcomes

Total possibilities =36

Therefore, the sums that are divisible by both by 2 and 3 =6/36=1/36

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?

Solution: The total number of balls in the bag are 2 red,3 green and 2 blue balls=7balls

Two balls can be drawn in 7c2 ways=21 ways

7C2=7!/2!(7-2)!=7!/2!5!=7\*6/2=21

Number of ways of drawing 2 balls none of the balls drawn is blue=number of ways of drawing 2 balls from 2 red and 3 green balls=10 ways

5C2=5!/2!(5-2)!=5!/2!3!=5\*4/2=10

Hence,Probability of drawing 2 balls such that none of balls are 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.

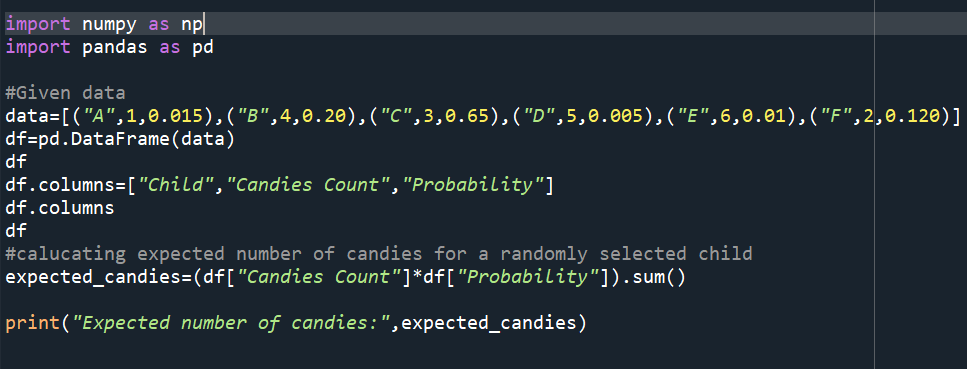
Child B – probability of having 4 candies = 0.20

Solution:To calculate the expected number of candies for a randomly selected child

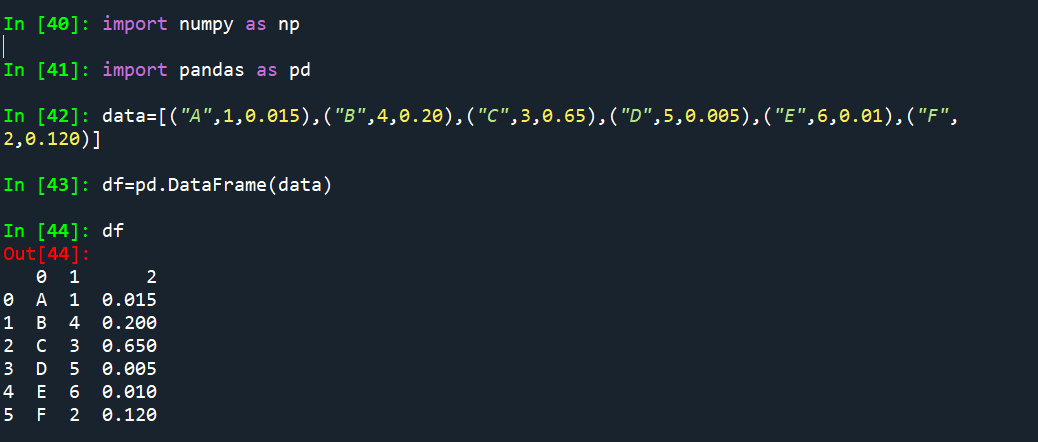
We use the formula **Expected Value**=∑ (Value× Probability)

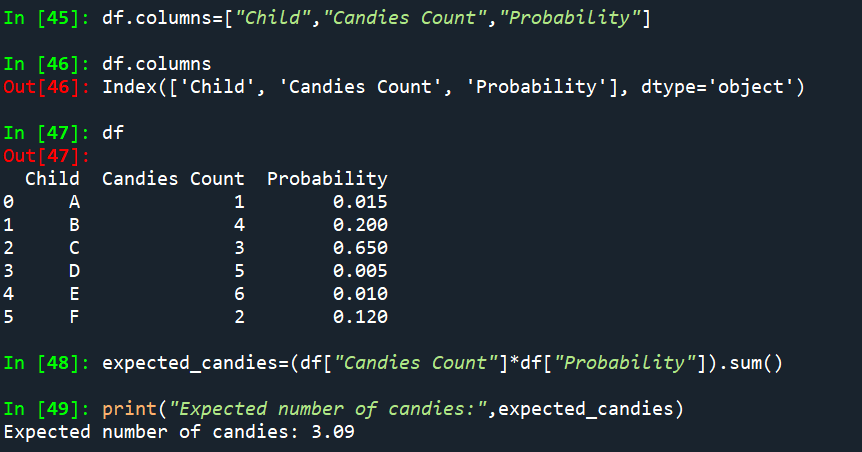
where value here is Candies count and Probability is equal to probability of having number of candies by each individual .this can be achieved by pandas. Firstly create a numpy data by importing numpy library and change numpy data into pandas and assign the column names to each column and apply the formula over the pandas data.Here is the code for above

