**Student Name:** Prajwal Suresh Sonawane

**Batch:** Data Science weekday Batch – Data science Weekday-Aishwarya-Pune-24th February 2023-01:00 PM to 03:00 PM

**Assignment Name:** Basic Statistics Level 1 (Assignment No.1)

***Solution:-***

**Q1) Identify the Data type for the Following:**

* Ans:

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

**Q2) Identify the Data types, which were among the following**

**Nominal, Ordinal, Interval, Ratio.**

* Ans:

|  |  |
| --- | --- |
| **Data** | **Data Type** |
| Gender | Nominal Data |
| High School Class Ranking | Ordinal Data |
| Celsius Temperature | Interval Scale Data |
| Weight | Interval, Ratio Scale Data |
| Hair Color | Nominal Data |
| Social Economic Status | Ordinal Data |
| Fahrenheit Temperature | Interval Scale Data |
| Height | Interval, Ratio Scale Data |
| Type of living accommodation | Nominal Data |
| Level of Agreement | Ordinal Data |
| IQ (Intelligence Scale) | Ordinal Data |
| Sales Figures | Ratio Scale Data |
| Blood Group | Nominal Data |
| Time Of Day | Interval Scale Data |
| Time on a Clock with Hands | Interval Scale Data |
| Number of Children | Interval, Ratio Scale Data |
| Religious Preference | Nominal Data |
| Barometer Pressure | Interval, Ratio Scale Data |
| SAT Scores | Interval, Ratio Scale Data |
| Years of Education | Interval, Ratio Scale Data |

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

* Ans:

If Three Coins are tossed possible combinations are:

S = { HHH,HHT,HTH,THH,HTT,THT,TTH,TTT }

n(S) = 23 = 8

i) Probability of getting two heads and one tail:

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

A = { HHT,HTH,THH }

n(A) = 3

P(A) = n(A)

n(S)

= 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**

* Ans:

S = { (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) }

n(S) = 6×6 = 36

a) Probability of getting sum is equal to 1 when two dice are rolled:

*Let, A be the event of getting sum is equal to one*

A = { 0 }

n(A) = 0

P(A) = n(A)

n(S)

= 0

36

= 0

b) Probability of getting sum is less than or equal to 4 when two dice are rolled:

*Let, B be the event of getting sum is less than or equal to 4*

B = { (1,1), (1,2), (1,3),

(2,1), (2,2),

(3,1) }

n(B) = 6

P(B) = n(B)

n(S)

= 06

36

= 0.166666667

c) Probability of getting sum is divisible by 2 and 3 when two dice are rolled:

*Let, C be the event of getting sum is divisible by 2 and 3*

C = { (1,5),

(2,4),

(3,3),

(4,2),

(5,1),

(6,6) }

n(C) = 6

P(C) = n(C)

n(S)

= 06

36

= 0.166666667

**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:

Total number of balls  
= (2 + 3 + 2)  
= 7  
Let, S be the sample space  
Then,

n(S) = Number of ways of drawing 2 balls out of 7

n(S) = 7C2

n(S) = (7×6)

(2×1)

n(S) = 21

Probability that none of the balls drawn is blue:

*Let, A = Event of 2 balls, none of which is blue*

∴ n(A) = Number of ways of drawing 2 balls out of (2 + 3) balls

n(A) = 5C2

n(A) = (5×4)

(2×1)

n(A) = 10

∴P(A) = n(A)

n(S)

= 10

21

= 0.476190476

**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**

* Ans:

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 × 0.12)

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

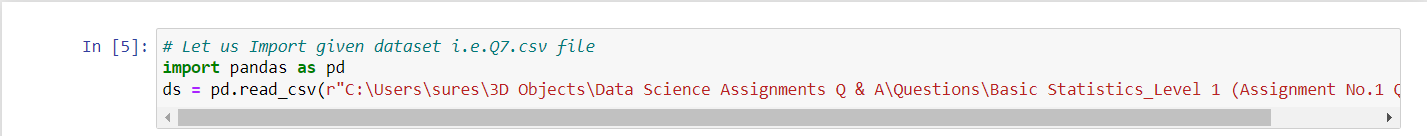
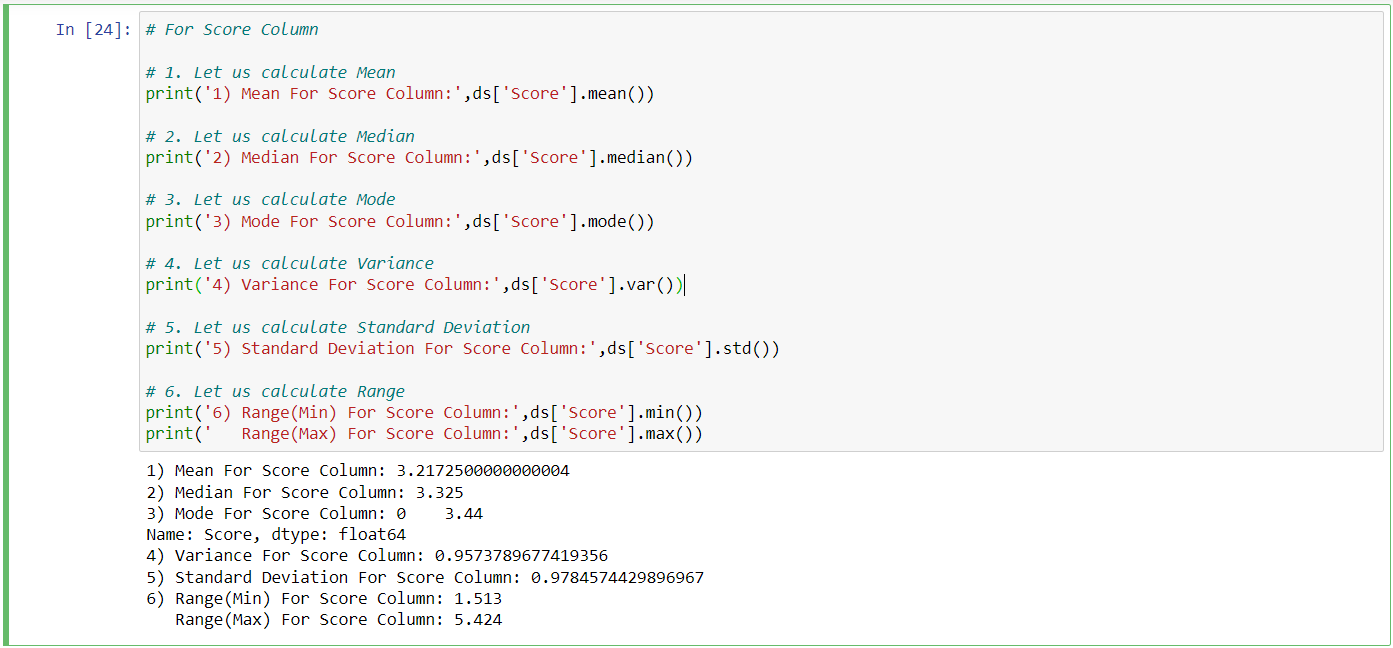
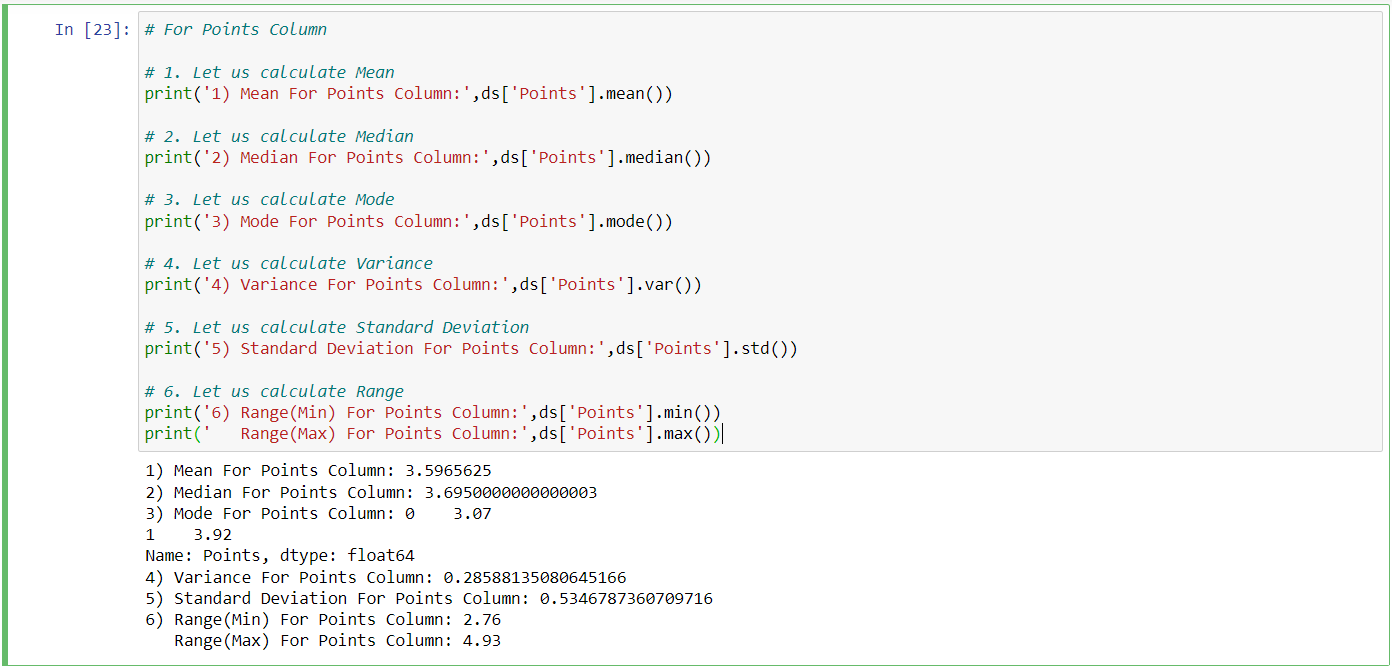
=  3.09

