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

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 possible outcomes when tossing three coins :2^3 = 8

Let A be the event of getting two heads and one tail favorable

Outcomes: HHT, HTH, THH,HHH,TTT,TTH,THT,HTT

Number of favorable outcomes (n(A)) :3

Probability of event A(p(A)) = n(A) / total possible outcomes = 3/8

Therefore, the probability of getting two heads and tails when three coins

are is :P(A) = 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: Total possible outcomes when rolling two dice: 6 sides \* 6 sides = outcomes

1. Equal to 1

There is only one way to get a sum of 1: rolling a 1 on the first die and a 1 on the second die

Number of favorable outcomes = 1

Total possible outcomes = 36

Probability = Number of favorable outcomes /Total possible outcome Probability = 1/36

1. Less than or equal to 4

Favorable outcomes: (1,1), (1,2), (1,3), (2,1), (2,2), (3,1)

Number of favorable outcomes =6

Total possible outcomes = 36

Probability = Number of favorable outcomes / Total possible outcomes Probability = 6/36 =1/6

C) Sum is divisible by 2 and 3

Favorable outcomes: (1,5), (2,4), (3,3), (4,2), (5,1), (6,6)

Number of favorable outcomes = 6

Total possible outcomes = 36

Probability = Number of favorable outcomes/Total possible outcomes

Probability = 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: Given

🡪Total red balls (R) = 2

🡪 Total green balls (G) = 3

🡪 Total blue balls (B) =2

We need to find the probability of drawing two balls that are not blue

Total ways to choose 2 balls out of 7:C (7,2) = 7! / (2!\*(7-2)!) =21

Now, let’s calculate the ways to choose 2 balls that are not blue:

🡪 we have 2 red and 3 green balls, so the number of ways to choose 2 non- blue is C (5,2) =5! / (2! \* (5-2)!) = 10

The probability of choosing 2 non-blue balls is given by the ratio of favorable outcomes to total outcomes:

Probability = Number of ways to choose 2 non-blue balls/ Total number of ways to choose 2 balls

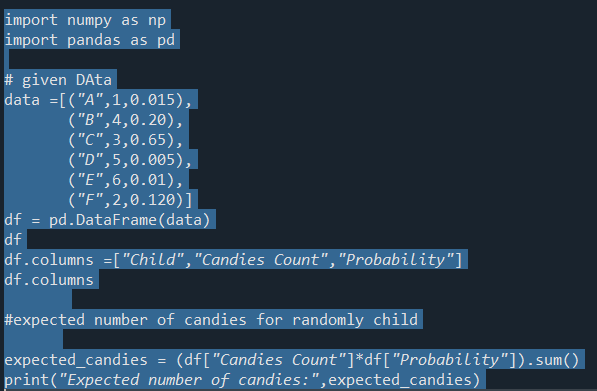
Probability = 10/21 = 0.476

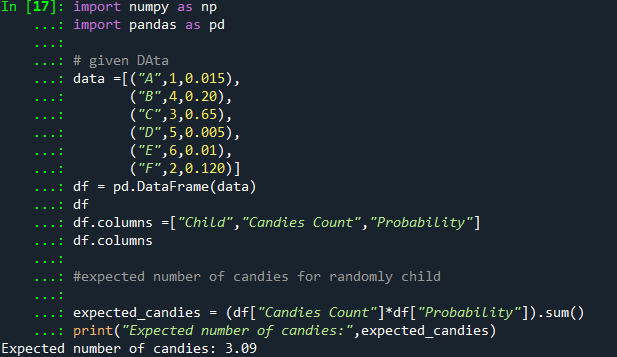
Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 |  |
| 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



