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

**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) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval/Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

**Answer)** Sample Space = {HHH, TTT, HHT, THH, TTH, HTT, HTH, THT}

Probability of getting 2 heads & 1 tail ={HHT,THH,HTH}

= 3/8

= 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

**Answer)** Sample Space = {(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. Probability that sum is equal to 1 = 0
2. Probability that sum is less than or equal to 4 = 6/36

= 1/6

= 0.166

1. Probability that 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?

**Answer)** Total number of balls = 2+3+2

= 7

Probability of drawing 2 balls out of 7 = 7*c*2

= 7!/2!\*(7-2)!

= 21

There are 5 balls which are not blue.

Therefore the probability that none of the balls drawn is blue = 5*c*2/7*c*2

= 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

**Answer)** E(x) = ∑x\*P(x)

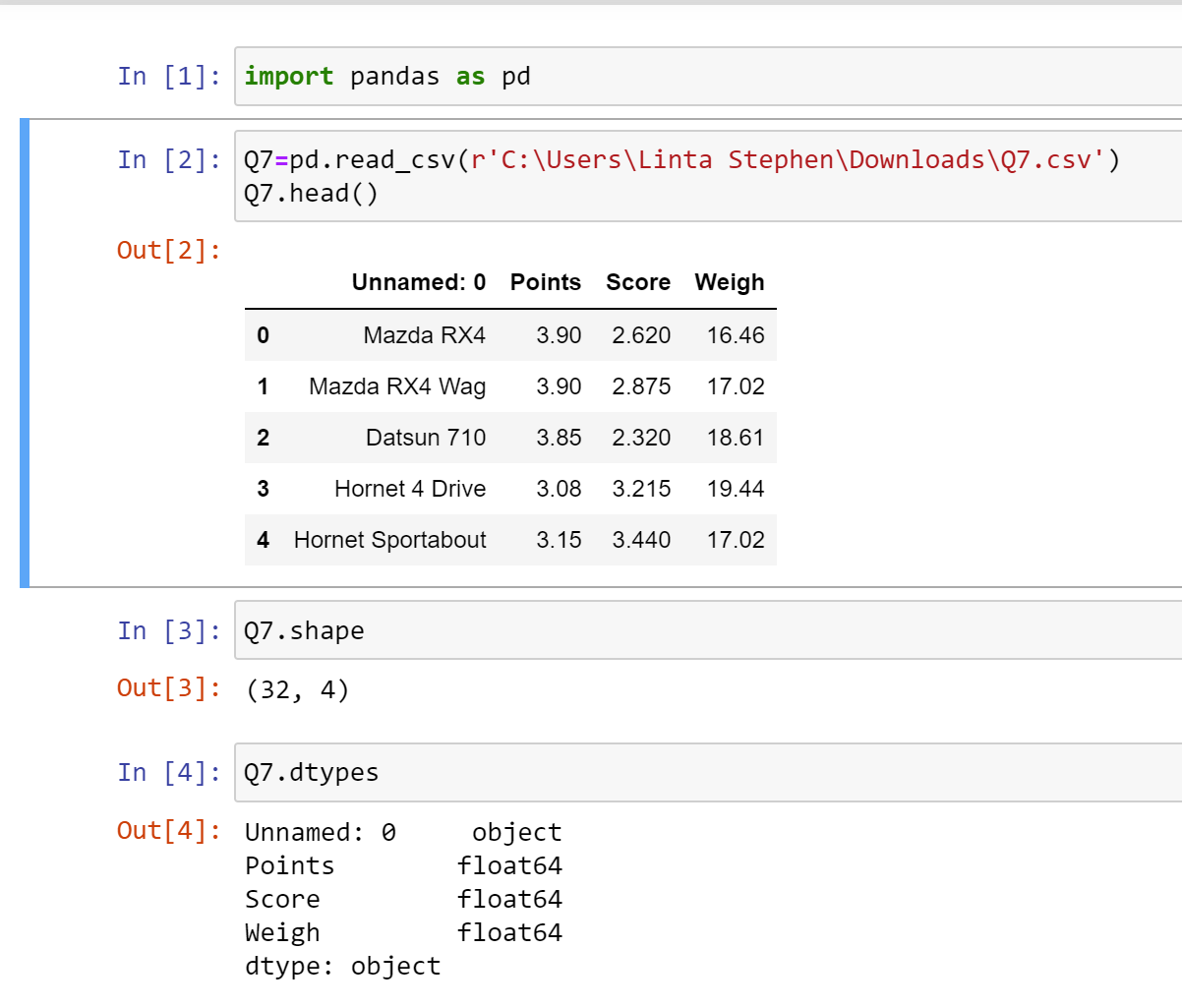
The expected number of candies for a randomly selected child = 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

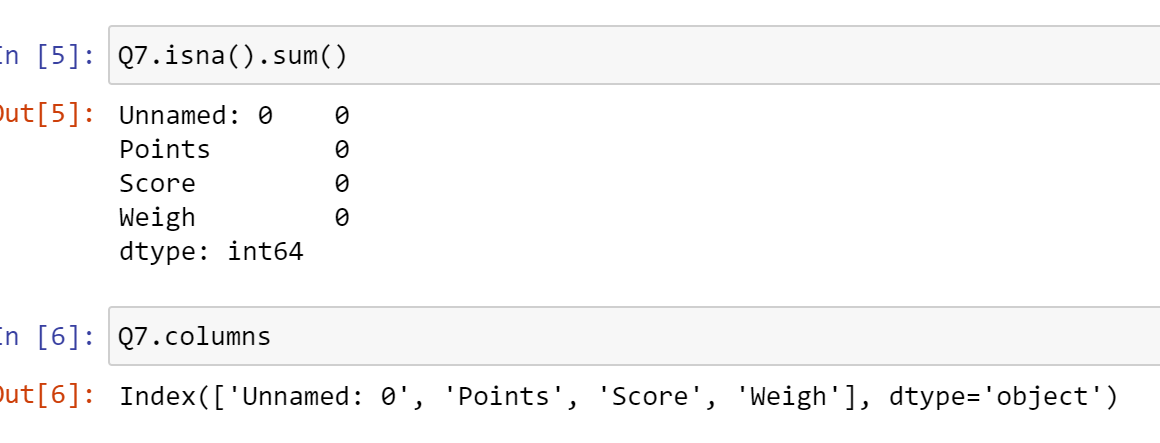
= 3.09

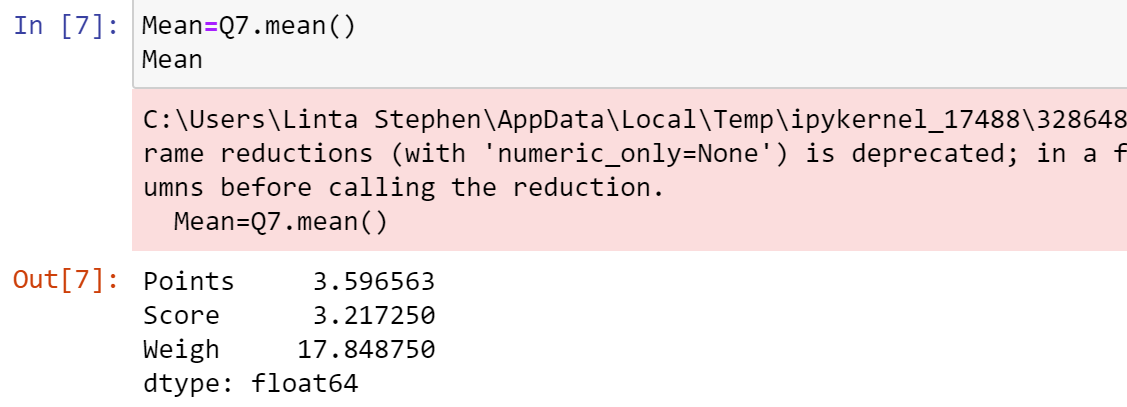
**Q7)** Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

