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

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 | **Ordinal** |
| Time on a Clock with Hands | **Interval** |
| Number of Children | **Ratio** |
| Religious Preference | **Nominal** |
| Barometer Pressure | **Ratio** |
| SAT Scores | **Interval** |
| Years of Education | **Ratio** |

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

**Ans. Three Coins are tossed**

**{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}**

**No of sample n(s) = 8**

**The probability of getting two heads and one tail on tossing three coins at once is equal to {HHT, HTH, THH}**

**P(E)=n(E)/n(S) P(E)=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. a) Equal to 1. = 0**

The minimum possible sum is (1, 1) = ( 2 ).

Therefore P( 1 ) = ( 0 )/( 36 ) = **0.**

**b) less than or equal to 4 = (1/6)**

A number less than or equal to 4 can be achieved with number combinations (1,1)(1,2)(1,3)( 2, 1)(2,2) and (3, 1), that is only with 3 combinations of numbers.

P(x>= 4) = ( 6/ 36 ) = **( 1 / 6).**

**c) Sum is divisible by 2 and 3**

A sum is divisible by 2 and 3 combination are (1,5)(2,4)(3,3)(4,2)(5,1)(6,6) that is only with 5 combination of numbers.

P(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?

**Ans.** **There are 7 balls originally with 2 of them blue so the probability of the first ball not being blue is 5/7.**

**This leaves 6 balls with 2 blue.**

**The probability of the second ball not being blue assuming that the first wasn’t is 4/6.**

**The probability that neither ball drawn was blue is (5/7)\*(4/6)=20/42=10/21.**

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. (0.015\*1)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120) = 3.09**

**Expected number of candies for a randomly selected child = 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>

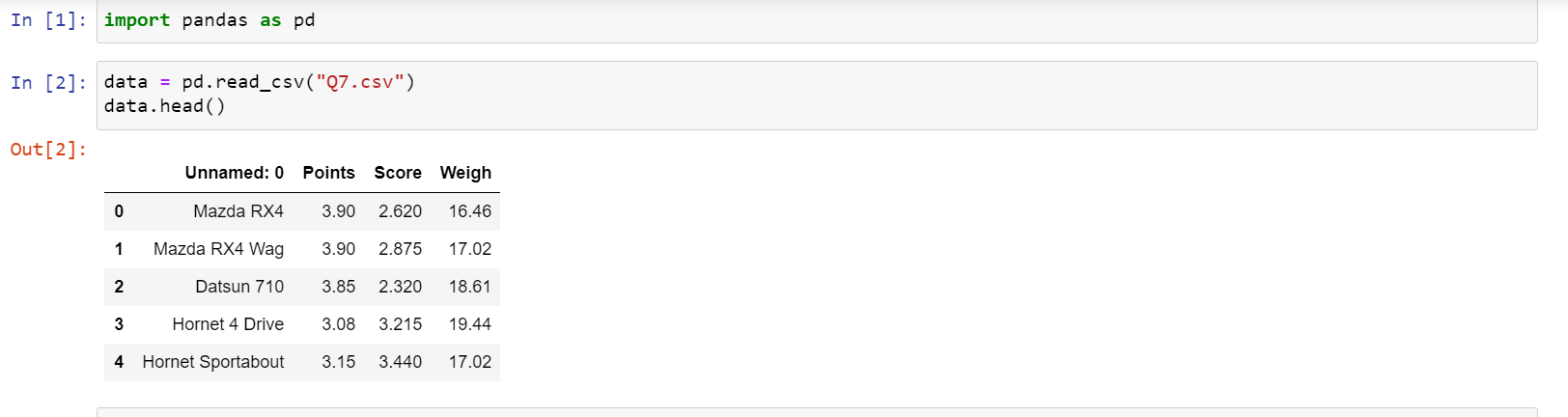
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

