|  |  |
| --- | --- |
| 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 | ordinal |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | interval |
| Sales Figures | ratio |
| Blood Group | nominal |
| Time Of Day | ordinal |
| Time on a Clock with Hands | interval |
| Number of Children | nominal |
| 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 :

N = 3

distribution = (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT)

2 head and one tail =>HHT, HTH, THH

Total = 8

Probability of two heads and one tail = 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

Answer :

Distribution = (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)

Total = 36

1. There is no distribution for one dice to fall as zero, thereby making the sum of two dice as one,

So, Probability of sum is equal to one --> 0/36 => 0

1. Sum Less than or equal to 4 = (1&1, 1&2, 1&3, 2&1, 2&2, 3&1)

So, Probability of sum less than or equal to 4 --> 6/36 => 1/6

1. Sum divisible by 2 and 3 = (1&5, 2&4, 3&3, 4&2, 5&1, 6&6)

So, Probability of sum divisible by 2 and 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 balls = 7

Probability of 2 balls which are not blue = 5C2 / 7C2

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 :

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) ]

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 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.

**Use Q7.csv file**

**Answer:**

q7 = pd.read\_csv("Q7.csv")

q7.mean()

q7.median()

q7.mode(numeric\_only=True)

q7.std()

q7.var()

q7.describe()

q7.corr()

range\_points=(max(q7['Points']-min(q7['Points'])))

range\_score=(max(q7['Score']-min(q7['Score'])))

range\_weigh=(max(q7['Weigh']-min(q7['Weigh'])))

**Inference:**

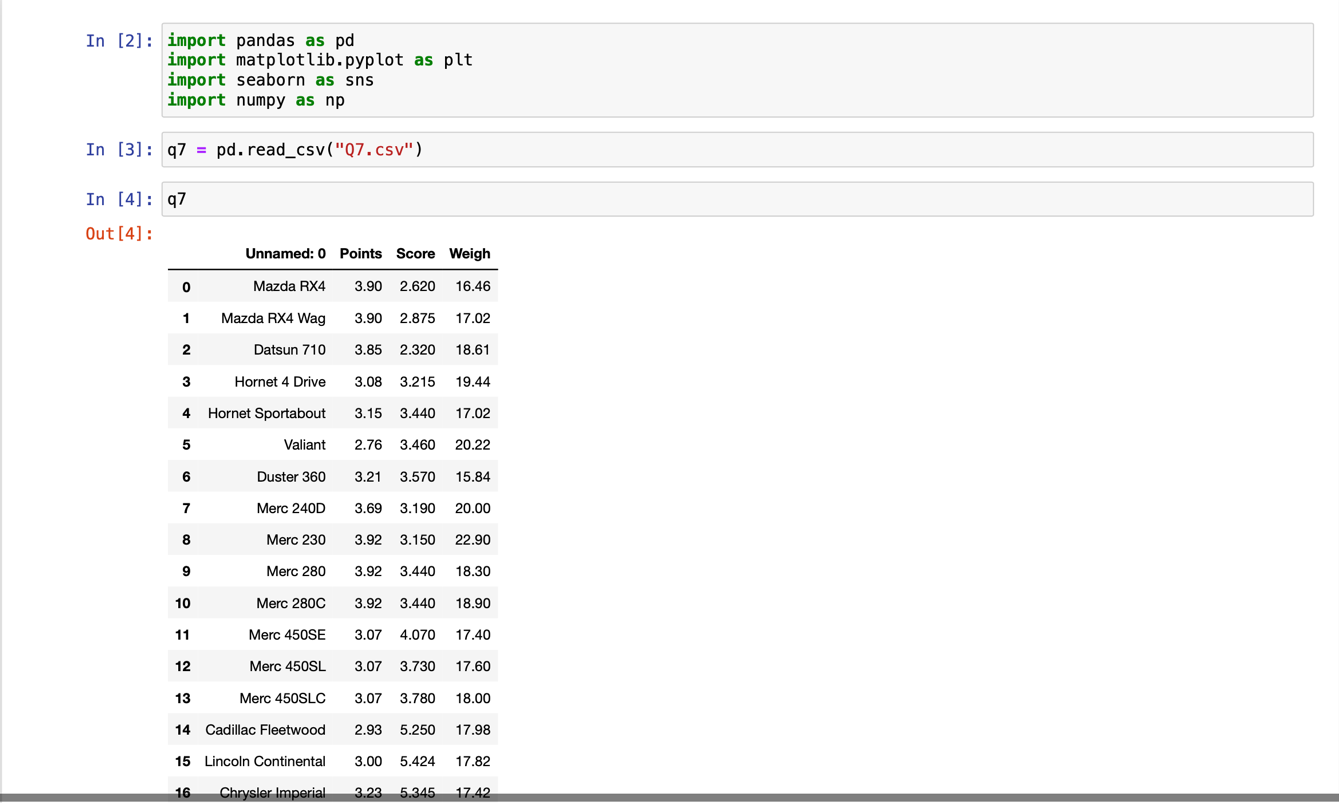
1. Mean / Average observation for Points among the data set is 3.59, similarly, Mean observation for Score among the data set is 3.21 and for weight is 17.84
2. Median is the centric value of the observation, for which: Points is 3.69, Score is 3.32 and for Weight is 17.71
3. Mode is the observation that occurs for the highest amount of times / frequency in the data.

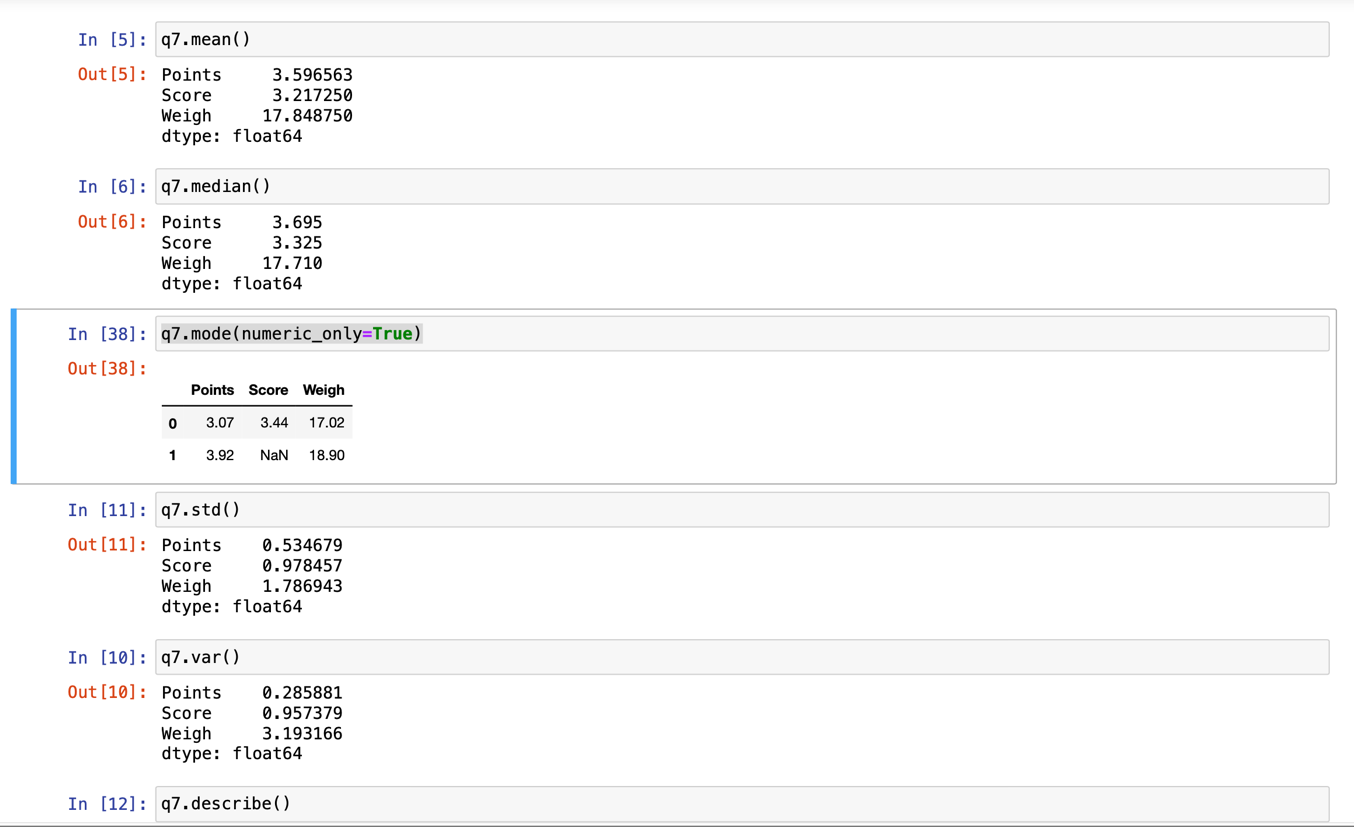
The mode for Points is 3.92, score is 3.44 and for Weigh is 18.90

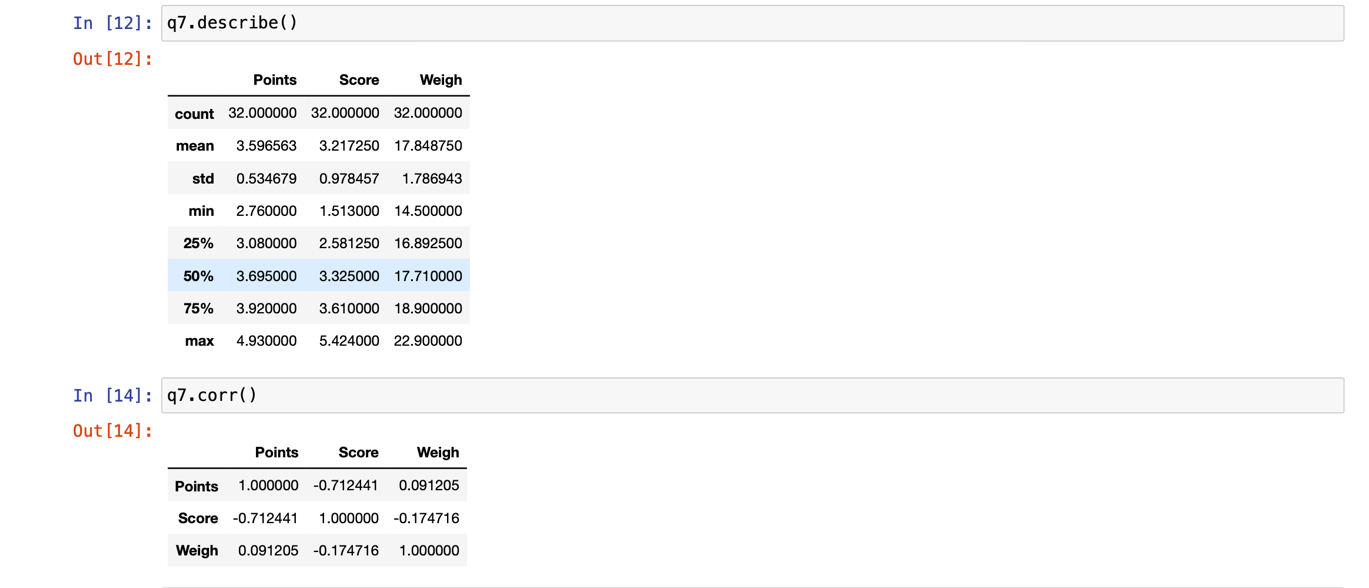
1. The standard deviation of the data is 0.53 for Points which mean that on an average the difference between the mean and the other observational data points is 0.53. So, the standard deviation of score is 0.97 and for Weigh is 1.78.