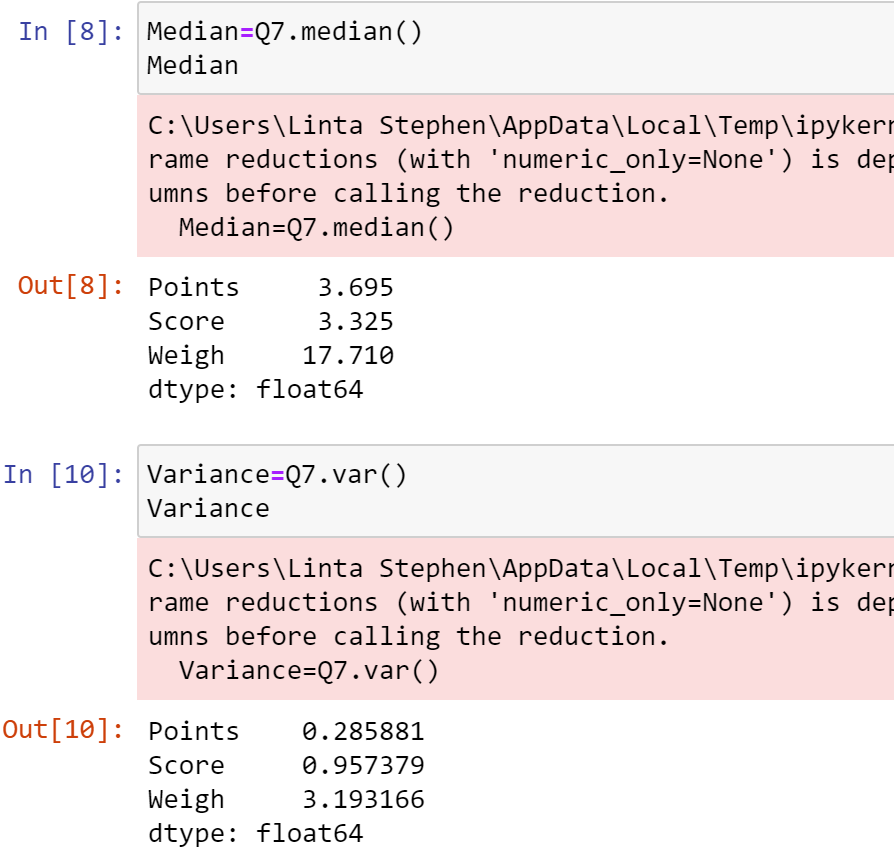
* For Points,Score,Weigh>

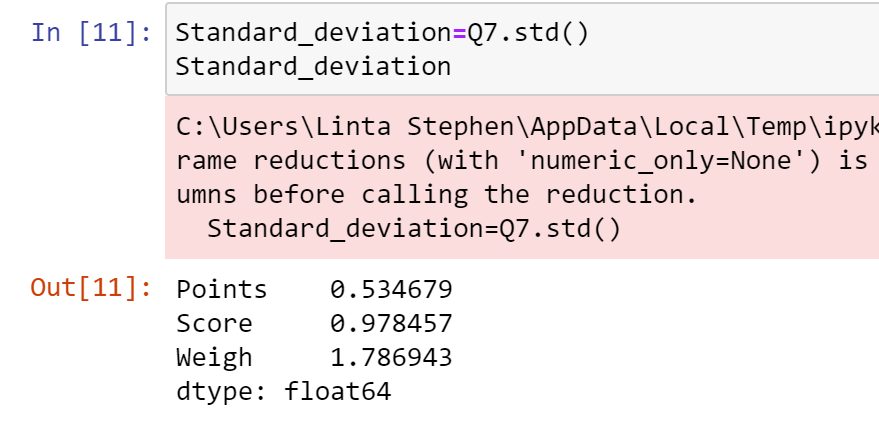
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

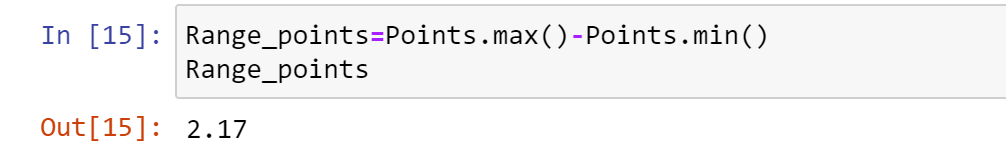
**Answer)** 

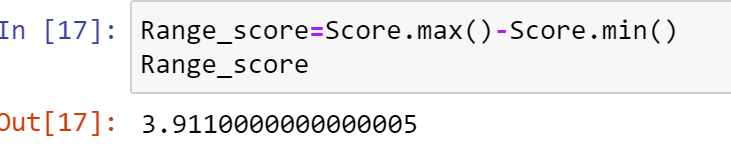


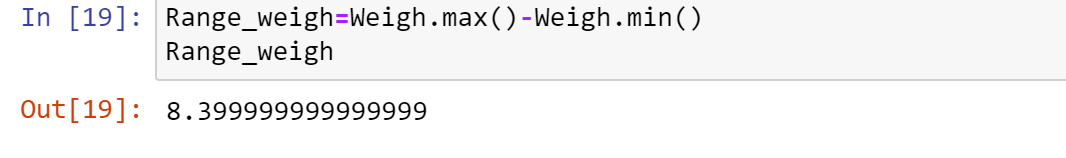


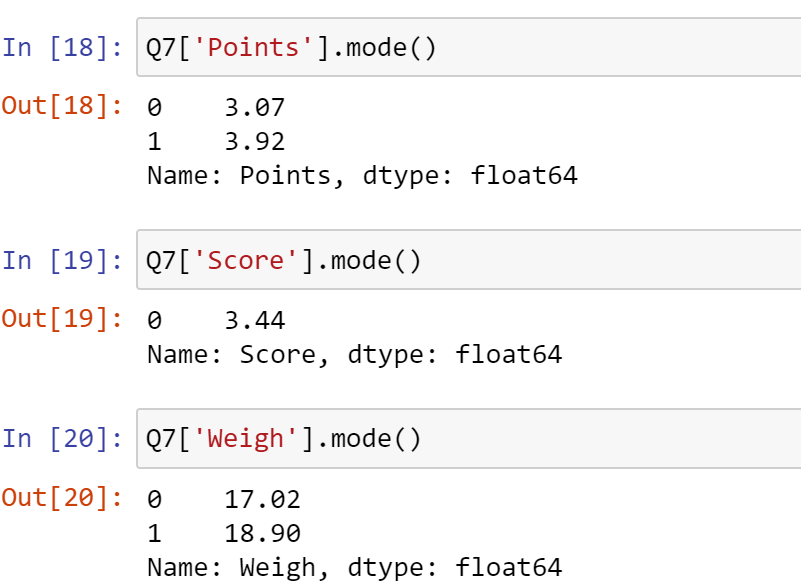












**Inference:** In points, score & weigh, the difference between mean & median values are less. Points & Weigh contain two mode. On an average each data points are less variated from the mean value.

