|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Numeric data(discrete) |
| Results of rolling a dice | Numeric data(discrete) |
| Weight of a person | Numeric data(continuous) |
| Weight of Gold | Numeric data(continuous) |
| Distance between two places | Numeric data(continuous) |
| Length of a leaf | Numeric data(continuous) |
| Dog's weight | Numeric data (continuous) |
| Blue Color | String data(nominal) |
| Number of kids | Numeric data(discrete) |
| Number of tickets in Indian railways | Numeric data(discrete) |
| Number of times married | Numeric data(discrete) |
| Gender (Male or Female) | String data(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 | ordinal |
| Level of Agreement | ordinal |
| IQ(Intelligence Scale) | ratio |
| Sales Figures | ratio |
| Blood Group | nominal |
| Time Of Day | ratio |
| Time on a Clock with Hands | interval |
| Number of Children | ratio |
| Religious Preference | Ordinal |
| Barometer Pressure | interval |
| SAT Scores | ratio |
| Years of Education | Nominal |

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

Solution:

Total combinations are HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

Possible outcomes = p(H,H,T) +p(H,T,H)+p(T,H,H)

P(HHT) = 3/8

Q4) Two Dice are rolled, find the probability that sum is

Solution:

Total possible outcomes n(s)=36  
a) the sum is equal to 1 is zero because they starts with (1,1) likewise . other than in the dice we are not having zero.  
b) the sum is equal to 4 the possible outcomes are (1,3),(2,2),(3,1) therefore n( b) = 3/36 = 1/12  
c) sum is divisible by 2 and 3 = 6/36 i.e 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?

Solution:

Total number of balls = (2 + 3 + 2) = 7  
 n(S) = Number of ways of drawing 2 balls out of 7  
7C2​  
=(7×6)/(2\*1)​  
=21  
Let E = Event of drawing 2 balls, none of which is blue.

n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.  
=5C2​  
=(5\*4)​/(2\*1)  
=10  
P(E)=n(E)/n(S)​=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.

Solution:

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

= 3.090

=  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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | mean | median | mode | variance | std | range |
| Points | 3.596563 | 3.695 | 3.07 | 0.285881 | 0.534679 | 2.17 |
| Score | 3.217250 | 3.325 | 3.443 | 0.957379 | 0.978457 | 3.911 |
| weigh | 17.848750 | 17.710 | 17.02 | 3.193166 | 1.786943 | 8.399 |

Score values are Right tail(mode<median<mean)

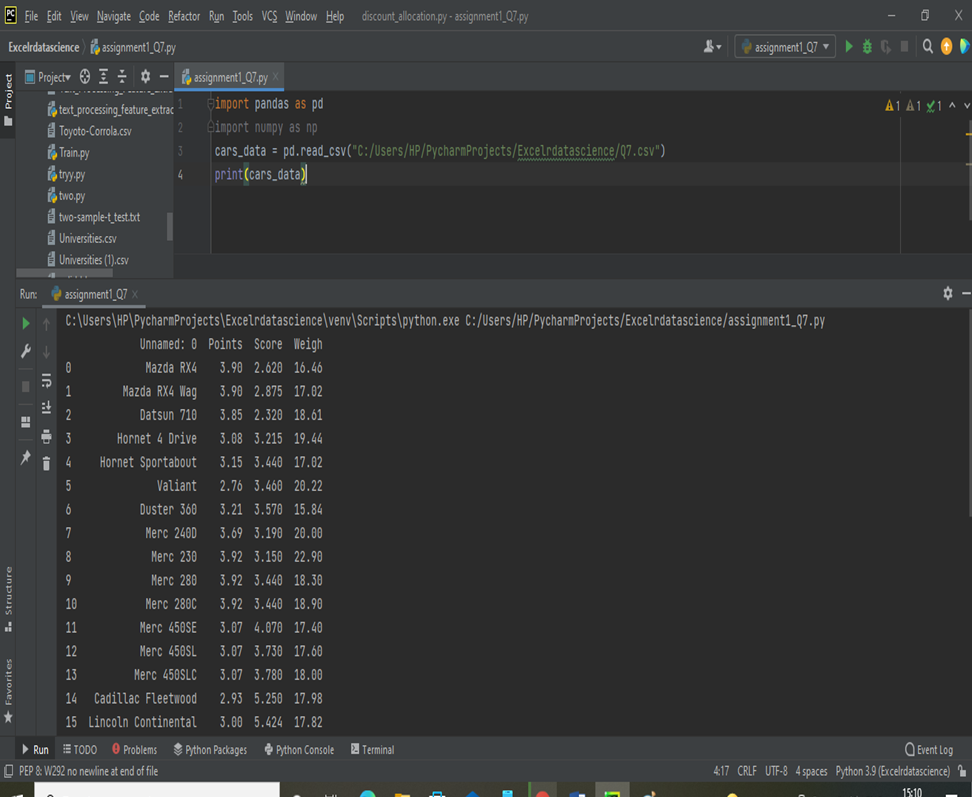
Weigh values are left tail(mean<median<mode)