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

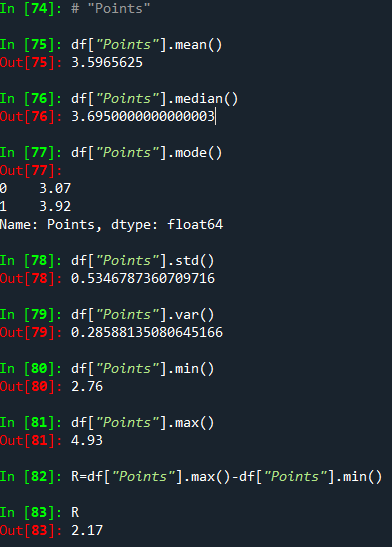
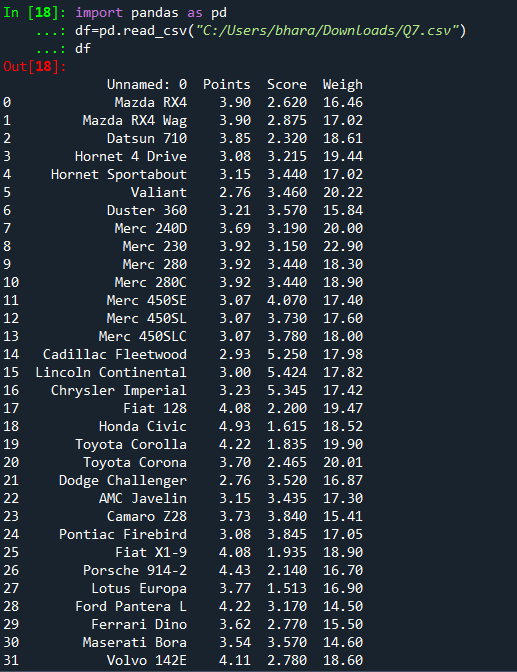
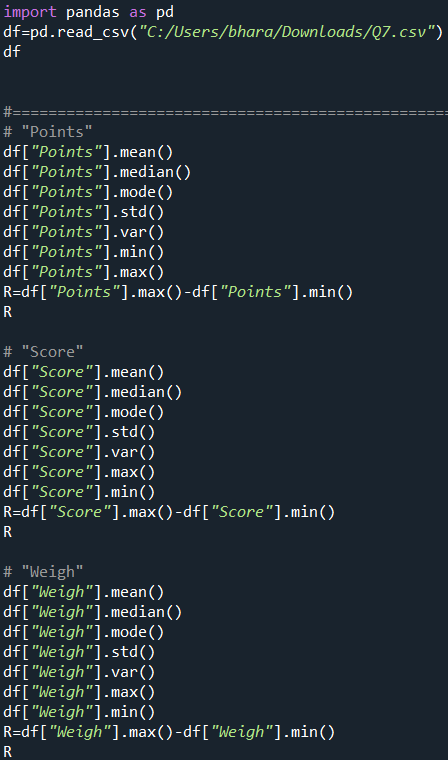
E = 0.015 + 0.80 +1.95 +0.025 + 0.06 + 0.240

E = 3.09

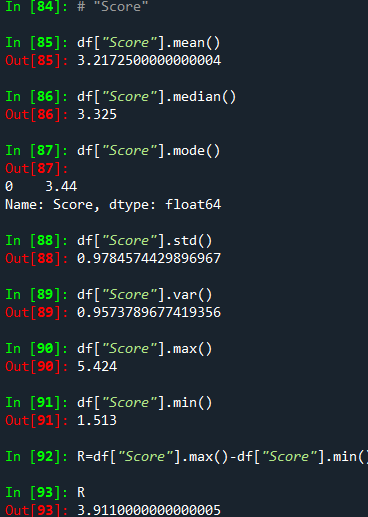
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

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



Ans: mean (point ,score ,weight) =



**Use Q7.c**

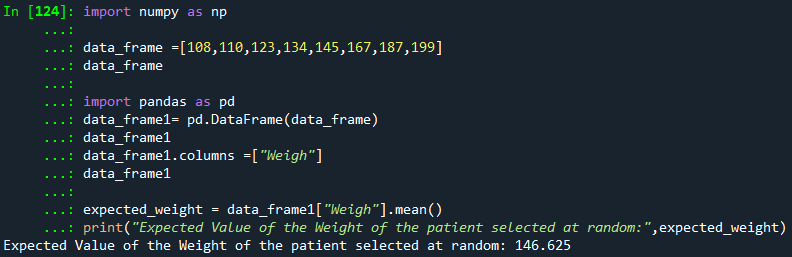
Q8) Calculate Expected Value for the problem below

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?

