|  |  |
| --- | --- |
| 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 | Classification(Nominal)data type |
| Number of kids | Countable data type |
| Number of tickets in Indian railways | Discrete(Countable datatype) |
| Number of times married | Discrete(Countable datatype) |
| Gender (Male or Female) | Classification(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?

Sol: When three coins are tossed there are 2^3=8 possible outcomes because each coin has two possible outcomes (head/tail) and the total possible number of outcomes

HHH, HHT, HTH, THH, TTT, TTH, THT, HTT🡪n(s)=8

Now we want to find the probability of getting two heads and one tail. There are three ways to achieve this : HHT,HTH and THH

So the probability is 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

Sol: When two dice are rolled the possible 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)

1. Equal to 1

There is no possibility of occurring the sum is equal to 1 when throwing two dice the probability is considered as 0. It is possible as the possibility always lies between 0<=P(A)<=1.

1. Less than or equal to 4

The possible ways to get a sum less than or equal to 4 are (1,1),(1,2),(2,1),(1,3),(2,2) and (3,1). There are 6 favorable outcomes.

Probability=(Number of favorable outcomes)/(Total number of outcomes)

P(A)=6/36=1/6

1. Sum is divisible by 2 and 3

The common multiples of 2 and 3 are 6 and 12

So we need to find the outcomes where the sum is 6 & 12

The possible outcomes for a sum of 6 are (1,5),(2,4),(3,3),(4,2),(5,1)

The possible outcomes for a sum of 12 is (6,6)

P(sum is divisible by 2 &3) = 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?

Sol: The total number of balls are 7

The total number of ways to draw 2 balls out of 7 is

C(7,2) = 7!/2!(7-2)!=7!/2!.5!=7.6/2=21

Here the number of ways to draw 2 balls that are not blue

C(5,2) = 5!/2!(5-2!)=5!/2!.3!=5.4/2=10

P(non blue)=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 |

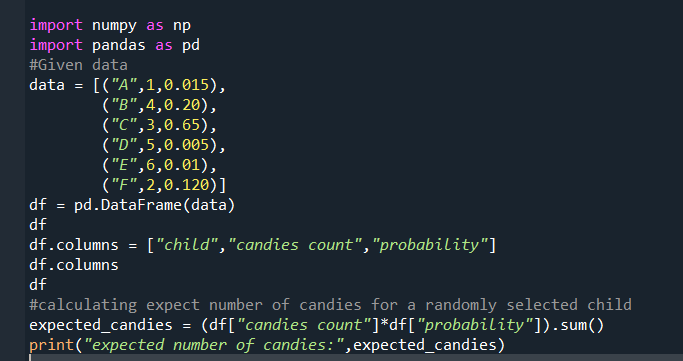
Child A – probability of having 1 candy = 0.015.

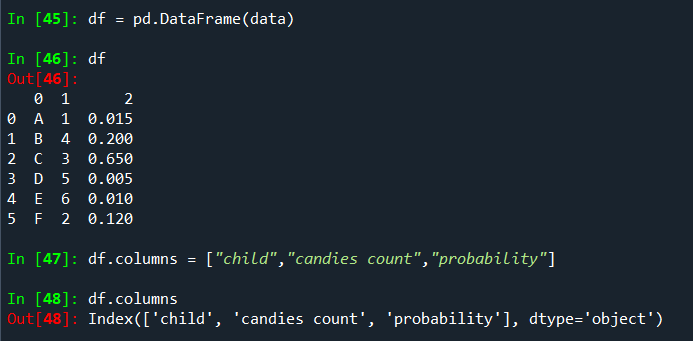
Child B – probability of having 4 candies = 0.20

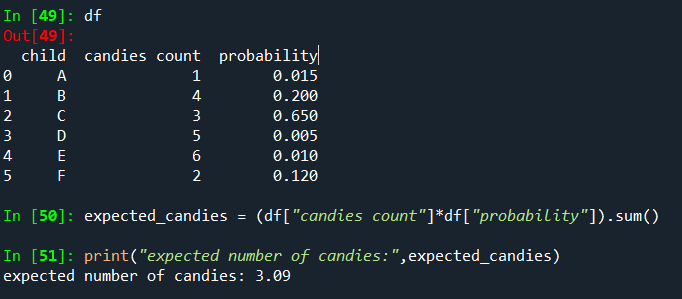
Sol: Here the formula for **Expected Value**=∑(Value× Probability)

Where the value is candies 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 question

CODE:







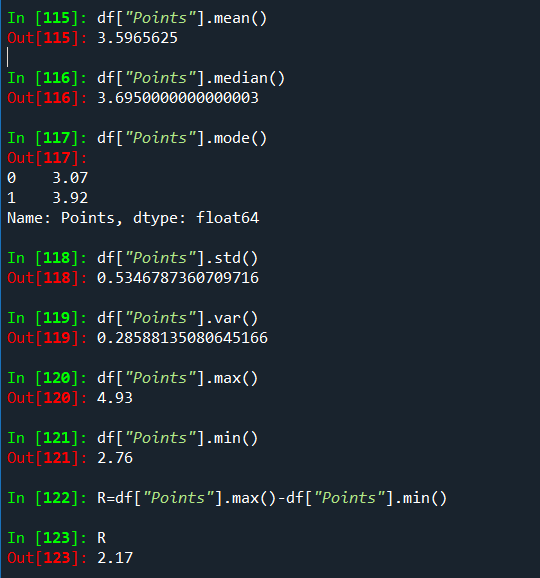
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

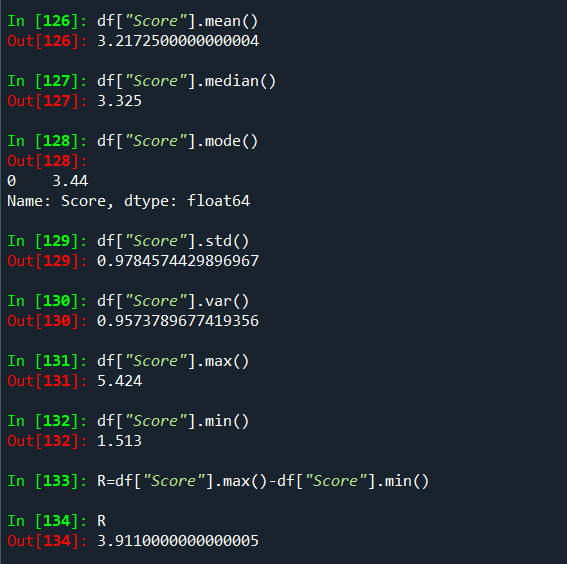
* For Points,Score,Weigh>

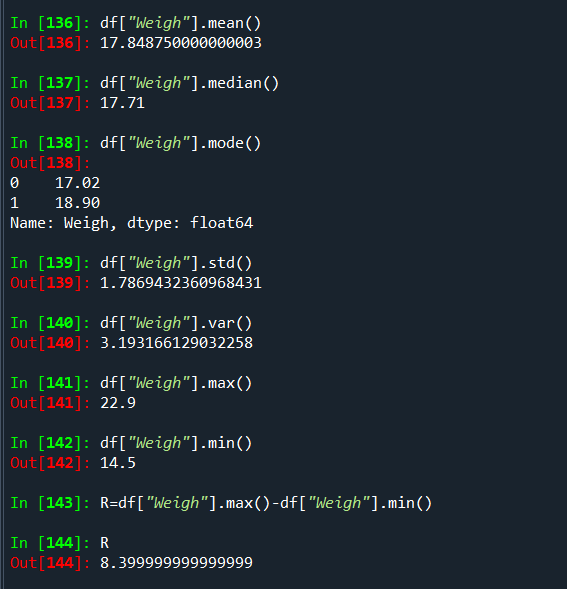
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**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** | 3.5965625 | 3.21725 | 17.84875 |
| **Median** | 3.695000 | 3.325 | 17.71 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Standard Deviation** | 0.534678 | 0.978457 | 1.786943 |
| **Variance** | 0.285881 | 0.957379 | 3.193166 |
| **Range** | 2.17 | 3.911 | 8.39 |

****

****

****

**Inference:**

**1**.**Mean**:

Points: The average points is around 3.5965625.

Score: The average score is approximately 3.21725.

Weigh: The average weight is about 17.84875

The mean is the average of a set of values. It is calculated by summing up all the values and dividing by the total number of values. The mean is sensitive to extreme values, also known as outliers.

**2**.**Median**:

Points: The median points is around 3.695000.

Score: The median score is approximately 3.325.

Weigh: The median weight is about 17.71.

The median is the middle value of a dataset when it is ordered. If there is an even number of observations, the median is the average of the two middle values. The median is less sensitive to extreme values than the mean and provides a measure of central tendency.

**3**.**Mode**:

Points: The mode of the Points is around 3.92.

Score: The mode of the score is around 3.44.

Weigh: The mode of the weight is around 17.02.

The mode is the value that appears most frequently in a dataset. A dataset may have no mode (no repeated values), one mode (unimodal), or multiple modes (multimodal). It is a measure of the most common or frequently occurring value.

**4**.**Standard** **Deviation**:

Points: The standard deviation for points is about 0534678.

Score: The standard deviation for score is around 0.978457.

Weigh: The standard deviation for weight is approximately 1.786943.