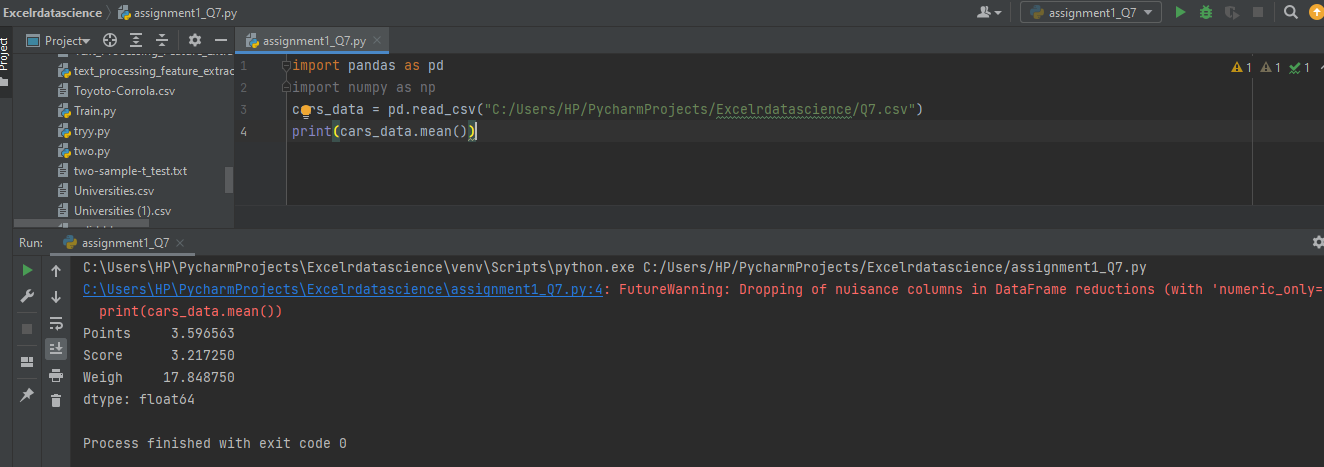
Point’s mean and median is almost same

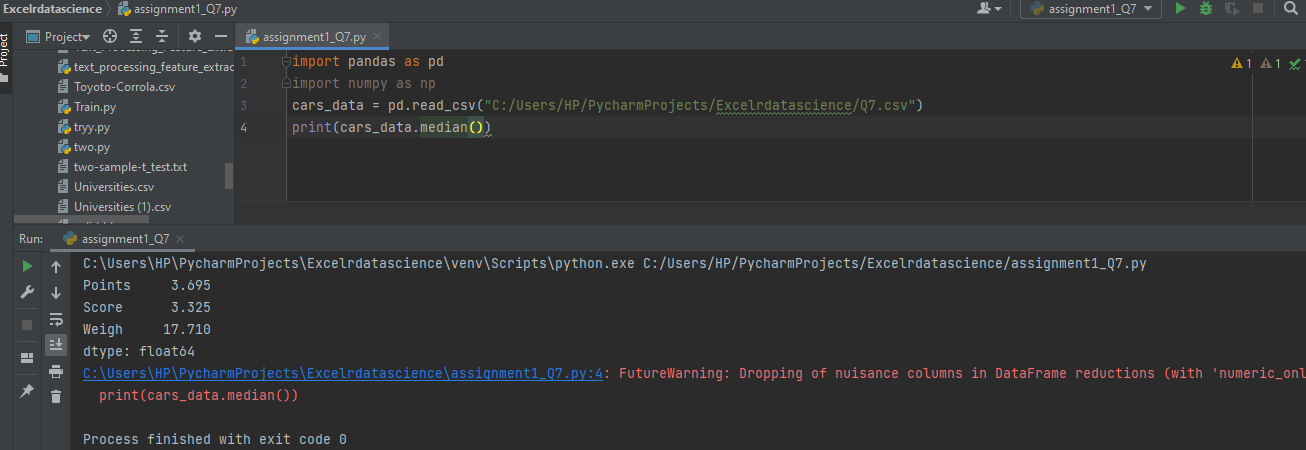
The points and score data is less spread.

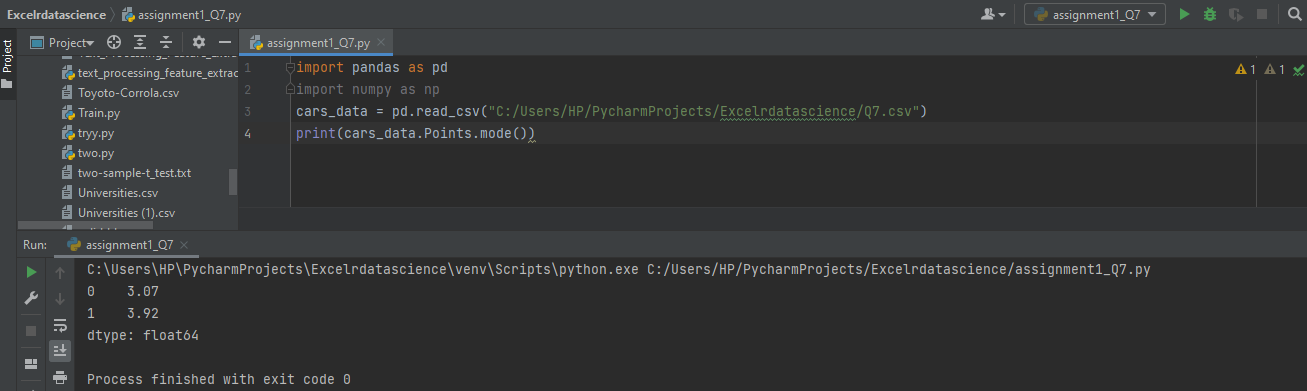
Solution:

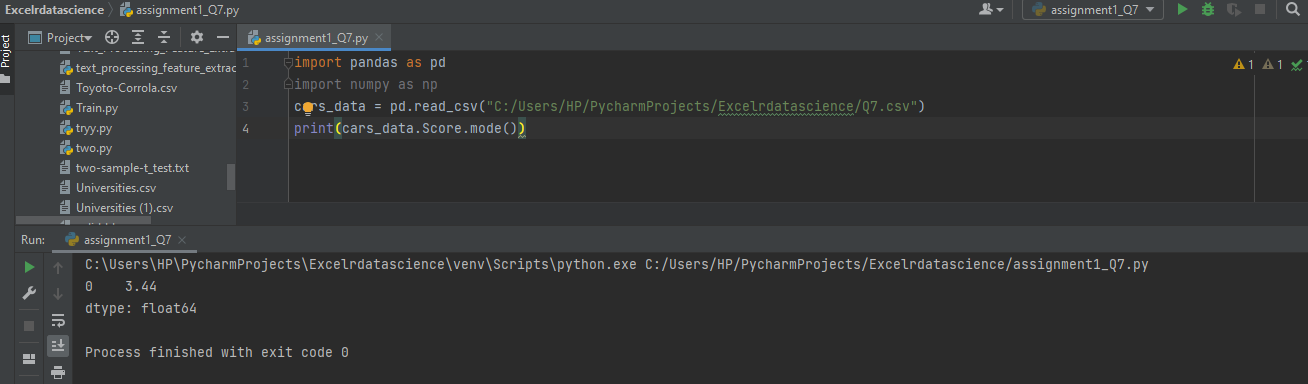
Reading csv in pycharm id………

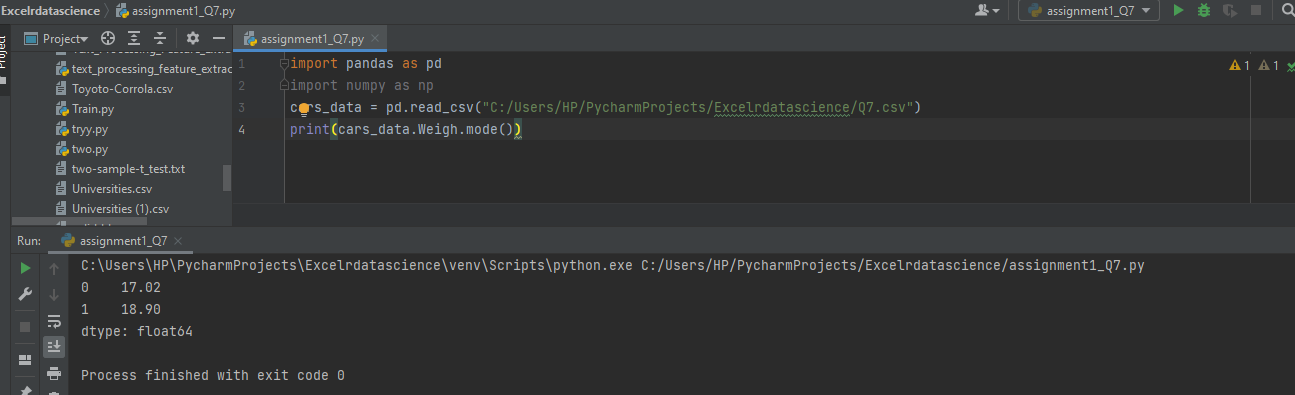


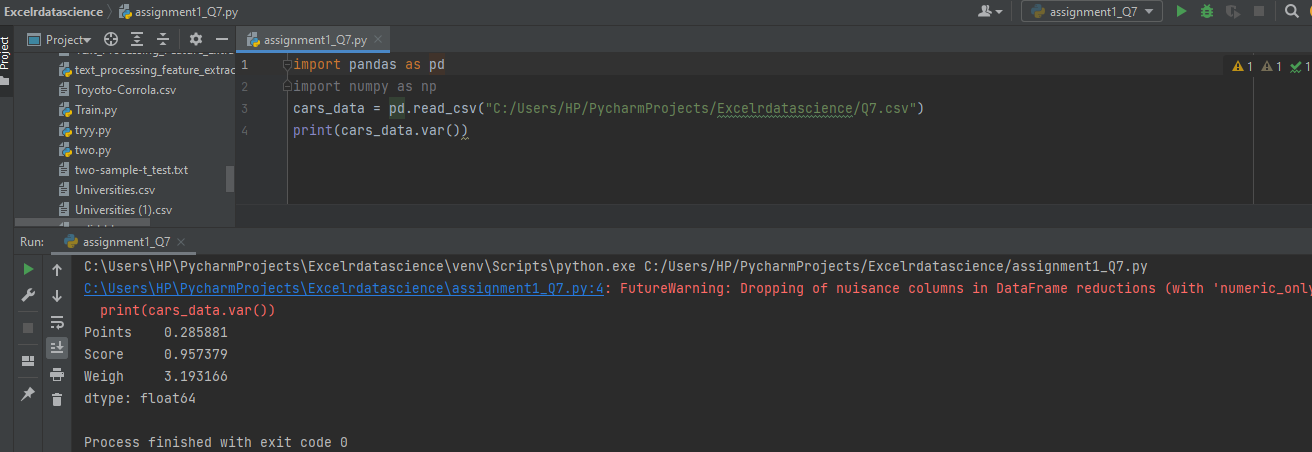
**MEAN**

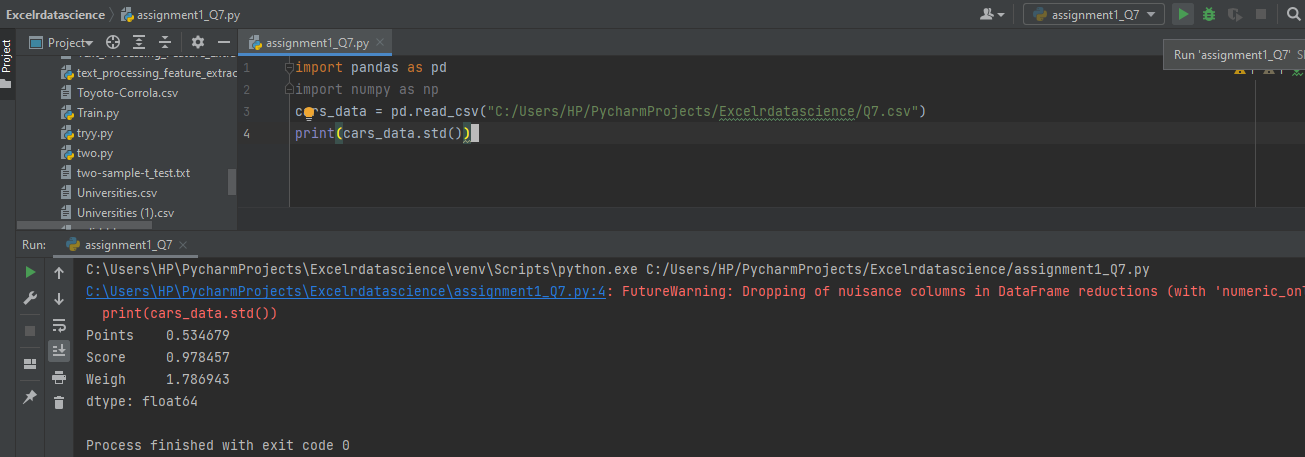
**MEDIAN**

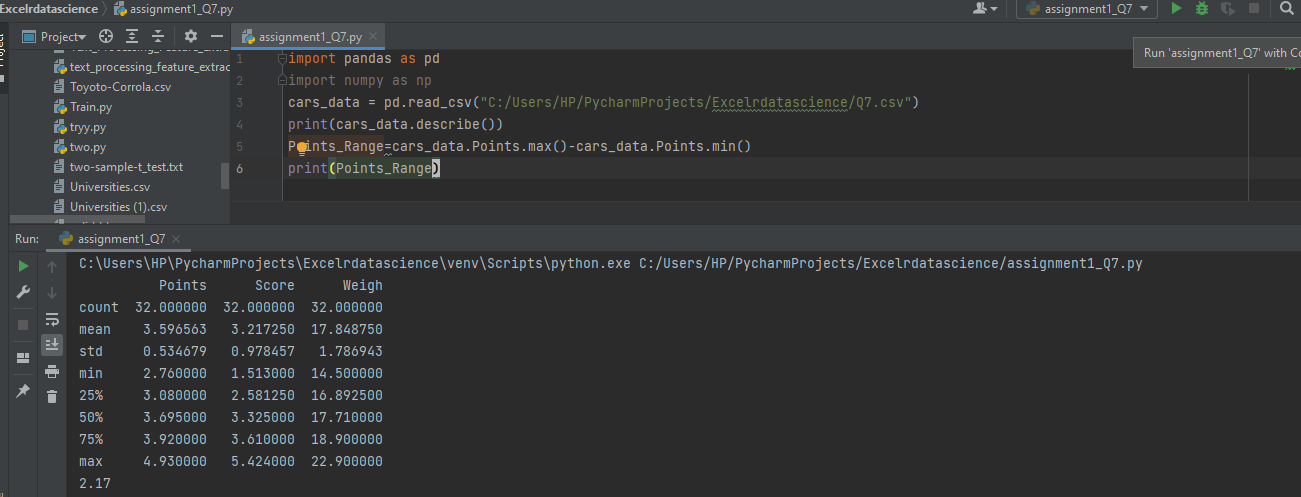
**MODE(points)**

**MODE(Score)**

**MODE(weigh)**

**VARIANCE**

**STANDERD DEVIATION**

**RANGE**

Q8) Calculate Expected Value for the problem below

(a)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?

Solution:Expected value = ∑(Probability \* Value)

∑P(X)\*E(X)

There are 9 patients

Probability of selecting each patient = 1/9

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**

|  |  |  |
| --- | --- | --- |
|  | **Skewness** | **kurtosis** |
| **Speed** | **-0.117510** | **-0.508994** |
| **distance** | **0.806895** | **0.405053** |
|  |  |  |