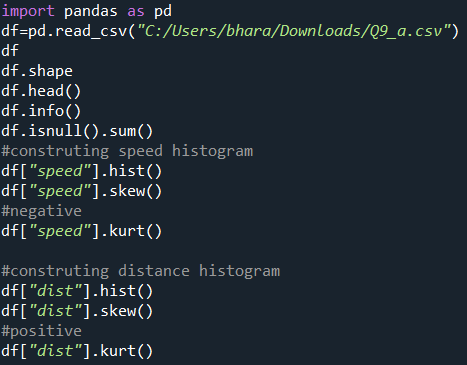
Number of patients = 9

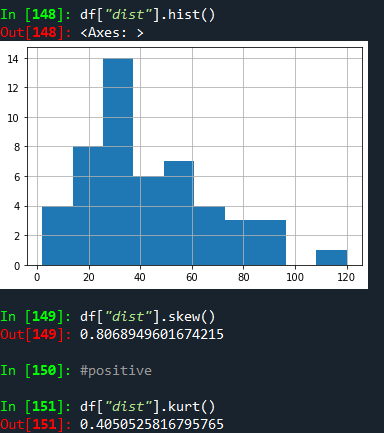
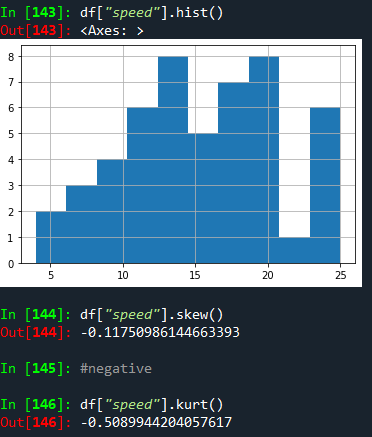
****

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

**Cars speed and distance**

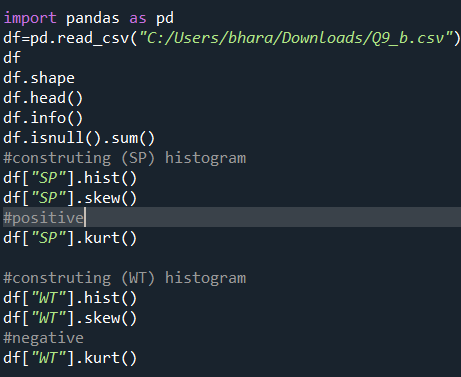
**Use Q9\_a.csv**

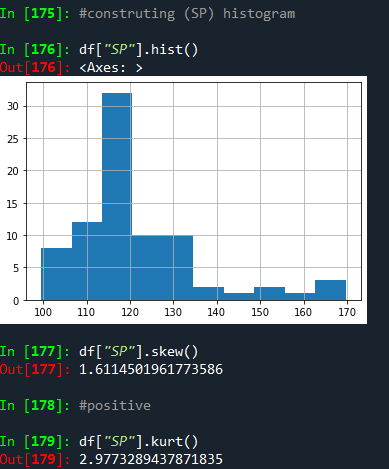
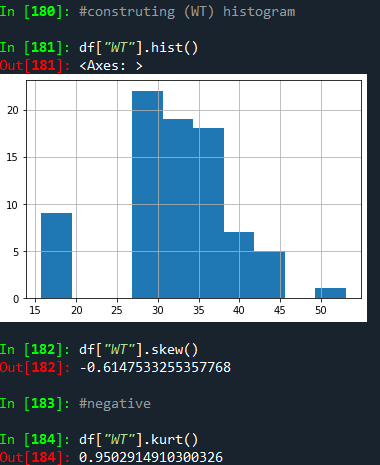
****