The standard deviation is a measure of the amount of variation or dispersion in a set of values. It quantifies how much individual values deviate from the mean. A higher standard deviation indicates greater variability, while a lower standard deviation suggests less variability.

**5**.**Variance**:

Points: The points have a variance of approximately 0.285881.

Score: The score has a variance of about 0.957379.

Weigh: The weight has a variance of roughly 3.193166.

Variance is the square of the standard deviation. It measures the average squared deviation of each number from the mean of a dataset. Like the standard deviation, it provides a measure of the spread or dispersion of values.

**6**.**Range**:

Points: The range of points is about 2.17.

Score: The range of score is approximately 3.911.

Weigh: The range of weight is around 8.39.

The range is the difference between the maximum and minimum values in a dataset. It provides a simple measure of the spread or dispersion of values but is sensitive to extreme values.

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

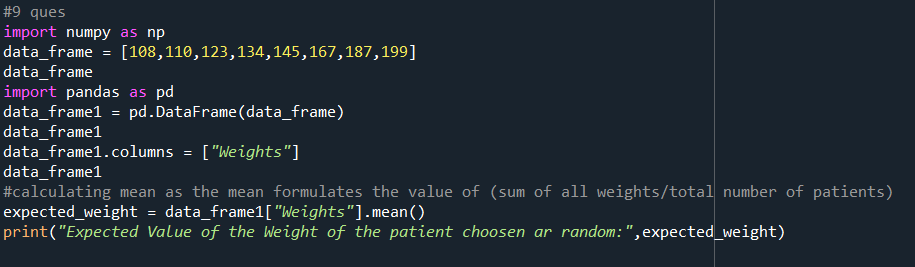
Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

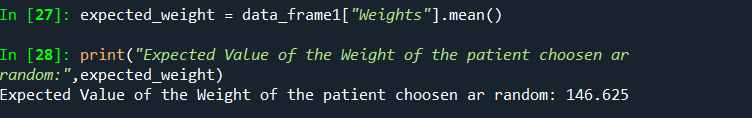
Sol: The expected value (or mean) of a set of values is calculated by summing up all the values and dividing by the number of values. In this case, you have the weights of patients as follows:

X={108,110,123,134,135,145,167,187,199}

Expected Value = ∑Weights / Number of Patients

Total number of patients = 9 (from the above provided data)