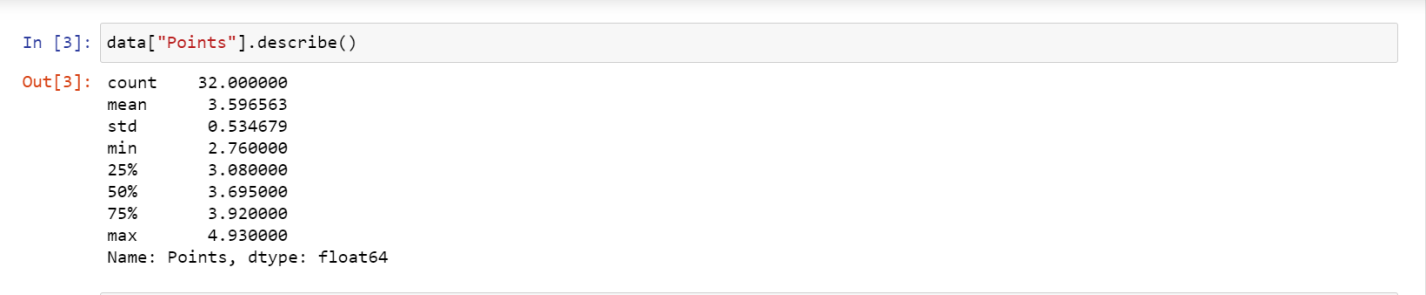
**Use Q7.csv file**

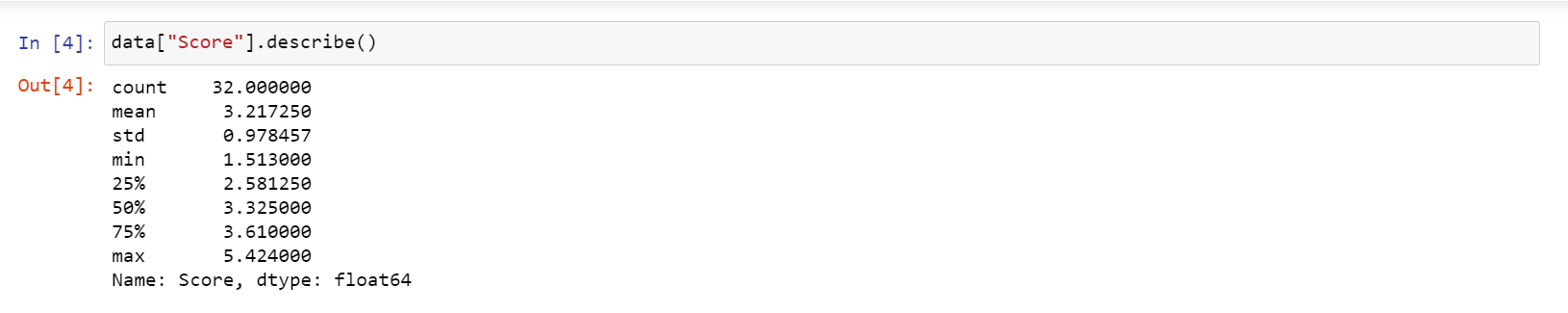
**Ans.** Ans:

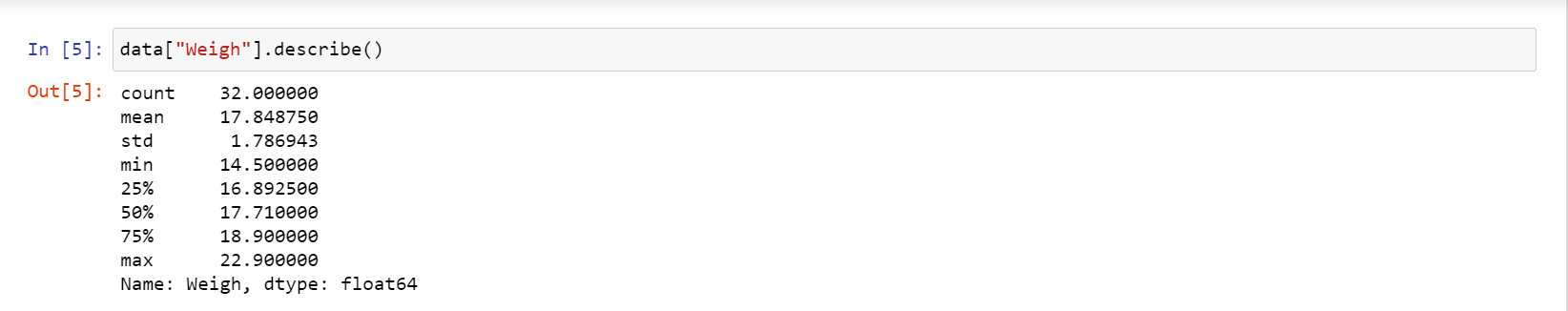
|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| MEAN | 3.596563 | 3.21725 | 17.84875 |
| MEDIAN | 3.695 | 3.325 | 17.71 |
| MODE | 3.92 | 3.44 | 17.02 |
| Variance | 0.285 | 0.957 | 3.19 |
| Std. Dev | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.17 | 3.911 | 8.4 |

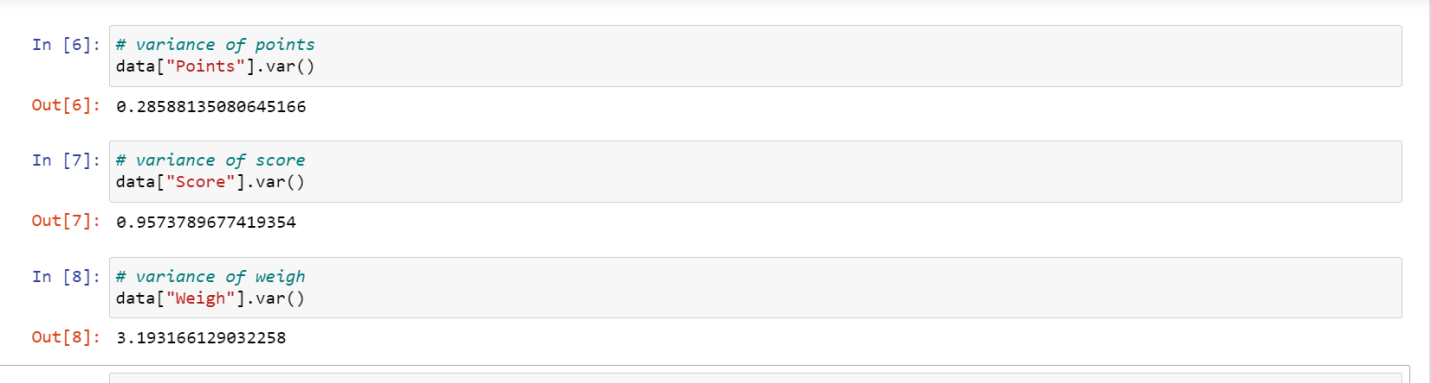
* In the above table the none of the variable are Mean = Median = Mode .
* Two of the columns i.e. “Score” and “Weigh” in the data set contains outliers.
* In “Points” and “Scores” Mean<Median<Mode so we can say that the data may be negatively skewed.
* In “Weigh” Mean>Median>Mode so we can say that the data is positively skewed.