****

**SP and Weight(WT)**

**Use Q9\_b.csv**

****

** **

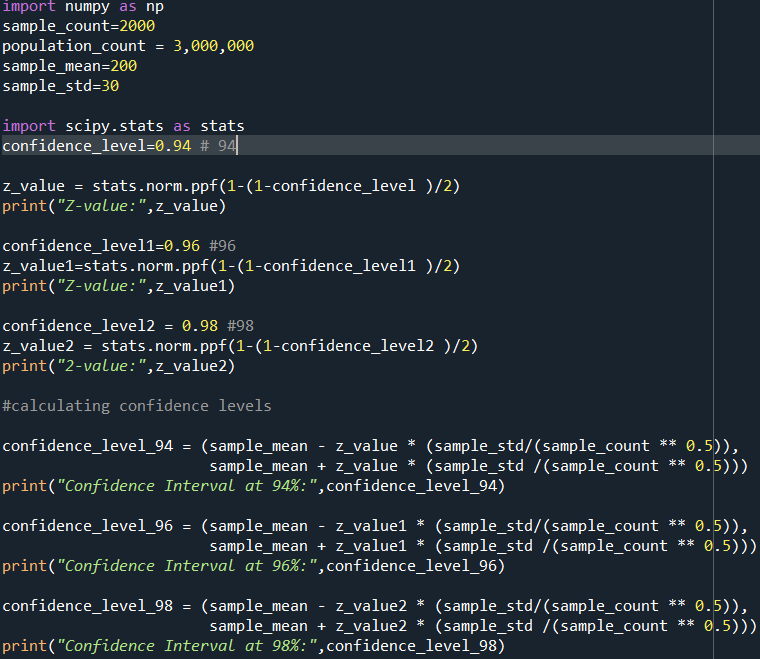
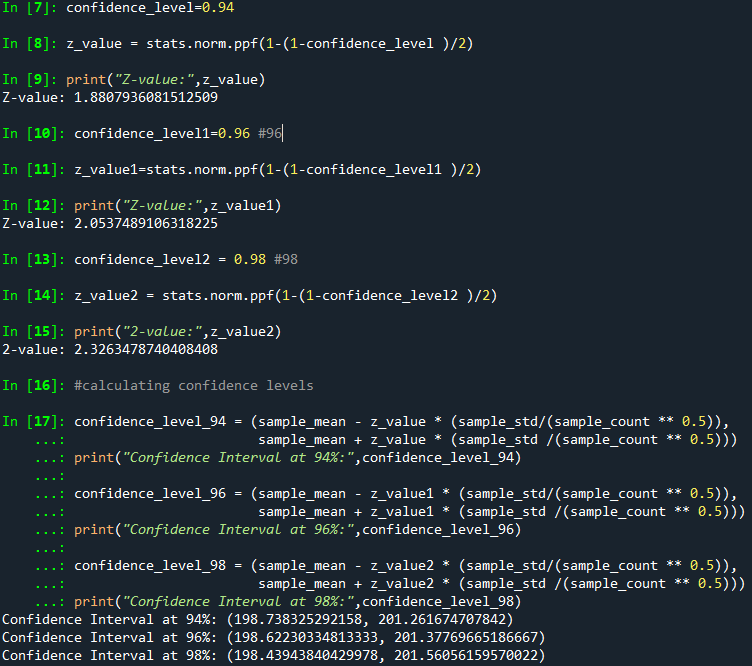
**Q10) Draw inferences about the following boxplot & histogram**



**Ans**: The data exhibits positive skewness, as indicated by the histogram (mode<median<mean).

**Ans:** 

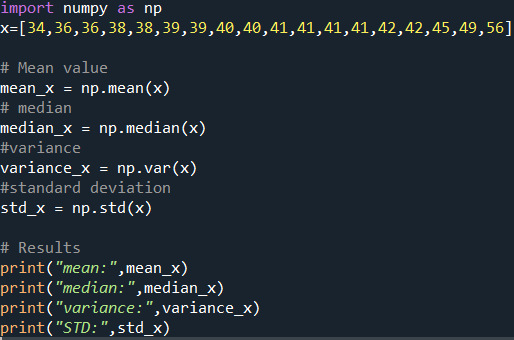
**We can infer that the entire outliers are on the top side of the box figure above.**

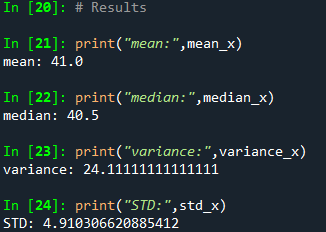
**Q11)** Suppose we want to estimate the averag weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate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





2)

**Mean**: The average of all student grades is the mean. In this instance, it is roughly 41.0. This suggests that the typical student grade is close to 41.0.