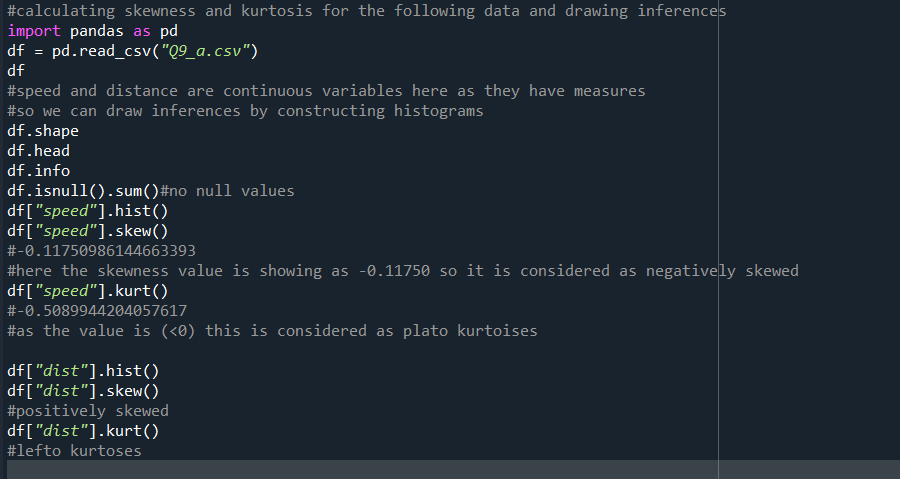


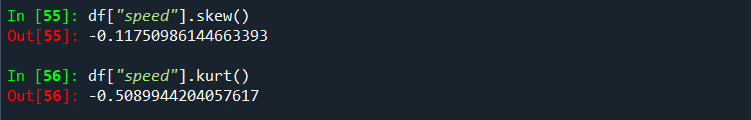


**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

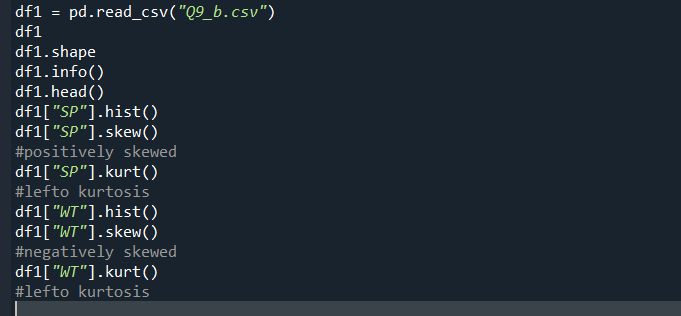
****

****

****

**SP and Weight(WT)**

**Use Q9\_b.csv**

****

****

****

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

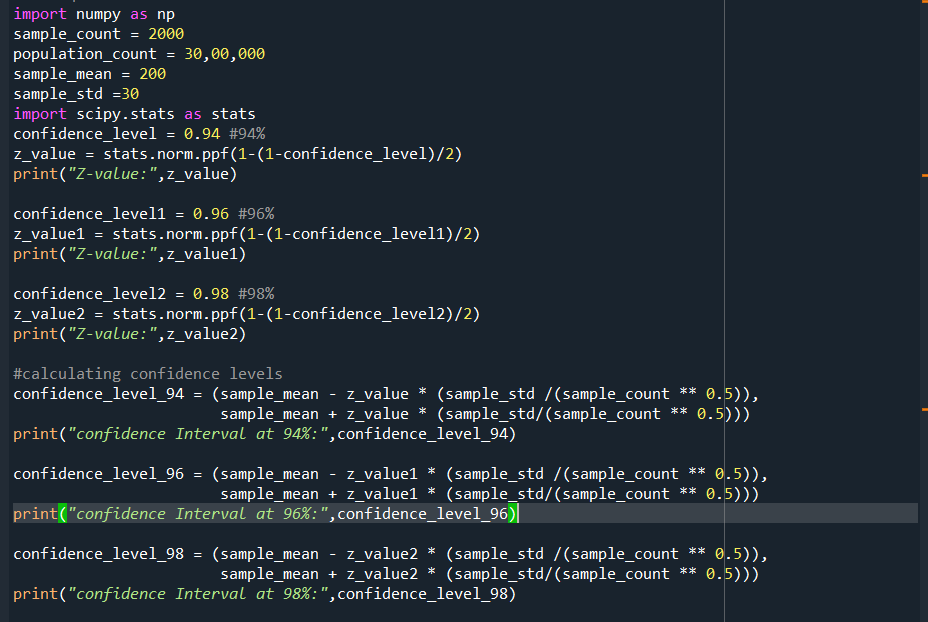


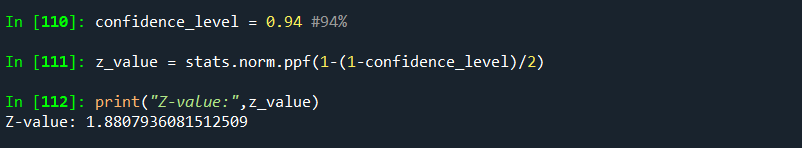
Sol:Here the majority of the data points cluster towards the lower end, and a smaller number of higher values cause the distribution to have a tail on the right side.In this distribution the mean is generally greater than the median and the skewness value is positive.

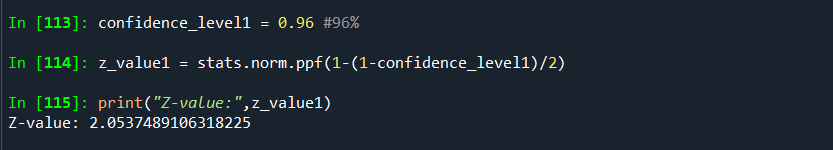


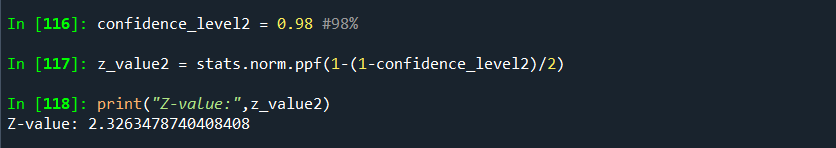
Sol: Boxplot is used to identify the outliers. It is is a graphical representation of the distribution of a dataset. It provides a summary of the central tendency, spread, and skewness of the data. The whiskers extend from the box to the minimum and maximum values within a certain range. Individual data points beyond the whiskers are often marked as dots and are considered potential outliers. Here the whiskers are unequal in length or there are many outliers present it suggests potential asymmetry or variability in the datasets.