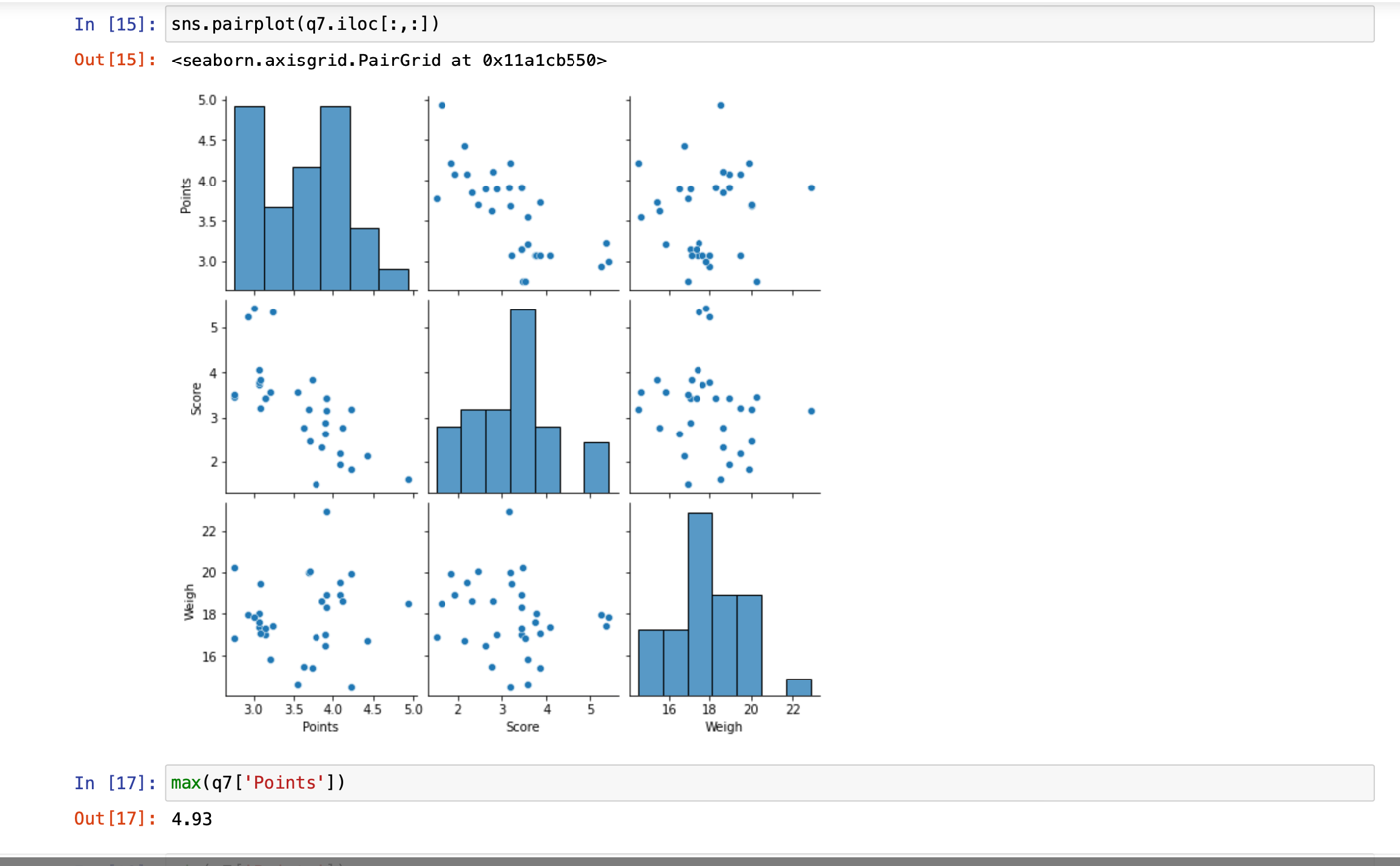
Which means that, the deviation of the data point from the mean is greater in the Weight column compared to Score and Points.

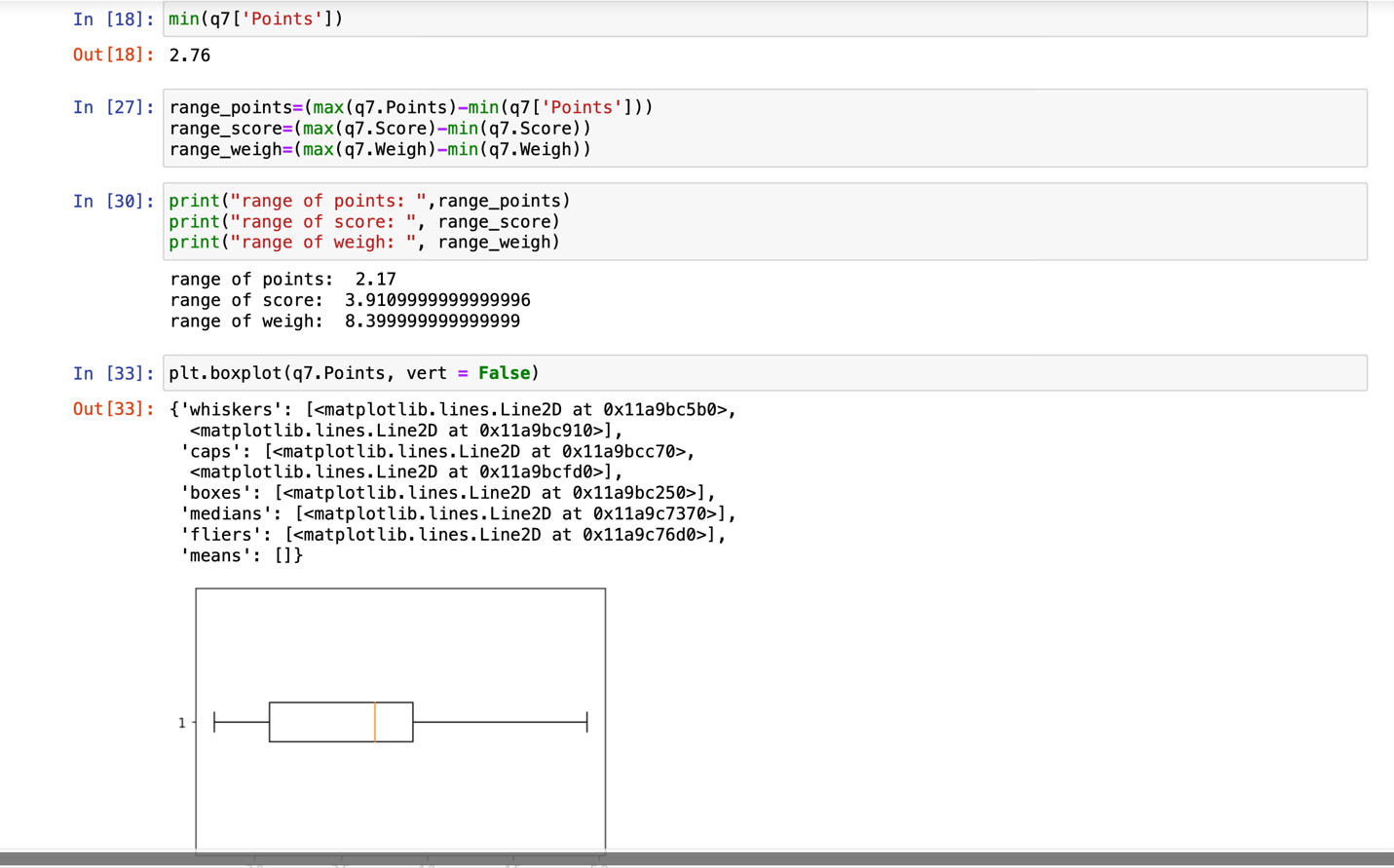
Std. weigh > Std. Score > Std. Points





****

****

****

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 :

Expected value --> mean of the list:

=>[ (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)/ 9 ]