**Median**: When data is sorted in ascending order, the median is the middle value. The median in this situation is 40.5. According to this, 50% of the students have grades below 40.5 and 50% have grades above 40.5.

**Variance**: is a measure of how far apart data points are from the mean. A higher variance indicates a wider distribution of the data points. The difference in this instance is roughly 24.11. The greater variance suggests that student marks are somewhat dispersed from the mean and that there is some variation in grades.

**standard deviation**: Another indicator of how widely apart the data points are from the mean is the standard deviation. It is the variance's square root. The standard deviation in this situation is roughly 4.91. The higher standard deviation suggests that the student grades deviate from the mean with a discernible level of variability.

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

Ans: The data is said to be symmetrically distributed around the centre when the dataset's mean and median are equal. The data's skewness in this instance is zero. Skewness is a metric for the asymmetry in a real valued random variable's probability distribution. It informs us of the direction of the data's skewness in relation to the mean: either left (negative skewness) or right (positive skewness).

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

Ans: Positive skewness is present.

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

Ans: Negative skewness is present.

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

Ans:Positive Kurtosis value suggests that the dataset exhibits leptokurtic distribution, heavier tails, or both, when calculated as (>0), positive kurtosis values always have the kurtosis value.

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

Ans: A dataset with lighter tails, less/flatter peaks, or a platykurtic distribution has a negative score for kurtosis.When calculated as (0), the kurtosis value is always present for positive kurtosis values.

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



What can we say about the distribution of the data?

• The data is spread out throughout a range from about 2 to more than 18.

• The upper whisker length exceeding 18 points to the potential for outliers or more dispersed data points.

• The fact that the quartile range 1 begins exactly at 10 demonstrates that the data set's minimum value begins at 10, and the lower 25% of the data is centred around a value near 10.

What is nature of skewness of the data?

Ans:

• The average of positions Q1 and Q3 can be used to compute position Q2 (median).

• Q1 = 10

• Let's evaluate Q3 as 18.5 because it exceeds 18.

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

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

• The IQR represents the variation between Q3 and Q1.

• Q1 = 10

Assuming Q3 is 18.5 and IQR is Q3 - Q1 which is 18.5 - 10 which is around 8.5.

Q19) Comment on the below Boxplot visualizations?