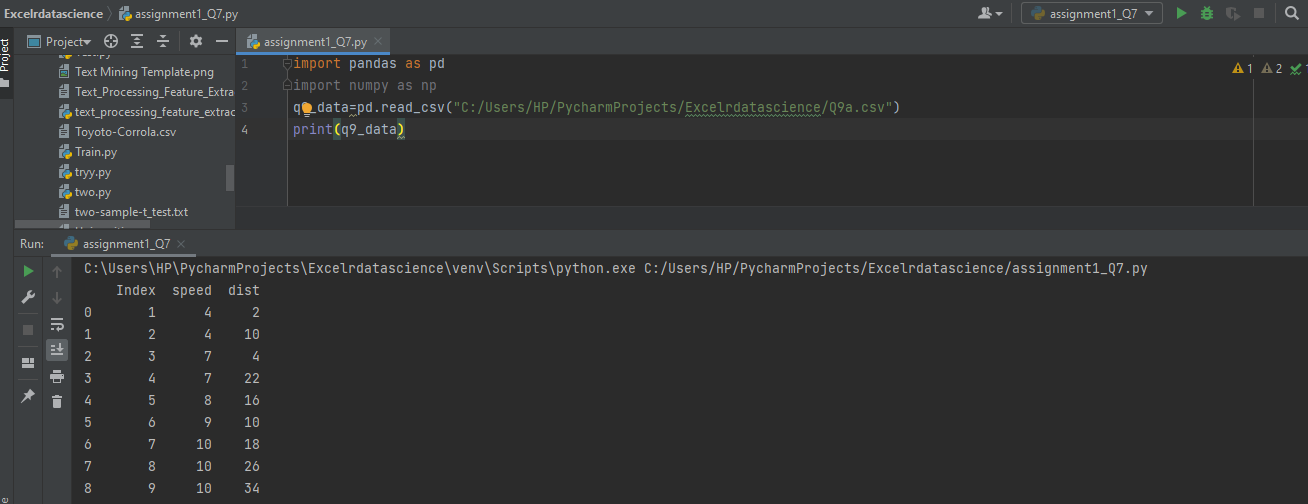
**COMMENTS**

**The values of speed is left skewed and values of distance right skewed.**

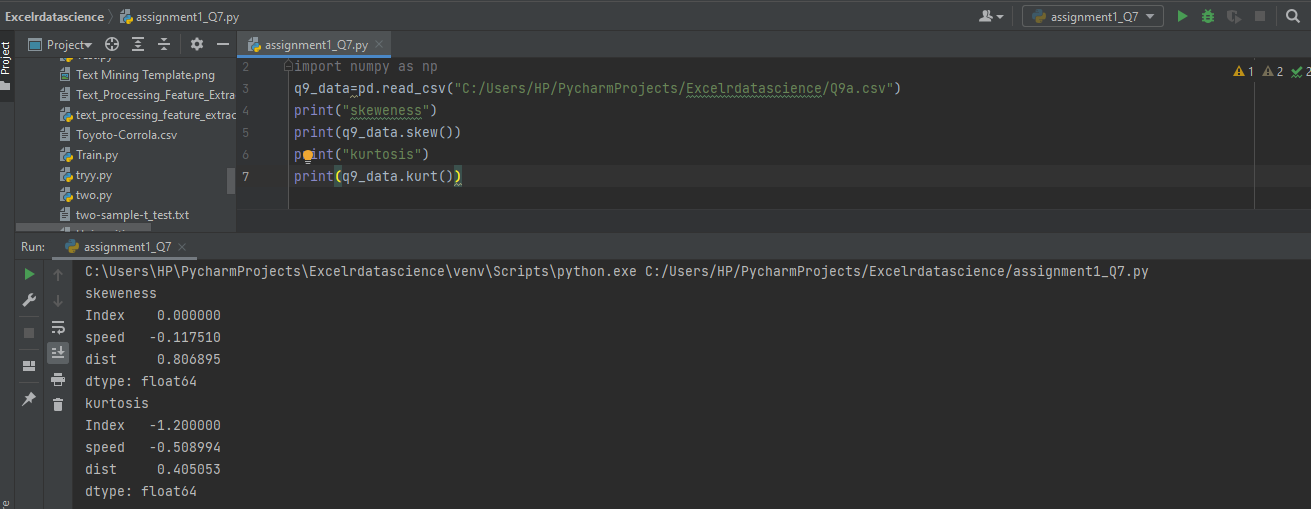
**The kurtosis points for distance is greater than zero means normal distribution.**

**SP and Weight(WT)**

**READ CSV DATA:**

****

**SKEWENESS AND KURTOSIS**

****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **skewness** | **kurtosis** |  |
| **SP** | **1.611450** | **2.977329** |  |
| **WT** | **0.614753** | **0.950291** |  |

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

**Histogram shows that the data is right skewed. i.e(mode<median<mean)**



Solution:

Histogram:

1. Chick weight data is positively skewed.
2. Most data of chick weight is between 50 and 100.
3. Chick weight is positively skewed.

Boxplot:

1. The data is right skewed
2. There are outliers 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?

Solution:

**94% Confidence Interval**

T = x+t σ/ **√** n

= 200+1.88\*30/√2000

= 200+1.2611

t = 201.2611

200-1.2611

= 198.73

t-score = (201.2611,198.73)

**98% Confidence Interval**

t = 200+2.326\*30/√2000

= 200+1.56032

t = 201.56032

200-1.56032

= 198.43968

t-score =(201.56032,198.43968)

96% Confidence Interval

t = 200+2.05\*30/√2000

= 200+1.375

t = 201.375

200-1.375

= 198.625

t-score =(201.375,198.625)

**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?

solution:

Mean=41, median=40.5,variance=25.52,std=5.052

There is two outlier that are 49 and 56

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

solution: symmetrical

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

solution: right tail or positive skewed

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

solution: left tail or negative skewed

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

solution: that a distribution is peaked and possess thick tails.

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

solution: that a distribution is flat and has thin tails

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



What can we say about the distribution of the data?

solution: it is not normal distribution. Left tail.

What is nature of skewness of the data?

solution: the distribution of data is left skewed.

What will be the IQR of the data (approximately)?   
solution: 8  
Q19) Comment on the below Boxplot visualizations?