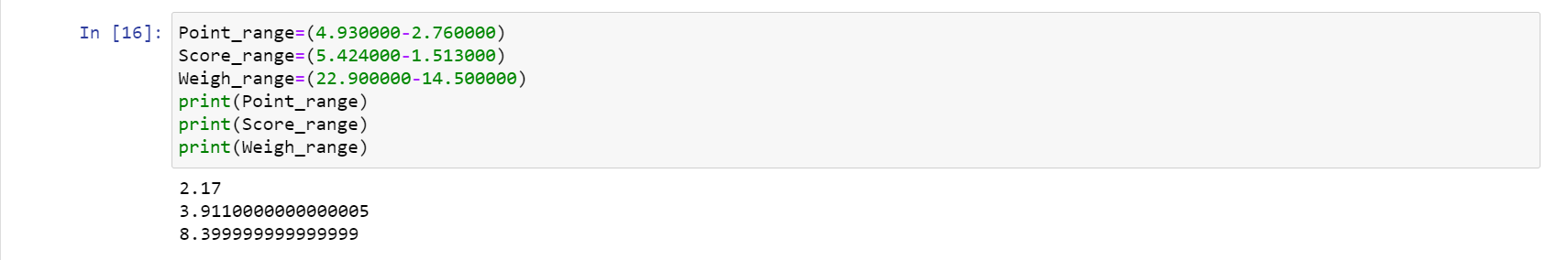












Q8) Calculate Expected Value for the problem below

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 = ∑ ( probability \* Value )**

**∑ P(x).E(x)**

**There are 9 patients**

**Probability of selecting each patient = 1/9**

**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

**Ans.**

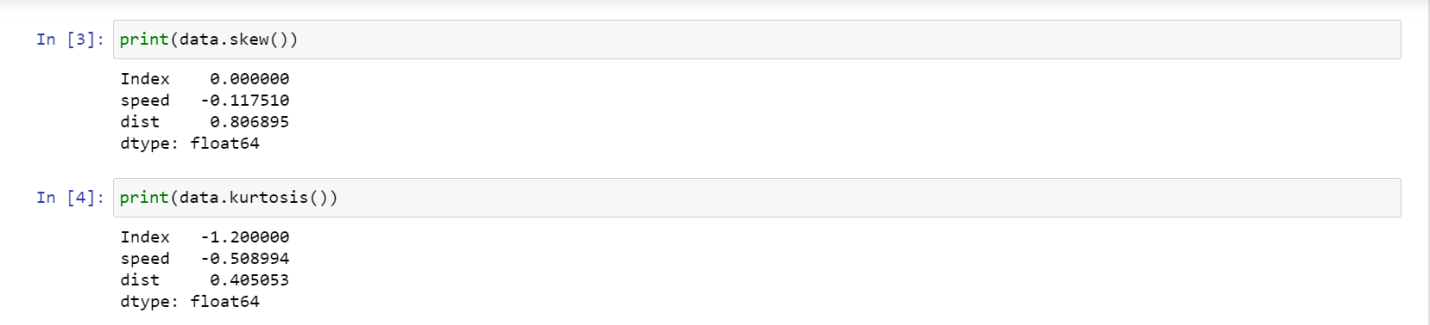
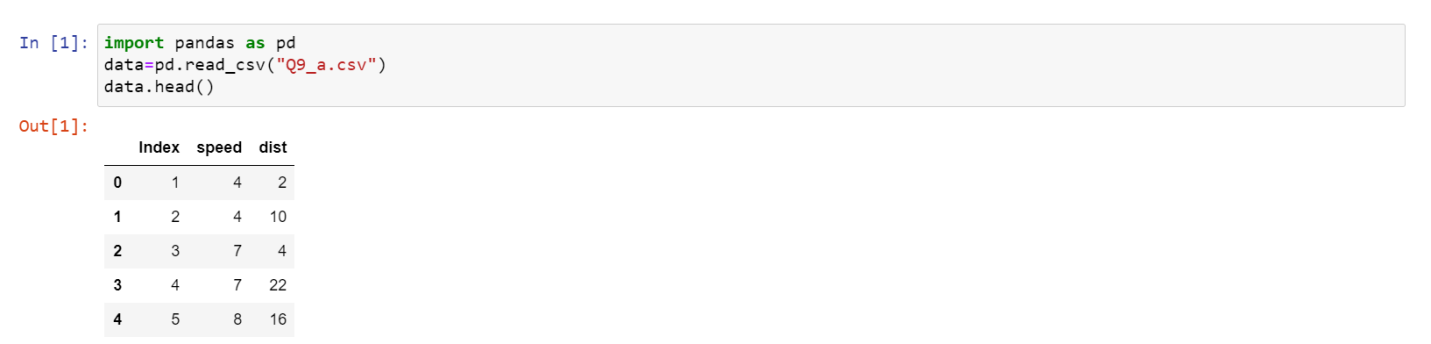
1. **Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
| **Cars** | Skewness | Kurtosis |
| Column(speed) | **-0.11751** | **-0.508994** |
| Column(dist) | **0.806895** | **0.405053** |