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

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

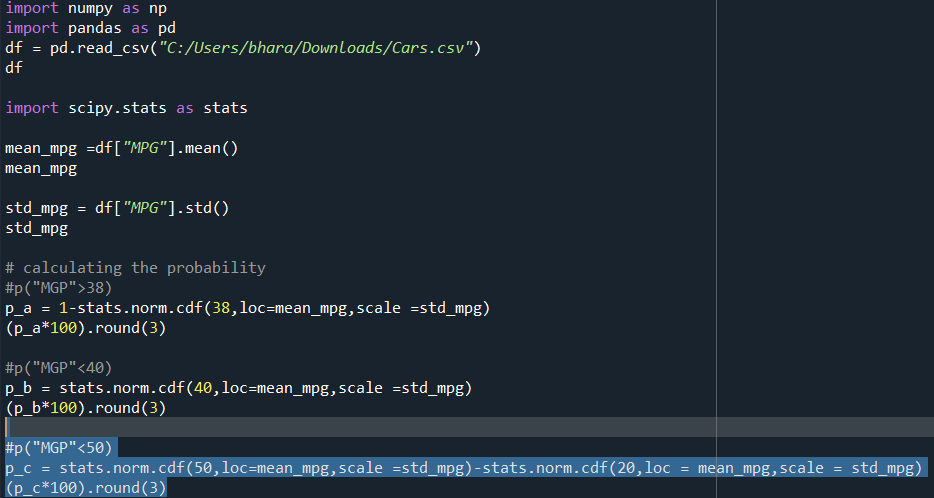
Data \_set: Cars.csv

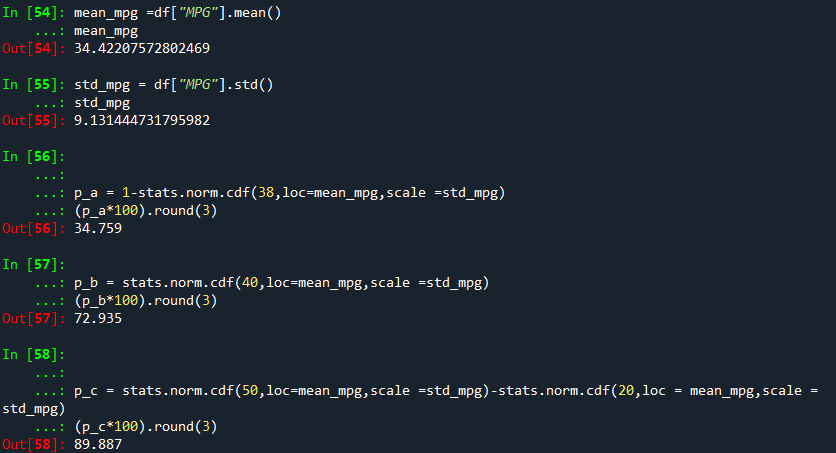
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

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

c. P (20<MPG<50)

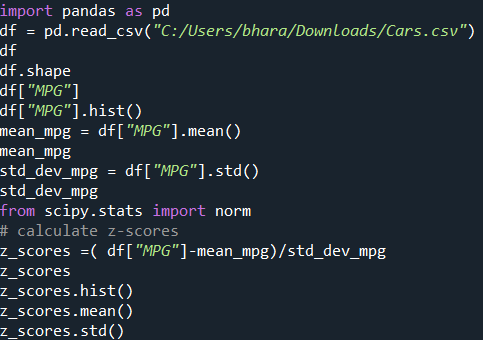


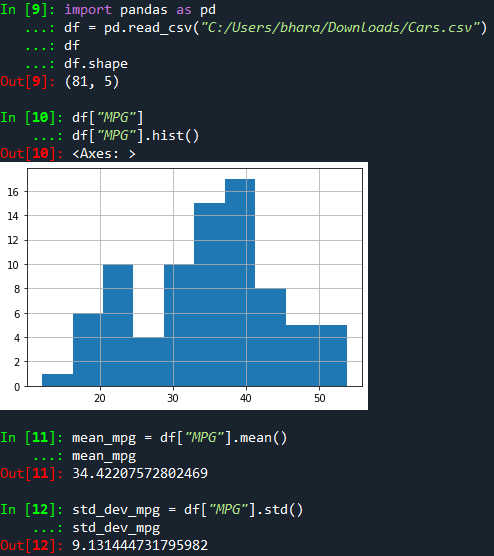
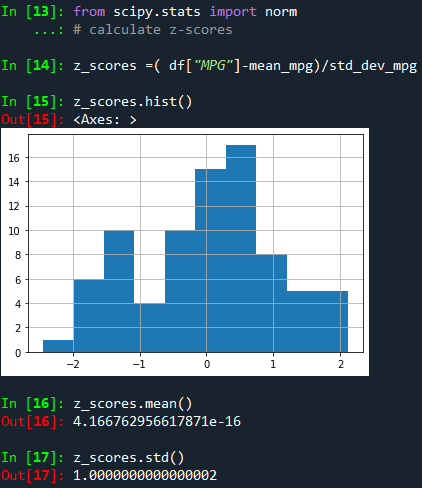