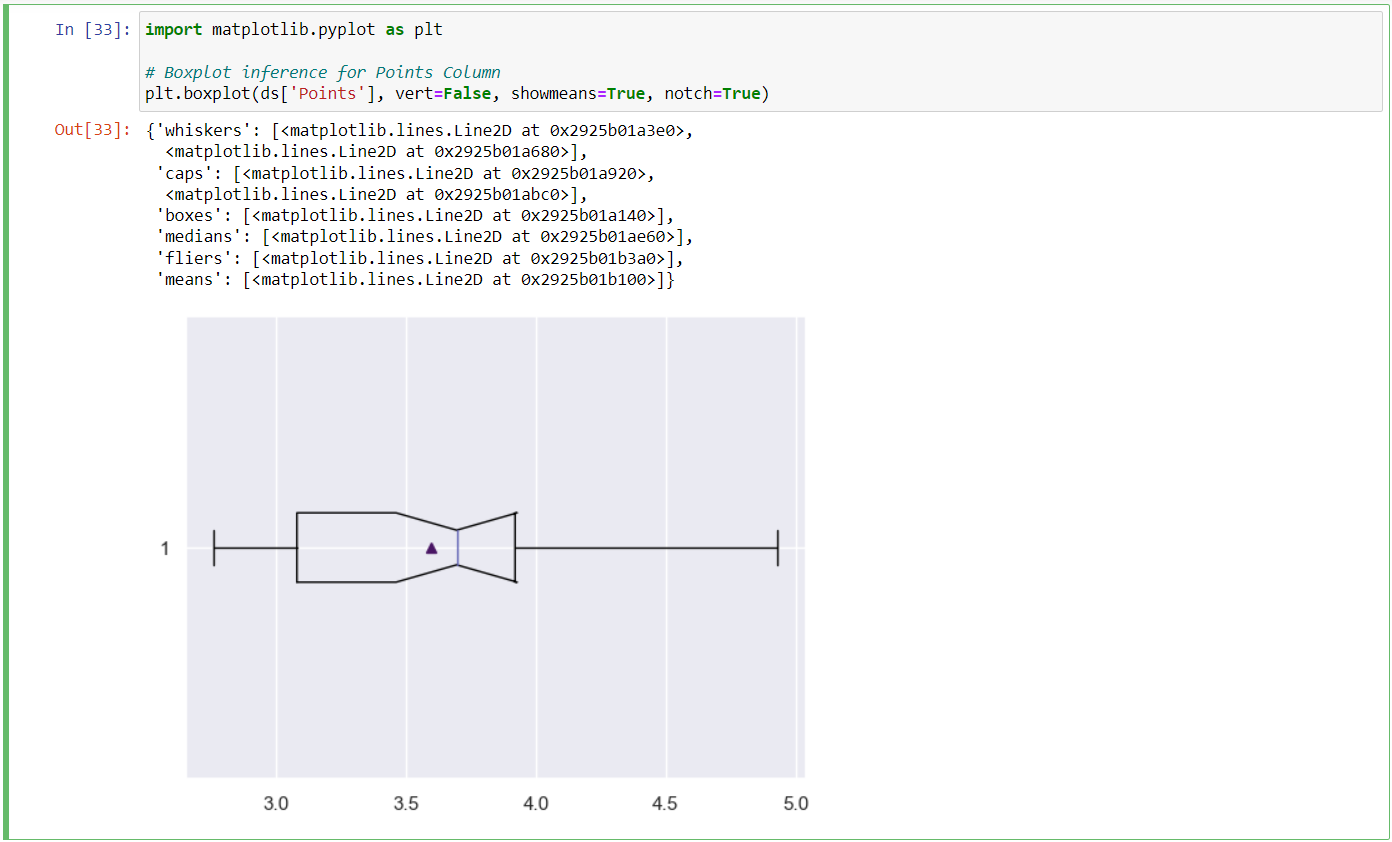
**Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset**