Code: Input



Output:





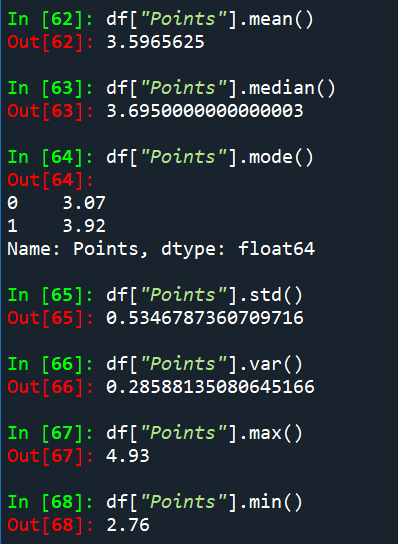
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

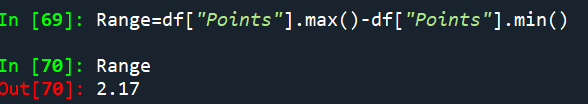
* For Points,Score,Weigh>

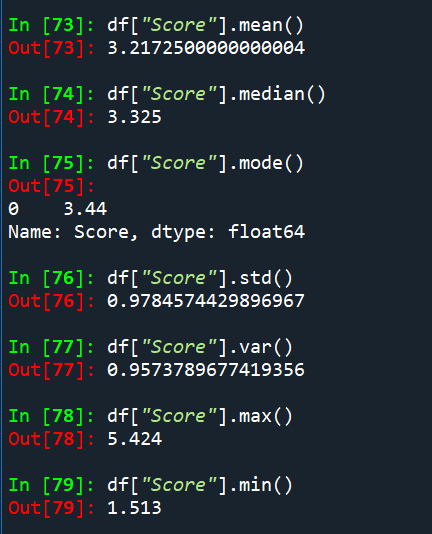
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

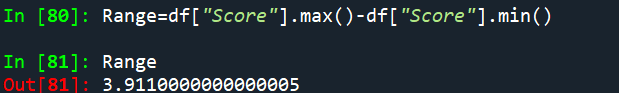
**Use Q7.csv file**

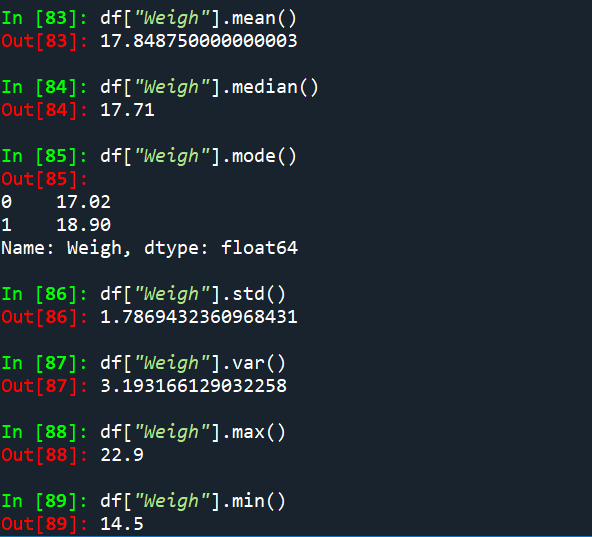
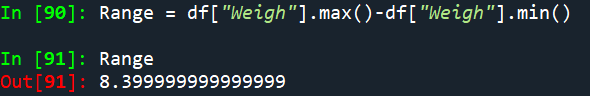
|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Points** | **Score** | **Weigh** |
| Mazda RX4 | 3.9 | 2.62 | 16.46 |
| Mazda RX4 Wag | 3.9 | 2.875 | 17.02 |
| Datsun 710 | 3.85 | 2.32 | 18.61 |
| Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| Hornet Sportabout | 3.15 | 3.44 | 17.02 |
| Valiant | 2.76 | 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** | 3.59 | 3.12 | 17.84 |
| **Median** | 3.69 | 3.32 | 17.71 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Standard Deviation** | 0.54 | 0.97 | 1.78 |
| **Variance** | 0.28 | 0.95 | 3.19 |
| **Range** | 2.17 | 3.91 | 8.39 |





****

****

**** ****

**Inference:**

1)The average value of the “Points” column is approximately 3.5965.