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

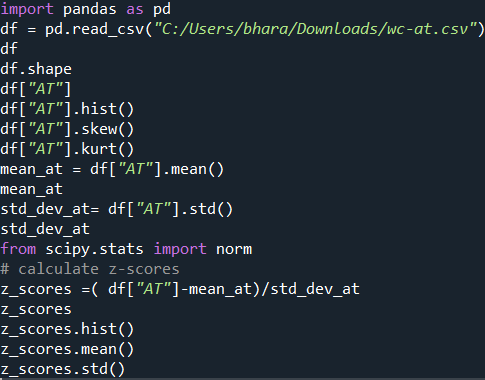


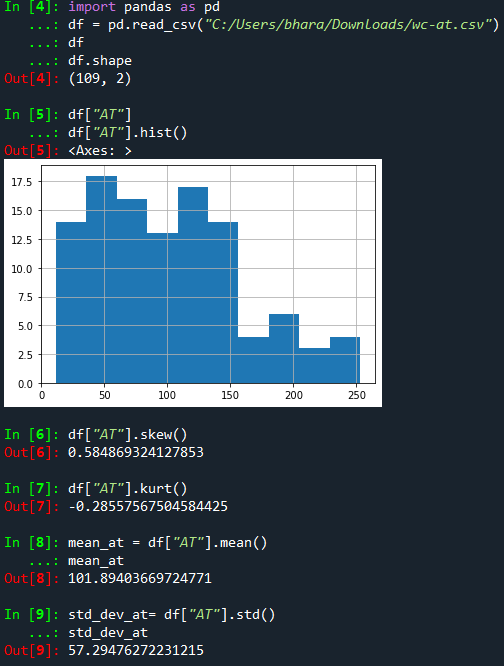
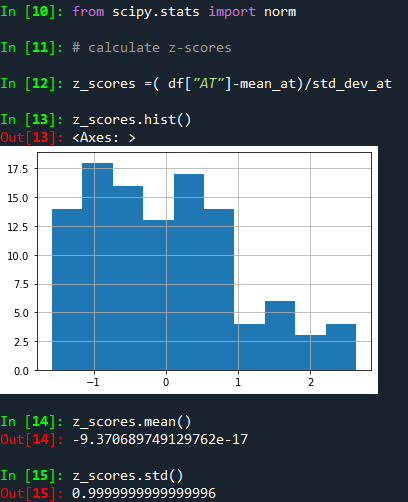


So after the Z-transformation we got the mean nearly equal to zero and standard deviation as equal to 1 so the “MPG” follows normal distribution

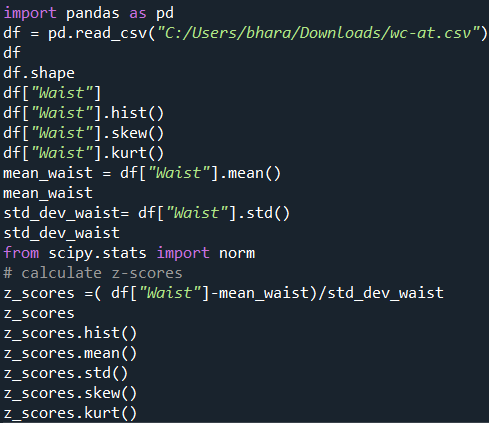
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

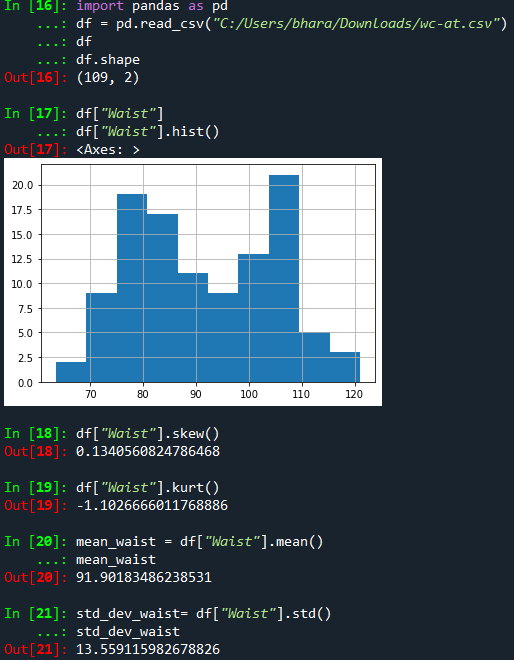
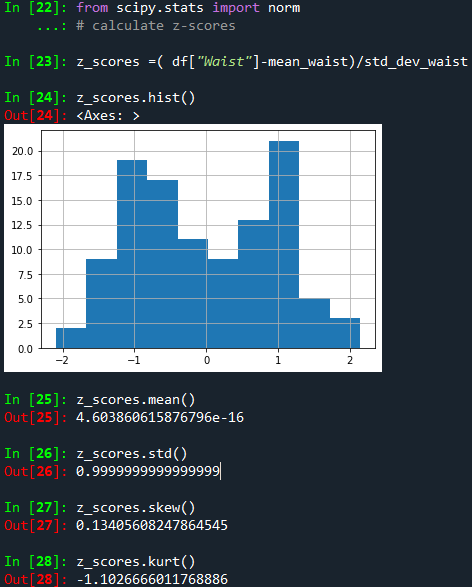
Dataset: wc-at.csv