**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?

**Answer)** E(x)= ∑x\*P(x)

|  |  |  |
| --- | --- | --- |
| X | P(X) | X\*P(X) |
| 108 | 1/9 | 12 |
| 110 | 1/9 | 12.22 |
| 123 | 1/9 | 13.66 |
| 134 | 1/9 | 14.88 |
| 135 | 1/9 | 15 |
| 145 | 1/9 | 16.11 |
| 167 | 1/9 | 18.55 |
| 187 | 1/9 | 20.77 |
| 199 | 1/9 | 22.11 |

E(X)= 12+12.22+13.66+14.88+15+16.11+18.55+20.77+22.11

= 145.3

Therefore the expected value of weight of that person is 145.3.

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

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

**Answer) (Inference)**

1. Skewness\_Speed: The skewness of the 'speed' column is approximately -0.114. Skewness measures the asymmetry of the distribution. A negative skewness value indicates that the distribution is slightly skewed to the left (tail is longer on the left side).

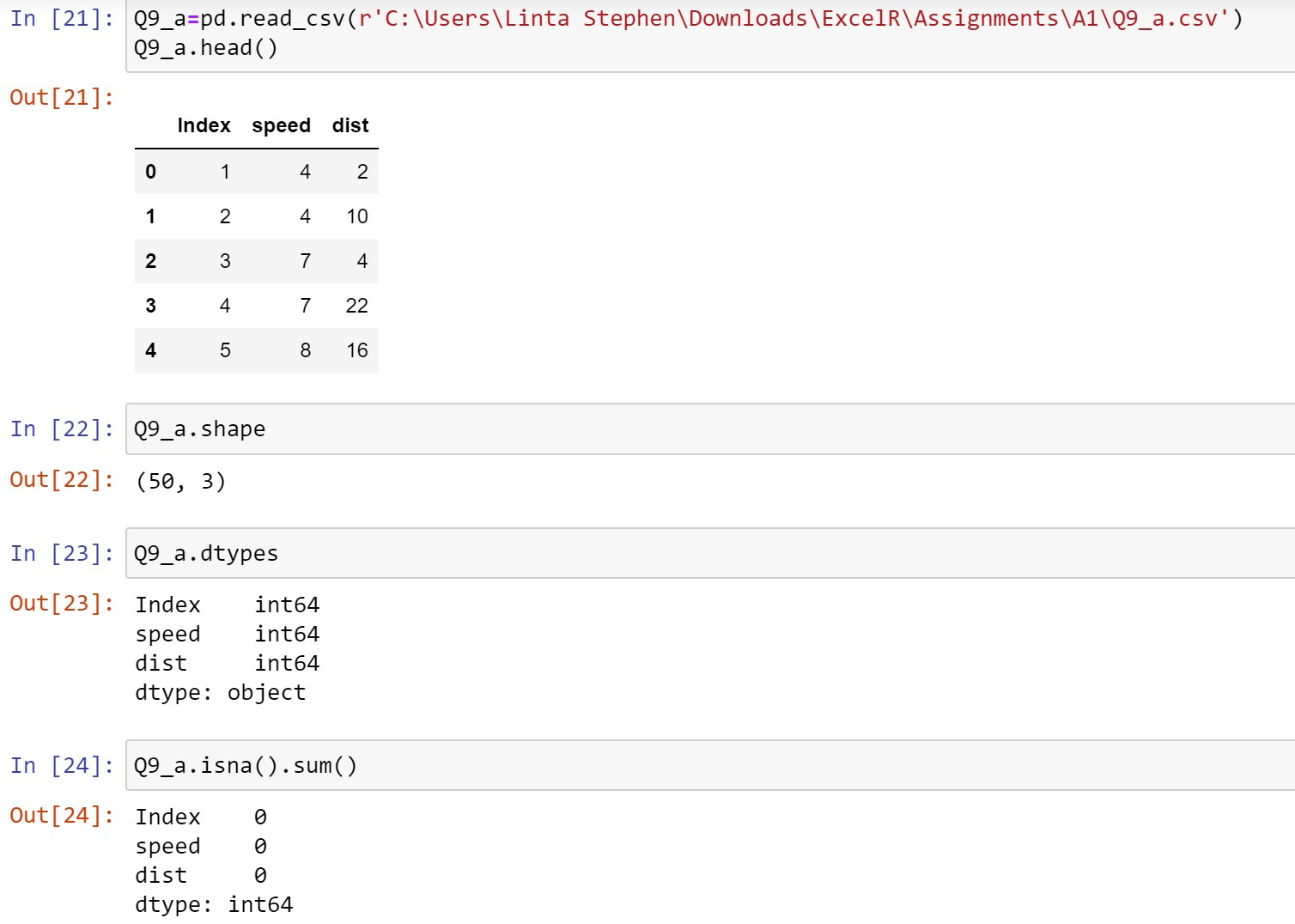
2. Skewness\_Distance: The skewness of the 'dist' column is approximately 0.782. This positive skewness value suggests that the distribution of the 'dist' column is skewed to the right (tail is longer on the right side).

3. Kurtosis\_Speed: The kurtosis of the 'speed' column is approximately -0.577. Kurtosis measures the peakedness of the distribution. A negative kurtosis value indicates a platykurtic distribution, which means the distribution has lighter tails and a flatter peak compared to a normal distribution.

4. Kurtosis\_Distance: The kurtosis of the 'dist' column is approximately 0.248. This positive kurtosis value suggests a mesokurtic distribution, which means the distribution has a similar peakedness and tail behavior compared to a normal distribution.

Therefore the 'speed' column appears to have a slightly left-skewed distribution with a relatively flat peak, while the 'dist' column shows a slightly right-skewed distribution with a similar peakedness compared to a normal distribution.

**Code**





SP and Weight (WT) (Use Q9\_b.csv)

**Answer) (Inference)**

1. Skewness\_SP: The skewness of the 'SP' column is approximately 1.581. This positive skewness value indicates that the distribution of the 'SP' column is skewed to the right (tail is longer on the right side).

2. Skewness\_Weight: The skewness of the 'WT' column is approximately -0.603. This negative skewness value suggests that the distribution of the 'WT' column is skewed to the left (tail is longer on the left side).

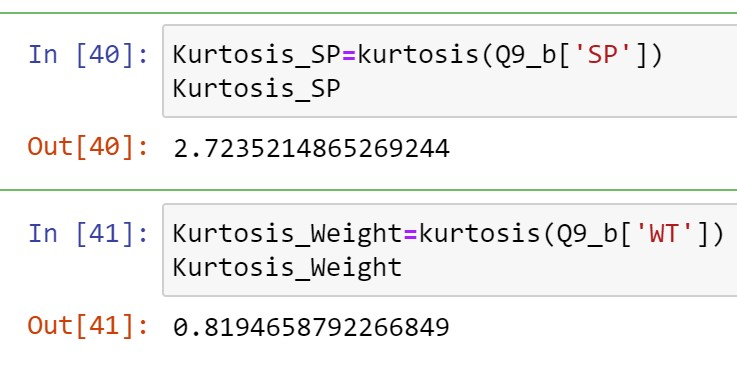
3. Kurtosis\_SP: The kurtosis of the 'SP' column is approximately 2.724. Kurtosis measures the peakedness of the distribution. A positive kurtosis value indicates a leptokurtic distribution, which means the distribution has heavier tails and a sharper peak compared to a normal distribution.