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

Solution:

Both are the symmetrical. IQR of boxplot2 is higher than boxplot1, and also the data is highly spread,

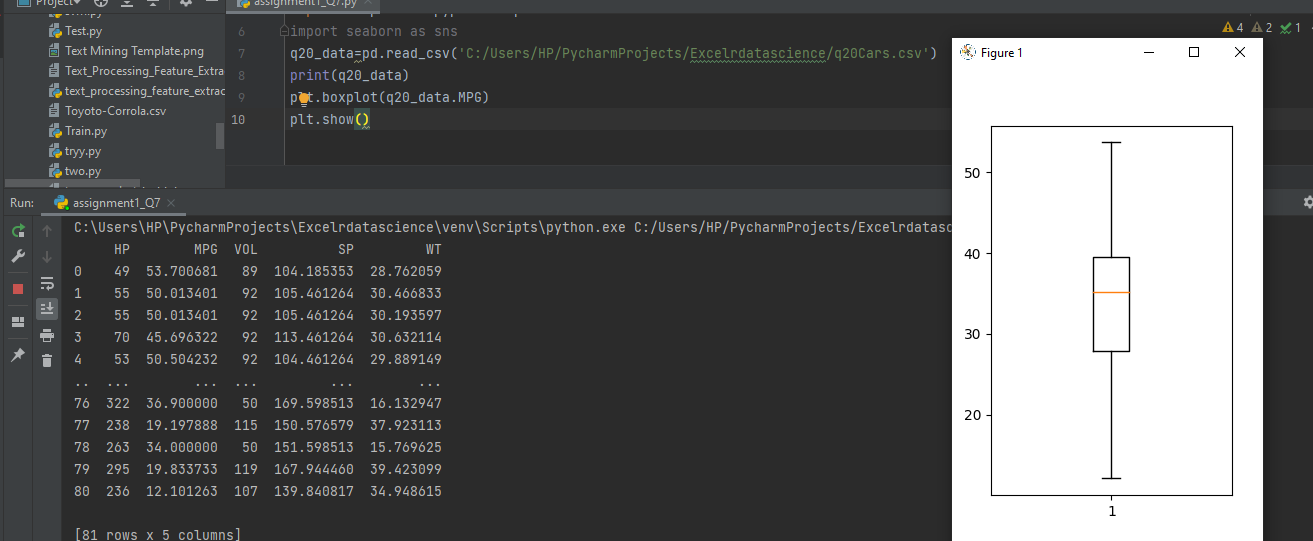
Q 20) Calculate probability from the given dataset for the below cases