For a standard normal distribution (one with mean 0 and standard deviation 1), the skewness is 0 and also excess kurtosis (kurtosis minus 3) is also 0. Therefore, "For a standard normal distribution (a normal distribution with mean 0 and standard deviation 1), the skewness is 0 and also excess kurtosis (kurtosis minus 3) is also 0."However, since the values for the aforementioned measures are not 0, the aforementioned variable is said to not follow "normal distribution" in this case.





After applying the z-transformation, the "Waist" column's mean and standard deviation are almost equal to zero and one, respectively, therefore we can conclude that this does not match the normal distribution. As you can see in the above image, we are also not receiving 0,1 values for skewness and kurtosis, but we can still use them as powerful and insightful insights for our understanding.The column is obviously not following a normal distribution, then.

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

We can utilise the standard normal distribution (Z-distribution) and its percentiles to get the Z-scores for various intervals. The following equation can be used to determine the Z-scores for a given confidence level:

Z = Z/2, where Z/2 is the Z-score that corresponds to the chosen alpha level of significance.

**90% Confidence Interval:**

In this instance, the confidence level is 90%, and alpha = 1-0.90(90%) = 0.10 as a result. When we check up the Z-score for the cumulative probability of 0.95(1-0.05) in the standard normal distribution table Z0.05 1.645, we see that half of this alpha is 0.05.

**94% Confidence Band:**

When the confidence level is 94%, alpha equals 1-0.94(94%), which equals 0.06 in this case.0.03 is the midpoint of this alpha.Whenever we check up the Z-score for the cumulative probability of 0.97(1-0.03) in the table of the standard normal distribution, Z0.05 1.880, it is the result.

**Interval of Confidence at 60%:**Since the level of confidence in this situation is 60%, alpha equals 1-0.60(60%) =0.40. This alpha's half is 0.20.Whenever we look up the Z-score for the cumulative probability of 0.80(1-0.20) in the table of the ordinary normal distribution, Z0.05 1.842, we find it to be 1.84.

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

When we get the required sample size (30), we will compute the t scores. Instead of the conventional normal distribution, we would employ the t-distribution.

Where t/2 is the t-score matching to the desired significance level alpha, the formula is t = t/2, df.

And df = n-1(sample size - 1), where n is the sample size.

**The confidence level is 96%:**

Which means that alpha = 1-0.96 = 0.04 and that the probability is 96% . Thus , 0.04/2 = 0.02 is the value of half of alpha. For a sample size of 25, the degrees of freedom (df) are 25-1 = 24. Therefore , we must now get the t-score, which equals 2.398 in the t-distribution table, that corresponds to the cumulative probability of 0.980(1-0.02) and df = 24.

**Confidence Level of 95%:**

Since the confidence level is 95%, alpha = 1-0.95 = 0.05.Therefore, 0.05/2 = 0.025 is the value of alpha at half.For a sample size of 25, the degrees of freedom (df) are 25-1 = 24.Therefore, we must now find the t-score, 2.064, in the t-distribution table that corresponds to the cumulative probability of 0.975(1-0.025) and df = 24.

**99% Confidence Level:**

With a 99% confidence level, alpha equals 1-0.99 = 0.01.0.01/2 = 0.005 is the value of half of alpha.For a sample size of 25, the degrees of freedom (df) are 25-1 = 24.Therefore, we must now find the t-score, which equals 2.797, in the t-distribution table for the cumulative probability of 0.995(1-0.005) and df = 24.

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