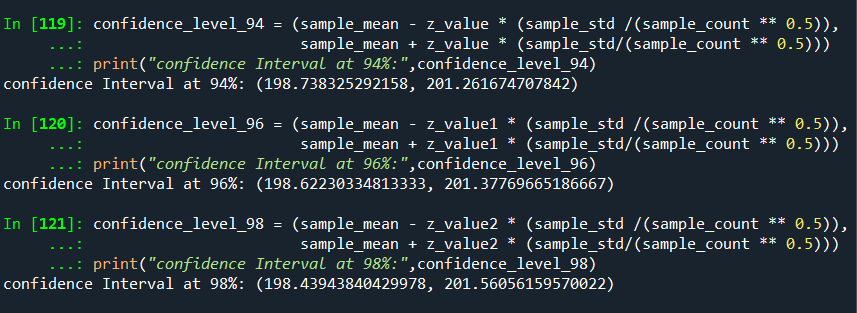
**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate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



****

****

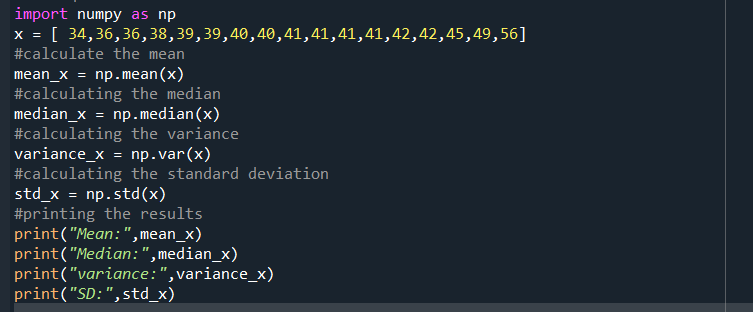
****

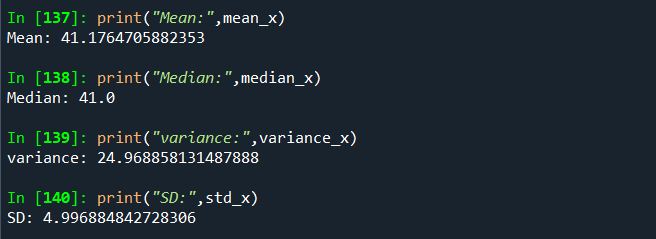
****

**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?





**Mean:**

the mean is a measure of central tendency that represents the average of a set of values. Here the mean is the average of all the student marks. In the case of it is approximately 41.1. This indicates that, on average, the student marks are close to 41.1.

**Median:**

The median is another measure of central tendency in statistics. The median is the middle value of the data when it is sorted in ascending order. In this case, the median is 41.0. This implies that half of the student marks are below 40.5 and other half are above 41.0.

**Variance:**

variance indicates the extent to which each data point differs from the mean of the entire dataset. In this case, the variance is approximately 24.96. 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 measure of the amount of variation or dispersion in a set of values. It is often used in statistics to quantify the degree to which individual data points in a dataset differ from the mean (average) of the dataset. In this case , the standard deviation is approximately 4.99. 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?

Sol:When the mean, median, and mode of a dataset are equal, the distribution is said to be perfectly symmetric, and the skewness is zero.In this case, the skewness of the distribution is zero. If mean = median, then the distribution is symmetric Skewness = 0. Symmetric distributions have a balanced shape, and the values are evenly distributed around the center. The tails on the left and right sides of the distribution are mirror images of each other.

Top of Form

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

Sol: When the mean is greater than the median in a dataset, the distribution is said to be positively skewed or right-skewed. In a positively skewed distribution.

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

Sol: When the median is greater than the mean in a dataset, the distribution is said to be negatively skewed or left-skewed. In a negatively skewed distribution.

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

Sol: Positive kurtosis indicates that a dataset has "heavy tails" or is "leptokurtic." In a distribution with positive kurtosis.

**Kurtosis>0:** Positive kurtosis indicates heavy tails and a pronounced peak. The distribution has more data points in the tails than a normal distribution.

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

Sol: A negative kurtosis value indicates that a dataset has "light tails" or is "platykurtic." In a distribution with negative kurtosis.

**Kurtosis<0:**Negative kurtosis (platykurtic) indicates lighter tails and a flatter peak compared to a normal distribution. The distribution has fewer extreme values.

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?

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?



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

Sol: First there are no outliers. Second both the box plot 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.