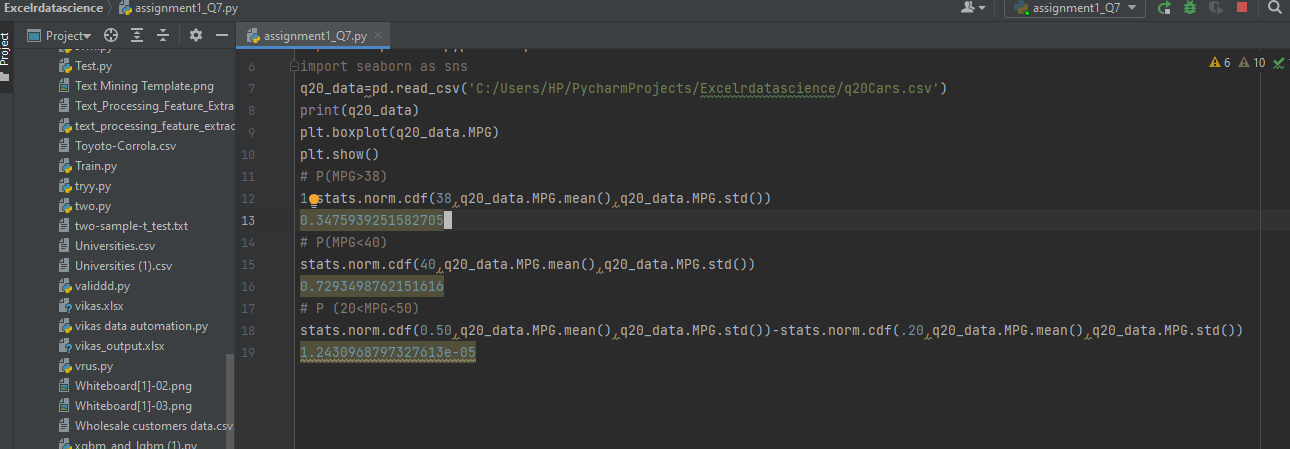
Data \_set: Cars.csv

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

MPG <- Cars$MPG

* 1. P(MPG>38) = 33/81
  2. P(MPG<40) = 61/81
  3. P (20<MPG<50) =69/81



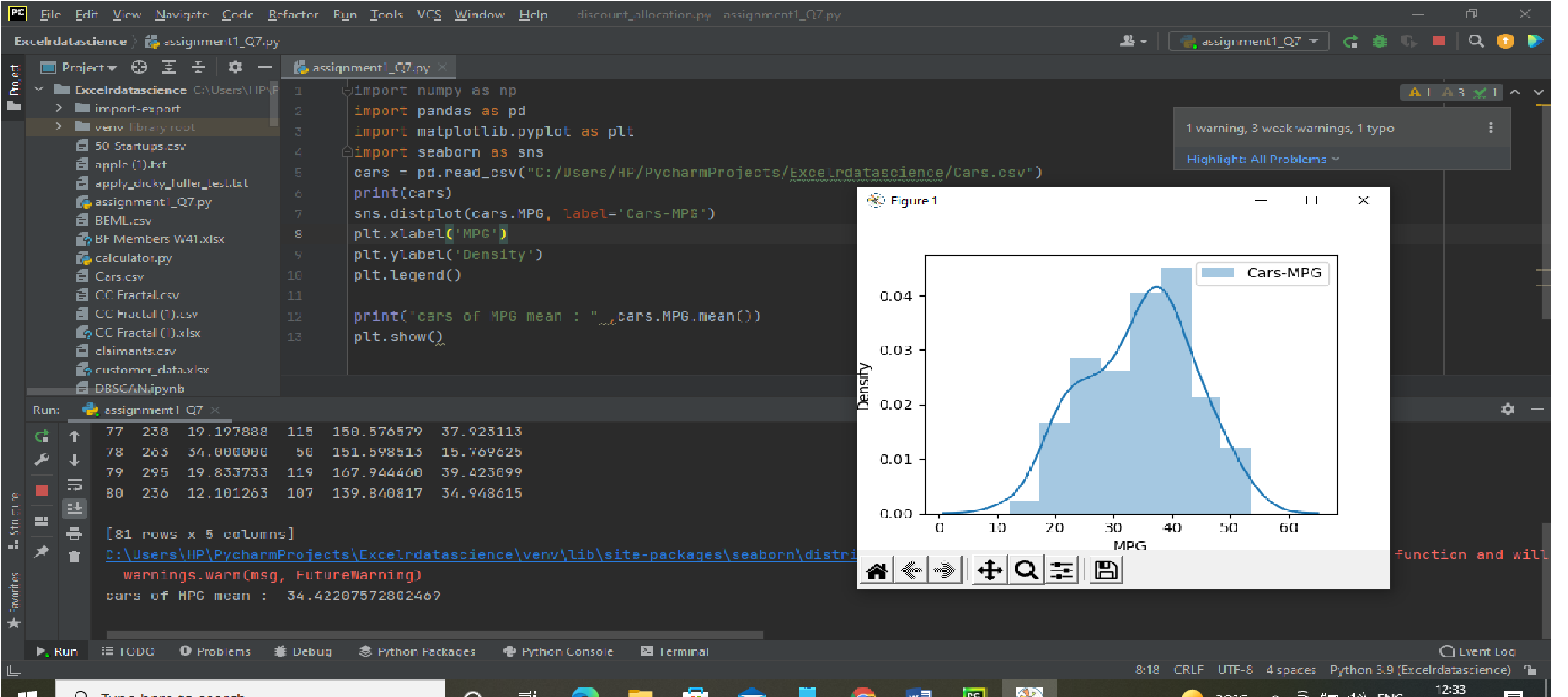


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

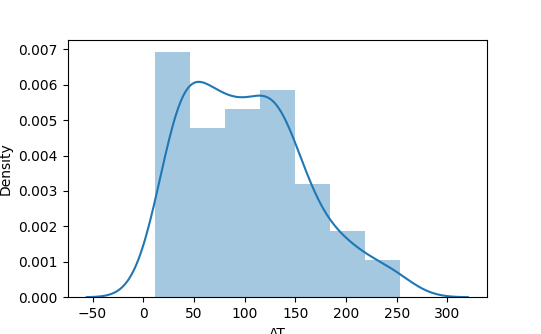
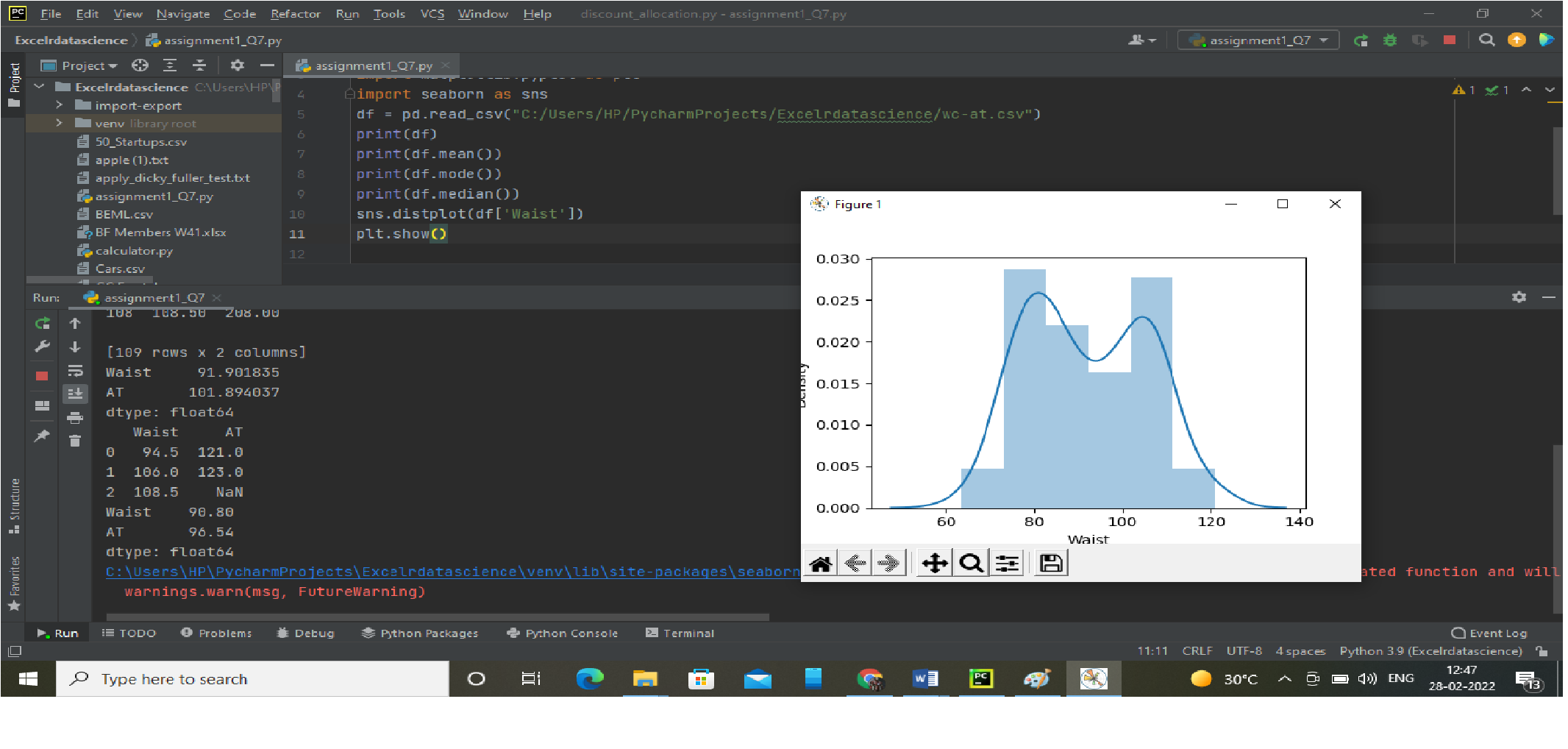
Dataset: Cars.csv