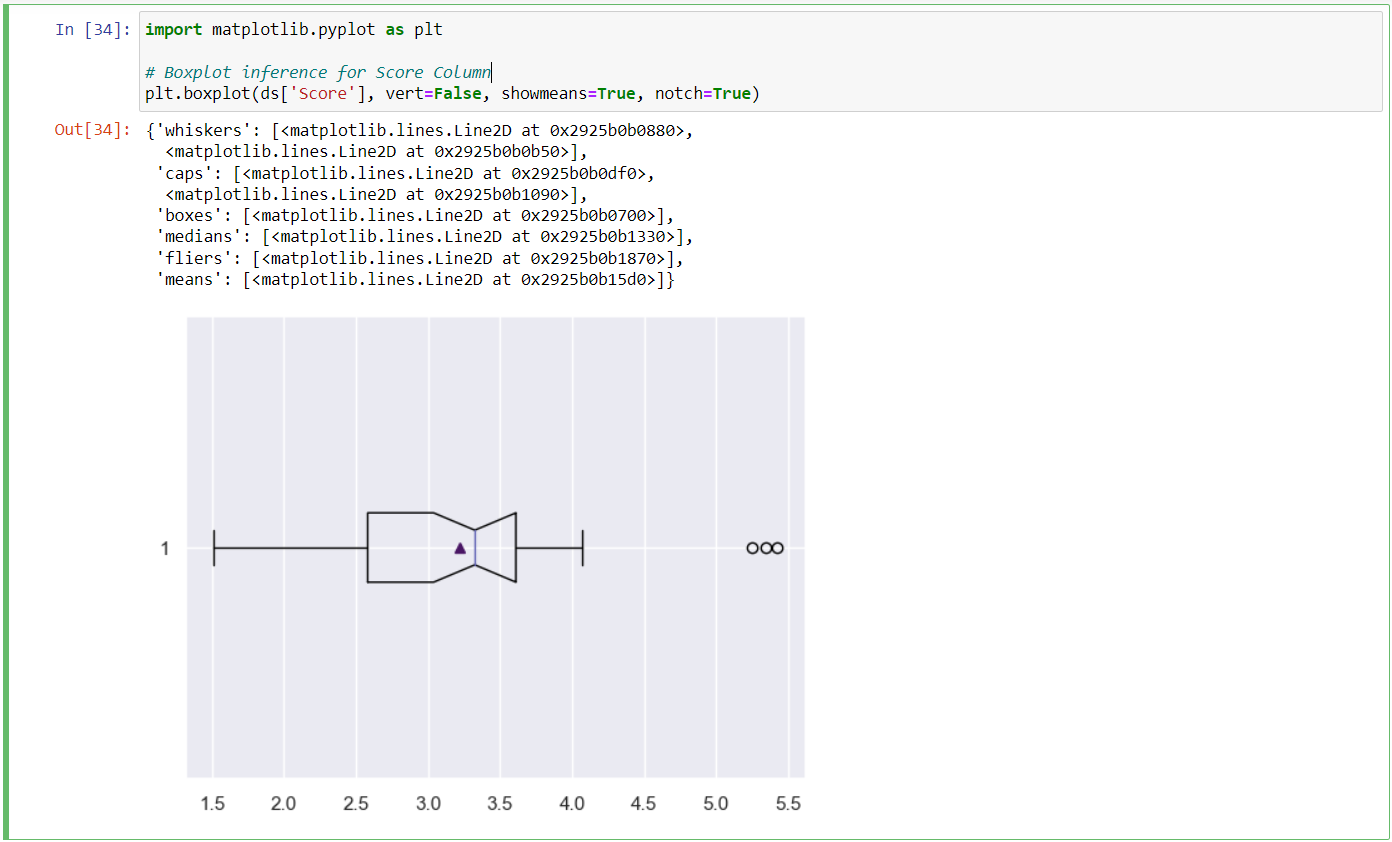
* **For Points,Score,Weigh>**

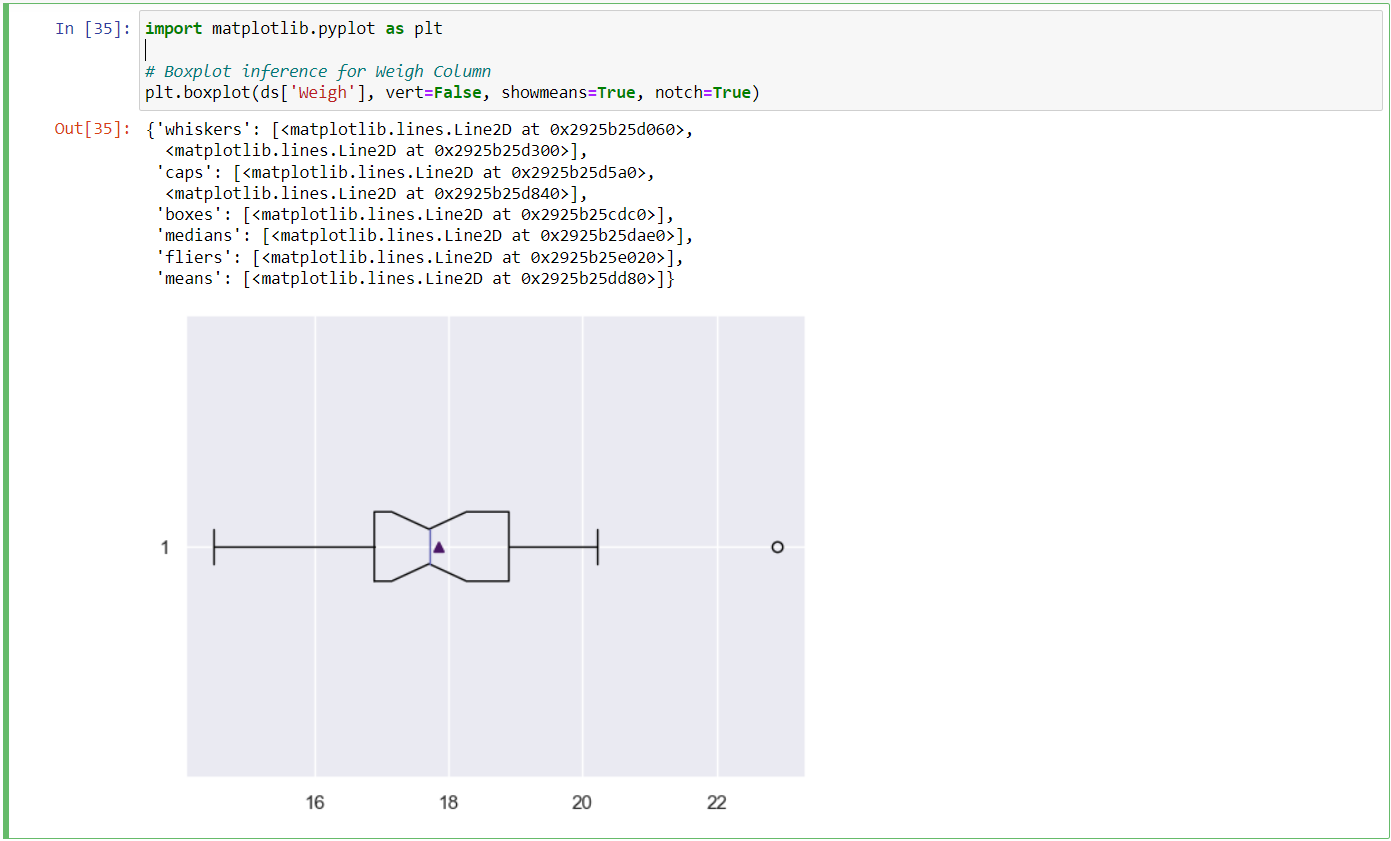
**Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.**

* Ans:









**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?**

* Ans:

Expected Value  =  [(1/9)×108] + [(1/9)×110]  + [(1/9)×123] + [(1/9)×134] + [(1/9)×135] + [(1/9)×145] + [(1/9)×167] + [(1/9)×187] + [(1/9)×199]

= (1/9) × ( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9) × ( 1308)