**Speed is Negatively skewed and plati kurtosis**

**Dist is Positively skewed and lepto kurtosis**



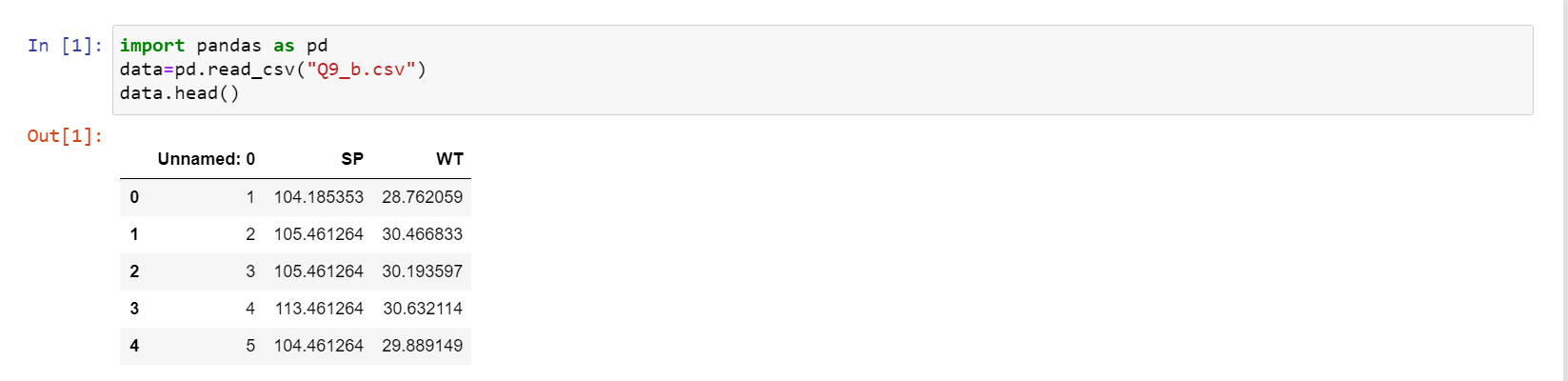
**B) SP and Weight(WT)**

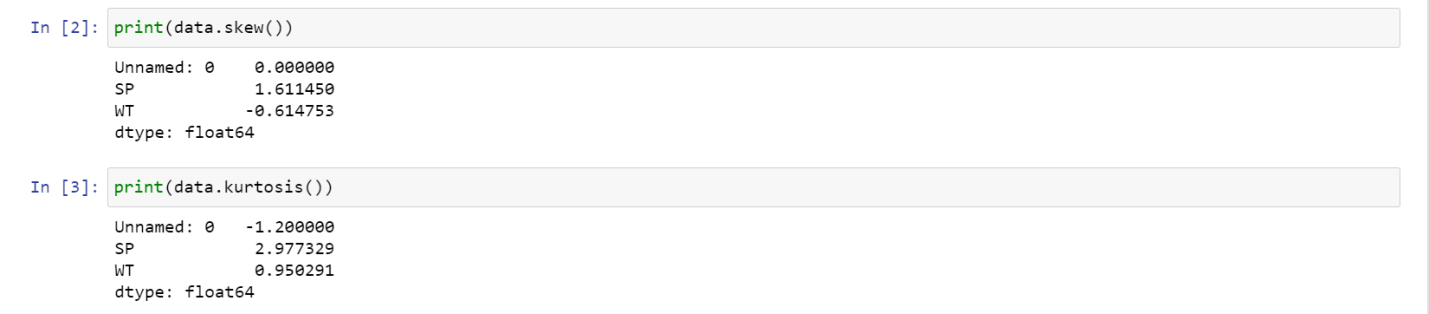
**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| Column(SP) | **1.611450** | **2.977329** |
| Column(Weight) | **-0.614753** | **0.950291** |

**Sp is positively skewed and lepto kurtosis**

**WT is negatively skewed and lepto kurtosis**





Q10) Draw inferences about the following boxplot & histogram



**Histogram:**

* This data is positively skewed.
* Many chickens are having weights in between 50-100 gm (unit)
* Very rare no of chickens are having weight more than 300gm
* After a certain point i.e. (100gm) the frequency of the chicken decreases with increase in their weight.



**Box Plot:**

* This boxplot represents the data is positively skewed
* The data contains outliers

**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:

**mean = 200, Std(s) = 30, n = 2000, df=1999**

**(1-α) =94%,98%,96%**

* **t-value of 94% is 1.55543**

**[200-1.55\*30/sqrt (2000,200+1.55\*30)/sqrt(2000)]**

**94% of confidence interval for weigh is [198.9569,201.04]**

* **t-value of 98% is 2.055089**

**[200-2.055\*30/sqrt(2000,200+2.055\*30)/sqrt(2000)]**

**98% of confidence interval for weigh is [198.621,201.37]**

* **t-value of 96% is 1.75157**

**[200-1.7515\*30/sqrt(2000,200+1.7515\*30)/sqrt(2000)]**

**96% of confidence interval for weigh is [198.824,201.175]**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 94% | 98% | 96% |
| Upper | 201.04 | 201.37 | 201.175 |
| Lower | 198.9569 | 198.621 | 198.824 |

****

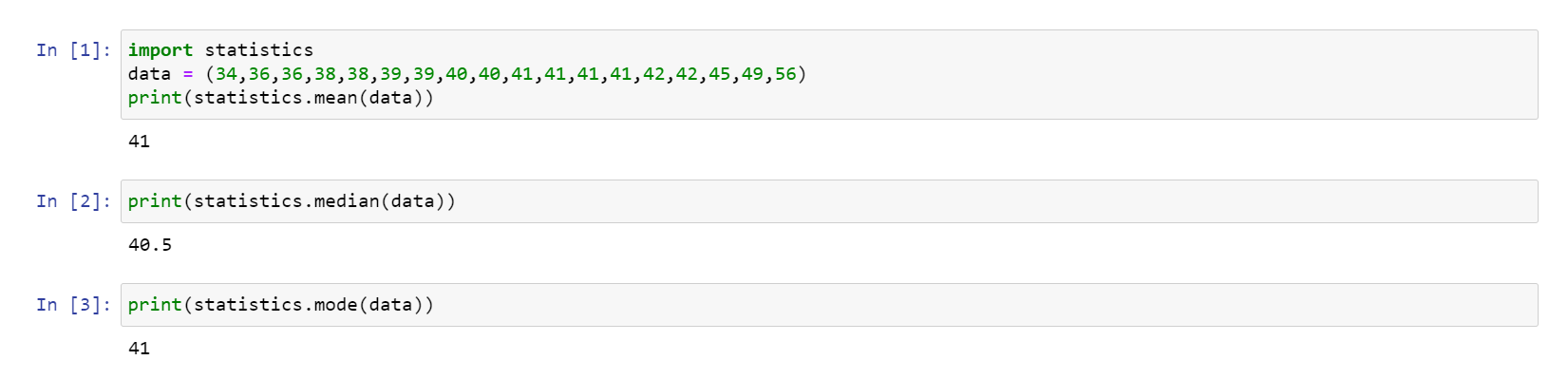
**Q12)** Below are the scores obtained by a student in tests

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?