4. Kurtosis\_Weight: The kurtosis of the 'WT' column is approximately 0.819. This positive kurtosis value suggests a mesokurtic distribution, which means thedistribution has a similar peakedness and tail behavior compared to a normal distribution.

Therefore, the 'SP' column exhibits a right-skewed distribution with a relatively sharp peak, while the 'WT' column shows a left-skewed distribution with a similar peakedness compared to a normal distribution.

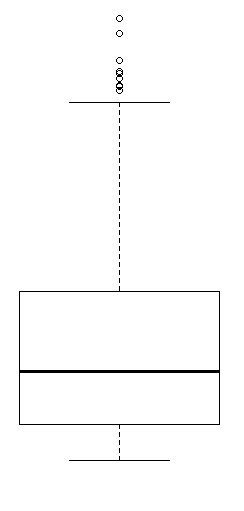
**Code**





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





**Answer)**

**Histogram:**

* The data is right skewed.
* More than 50% of the weight is between 50 to 150.
* And most weight is in between 50 to 100.

**Bar chart:**

* Data is right skewed.
* The outliers are at upper side.

**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?

**Answer)** Given: n=2000, N=3000000, Sample mean=200, s=30

* *For 94% confidence interval*
* α=1-(94/100)

= 0.06

* Criticial probability=1-(α/2)

=1-(0.06/2)

=1-0.03

=0.97

* Degrees of freedom=n-1

=2000-1

=1999

* t value=2.174
* Confidence interval=sample mean ± t value\*(s/√n)

= 200 ± 2.174\*(30/√2000)

= 200 ± 2.174\*0.6708

= 200 ± 1.456

= **(198.544, 201.456)**

* *For 98% confidence interval*
* α=1-(98/100)

= 0.02

* Criticial probability=1-(α/2)

=1-(0.02/2)

=1-0.01

=0.99

* Degrees of freedom=n-1

=2000-1

=1999

* t value=2.602
* Confidence interval=sample mean ± t value\*(s/√n)

= 200 ± 2.602\*(30/√2000)

= 200 ± 2.602\*0.6708

= 200 ± 1.7454

= **(198.2546 , 201.7454)**

* *For 96% confidence interval*
* α=1-(96/100)

= 0.04

* Criticial probability=1-(α/2)

=1-(0.04/2)

=1-0.02

=0.98

* Degrees of freedom=n-1

=2000-1

=1999

* t value=2.326
* Confidence interval=sample mean ± t value\*(s/√n)

= 200 ± 2.326\*(30/√2000)

= 200 ± 2.326\*0.6708

= 200 ± 1.560

= **(198.44, 201.56)**

**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?

**Answer) (1)**

* Mean=(34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+4 9+56)/18

=738/18

=41

* Median

Arrange in order: 34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

Since this contain 18 numbers, which is an even value , there contain 2 median ie 40 & 41

Therefore median= (40+41)/2

= 81/2

= 40.5

* Vaiance= 1/n[∑(xi-mean)2]

Squared difference:

(34 - 41)^2 = 49

(36 - 41)^2 = 25

(36 - 41)^2 = 25

(38 - 41)^2 = 9

(38 - 41)^2 = 9

(39 - 41)^2 = 4

(39 - 41)^2 = 4

(40 - 41)^2 = 1

(40 - 41)^2 = 1

(41 - 41)^2 = 0

(41 - 41)^2 = 0

(41 - 41)^2 = 0

(41 - 41)^2 = 0

(42 - 41)^2 = 1

(42 - 41)^2 = 1

(45 - 41)^2 = 16

(49 - 41)^2 = 64

(56 - 41)^2 = 225

Sum of squared difference= 49+25+25+9+9+4+4+1+1+1+1+16+64+225

= 434

Therefore variance= 434/18

= 24.11

* Standard deviation= √variance

= √24.11

= 4.910

**(2)**

* The mean score is 41. This indicates that, on average, the student scored around 41 marks in the tests.
* The median score is 40.5. This implies that half of the test scores are below 40.5, and the other half are above 40.5.
* The variance is 24.11, indicating that there is some variability or spread in the scores.
* The standard deviation is 4.910, which further confirms the dispersion or spread in the scores.
* Overall, the student's marks range from 34 to 56, with a relatively average performance centered around 41. The scores show some variability.

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

**Answer)** When the mean and median of a dataset are equal, it implies that the data distribution is symmetric. In such a case, the skewness of the data is zero.

Skewness is a measure of the asymmetry of a probability distribution. It quantifies the extent to which the data values are skewed or pulled towards one tail of the distribution. Skewness can be positive, negative, or zero.

* Zero skewness: When the distribution is symmetric, meaning that it is evenly distributed around its center, the mean and median are equal, and the skewness is zero.

So, when the mean and median are equal, it indicates that the data is symmetrically distributed, and there is no skewness present in the data**.**

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

**Answer)** When the mean is greater than the median, it indicates a positive skewness in the data distribution**.**

* Positive skewness: If the distribution has a long tail on the right side (the right tail is longer), it is said to be positively skewed. The mean will be greater than the median in this case.

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

**Answer)** When the median is greater than the mean, it indicates a negative skewness in the data distribution.

* Negative skewness: If the distribution has a long tail on the left side (the left tail is longer), it is said to be negatively skewed. The mean will be less than the median in this case.

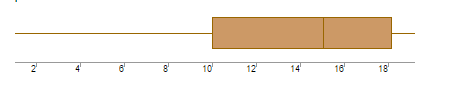
**Q16)** What does positive kurtosis value indicates for a data ?

**Answer)** A positive kurtosis value indicates that a data has heavier tails and a sharper peak compared to a normal distribution. It suggests that the data has more extreme values (outliers) and a higher concentration of values around the mean.

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

**Answer)** A negative kurtosis value indicates that a data has lighter tails and a flatter peak compared to a normal distribution. It suggests that the data has fewer extreme values (outliers) and a more dispersed spread of values.

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



What can we say about the distribution of the data?

**Answer)** Boxplot is normally distributed and median is towards higher value.

What is nature of skewness of the data?

**Answer)** The data is left skewed.

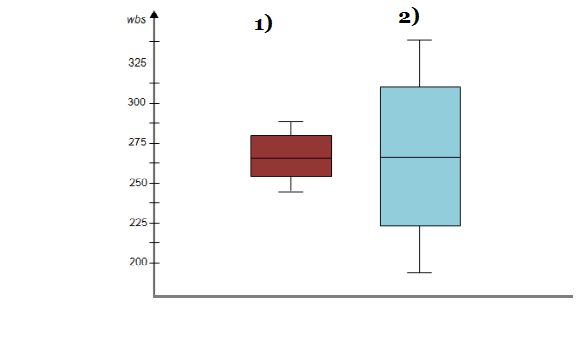
What will be the IQR of the data (approximately)?