=> 1308 / 9

=> 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

1. Answer:

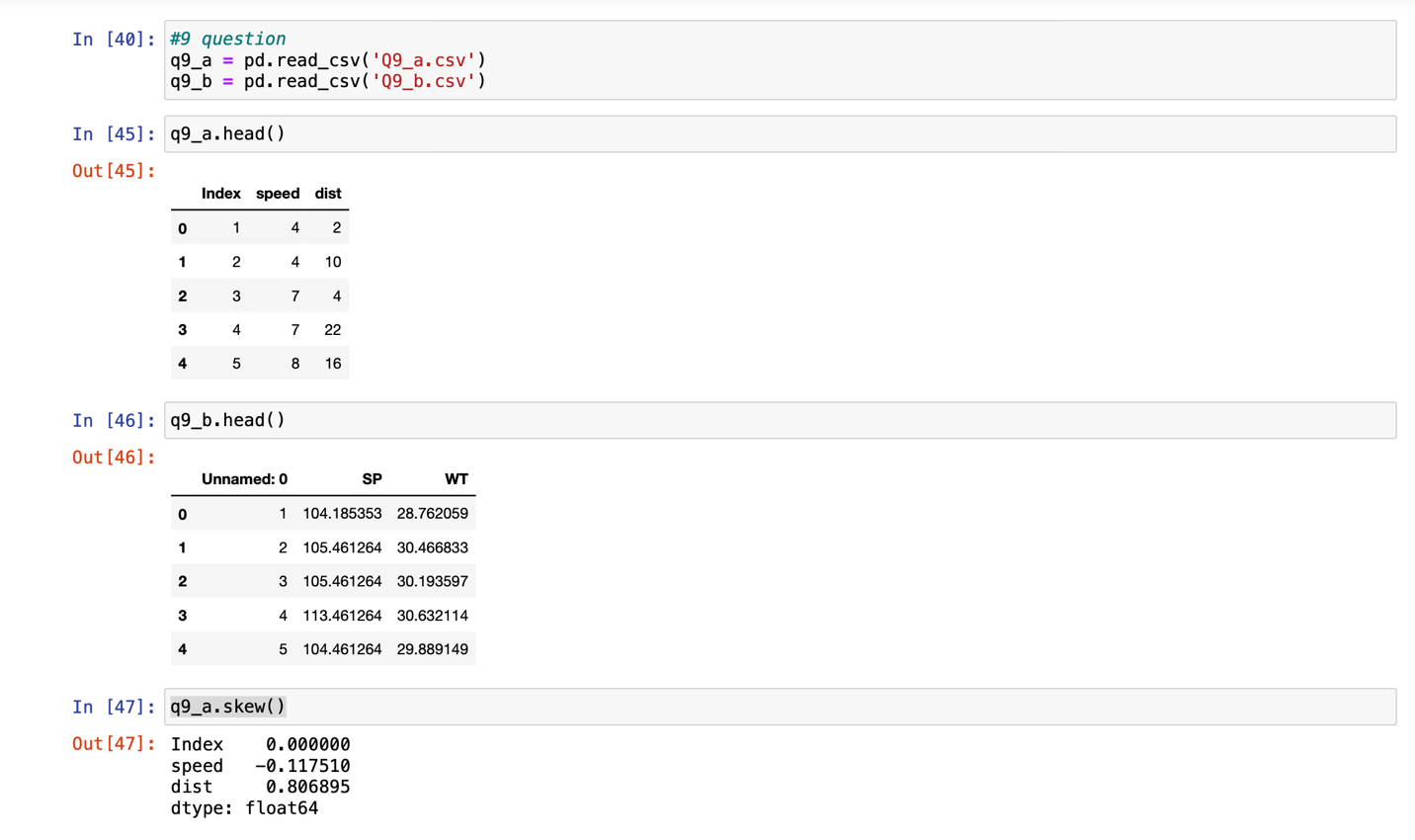
Skewness: - Measure of asymmetry

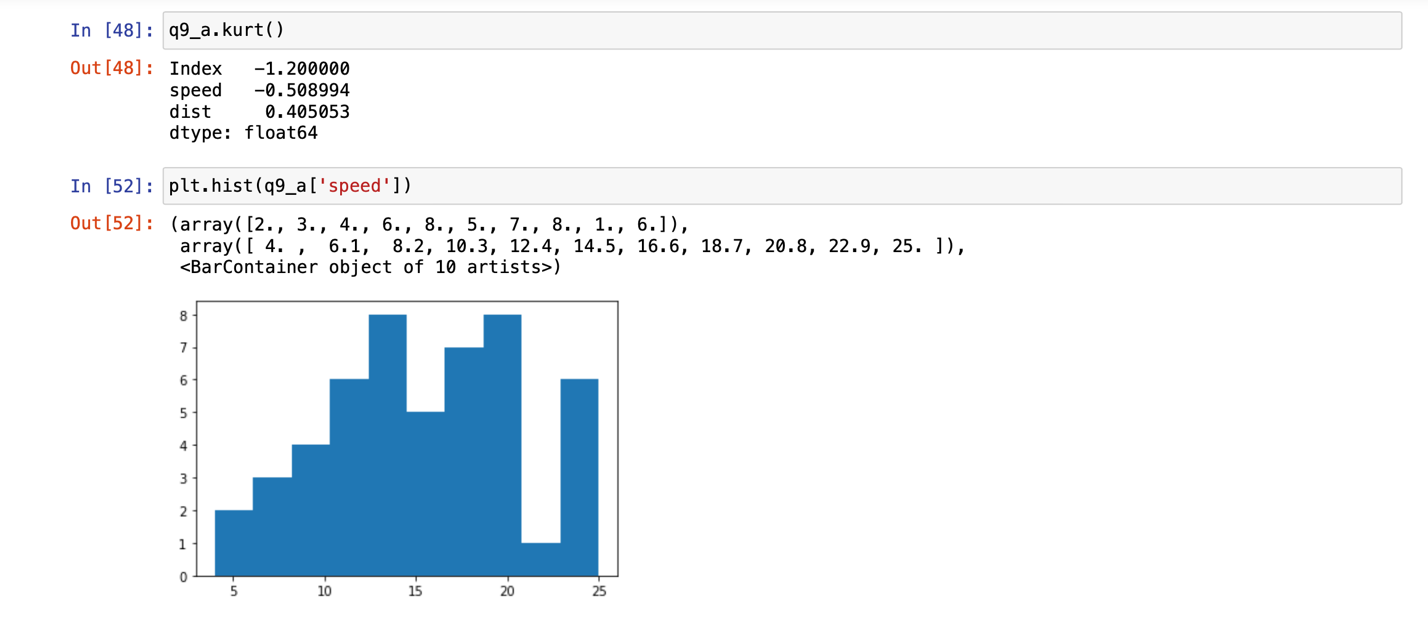
1. Skewness of Index is 0 , which means it is symmetrical with respect to mean
2. skewness of speed is -0.117, which means it is left skewed and there are more observation that lie towards the range in the right side.
3. Skewness if distance is 0.80, which means it is right skewed and a mass of the observation lies on the left side.

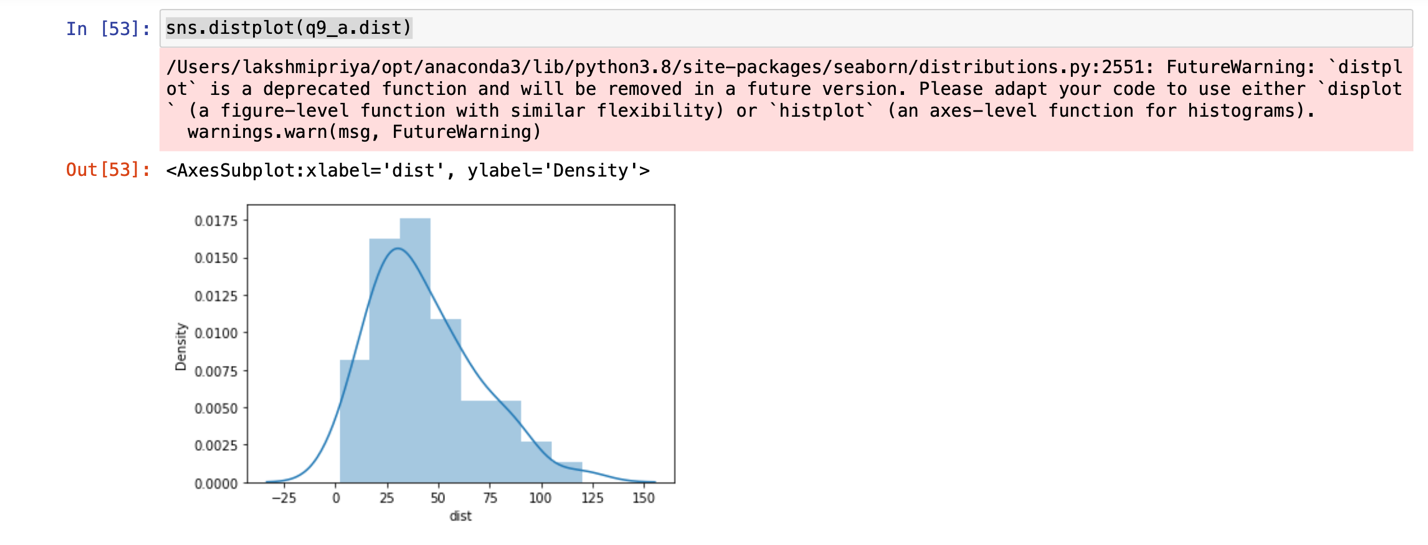
Kurtosis: - Measure of peakness of the distribution

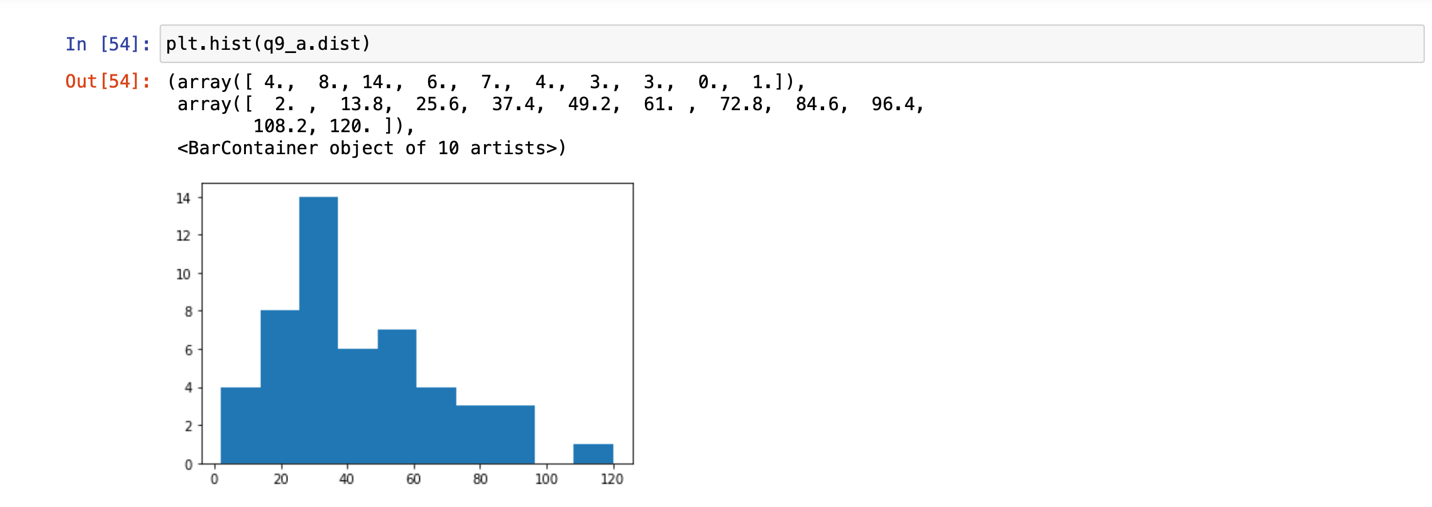
1. Kurtosis of Index is -1.20, which mean that the peak of the curve from the distribution of data is very flat.
2. Kurtosis of speed is -0.5, which mean that the peak of the curve from the distribution of data is flat, but not as much as the Index.
3. Kurtosis of distance is 0.405, which mean that the peak of the curve from the distribution of data is sharper than Index and speed