**1)Ans.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mean** | **Median** | **Mode** | **Variance** | **Std Deviation** |
| **41** | **40.5** | **41** | **25.5294** | **5.0526** |

****

****

2)Ans:

* Marks are not normally distributed.
* Person with mark 56 can be the outlier in our data.
* Many of the students are having 41 mark.

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

**Ans.**

**If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.**

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

**Ans.**

**If the mean is greater than the median, the distribution is positively skewed**.

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

**Ans.**

**If the median is greater than the mean, the distribution is negatively skewed**.

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

**Ans.**

**Positive values of kurtosis indicate that a distribution is peaked and possess thick tails. Leptokurtic distributions have positive kurtosis values.**

**A leptokurtic distribution has a higher peak and taller tails than a normal distribution.**

**An extreme positive kurtosis indicates a distribution where more of the values are located in the tails of the distribution rather than around the mean.**

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

**Ans.**

**Negative values of kurtosis indicate that a distribution is flat and has thin tails. Platykurtic distributions have negative kurtosis values. A platykurtic distribution is flatter when compared with the normal distribution, with fewer values in its shorter tails.**

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



What can we say about the distribution of the data?

**Ans:**

The data is not symmetric. Mass of the data is concentrated towards right side.

There is no outliers**.**

What is nature of skewness of the data?

**Ans:**

**When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).**

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

**Ans. Q3 = 18, Q1 = 10,**

**Q3 – Q1 = 18-10 = 8**

The Inter Quartile Range of the data is 8.  
  
  
  
  
Q19) Comment on the below Boxplot visualizations?



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

**Ans:**

* **Both the boxplot doesn’t have outliers.**
* **Both have same median value as (262.5).**
* **Comparing Boxplot\_1 (Red) has less variability, less variation, less standard deviation, less range, less Inter-quartile-range value as compared to Boxplot\_2(blue)**