**Answer)** IQR= Q3 -Q1

= 18-10

=8

**Q19)** Comment on the below Boxplot visualizations?



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

**Answer)**

* Both the data does not contain any outliers.
* Both have same median value.
* They are symmetric ie Normally distributed and zero skewness.

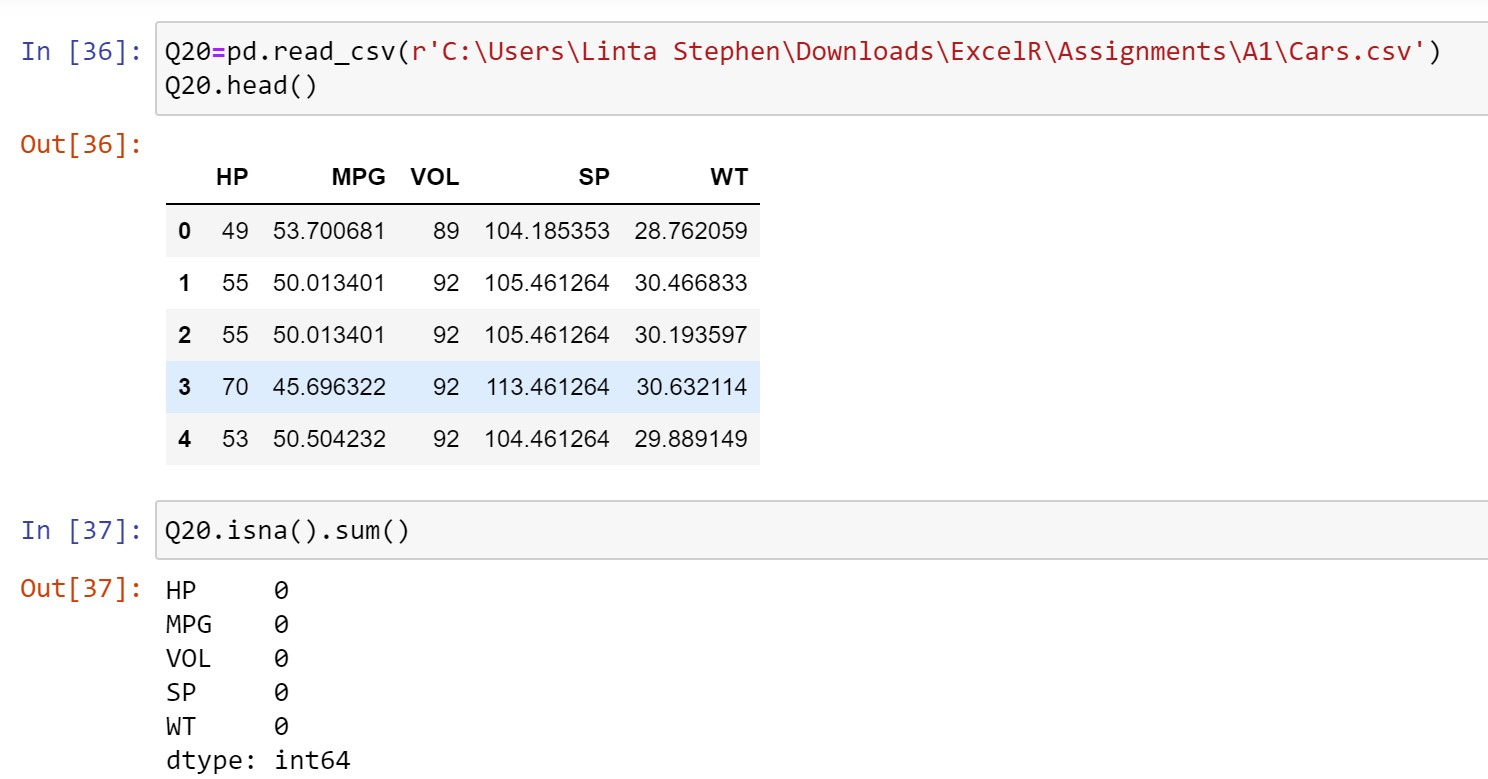
**Q 20)** Calculate probability from the given dataset for the below cases

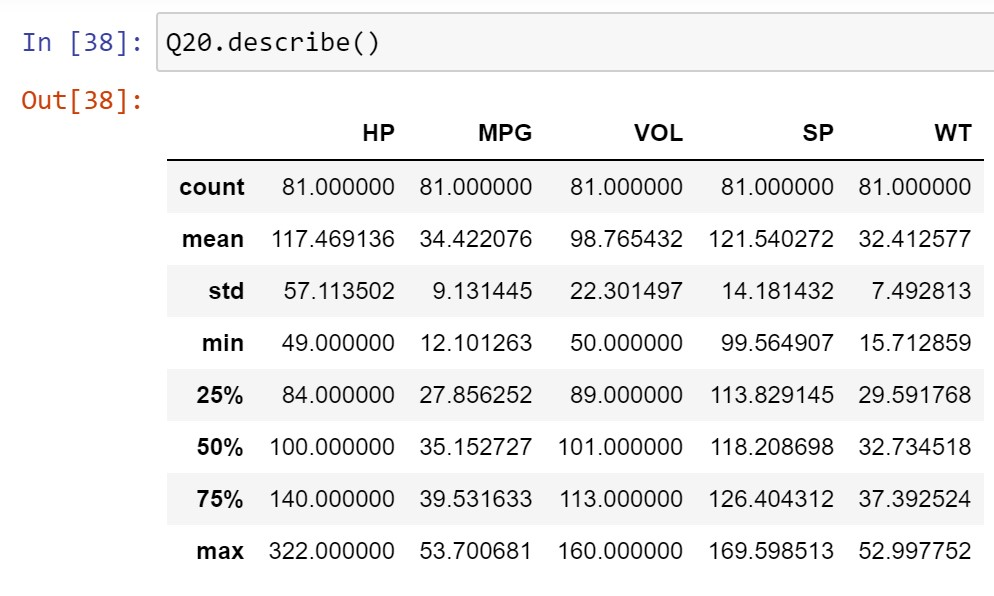
Data \_set: Cars.csv

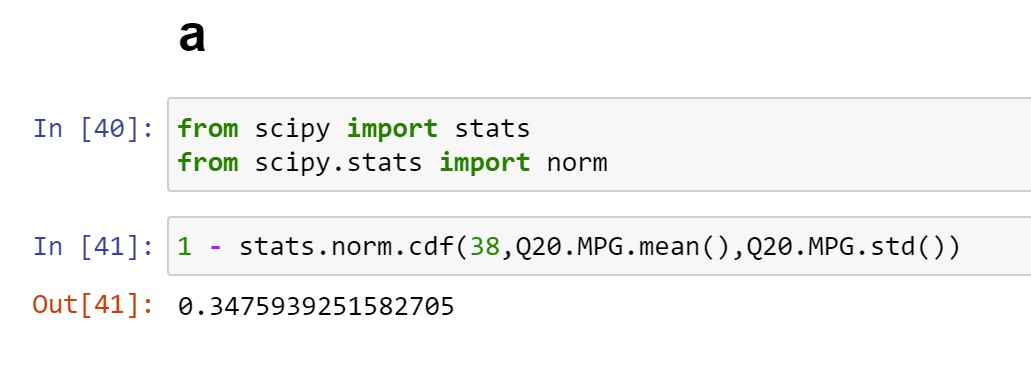
Calculate the probability of MPG of Cars for the below cases.