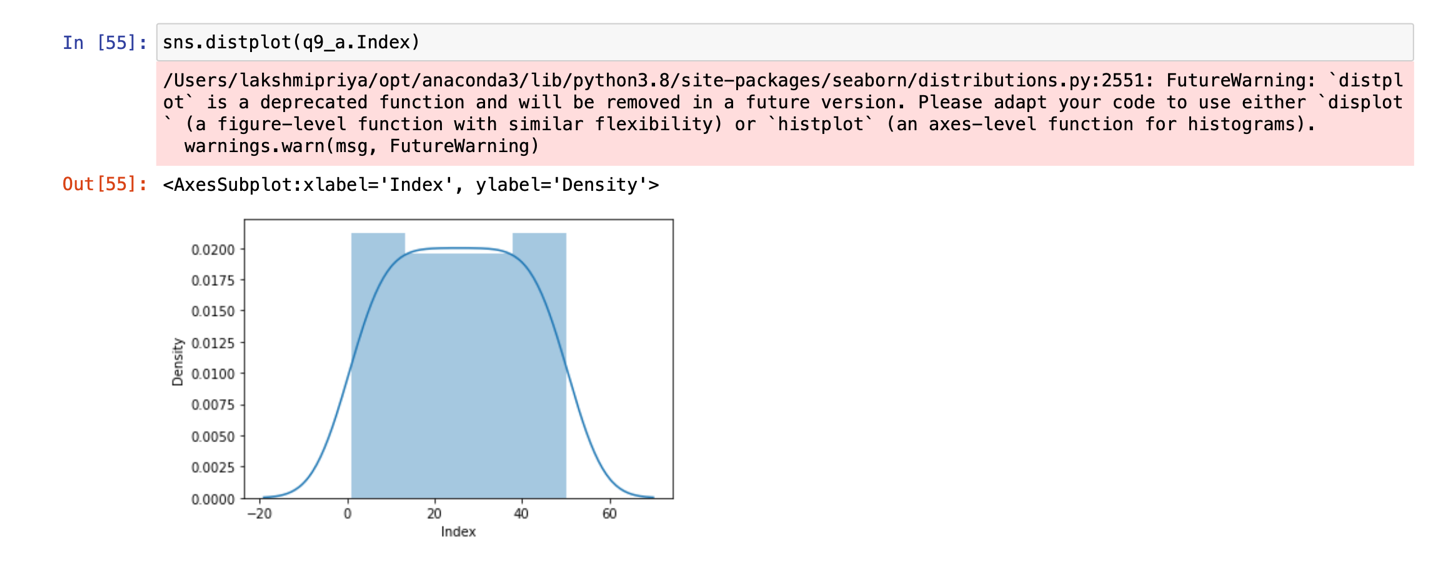
Thus the kurtosis of distance>speed>Index

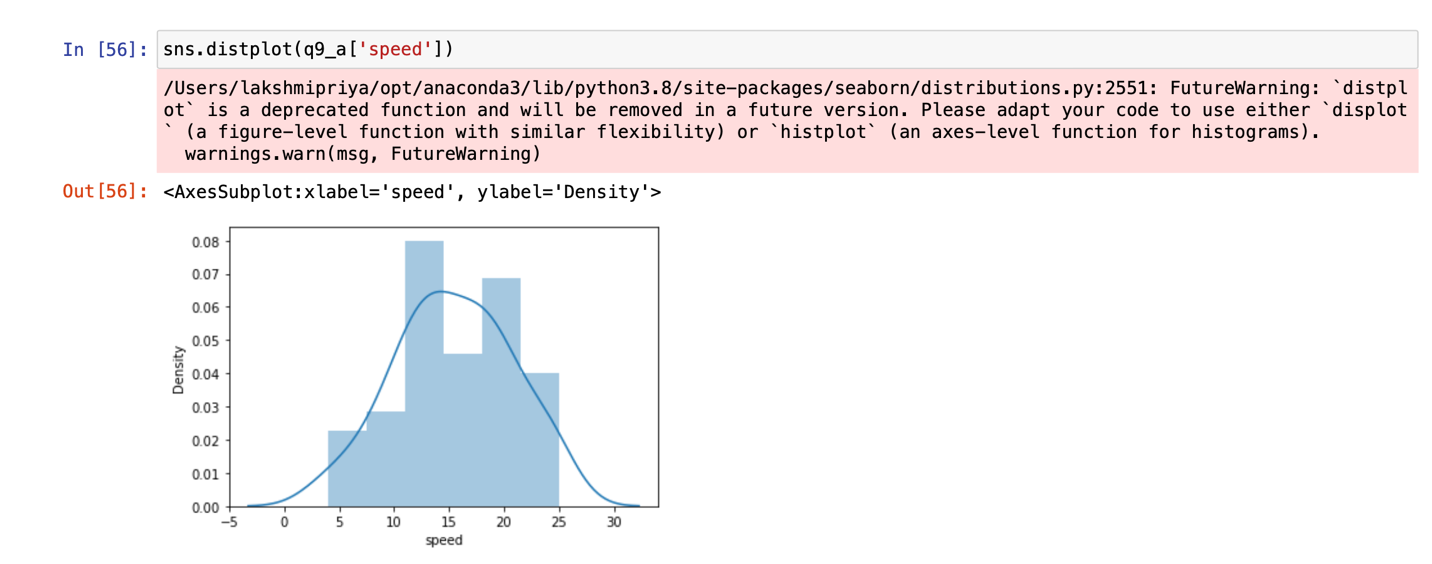












**SP and Weight(WT)**

**Use Q9\_b.csv**

1. Answer

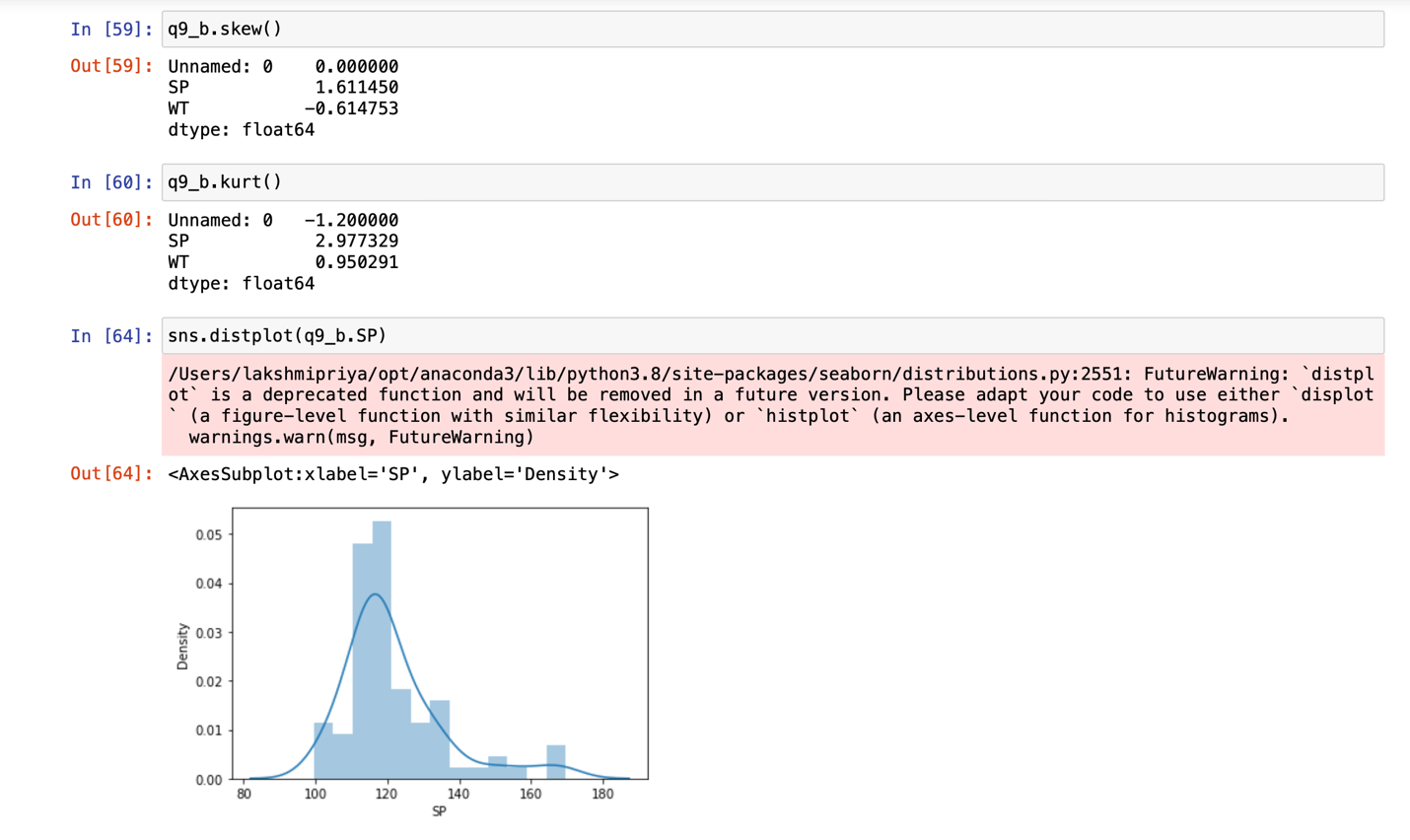
Skewness: - Measure of asymmetry

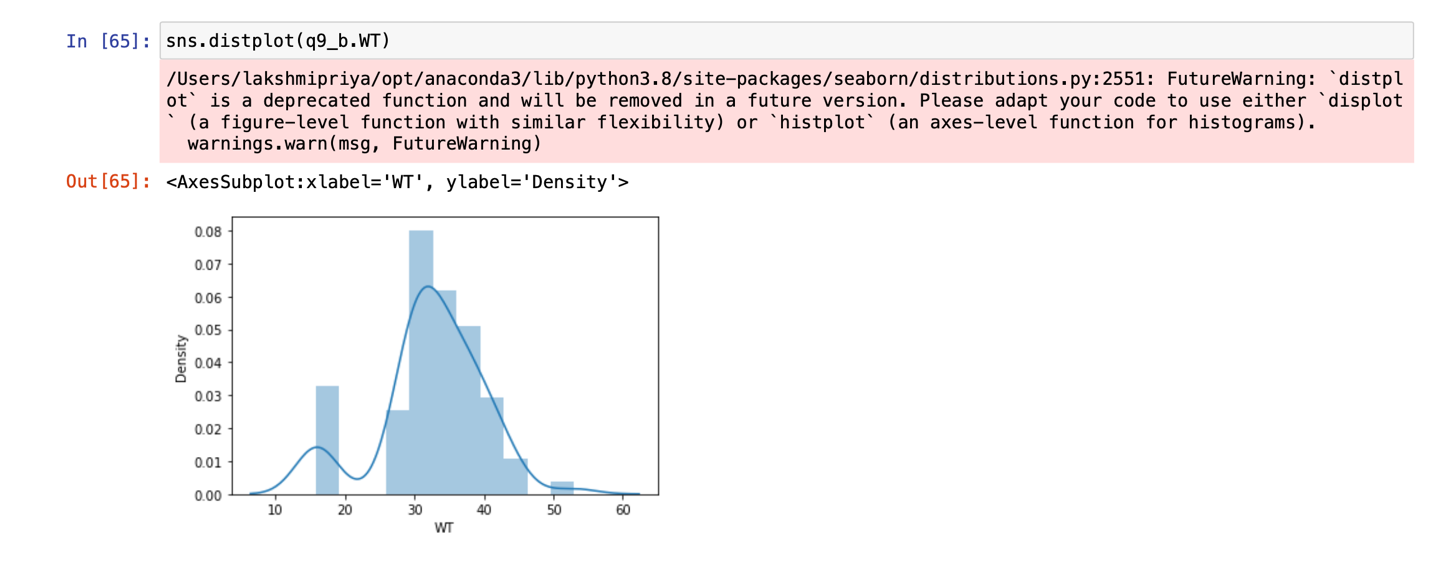
1. Skewness of the column Unnamed is 0 , which means it is symmetrical with respect to mean
2. skewness of SP is 1.611, which means it is right skewed and there are more observation that lie towards the range in the left side.
3. Skewness if WT is -0.614, which means it is left skewed and a mass of the observation lies on the right side.