= 145.33

Expected Value of the Weight of that Patient = 145.33

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

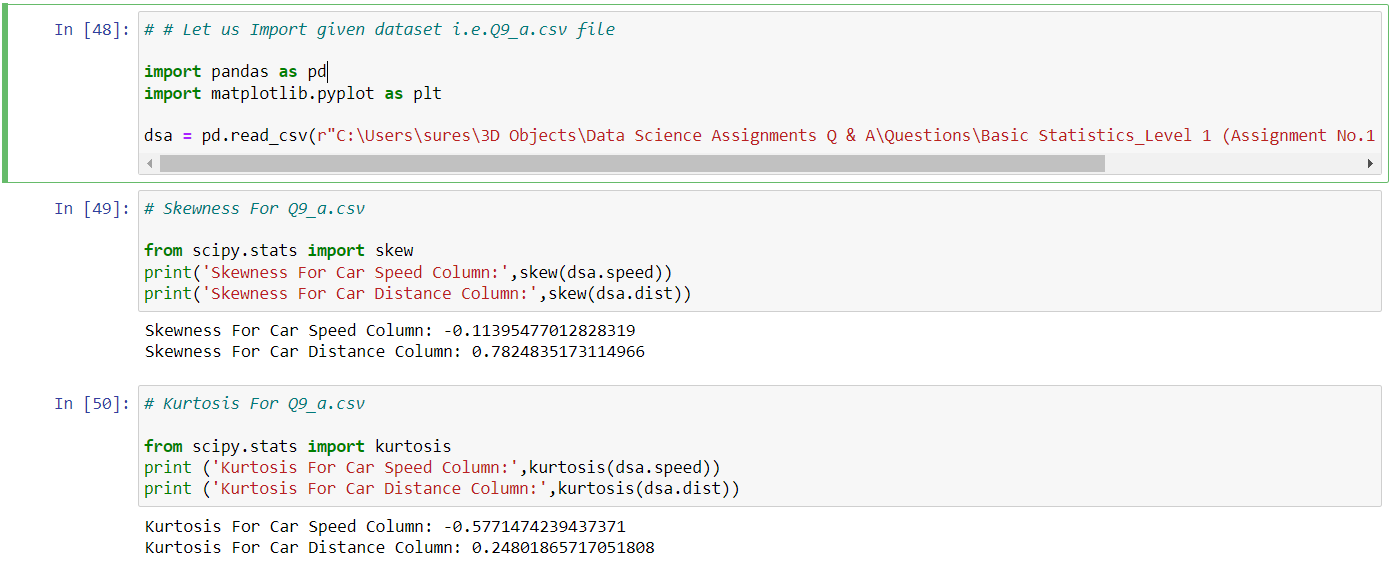
**Cars speed and distance**

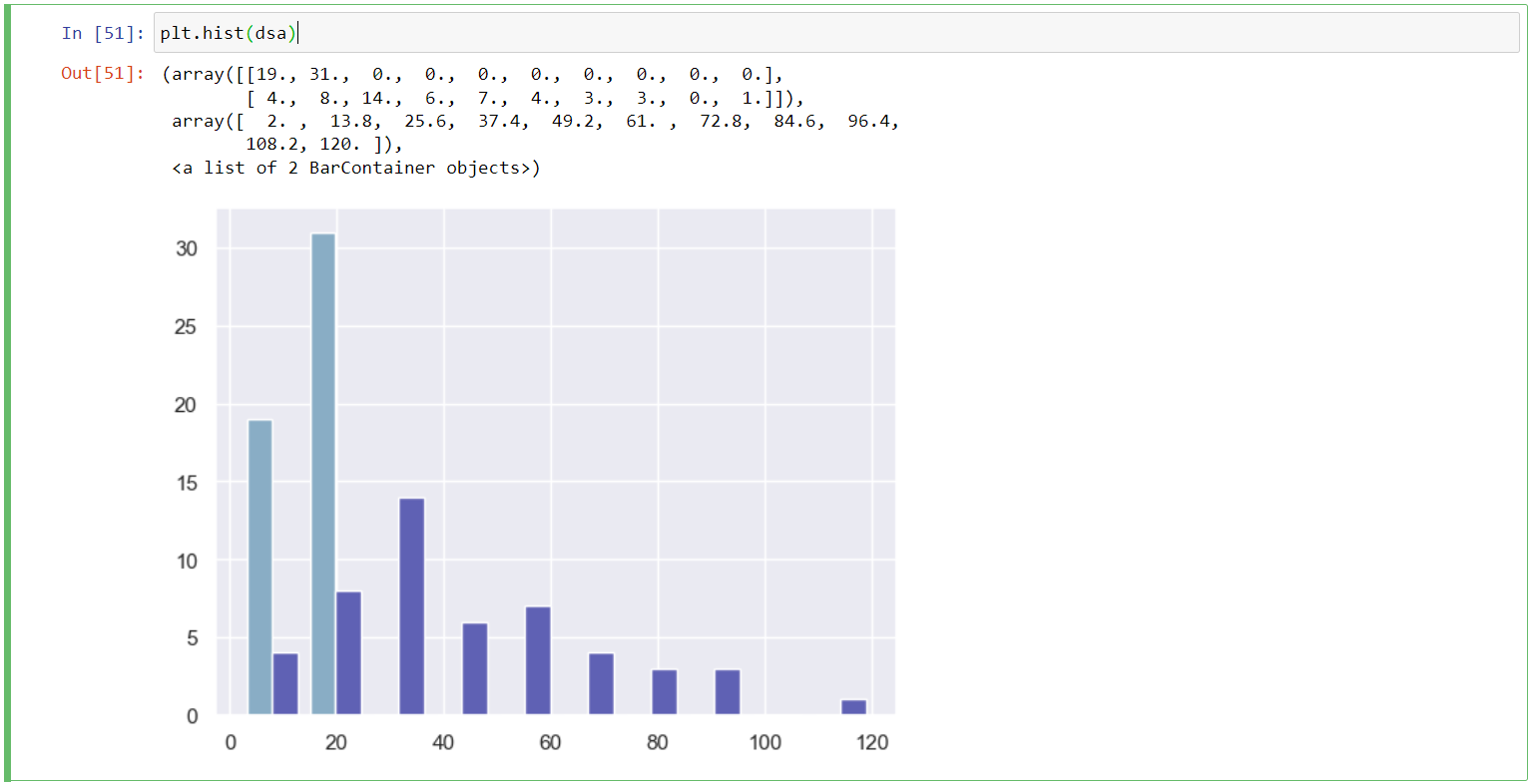
**Use Q9\_a.csv**

**SP and Weight(WT)**

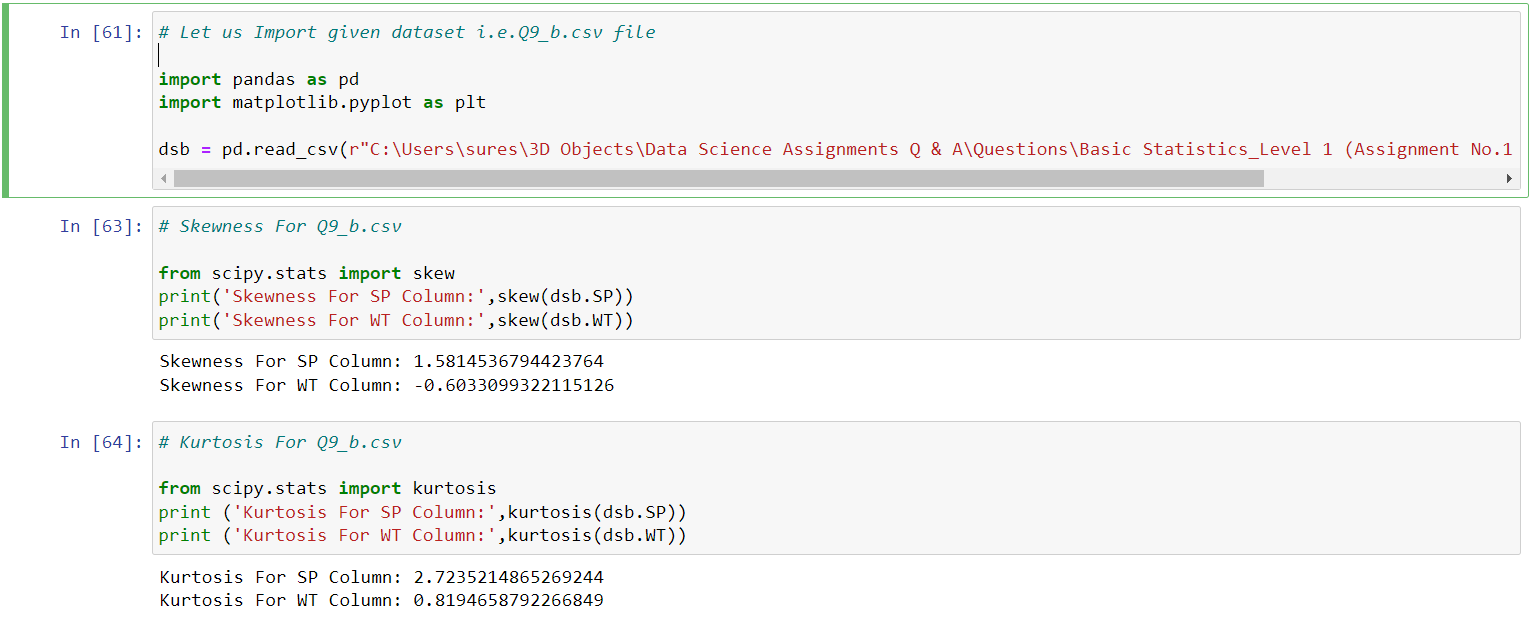
**Use Q9\_b.csv**

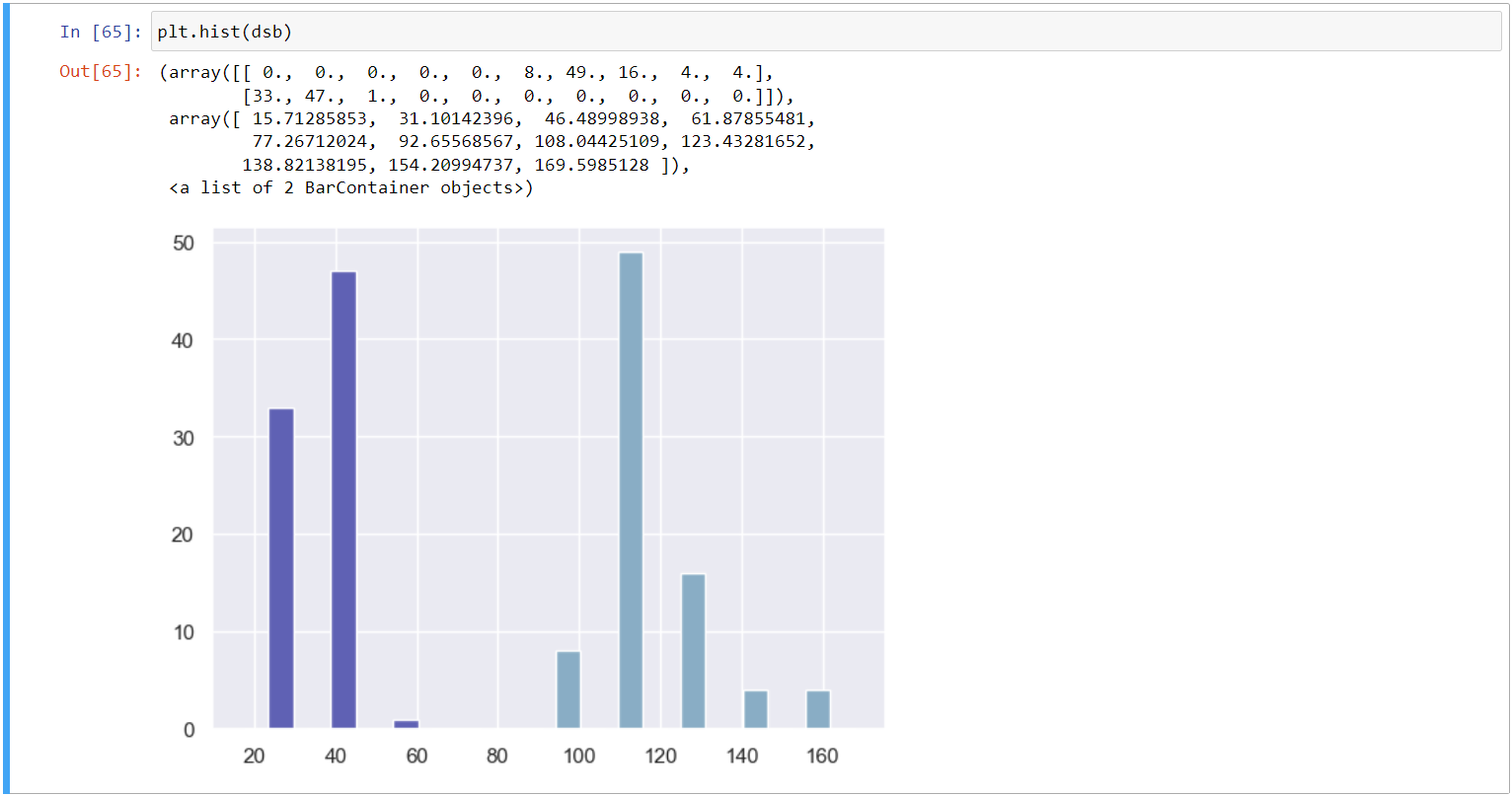
* Ans:





For Q9\_b.csv:





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



* Ans:

Histogram is positively Skew. We can predict that there is outlier above upper extreme. Most of data point is present in between 50-150 wt .

-From the box plot we see that the outlier is actually present above upper extreme and mean>median.

**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?**

* Ans:

Given data:

Sample Size(n) = 2000

Sample Mean(X̅) = 200

Sample SD(S) = 30

α = 0.06

Degree Of Freedom(n-1) = 2000-1

(n-1) = 1999

*(Note: Population parameter is not given therefore we are going to use T-Test)*

Check Inside: T-Table🡪 1999 & 0.06 = 1.89

C.I = X̅ ± t1-n(S/ √n)

C.I(Lower Value) = 200 – 1.89 × (30/ √2000)

= 198.7321

C.I(Upper Value) = 200 + 1.89 × (30/ √2000)