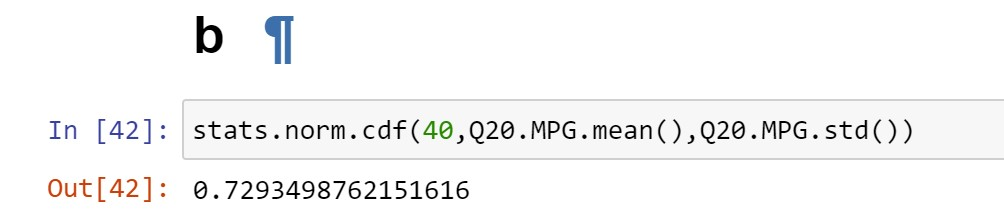
MPG <- Cars $ MPG

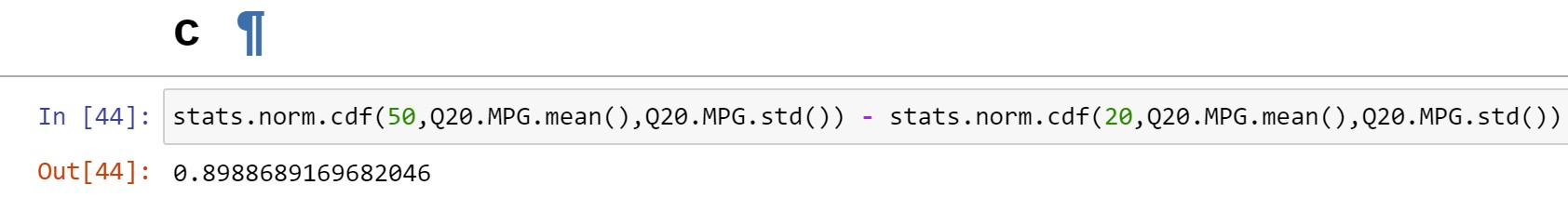
* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)







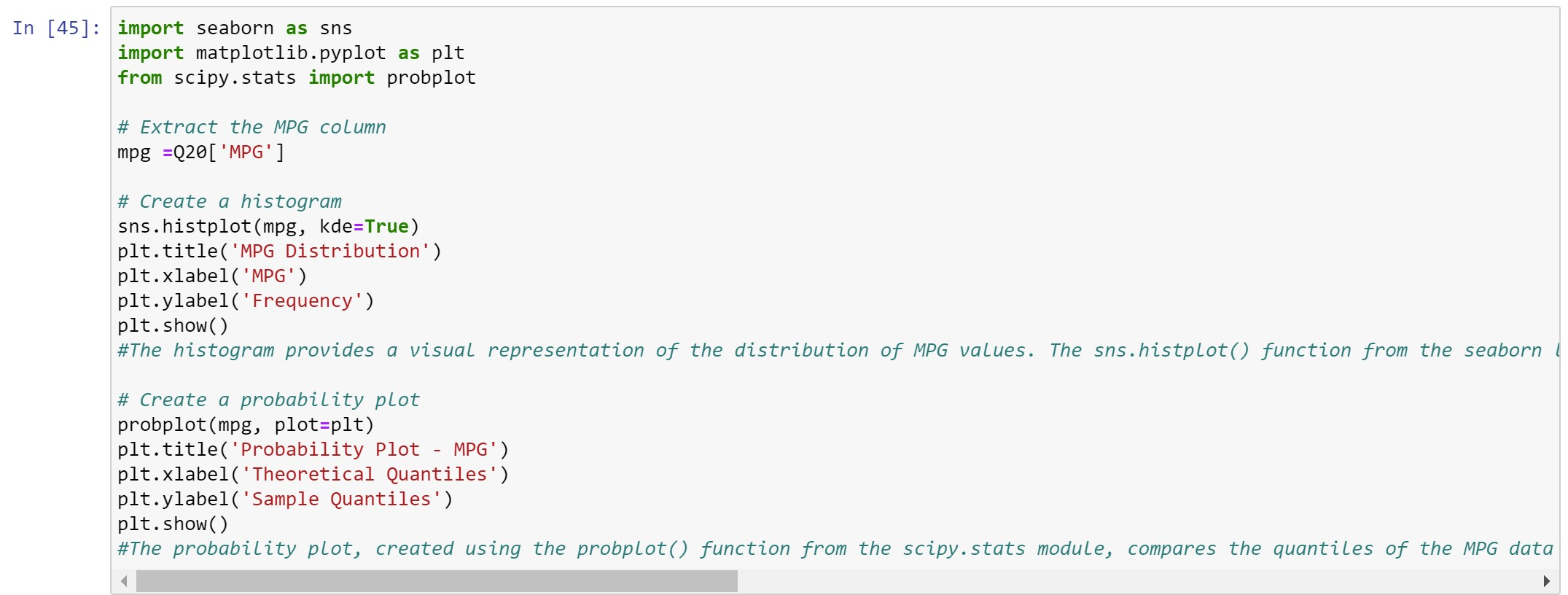




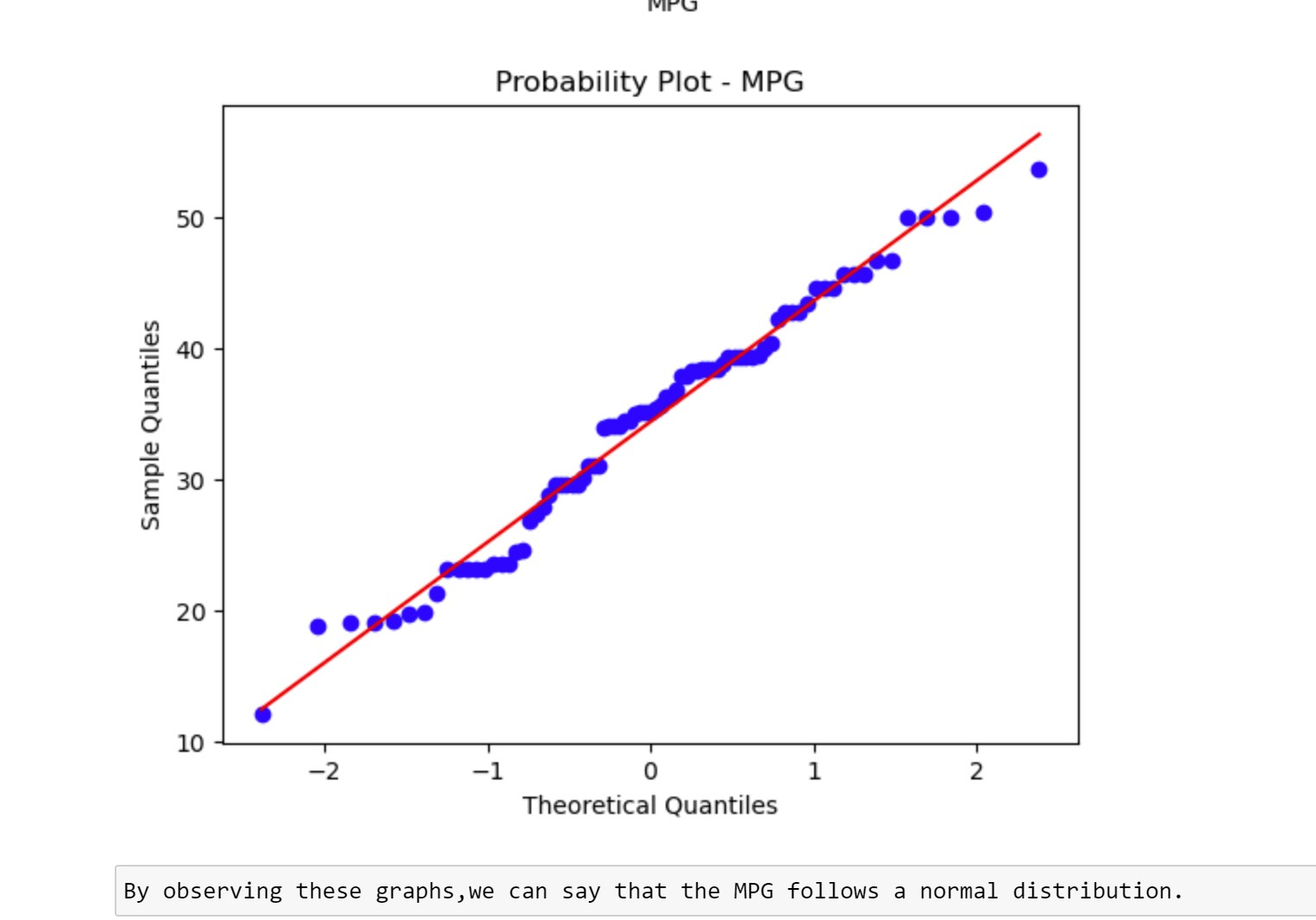
**Q 21)** Check whether the data follows normal distribution