“Score” column is approximately 3.2172.

“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 is approximately 0.9784

“Weigh” column is approximately 1.786

Standard deviation is a measure of the amount of variation or dispersion in a set of values.It quantifies how much data points differ from the mean(average) of the data set

3. The Variance value of the “Points” column is approximately 0.285.

“Score” column is approximately 0.9573.

“Weigh” column is approximately 3.1931.

Variance is a statistical measure that quantifies the spread or dispersion of a set points in a dataset.It provides a numerical value that describes how much each number in a set differs from the mean(average) and,therefore from other number in the set.

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

“Score” column is approximately 3.911.

“Weigh” column is approximately 8.399.

The range is a measure of the spread or dispersion of a set of values in a dataset.It is calculated as the difference between the maximum and minimum values in a datset

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 statistical measure that represents the middle value of a dataset when it is ordered from least to greatest.

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 is a statistical that represents the most frequently occuring value in a dataset

Q8) Calculate Expected Value for the problem below

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

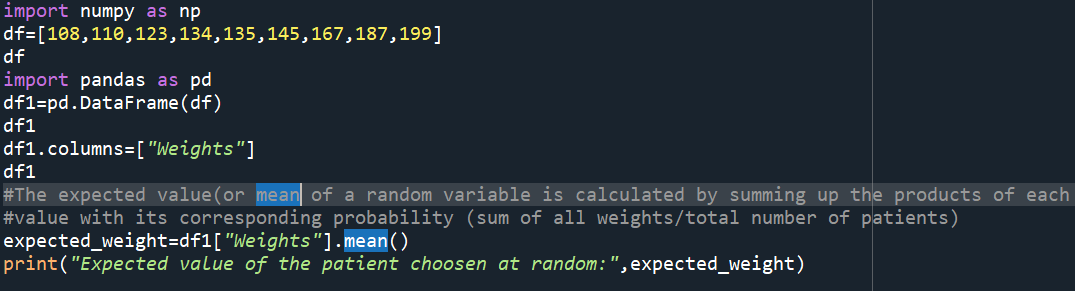
108, 110, 123, 134, 135, 145, 167, 187, 199

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

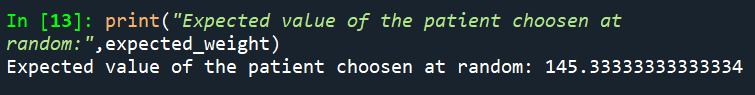
Solution:The expected value(or mean)of a random variable is calculated by summing up the products of each value with its corresponding probability.since each patient is equally likely to be chosen at a random ,the probability of each patient is 1/9(since there are 9 patients)

Therefore,Expected Value = ∑Weights / Number of Patients

Total number of patients at a clinic are 9

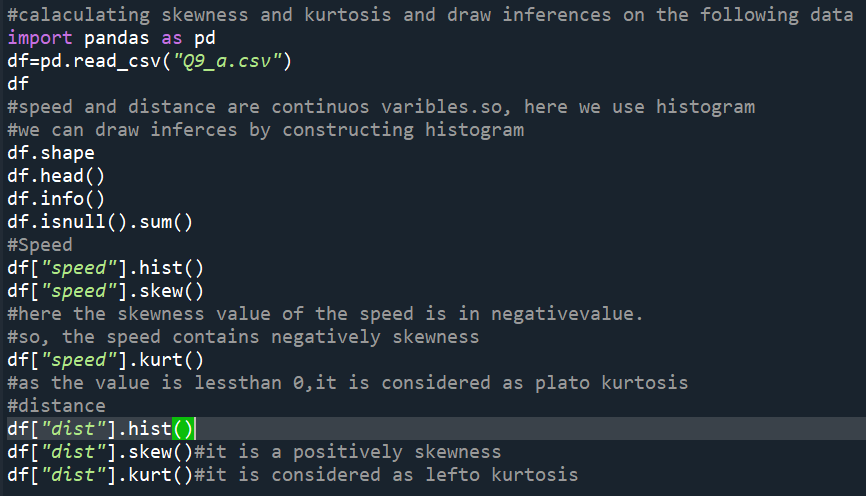


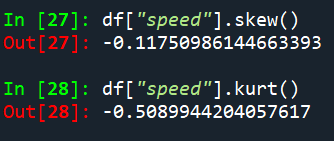
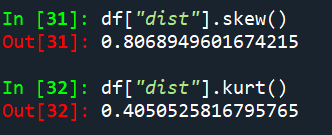
Output:



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

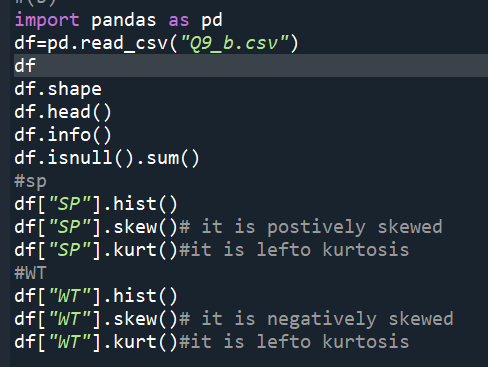
**Cars speed and distance**

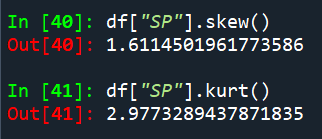
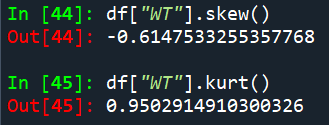
**Use Q9\_a.csv**

** **

**SP and Weight(WT)**

**Use Q9\_b.csv**

****

** **

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

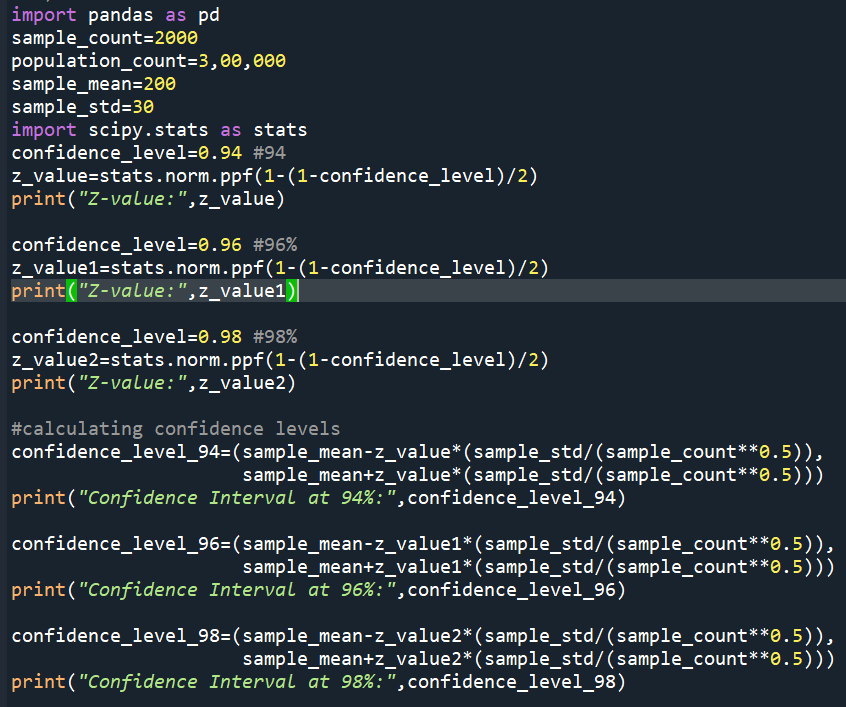


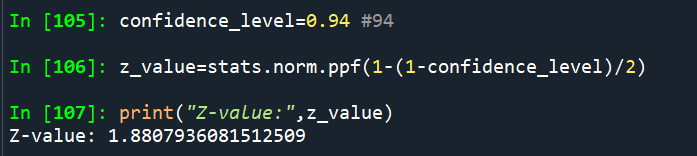
The histogram of chickweight and weight is positively skewed,the majority of the datapoints cluster on the leftside of the histogram.As we know a distribution is said to be positively skewed when the tail on the rightside of the histogram is longer than the left side.

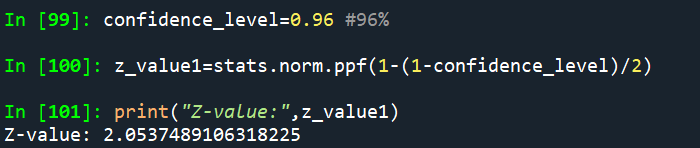


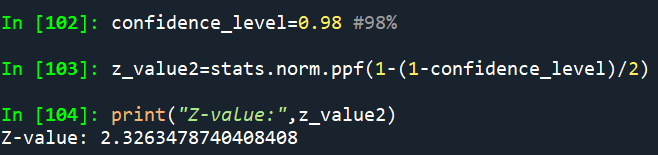
The above diagram shows that whiskers in a boxplot are not of the same length and outliers are present ,it indicates potential symmetry or skewness in the data.So,unequal whiskerlengths tells that spread of the data is not symmetrical.The presence of outliers,points beyond the whiskers,indicates that there are individual data points that are significantly different from the rest of the dataset.The combination of unequal whisker lengths and the presence of outliers represents the distribution is not symmetric.