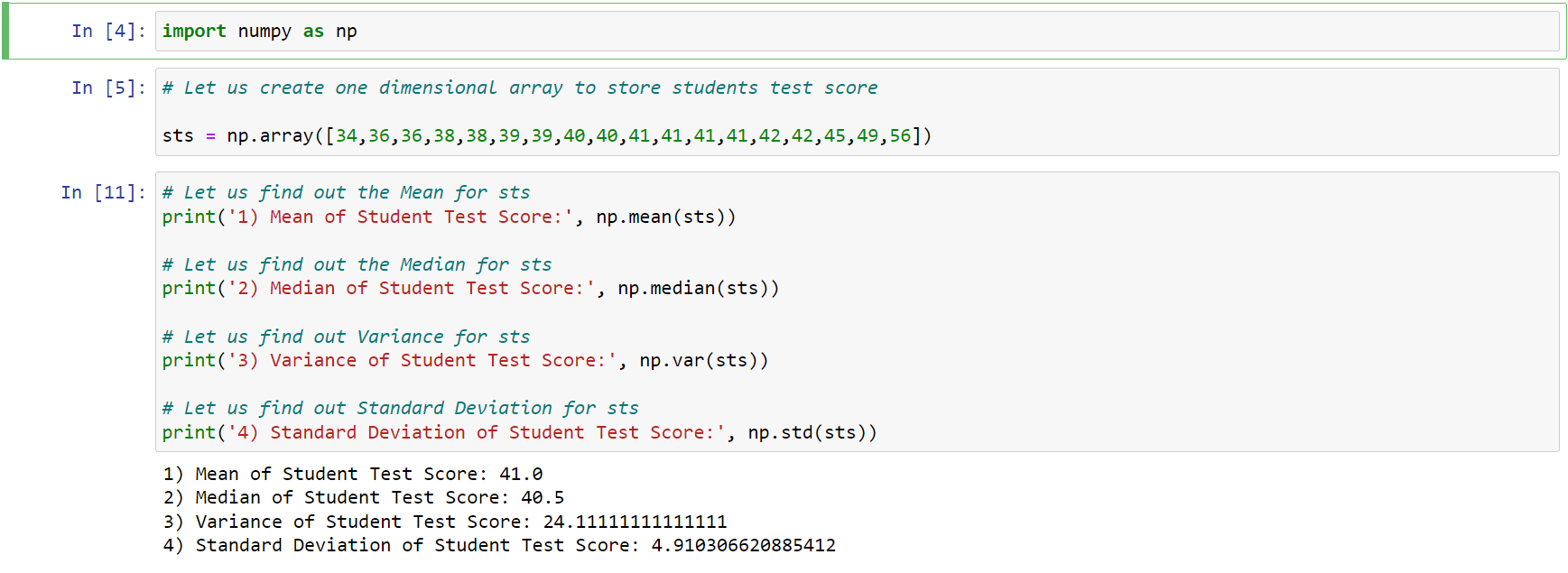
= 201.2678

**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?**

* Ans: 1)🡪



2)🡪 We don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

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

* Ans:

The data is normally distributed and there is perfect symmetry. Thus, there is no skewness is present.

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

* Ans:

It is Positively Skewed and also called as Right Skewed.

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

* Ans:

It is Negatively Skewed and also called as Left Skewed.

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

* Ans:

It indicates the measure of peakedness of the distribution.

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

* Ans:

It indicates the data has lighter tails and a flatter peaks than the normal distribution.

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

****

**a) What can we say about the distribution of the data?**

**b) What is nature of skewness of the data?**

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

* Ans:

a) What can we say about the distribution of the data?

It is not normally distributed, the median is towards the higher value.

b) What is nature of skewness of the data?

It is Negatively Skewed and also called as Left Skewed and the length of whisker of min. value is greater than max. value.

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

⸫ IQR = Q3 – Q1

= 18 - 10

= 08

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

****

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

* Ans:

i) The data is normally distributed because the whisker of both of the sides are symmetrical. Thus, there is no skewness (both boxplot 1 & 2).

ii) There is no outliers in the data (both boxplot 1 & 2)..

iii) Boxplot 1 & 2 has same median.

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

**Data \_set: Cars.csv**

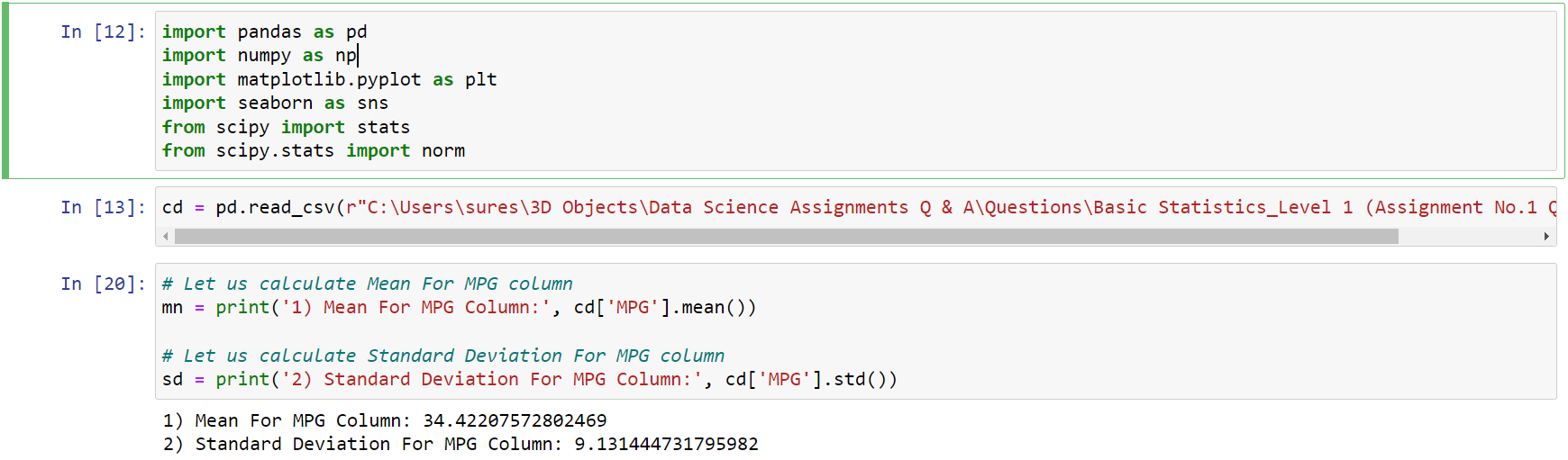
**Calculate the probability of MPG of Cars for the below cases.**

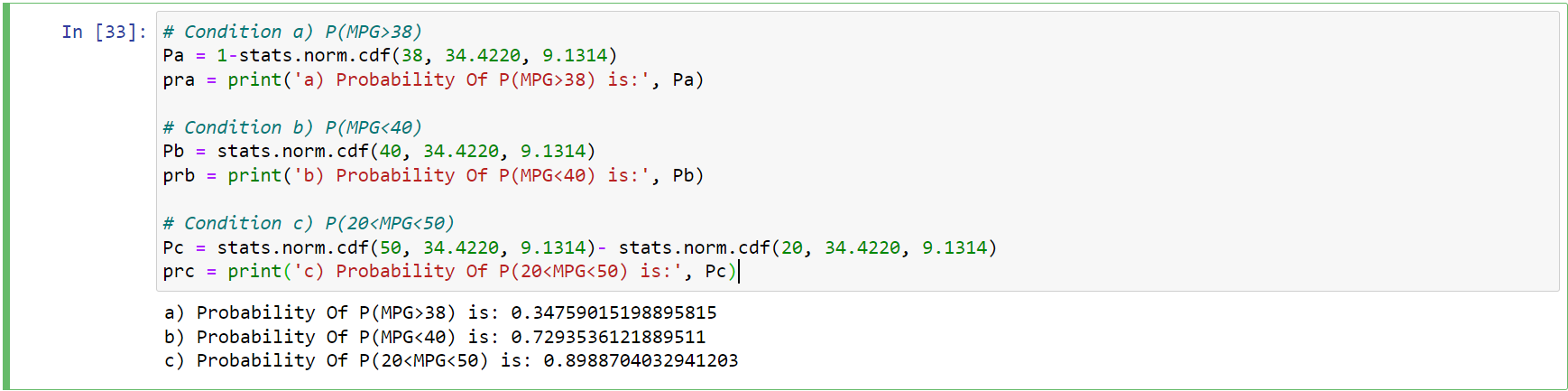
**MPG <- Cars$MPG**

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

**c. P (20<MPG<50)**

* Ans:





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

1. **Check whether the MPG of Cars follows Normal Distribution**

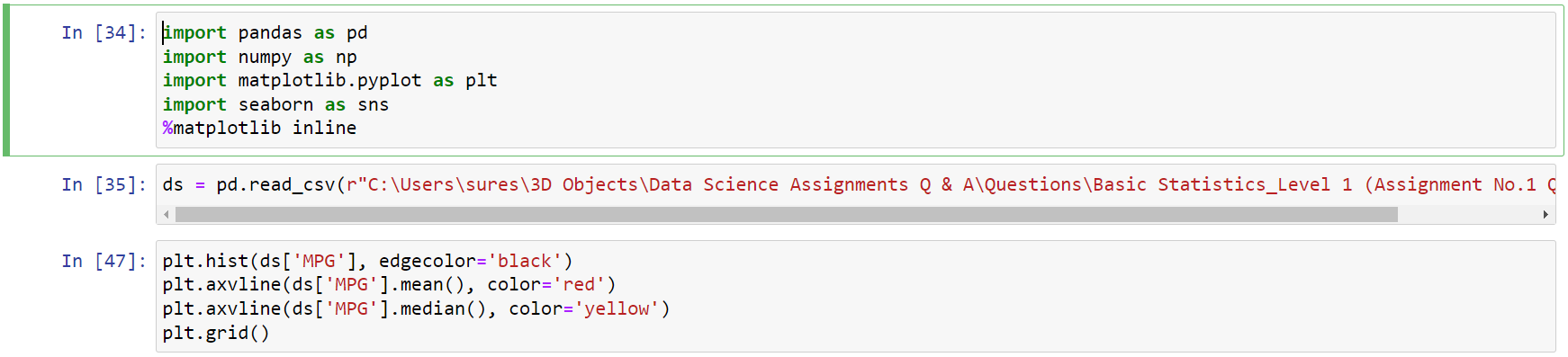
**Dataset: Cars.csv**

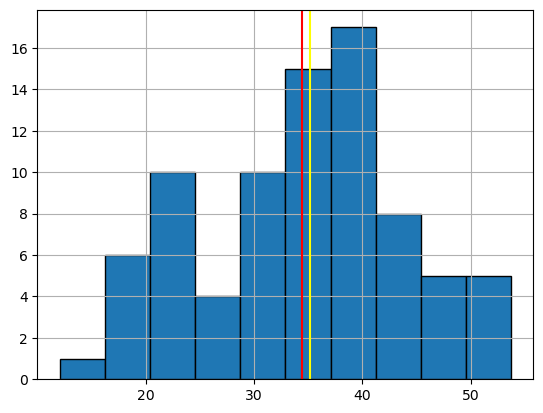
1. **Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution**

**Dataset: wc-at.csv**

* Ans:

a) Check whether the MPG of Cars follows Normal Distribution:

****

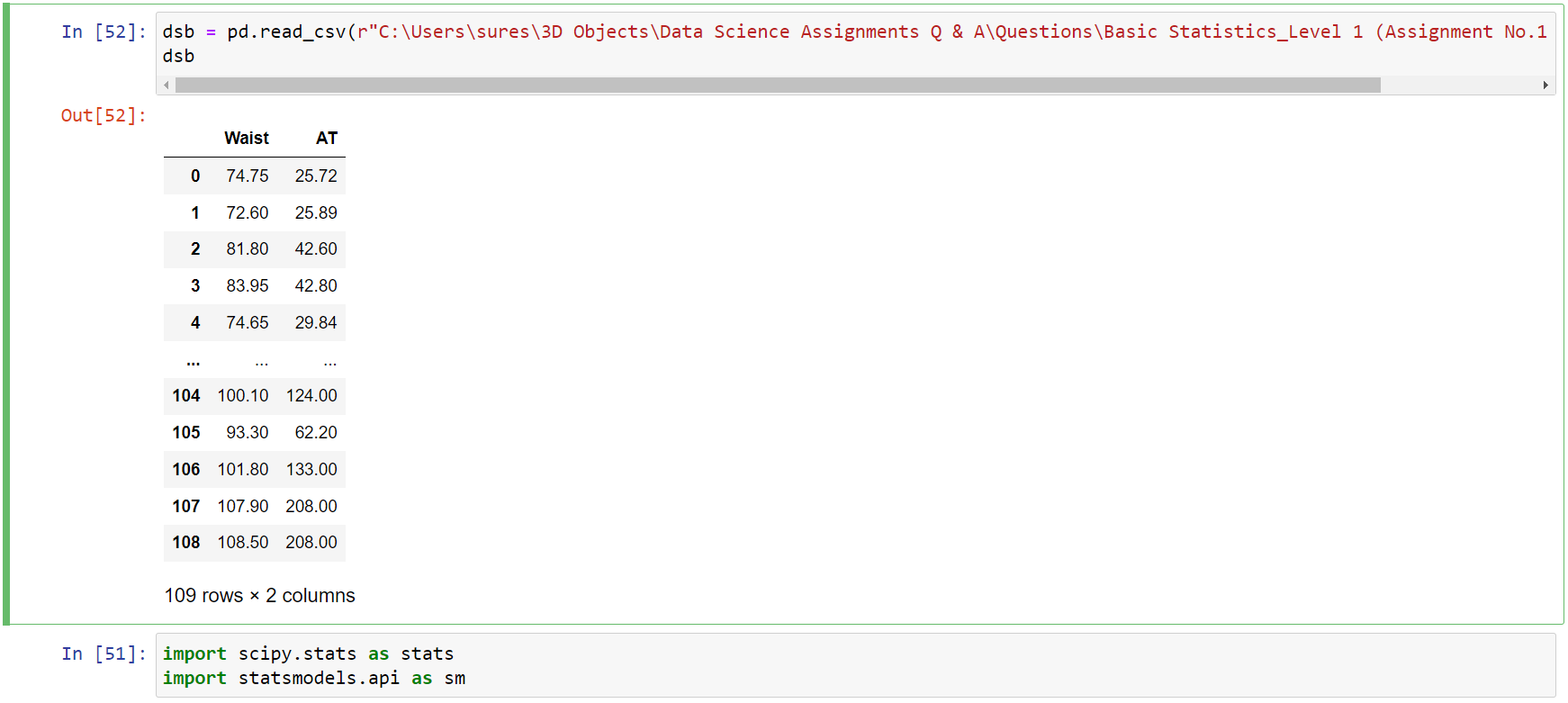
****

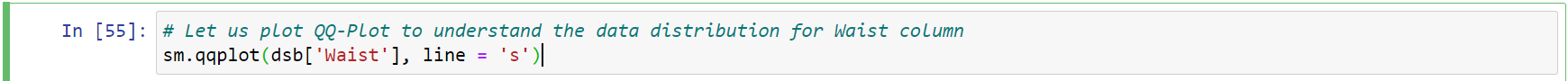
**1) Red Line Indicates Mean**

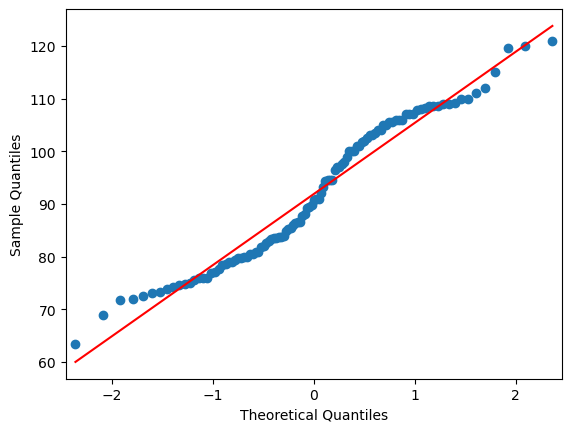
**2) Yellow Line Indicates Median**

Explanation: Mean and Median are nearly equal therefore this MPG column follows normal distribution. And if we draw curve on the bins, the curve has bell shaped that means the MPG column is normally distributed.