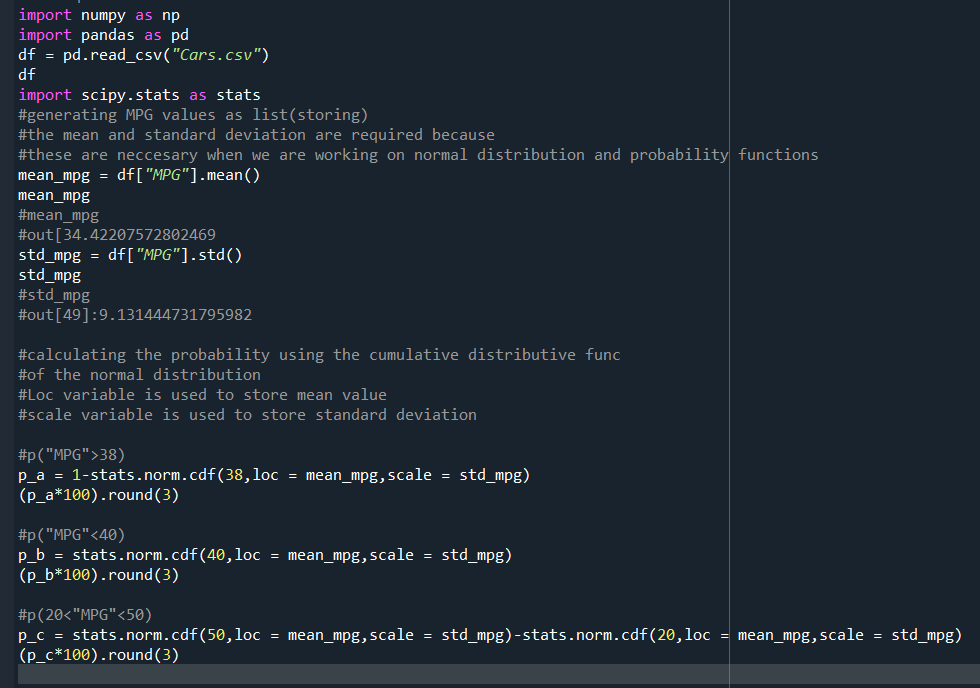
Q 20) Calculate probability from the given dataset for the below cases

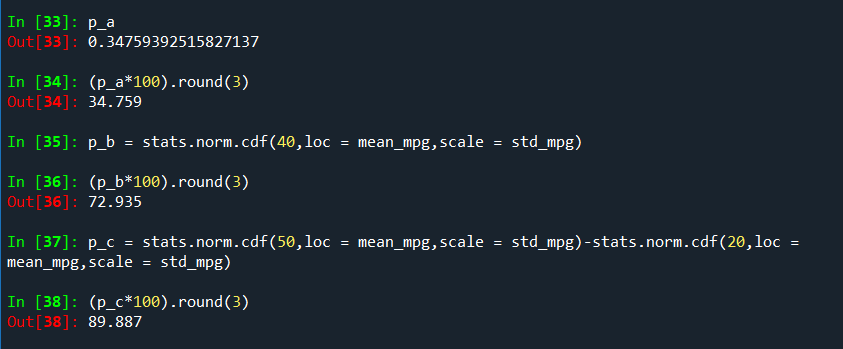
Data \_set: Cars.csv

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)

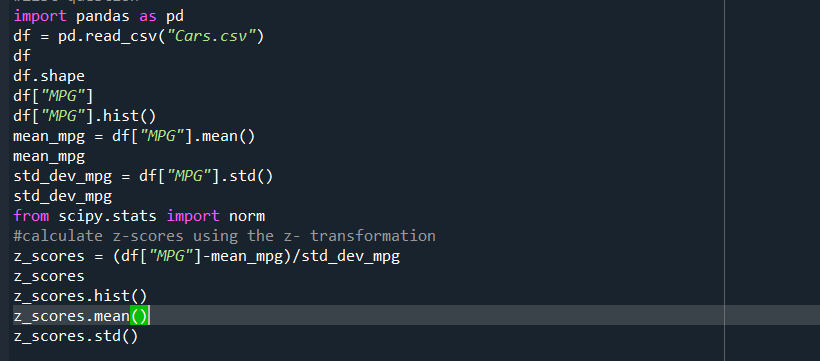


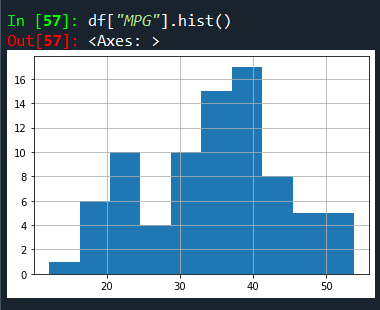
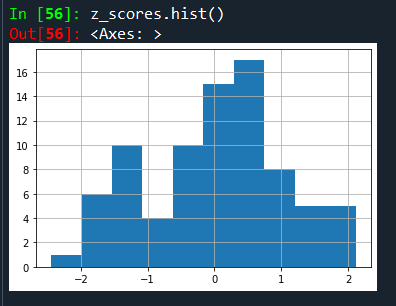