Ans: 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





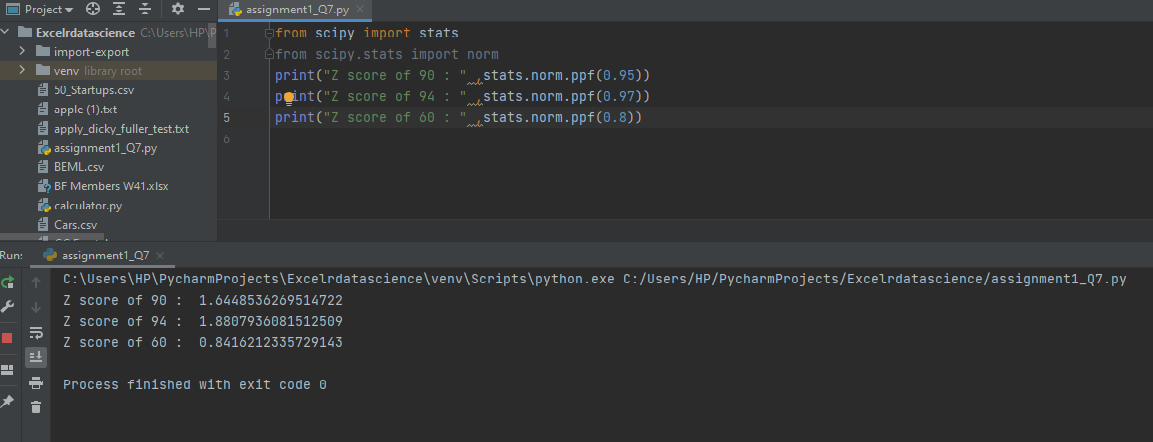
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

Ans:

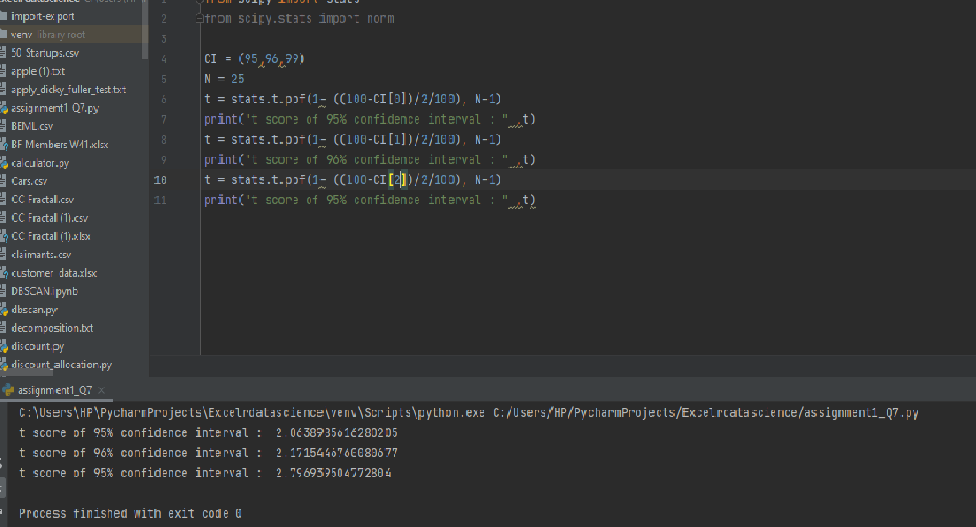
z-score of 90% confidence interval stats.norm.ppf(0.95)

z-score of 94% confidence interval stats.norm.ppf(.97)

z-score of 60% confidence interval stats.norm.ppf(.80)



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

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

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-10%7D%7B%5Cfrac%7B90%7D%7B3%5Csqrt%202%7D%7D

https://tex.z-dn.net/?f=t%20%3D%20%5Cdfrac%7B-1%20%5Ctimes%20%5Csqrt%202%7D%7B3%7D

t = - 0.471

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 0.3218 assuming the mean life of the bulbs is 300 days.