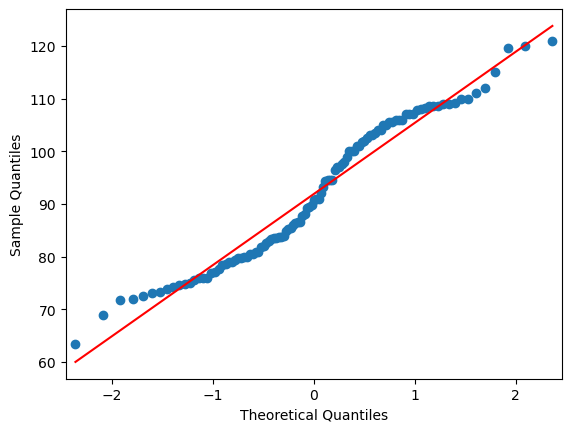
b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution:





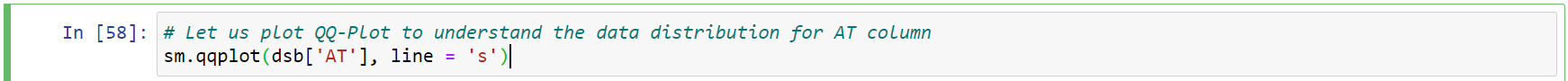


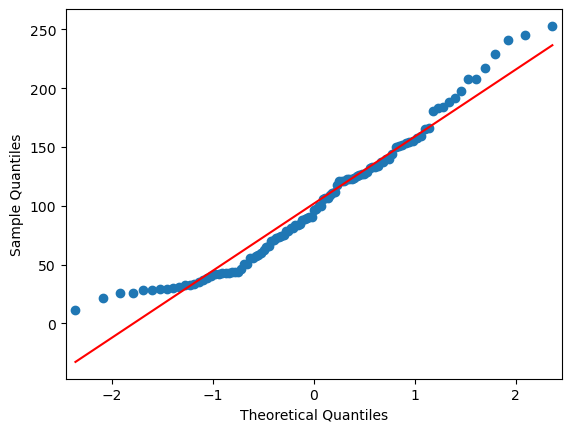
Waist 1



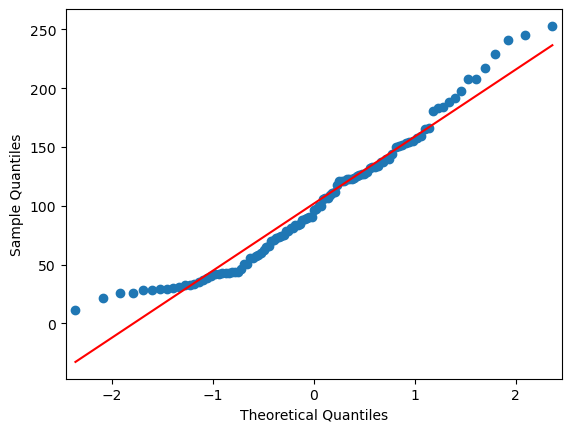
Waist 2

Explanation: The Waist column does not follows the normal distribution.





AT 1



AT 2

Explanation: The AT column does not follows the normal distribution.

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

* Ans:

1) Confidence Interval For 90% 🡪 Z-Score = 1.645

2) Confidence Interval For 94% 🡪 Z-Score = 1.88

3) Confidence Interval For 60% 🡪 Z-Score = 0.84

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

* Ans:

1) Confidence Interval For 95% 🡪 T-Score = 2.063

2) Confidence Interval For 96% 🡪 T-Score = 2.17

3) Confidence Interval For 99% 🡪 T-Score = 2.796

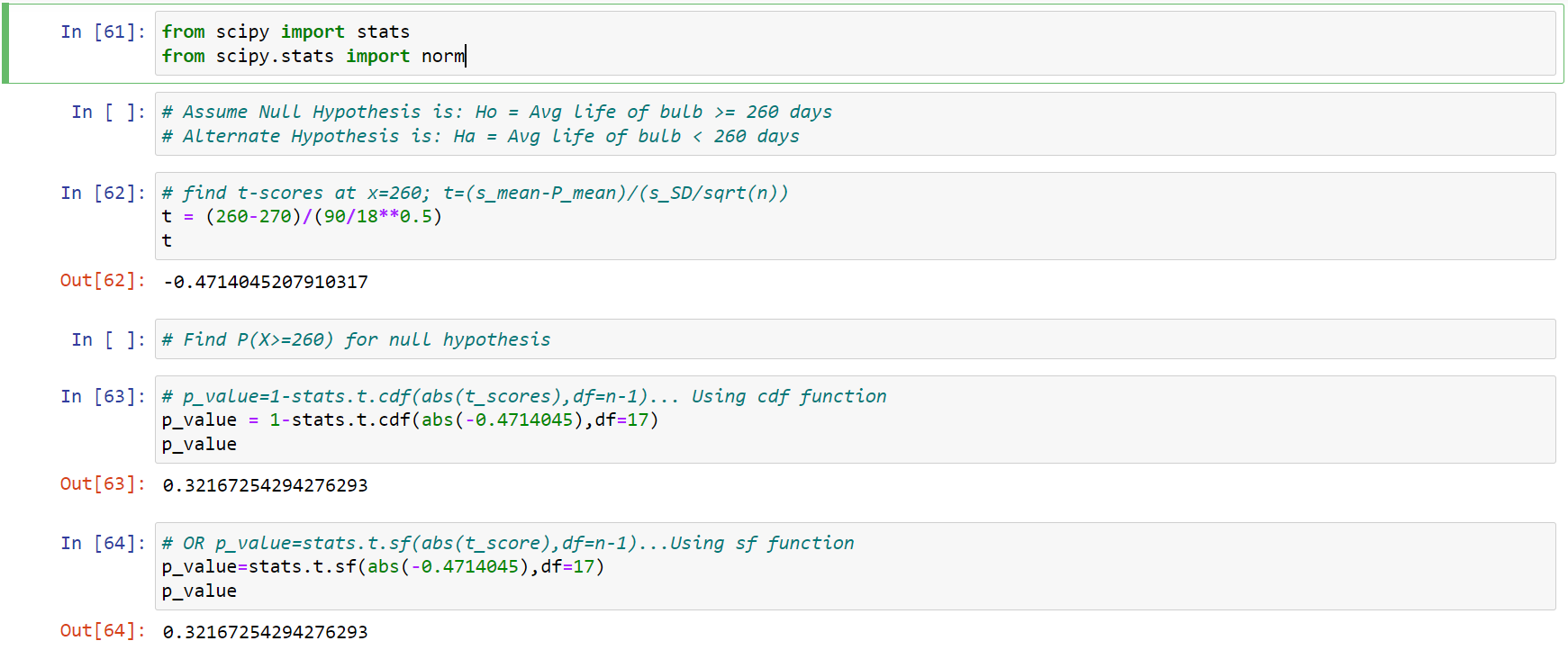
**Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs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

**Hint:**

**rcode 🡪 pt(tscore,df)**

**df 🡪 degrees of freedom**

* Ans:

****

**\*\*\*THANK YOU\*\*\***