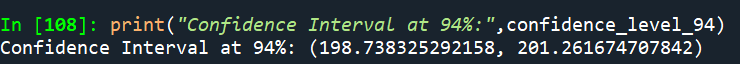
**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate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

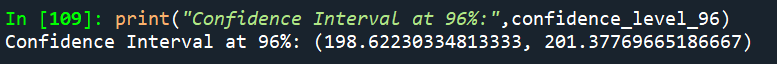












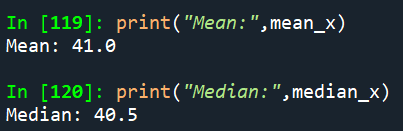
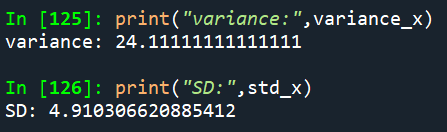


**Q12)** Below are the scores obtained by a student in tests

**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 calculated by adding up all the values in a dataset and then dividing the sum by the total number of values.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 a statistical measure that represents the middle value of a dataset when it is ordered from least to greatest. So,In this case, the median is 40.5. This implies that 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 extent to which individual data points in a dataset differ from the mean of the dataset. 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:

The standard deviation is a statistical measure of the amount of variation or dispersion in a set of values.The standard deviation 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?

When the mean,median of a data are equal,it implies that the distribution of the data is symmetric.In a symmetric distribution,the skewness is zero.skewness is a measure of the asymmetry of a probability distribution.If the mean and median are equal, it suggests that the data is centered around the middle,and the distribution has a balanced shape.Therefore distribution is symmetric ,the skewness is zero.

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

When the mean is greater than the median,the distribution is said to be positively skewed or right-skewed.The skewness value in such a distribution is positive. Skewness is a measure of the asymmetry of a probability distribution and a positive skewness indicates a longer or fatter tail on the right side.

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

When the median is greater than the mean,the distribution is said to be negatively skewed or left-skewed.The skewness value in such a distribution is negative. Skewness is a measure of the asymmetry of a probability distribution and a negative skewness indicates a longer or fatter tail on the left side.

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

Positive kurtosis indicates that the distribution of a dataset has heavier tails and a sharper peak than a normal distribution.Commonly,positive kurtosis is associated with a leptokurtic distribution. Positive kurtosis value always have the kurtosis value when calculated as (>0).

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

A negative kurtosis value indicates that the dataset has lighter tails and a flatter peak than a normal distribution. Negative kurtosis is associated with a platykurtic distribution . Negatively kurtosis value always have the kurtosis value when calculated as (<0).

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



What can we say about the distribution of the data?

Solution:The data is distributed ranging from approximately 2 to above 18,is visually represented through a boxplot analysis.

* The upper whiskers length extending beyond 18 implies the potential presence of outliers or datapoints that are more widely spread.
* The median line,positioned between 14 and 16,indicates that the central value of the data is slightly skewed toawards the higher end of the range.Additionally,the starting point of quartile range 1 at exactly 10 suggests that the minimum value of the dataset is 10,and the lower 25% of the data is concentrated around this value.
* Similarly,the slight extension of Q3 beyond 18 signifies that the upper 25% of the data surpasses this threshold.

What is nature of skewness of the data?

Solu:As we know,The position of Q2 can be calculated as the average of Q1 and Q3

Q1= 10[from boxplot]

Q3 exceeds the 18 . So let us consider Q3 as 18.5

Q2 = (Q1+Q3)/2 =10+18.5/2=28.5/2=14.25

The nature of skewness in the given data is difficult to conclusively determine without the actual distribution in the above case. However,with the information of what we could have the skewness is mild positive . This is because theupper whisker extending beyond 18 suggests that presence of a few higher values,potentially causing a longer right tail .

What will be the IQR of the data (approximately)?   
sol:The IQR is calculated as the difference between Q3 and Q1

IQR=Q3-Q1

Where Q1=10, Q3=18.5(assumed)

Therefore, IQR=18.5-10=8.5

Q19) Comment on the below Boxplot visualizations?