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

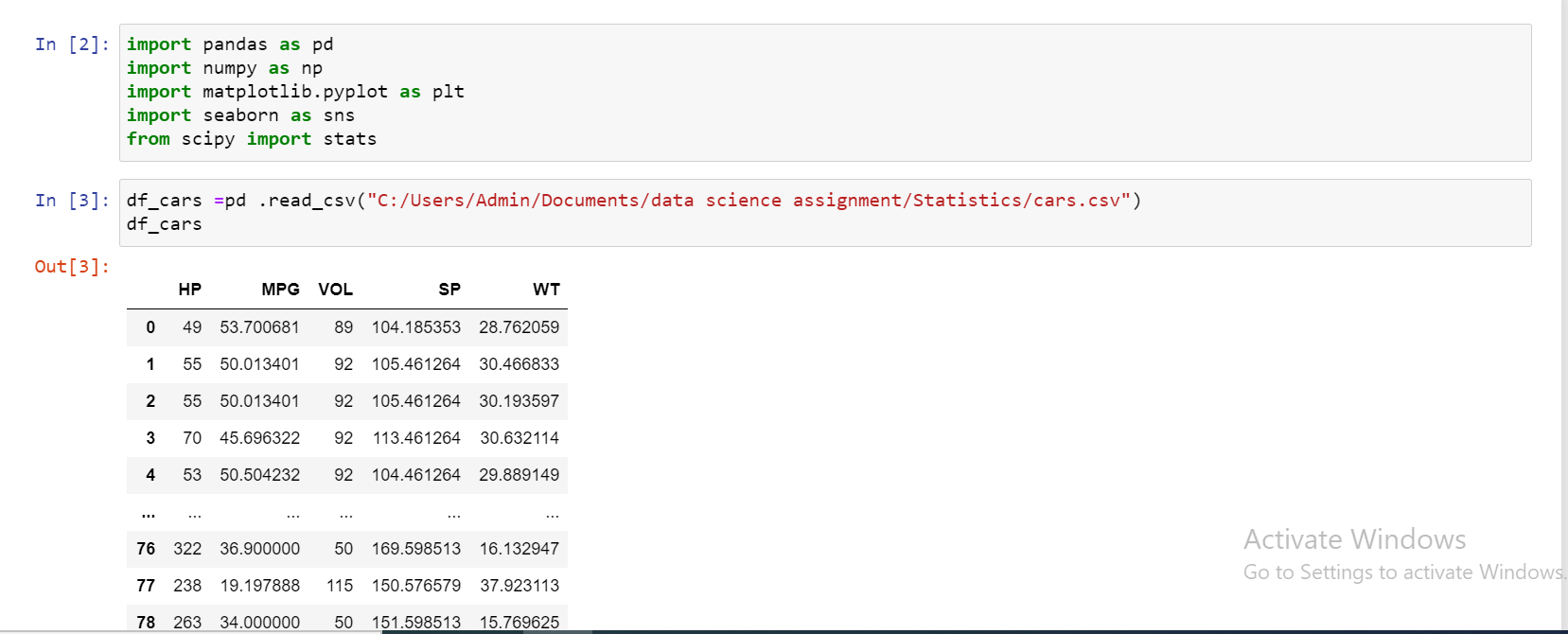
Data \_set: Cars.csv

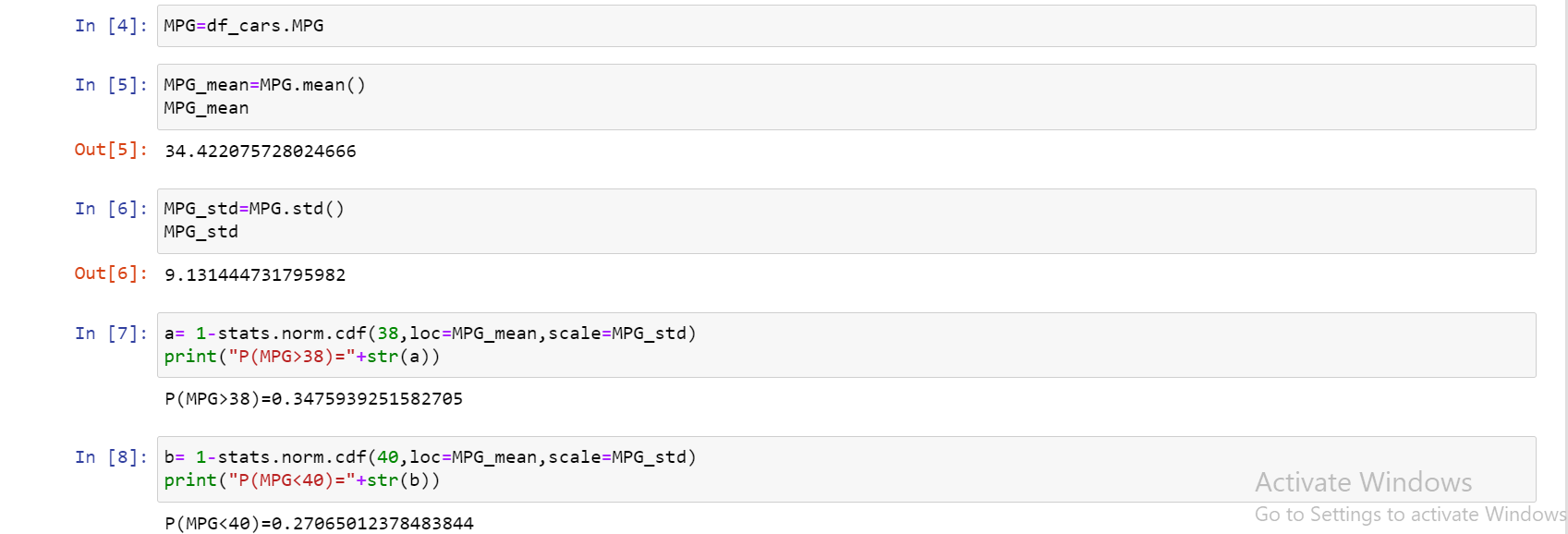
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38) Ans: 0.34759
  2. P(MPG<40) Ans: 0.27065