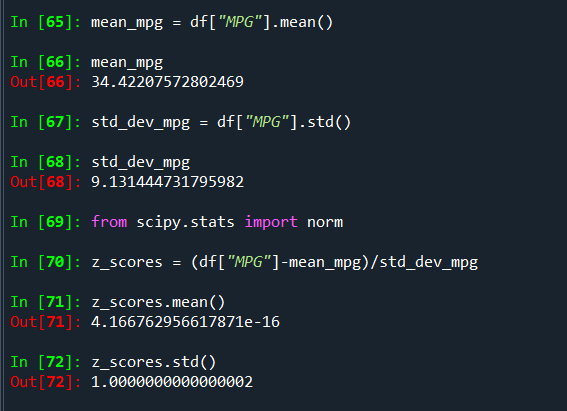
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv





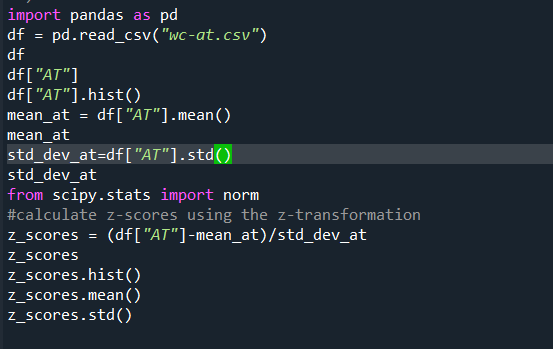


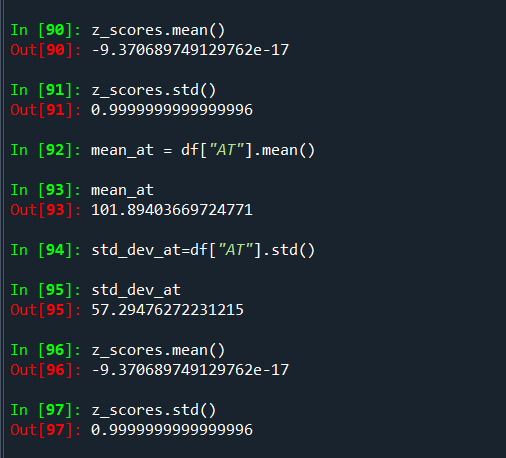
Therefore after doing Z-transformation we got the mean nearly equal to zero and standard deviation is equal to 1 so the “MPG” follows normal distribution.

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

Dataset: wc-at.csv

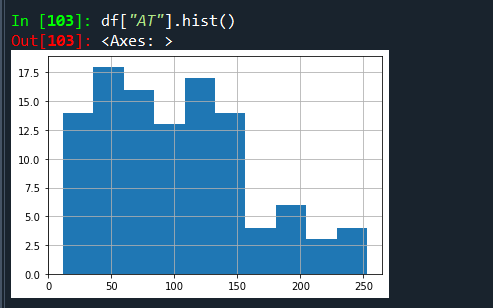
From Adipose Tissue





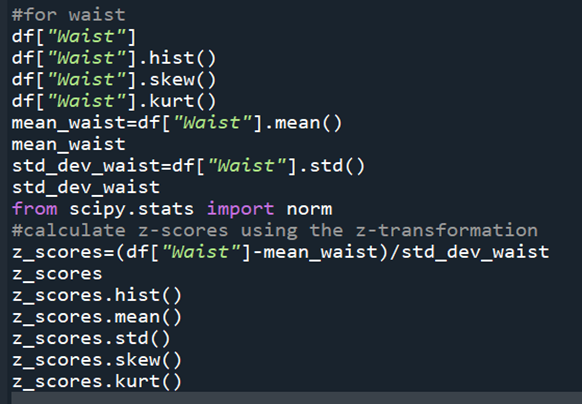
Therefore after the z transformation we have got the values of mean and standard deviation nearly equal to 1 but not exactly equal to 1 so then we also observe skewness and kurtosis value for that variable so “For a standard normal distribution (a normal distribution with mean 0 and std 1 )the skewness 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”.

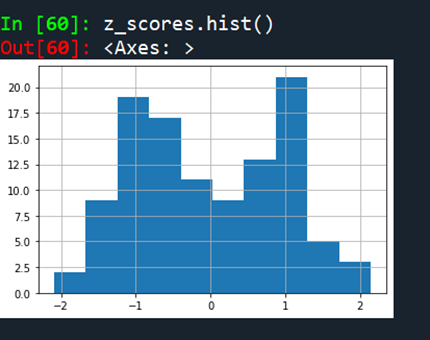
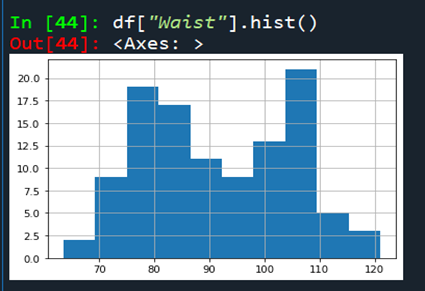


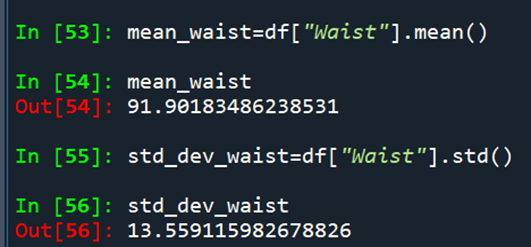


The skewness and kurtosis values are only valuable insights.