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

Solu:First there are no outliers.Second both the boxplot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

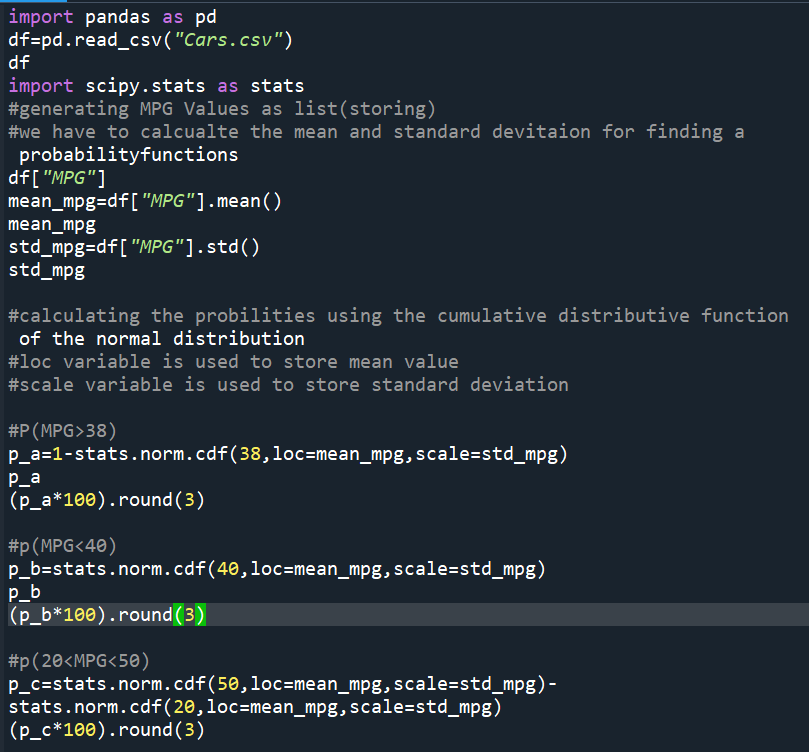
Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

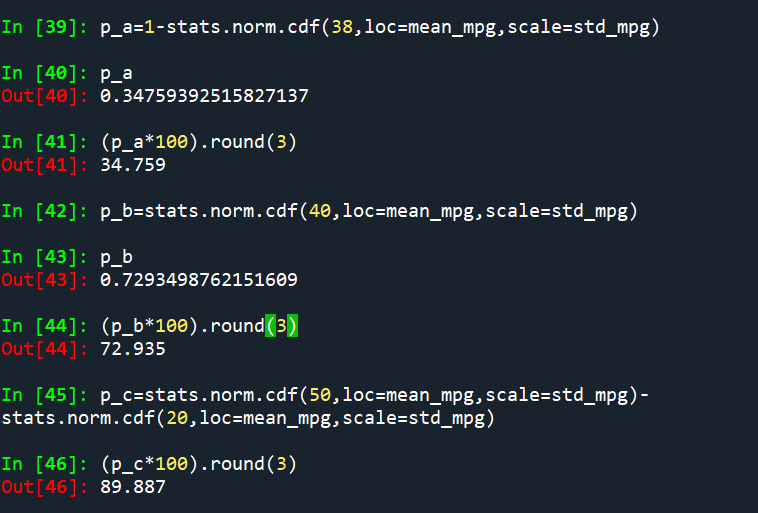
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

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



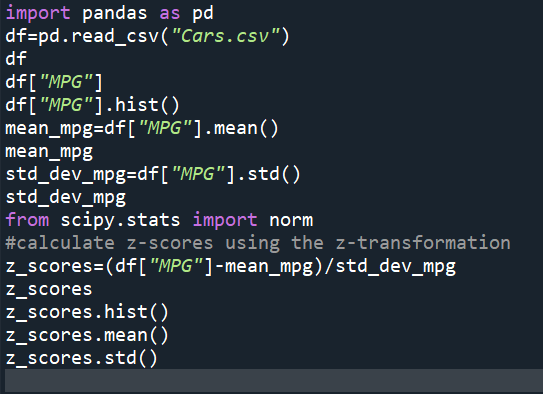


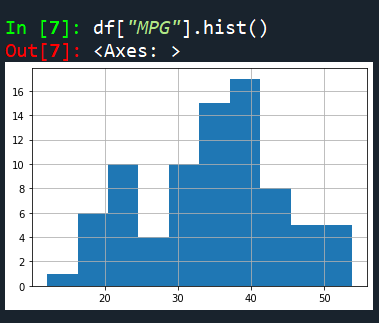
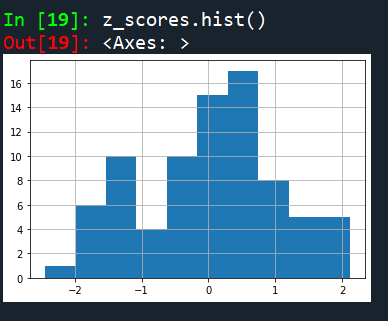