Kurtosis: - Measure of peakness of the distribution

1. Kurtosis of the column unnamed is -1.20, which mean that the peak of the curve from the distribution of data is very flat.
2. Kurtosis of SP is 2.977, which mean that the peak of the curve from the distribution of data is very sharp than unnamed column and that of WT.
3. Kurtosis of WT is 0.95, which mean that the peak of the curve from the distribution of data is sharp but not as sharp as that of SP.

Thus, Kurtosis of SP>WT>unnamed





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



1. The data is right skewed with a positive value, and thus, most of the collected data / observation lies on the left side at the range of 50 to 100.
2. There are comparatively more observations in the range of 50 to 100 in ChichkWeight$Weight (X axis range).

25% of the data

25% of the data

Whisker

Lower Quartile

Upper Quartile

Median

Lower extreme

Upper extreme

Outliers



3

**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 :

Random sample (n) = 2000

Population = 3,000,000

Sample mean (X) = 200 pounds

Standard deviation = 30 pounds

Confidence interval – 94%, 98% and 96%

X **±** (standard deviation /square root of n) x z score

1. For 94%, confidence interval:

200 ± (30 / √2000) \* [stats.norm.ppf(0.97)]

= 200 ± (30 / √2000) \* 1.88

* 200 ± 1.2596 => 198.7404 to 201.2596

1. For 96%, confidence interval:

200 ± (30 / √2000) \* [stats.norm.ppf(0.98)]

= 200 ± (30 / √2000) \* 2.05

* 200 ± 1.375 => 198.625 to 201.375

1. For 98%, confidence interval:

200 ± (30 / √2000) \* [stats.norm.ppf(0.99)]

= 200 ± (30 / √2000) \* 2.326

* 200 ± 1.56 => 198.44 to 201.56



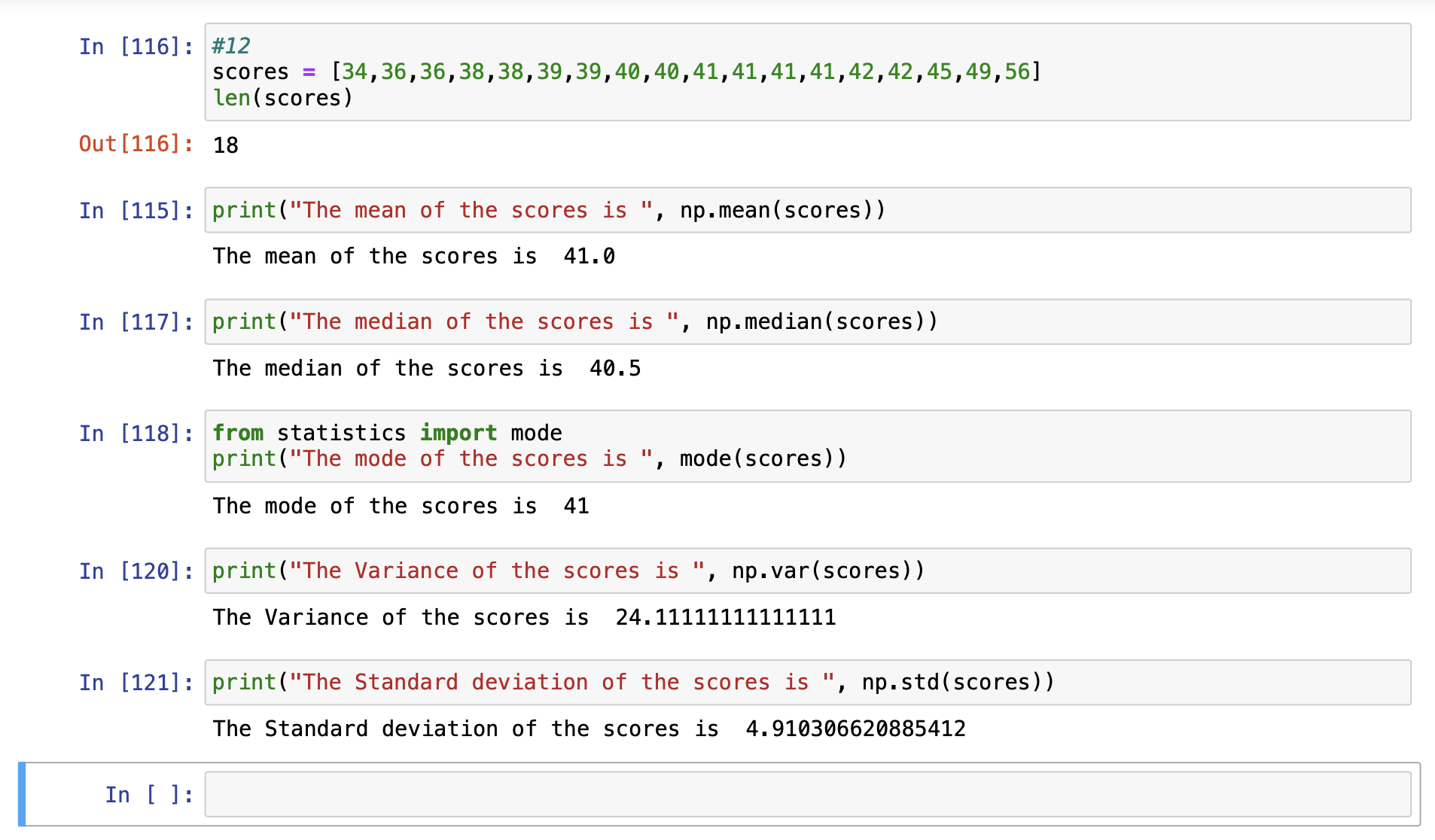
**Q12)** Below are the scores obtained by a student in tests

**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 = 41
  2. Median = 40.5
  3. Mode = 41
  4. Variance = 24.11
  5. Standard deviation = 4.91



1. A) The average score that the student has got in all the subjects is about 41.

B) The student has scored marks in the range of 34 to 39 in 7 subjects, 40 to 45 in 9 subjects, 46 to 51 in 1 subject and 52 to 57 in 1 subject.

C) The student’s highest mark is 56 and the lowest mark is 34

D) The subjects with common highest frequency score is 41

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

Answer : Symmetrical and the skewness = 0

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

Answer : Positive value and right skewed

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

Answer : When median > mean, it is Negative and left skewed

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

Answer : A positive kurtosis value in the data indicates a sharper peak

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

Answer : A negative kurtosis value in the data indicates a Flatter peak

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



What can we say about the distribution of the data?

Answer : The data is left skewed

What is nature of skewness of the data?

Answer : From the above boxplot picture, the value from the data is negative and Left skewed. More observations are towards the left side range.

What will be the IQR of the data (approximately)?   
  
Answer : IQR = Q3 – Q1 => (18-10)

Thus, IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

Answer:

1. Boxplot number 1 has comparatively less number of observation spread than that of the boxplot number 2
2. The observation is centered at 262.5 wbs for both boxplots, 1 and 2
3. The whiskers of boxplot 1 is smaller than its box, while the whiskers of the boxplot 2 is bigger.
4. The skewness is almost 0 for both the boxplots. Thus the observation is from a symmetric population and also the whiskers length are same for both boxplots.
5. There are no outliers for both boxplots.

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

Data \_set: Cars.csv

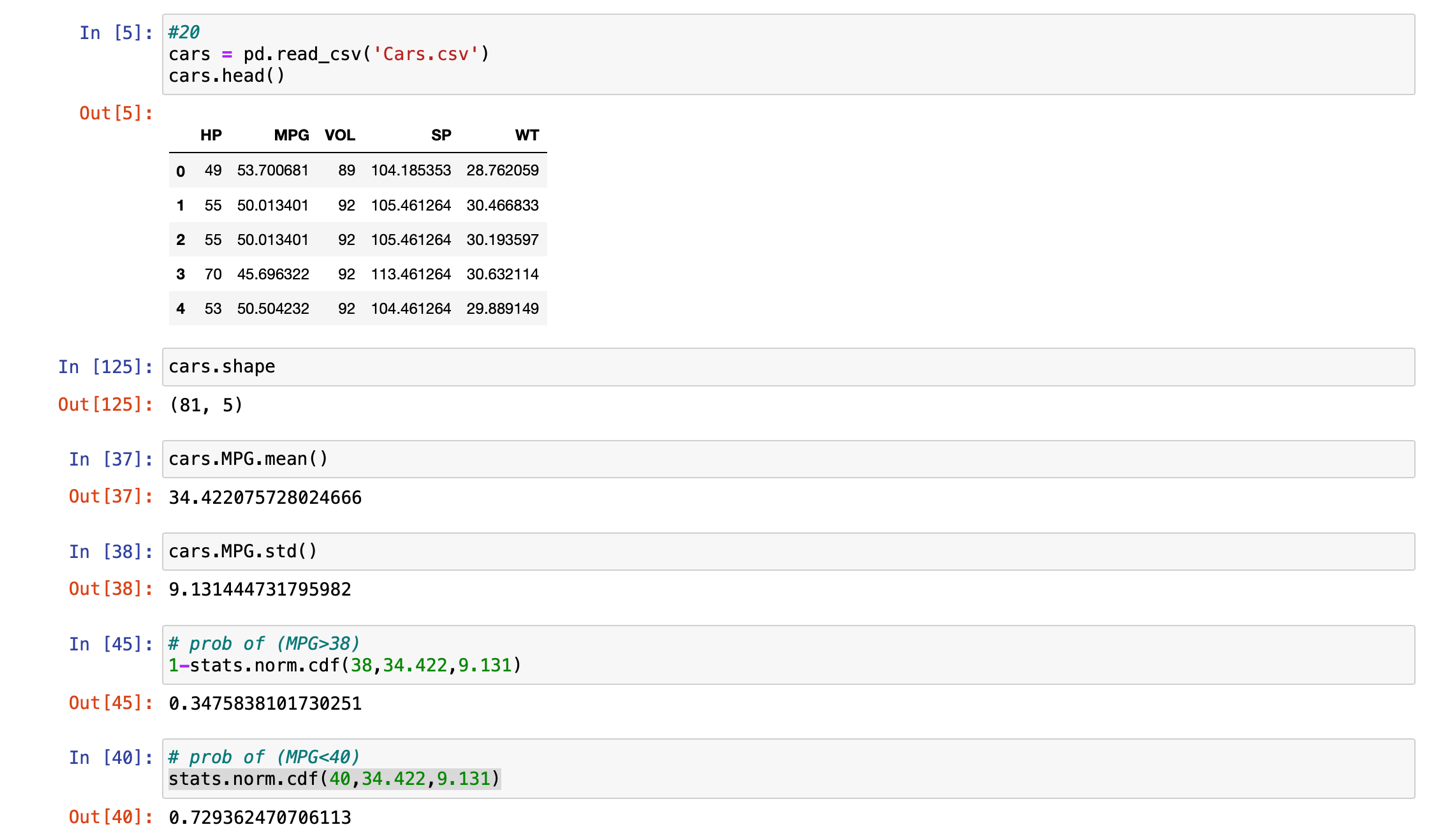
Calculate the probability of MPG of Cars for the below cases.