For Waist Circumstance (Waist):









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

Sol: To calculate the Z-scores for a given confidence interval, you can use the standard normal distribution table or a calculator

The formula to calculate the Z-scores for a specific confidence level is:

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

1. **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.

1. **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 ~1.881.

1. **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

Sol: The t-score for a confidence interval is determined by the confidence level and the degrees of freedom (df) which is related to the sample size. For small sample sizes, it's common to use the t-distribution for confidence intervals rather than the normal distribution.

The formula for the t-score in a confidence interval is given by:

*t*=±*tα*/2,*df*​  
where t is the t-score, *α* is the significance level (1 - confidence level) and df is the degrees of freedom.

**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

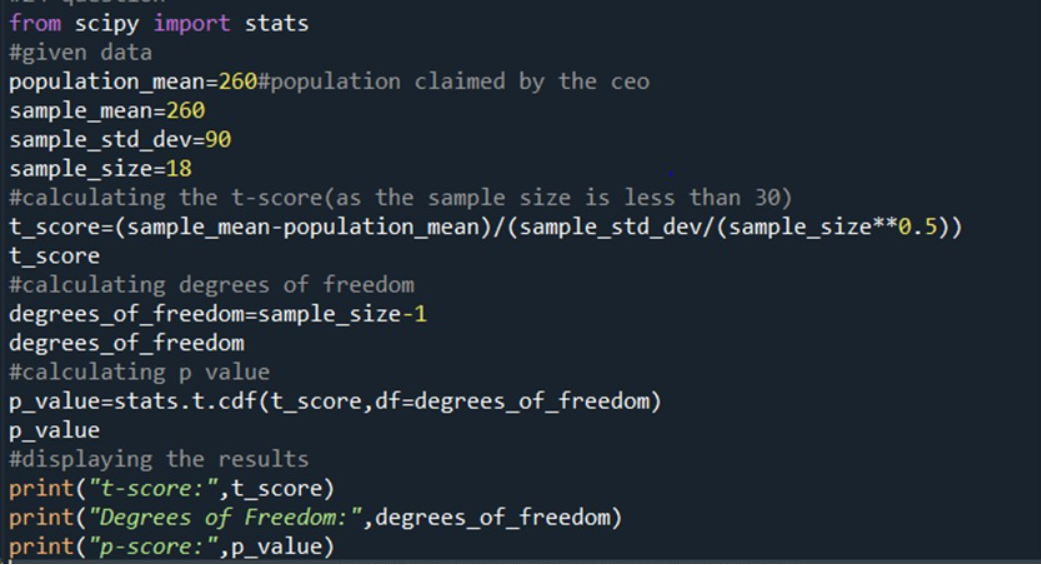
2.797.

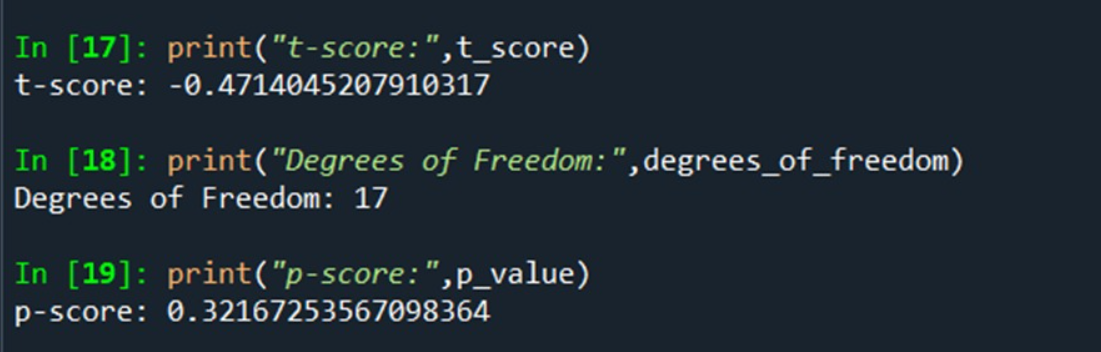
Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs 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.