c. P (20<MPG<50) Ans : 0.89886





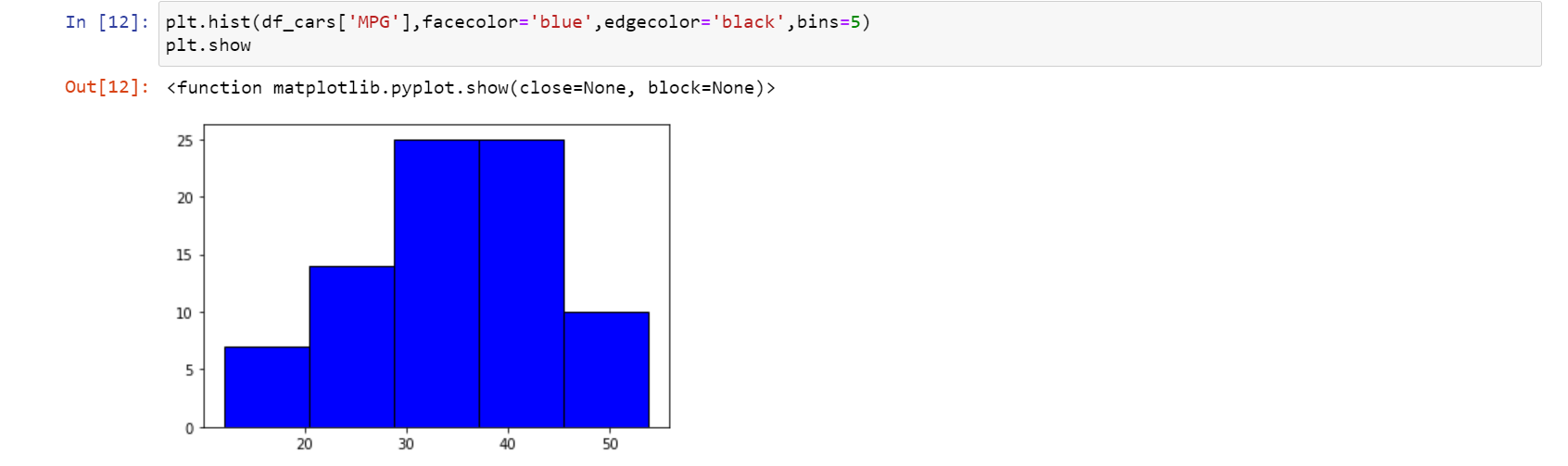


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

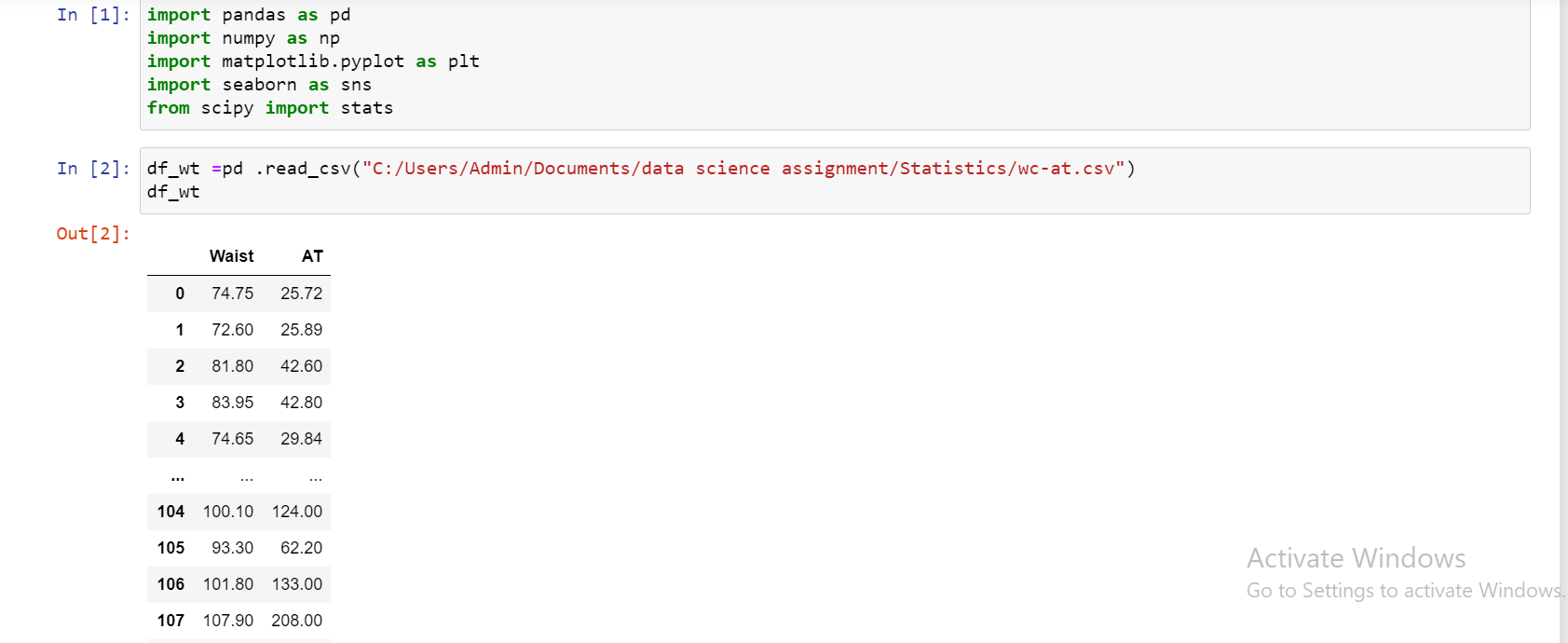
Ans : Yes MPG of cars follows Normal Distribution