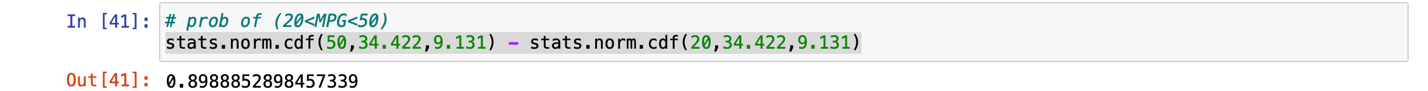
MPG <- Cars$MPG

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

c. P (20<MPG<50)

Answer:





1. P(MPG>38) = 0.347
2. P(MPG<40) = 0.729

c. P (20<MPG<50) = 0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer :

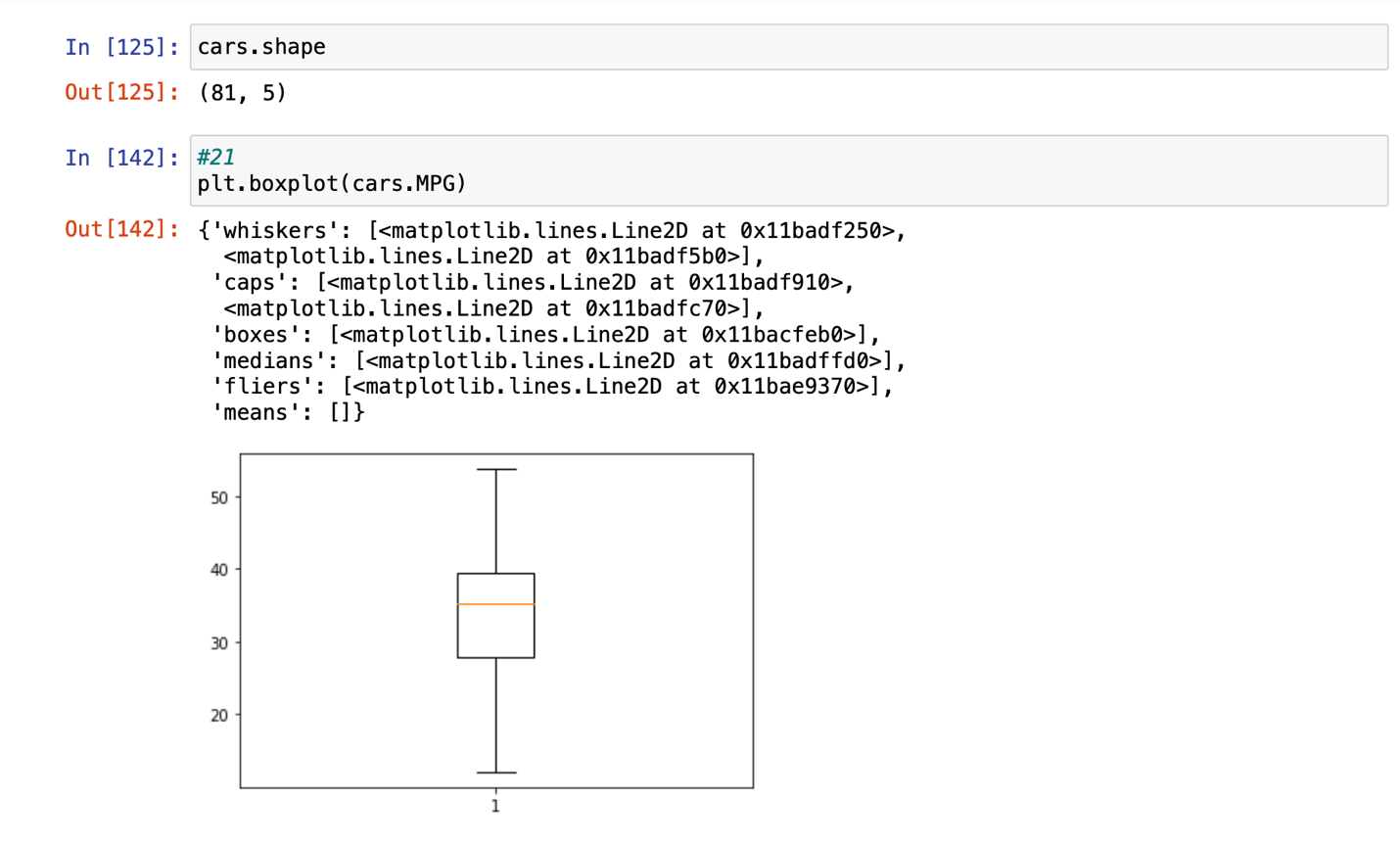
MPG follows a normal distribution because:

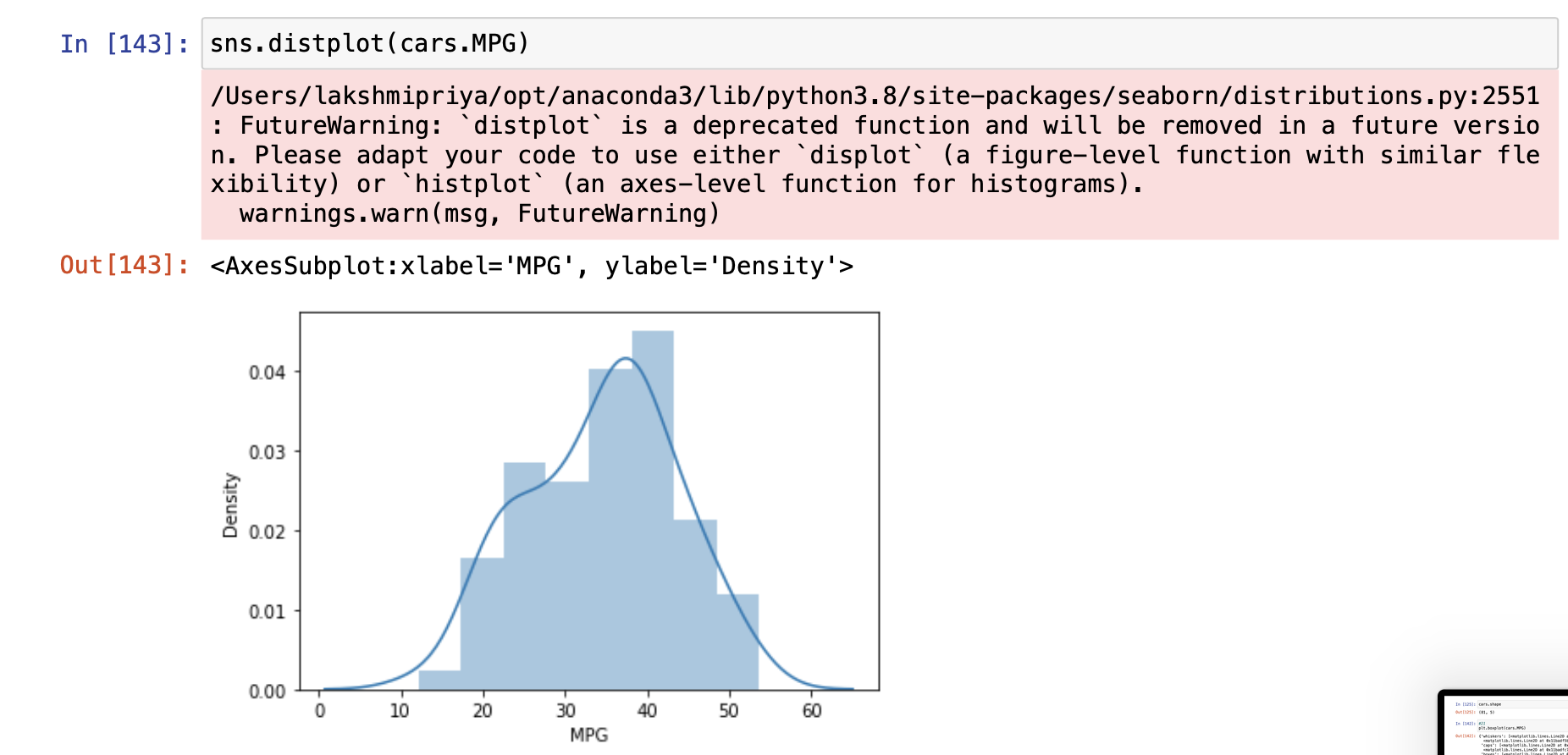
The probability value of MPG cars less than 54 is 0.98. thus pvalue is greater than 0.05.

Also the mean and median are almost close. Mean = 34.4, median = 35.15

USing a boxplot and distplot for the column MPG. Since, the number of observations are also greater than 30 and the box plot is with symmetrical observations.

It also has more observation/density towards the center compared to the corners with a bell shaped curve.





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

Dataset: wc-at.csv

**Answer:**

For waist circumference, the mean and median values are close – 91.9 and 90.8.