Q 21) Check whether the data follows normal distribution

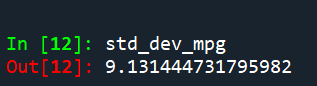
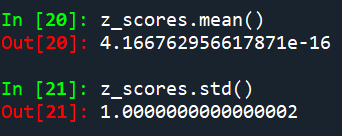
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



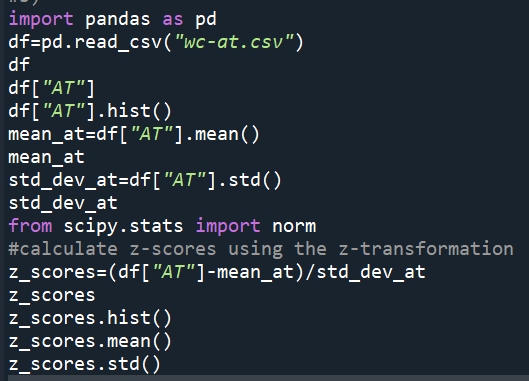


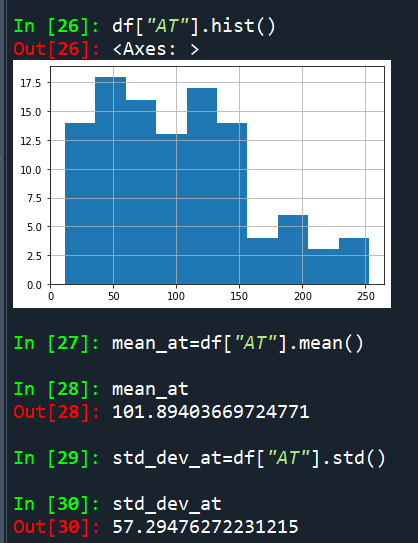
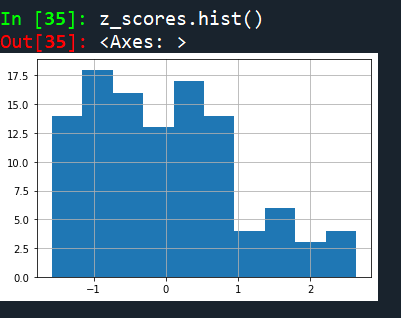
 

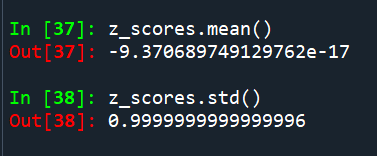
After doing Z-transformation we got the mean nearly zero and standard deviation is equal to 1 .Therefore, MPG follows the normal distribution.

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

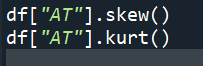
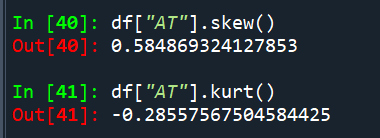
Dataset: wc-at.csv

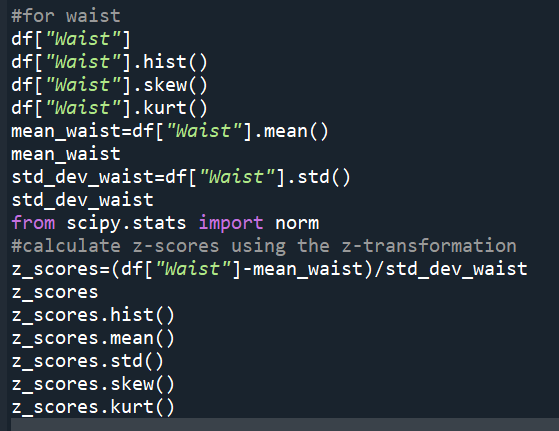


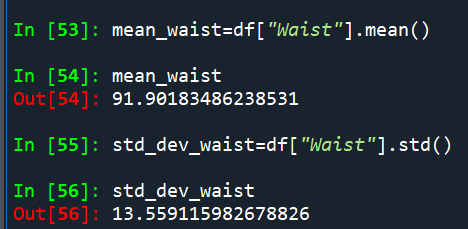
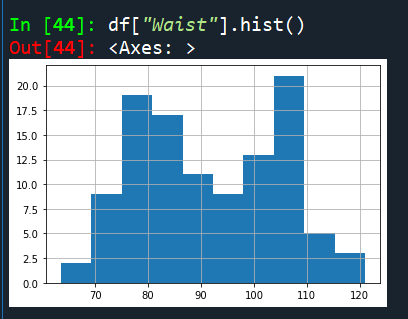
 