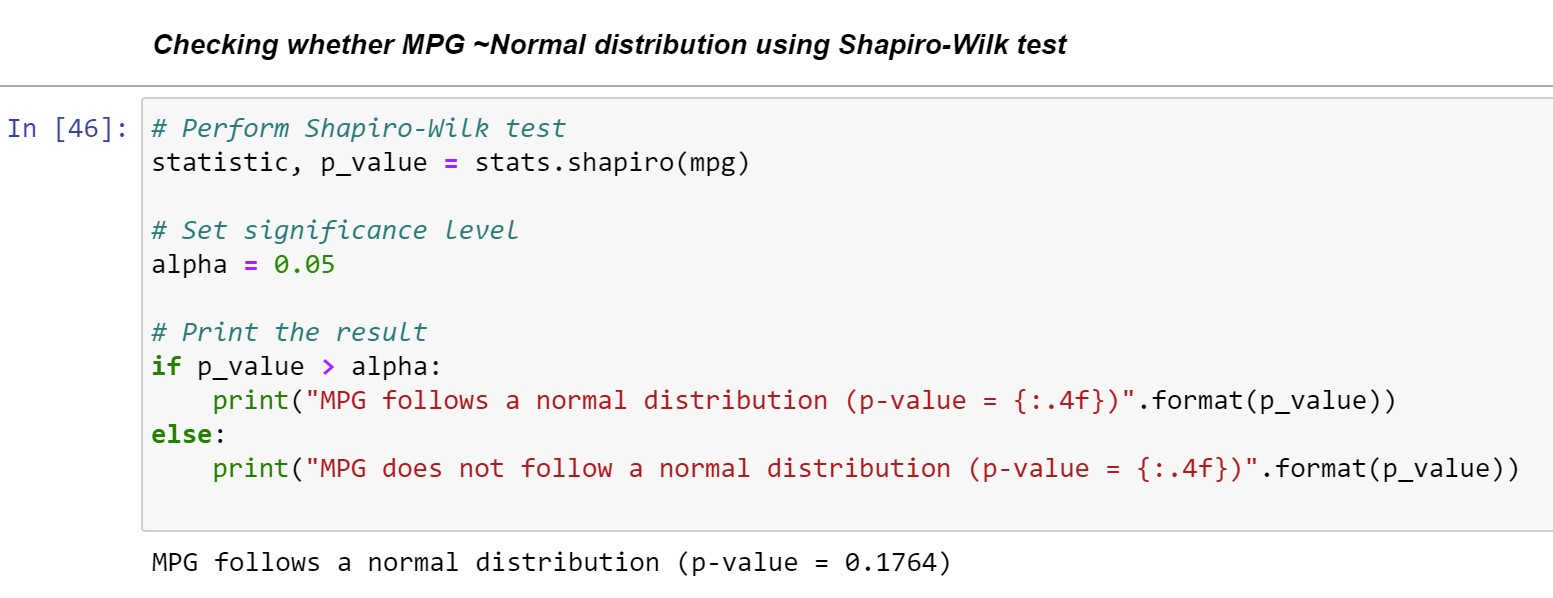
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv







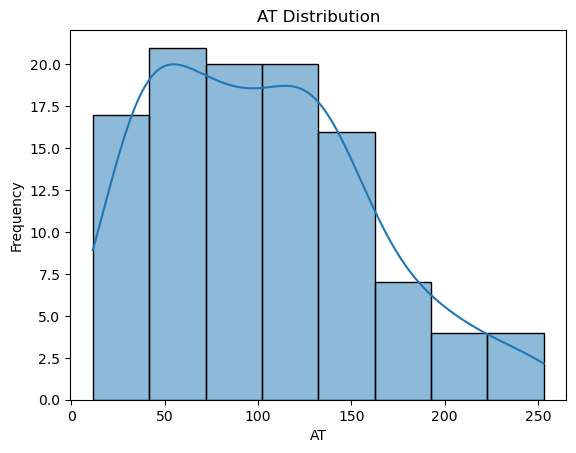


1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

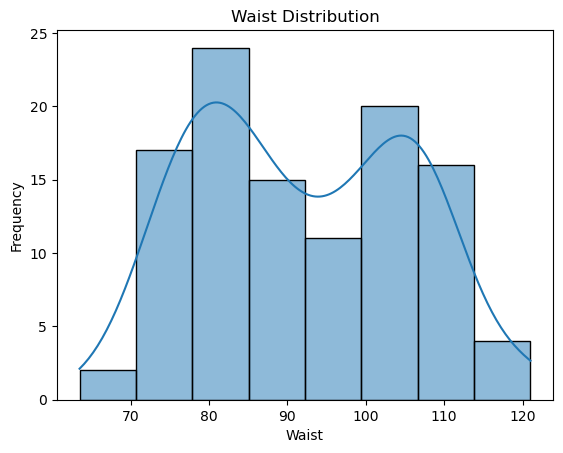
Dataset: wc-at.csv

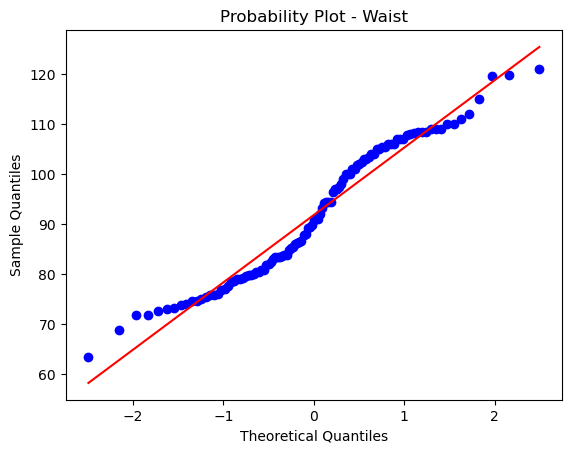


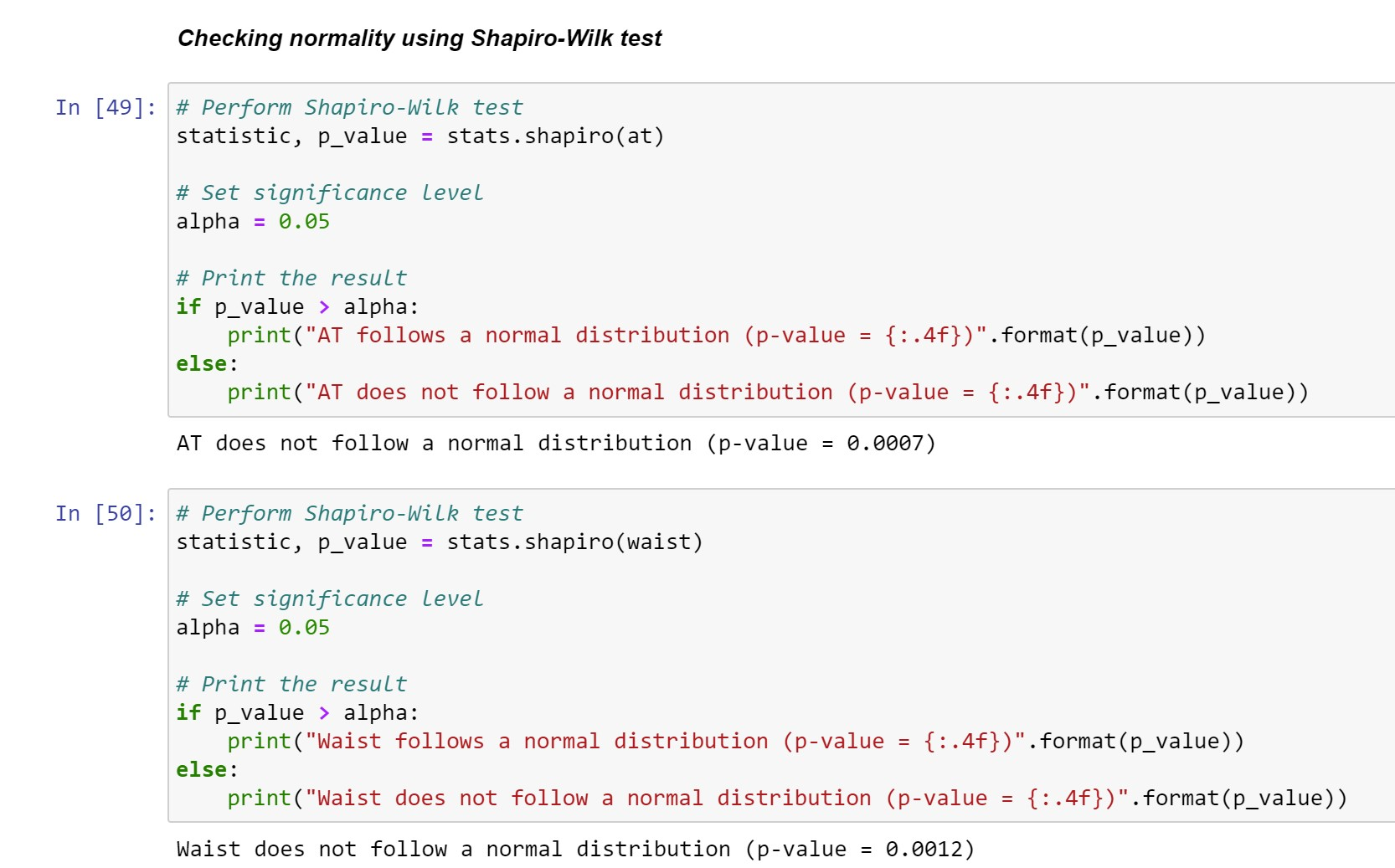






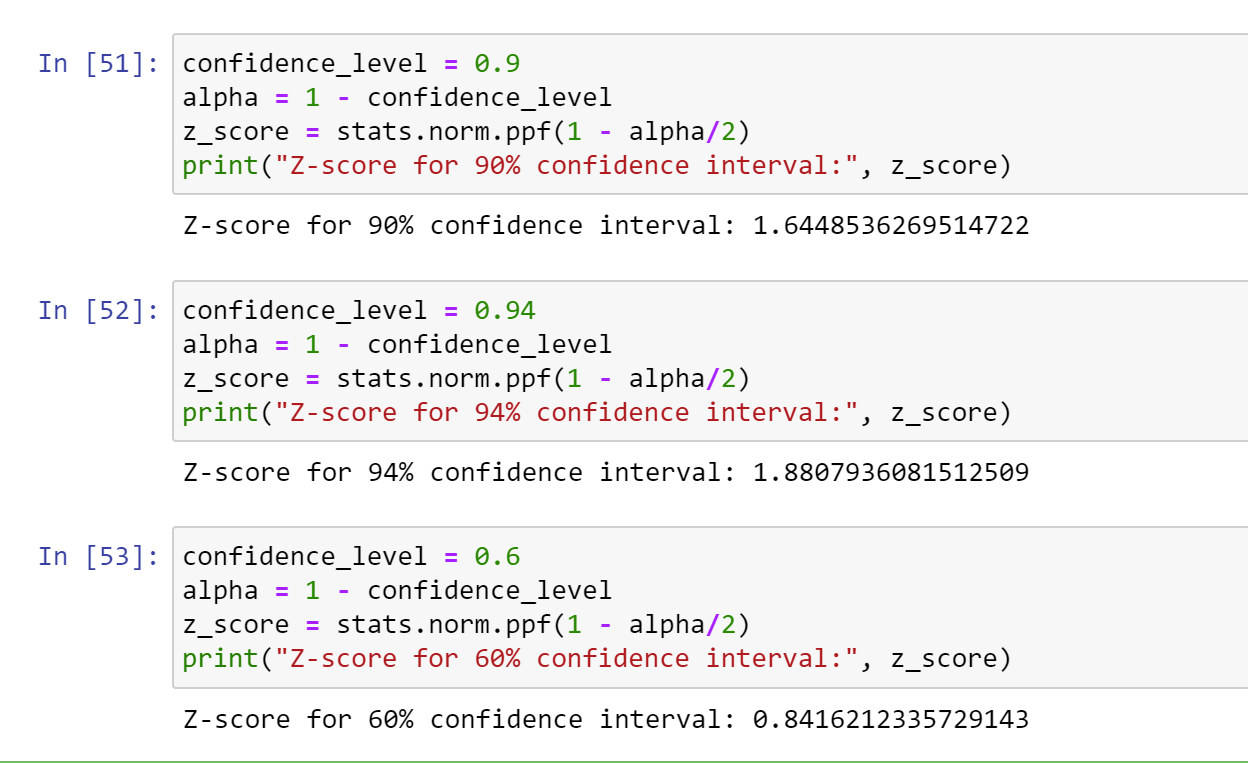






**Q 22)** Calculate the Z scores of 90% confidence interval,94% confidence

interval, 60% confidence interval



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



**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