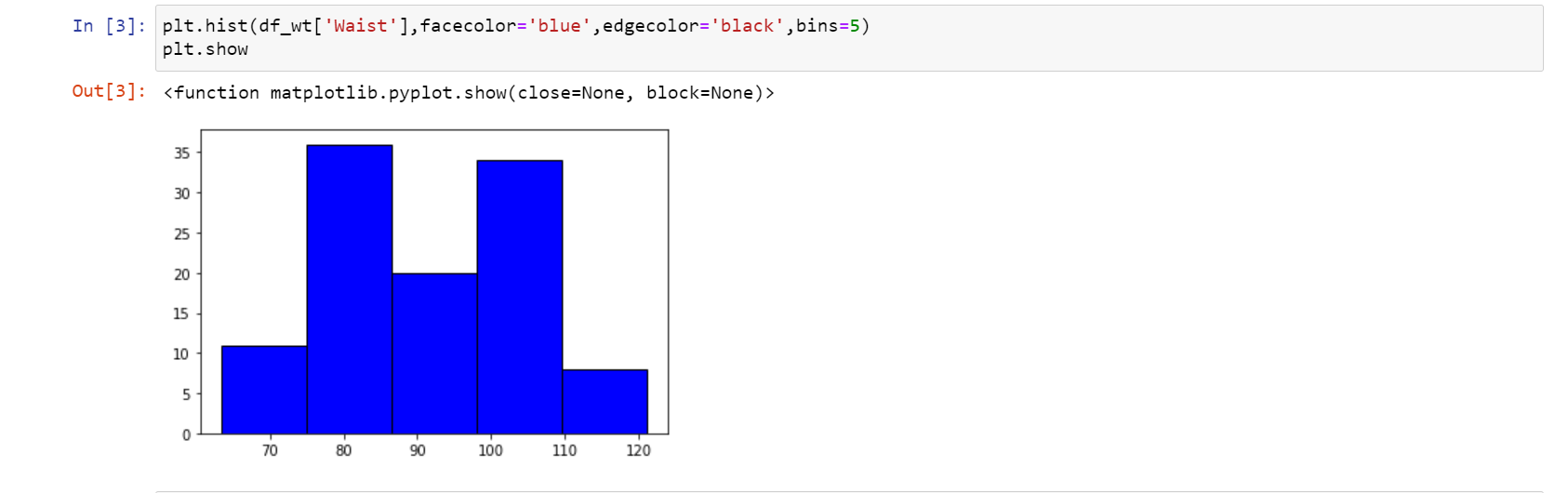


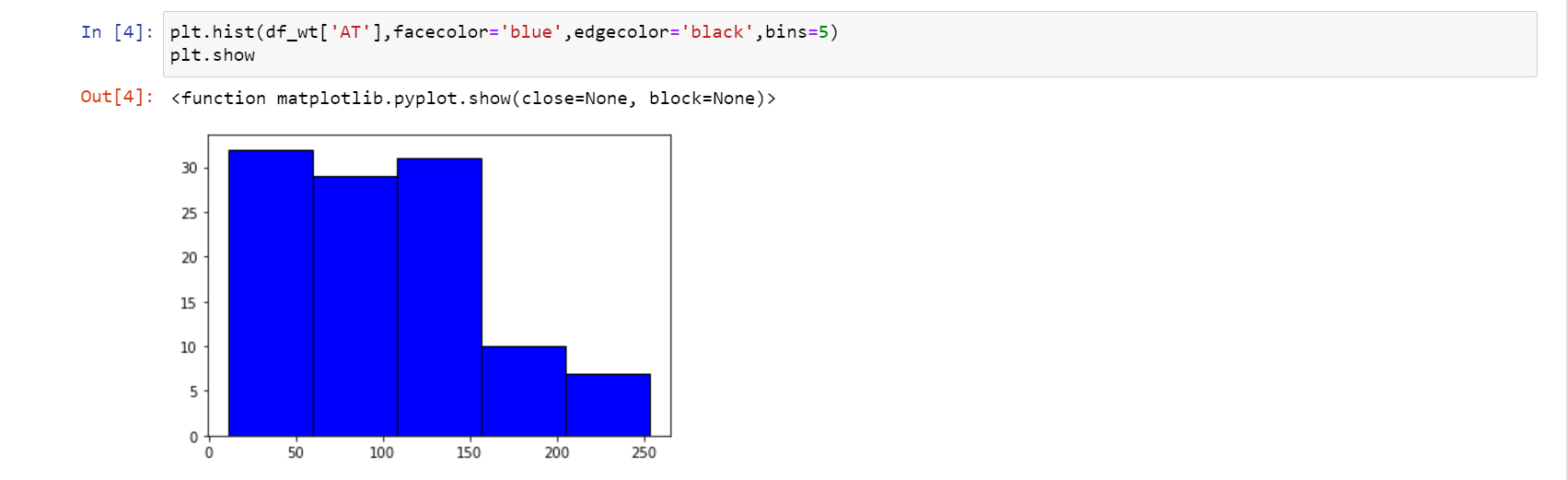
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Ans : None of the data follows Normal distribution







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

**Ans.**

|  |  |  |
| --- | --- | --- |
| 90% | 94% | 60% |
| ± 1.644854 | ± 1.880794 | ± 0.8416212 |

**stats.norm.ppf(0.95) = 1.644 for 90%**

**stats.norm.ppf(0.97) = 1.88 for 94%**

**stats.norm.ppf(0.80) = 0.841 for 60%**

****

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

**Ans.**

|  |  |  |
| --- | --- | --- |
| 95% | 96% | 99% |
| ± 2.06389 | ± 2.17154 | ± 2.79693 |
| 1.710882 | 1.928051 | 2.492159 |

**stats.t.ppf(0.975,df=24) = 2.06389**

**stats.t.ppf(0.98,df=24) = 2.17154**

**stats.t.ppf(0.995,df=24) = 2.79693**

****

****

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.**

**https://tex.z-dn.net/?f=t%3D%5Cdfrac%7Bx-%5Cmu%7D%7B%5Cfrac%7Bs%7D%7B%5Csqrt%20n%7D%7D**

**x = mean of the sample of bulbs = 260**

**μ = population mean = 270**

**s = standard deviation of the sample = 90**

**n = number of items in the sample = 18**

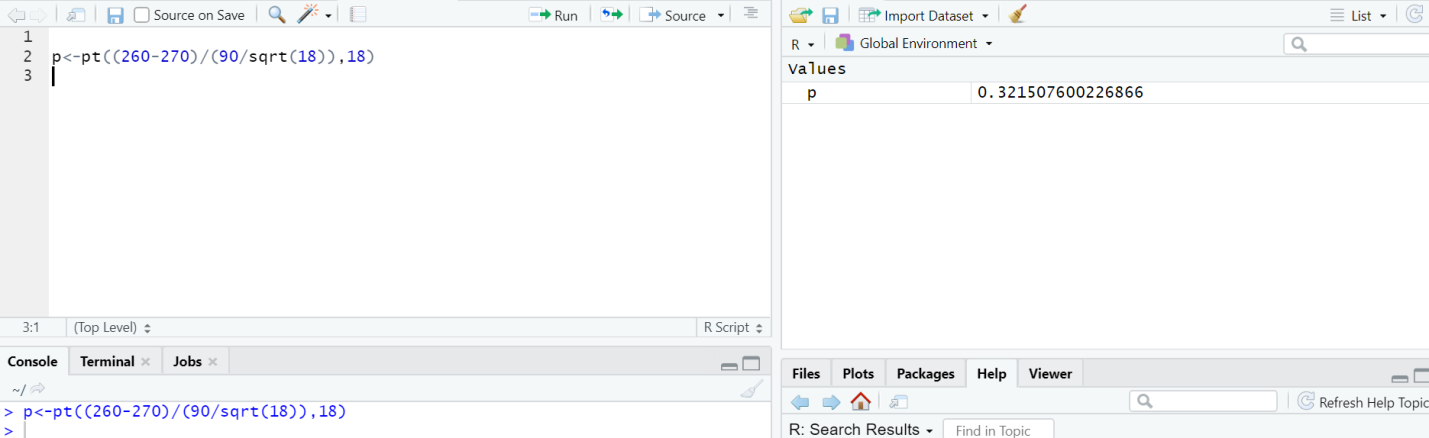
**https://tex.z-dn.net/?f=t%3D%5Cdfrac%7B260-270%7D%7B%5Cfrac%7B90%7D%7B%5Csqrt%2018%7D%7D**

**t = - 0.471**

**For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.**

**The probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of 0.3215 assuming the mean life of the bulbs is 300 days.**

P(t) = 0.3215

****