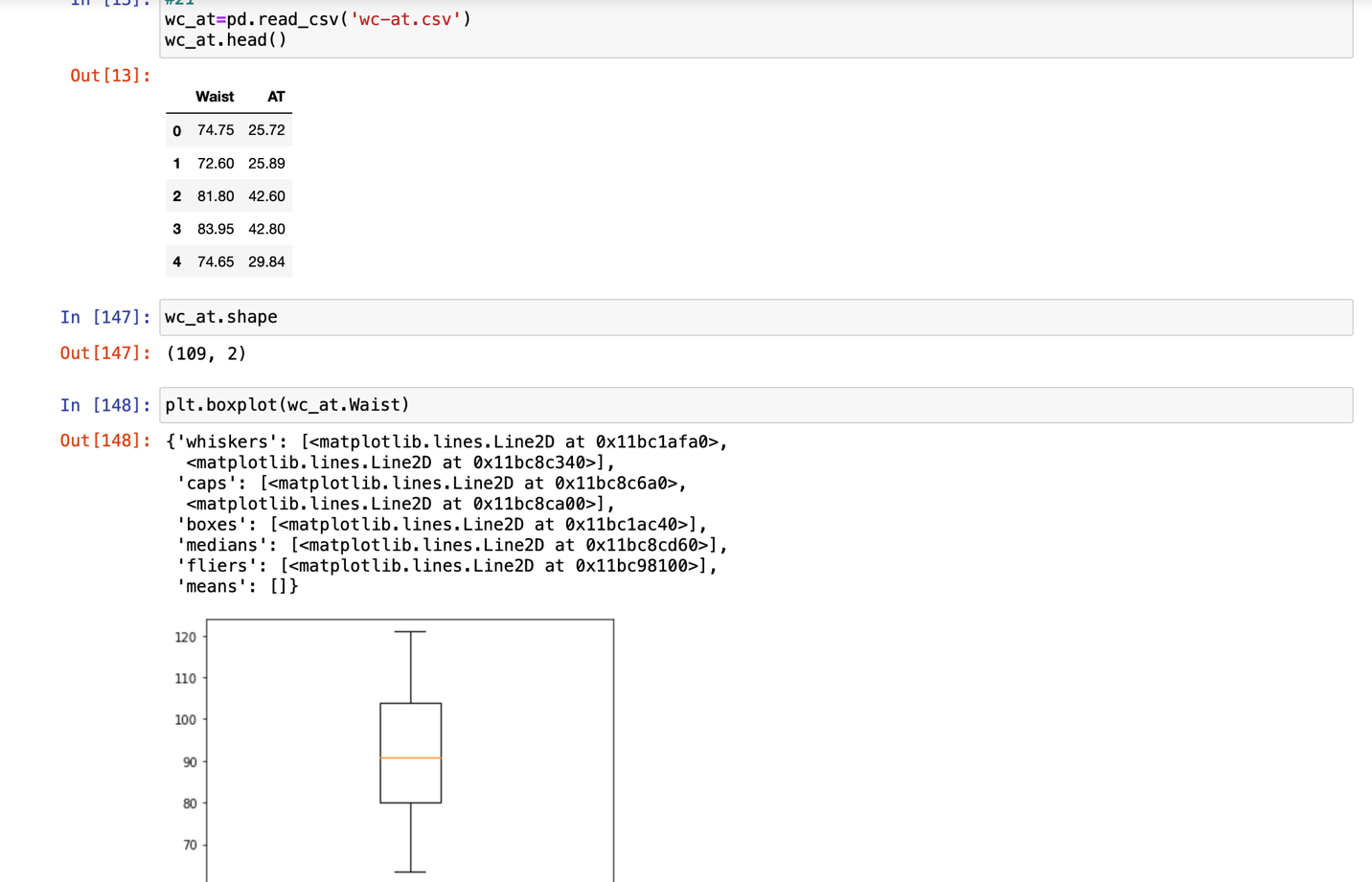
Moreover, the distplot and boxplot show that the data is symmetrical of the mean. The length of the whiskers in boxplot is same too.

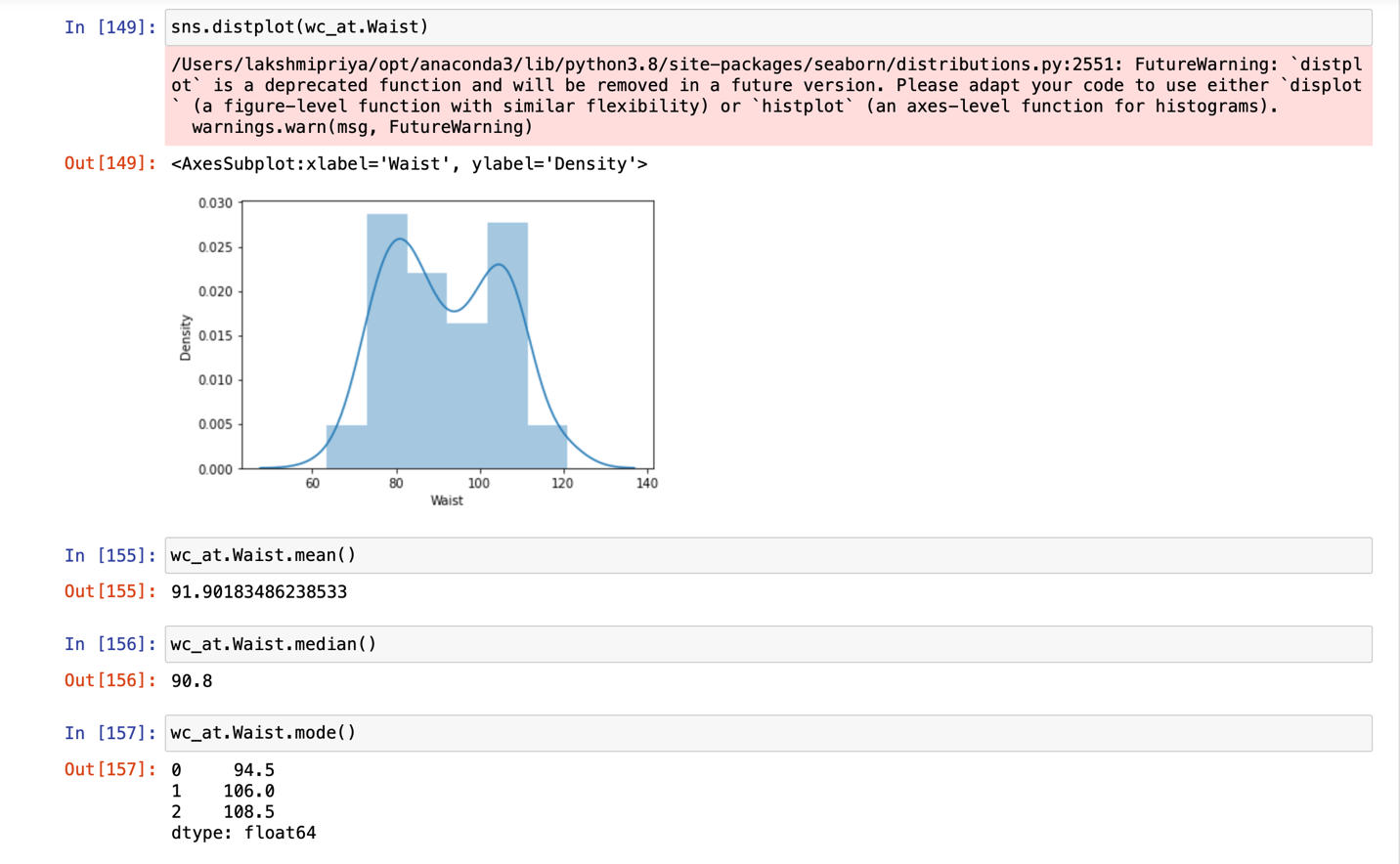
However, there is a significant difference between mean and median of AT.

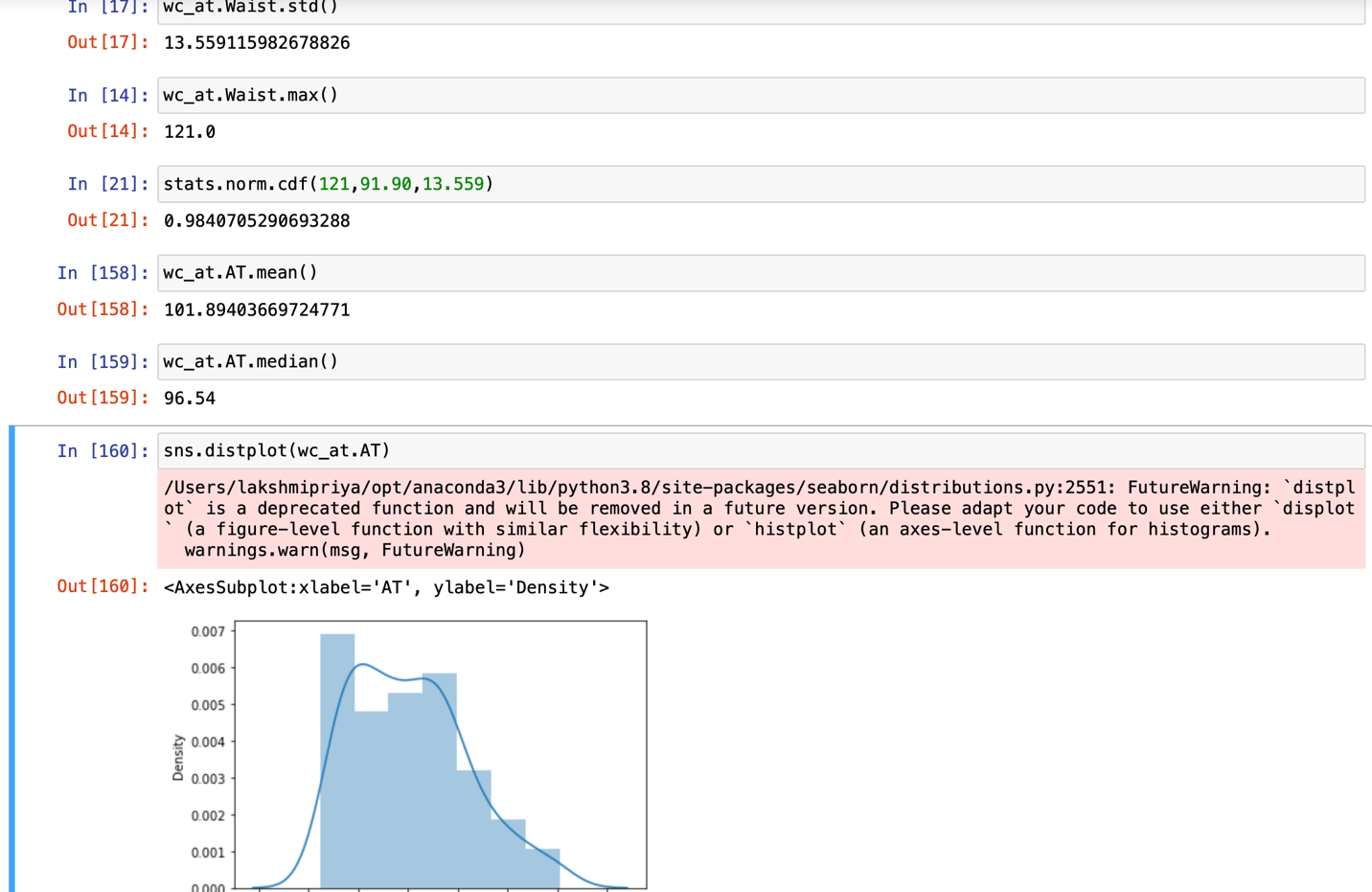
Mean(AT) = 101.89 and median (AT) = 96.54