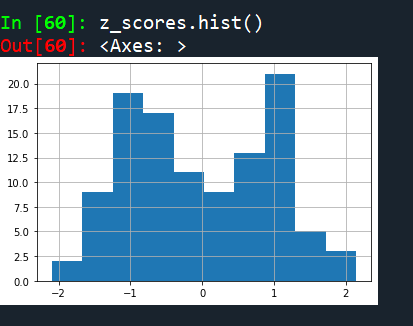
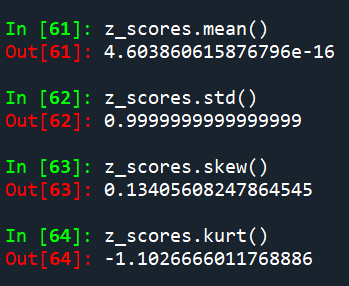


After doing Z transformation we got a values of mean and standard deviation nearly equal to 1 but not exactly as 1 .so we have to calculate skewness and kurtosis





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

Solu: To calculate the Z-scores for different intervals, we can use the standard normal distribution (Z-distribution) and its percentiles. The formula to calculate the Z-scores for a specific confidence level is:

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

**90% Confidence Interval:**

Here the confidence level is 90% ,which means alpha = 1-0.90(90%)=0.10.Half of this alpha is 0.05.when we look up the Z-score corresponding to the cumulative probability of 0.95(1-0.05) in the standard normal distribution table Z0.05 ~1.645.

**94% Confidence Interval:**

Here the confidence level is 94%, which means alpha = 1-0.94(94%)=0.06.Half of this alpha is 0.03.when we look up the Z-score corresponding to the cumulative probability of 0.97(1-0.03) in the standard normal distribution table Z0.05 ~1.880.

**60% Confidence Interval:**

Here the confidence level is 60%, which means alpha = 1-0.60(60%)=0.40.Half of this alpha is 0.20.when we look up the Z-score corresponding to the cumulative probability of 0.80(1-0.20) in the standard normal distribution table Z0.05 ~1.842.

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

Solu: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

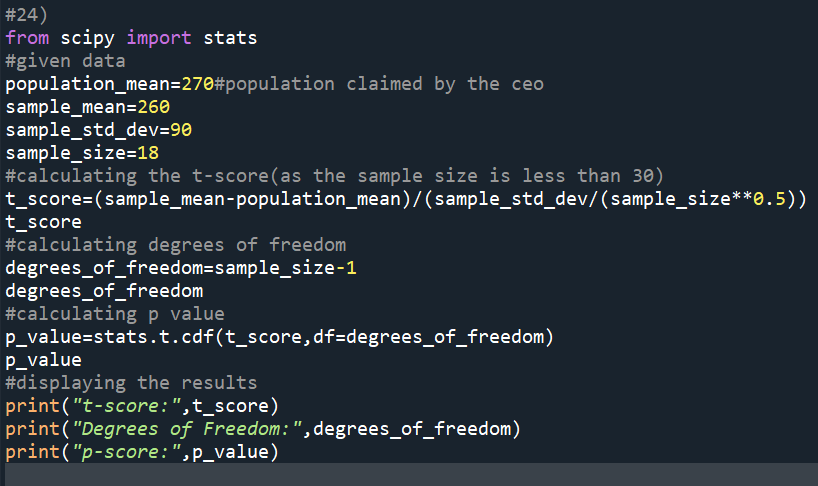
2.797.

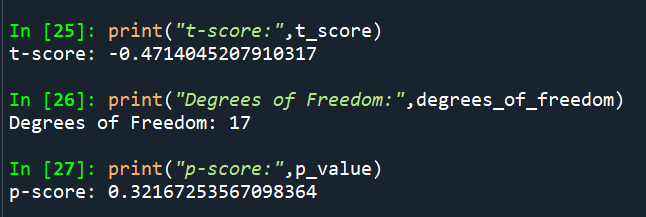
Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom





1. Formulate Hypothesis:

Null hypothesis(H0): the average bulb life is 270 days (CEO’S claim)

Alternative hypothesis(H1):

The average bulb life is less than 270 days.

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

1. 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…