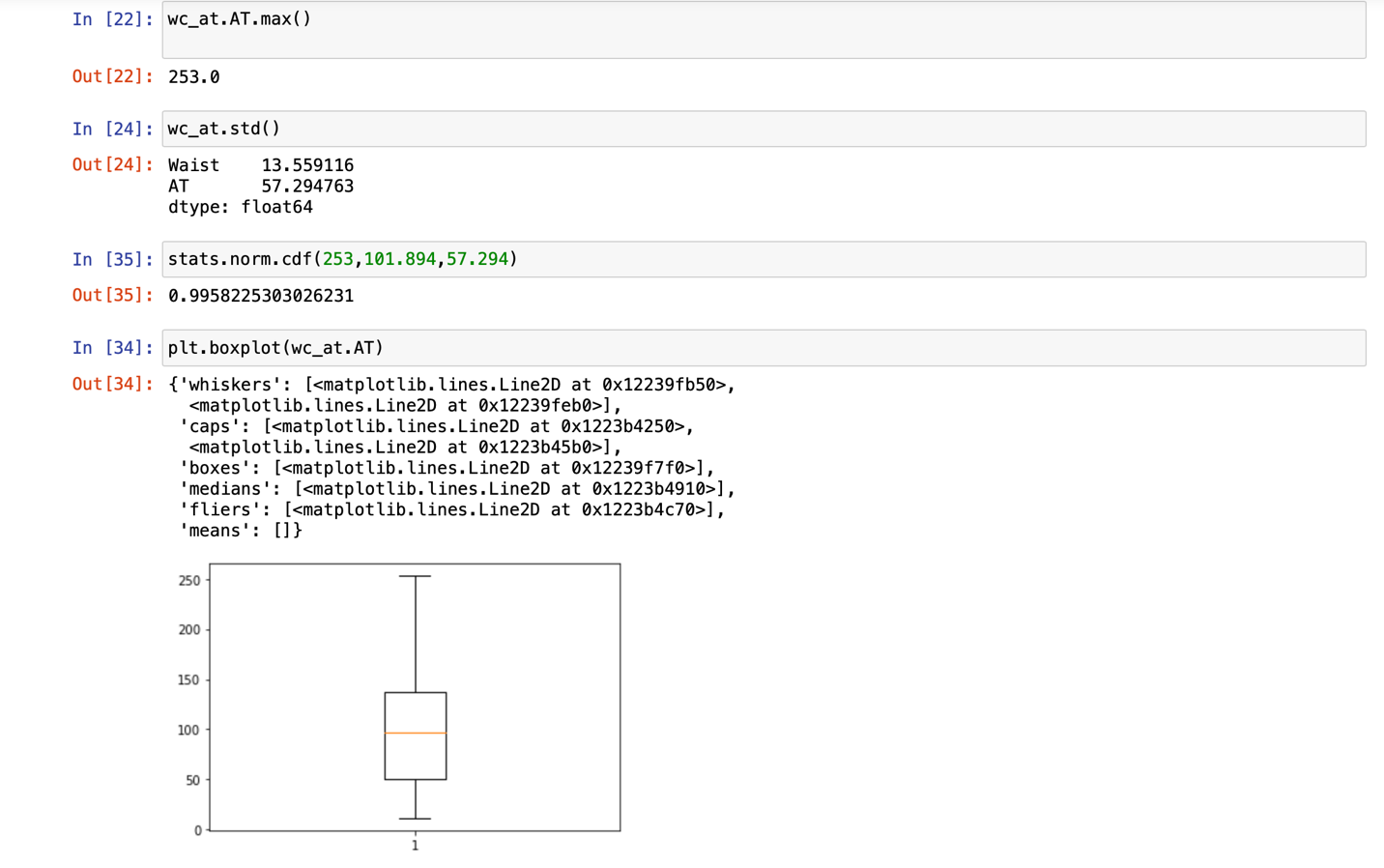
Also, the displot and boxplot visualization clearly show that the data distribution is not symmetrical of the median and is right skewed.

Thus, Waist circumference is normally distributed and Adipose tiisue is not normally distributed.







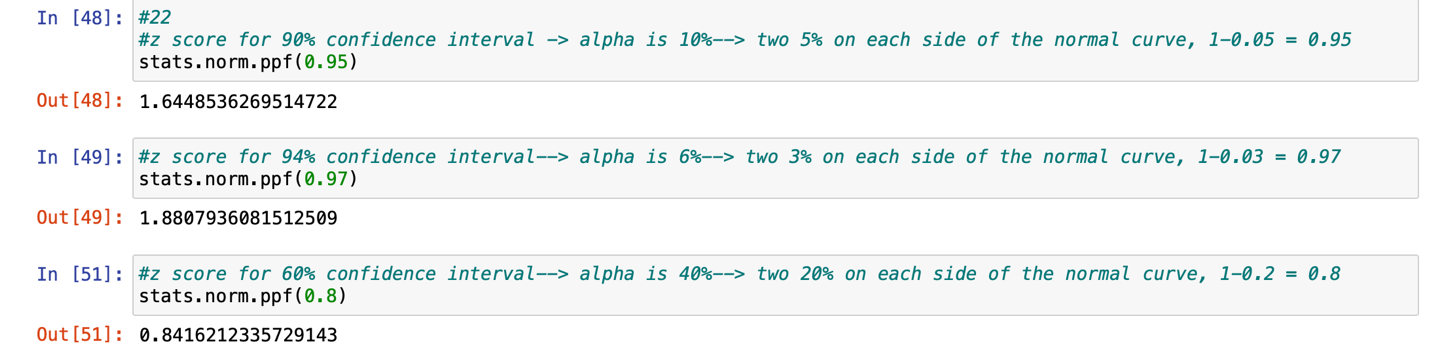


Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

Z score at 90% confidence interval = 1.6448

Z score at 94% confidence interval = 1.88

Z score at 60% confidence interval = 0.841

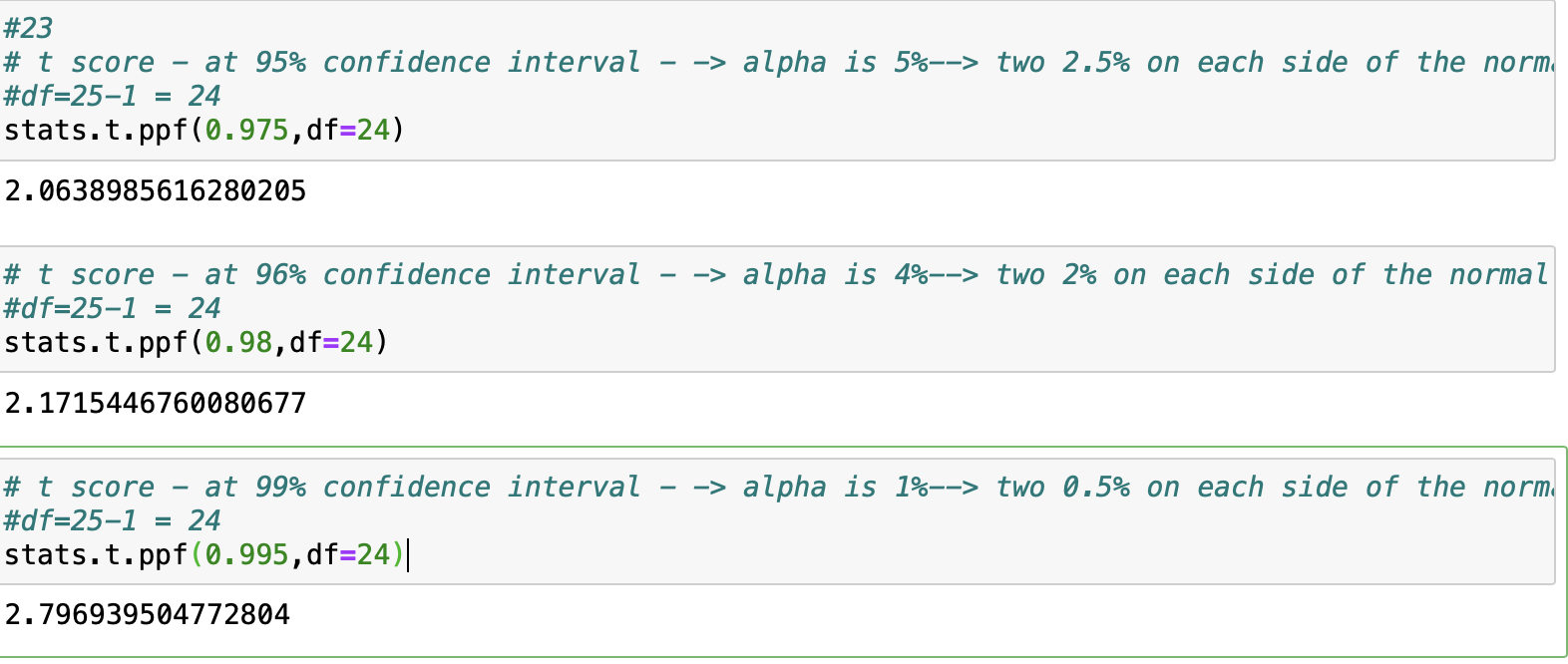


Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

T score at 95% confidence interval = 2.063

T score at 96% confidence interval = 2.171

T score at 99% confidence interval = 2.79



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

Answer :

Mean = 270 days

Random sample = 18

Sample mean = 260 days

Standard deviation = 90 days

T statistic, since standard deviation of the population is unknown.

t\_score = (270-260) / (90/np.sqrt(18)) => 0.471

stats.t.pdf(t\_score, df= 17) => 0.349

