|  |  |
| --- | --- |
| 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) | Discrete Data |

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

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

Sol:

P(H H T) + P(H T H) + P(T H H)

=1/8 +1/8+1/8

= 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

Sol :

1. There are no outcomes which corresponds sum equal to one because when two dice are rolled the sum will be greater than 0.

Hence Probability = 0/36 = 0.

1. Possible Outcomes = (1,3) (2,2) (3,1) = 3 outcomes,

Probability = 3/36 i.e. 1/12

1. Total Possible outcomes are 6.

Hence Probability = 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?

Sol: When First ball is drawn ( any ball except blue can be drawn) so we have 5 options out of 7 balls . Probability = 5/7

When second ball is drawn ( any of the remaining balls except blue can be drawn) so we have 4 options out of 6 remaining balls. Probability= 4/6=2/3

Therefore P(neither ball is blue) = (5/7) \* (2/3) = ( 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

Sol:

Expected number of candies for 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

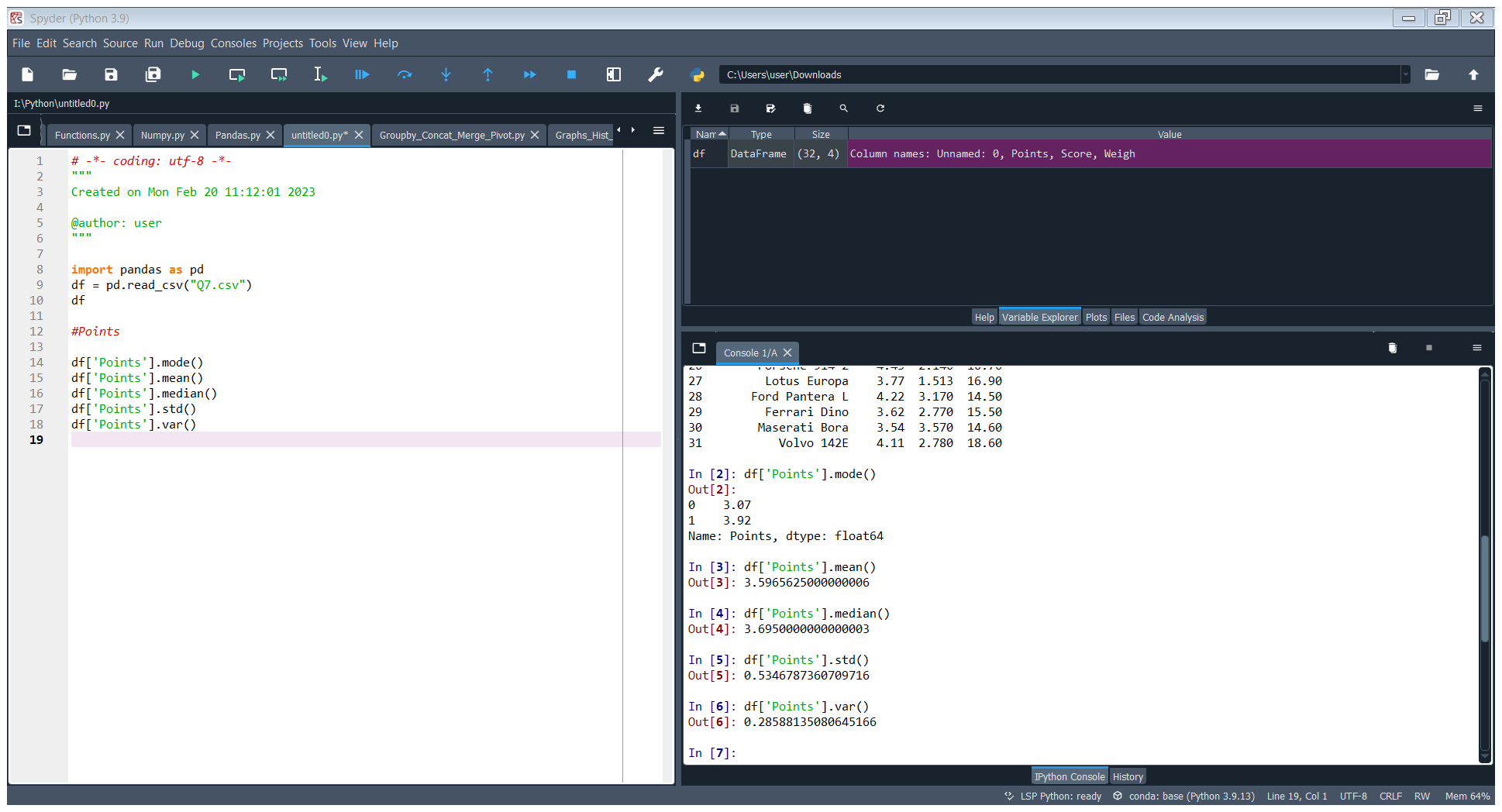
* For Points,Score,Weigh>

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

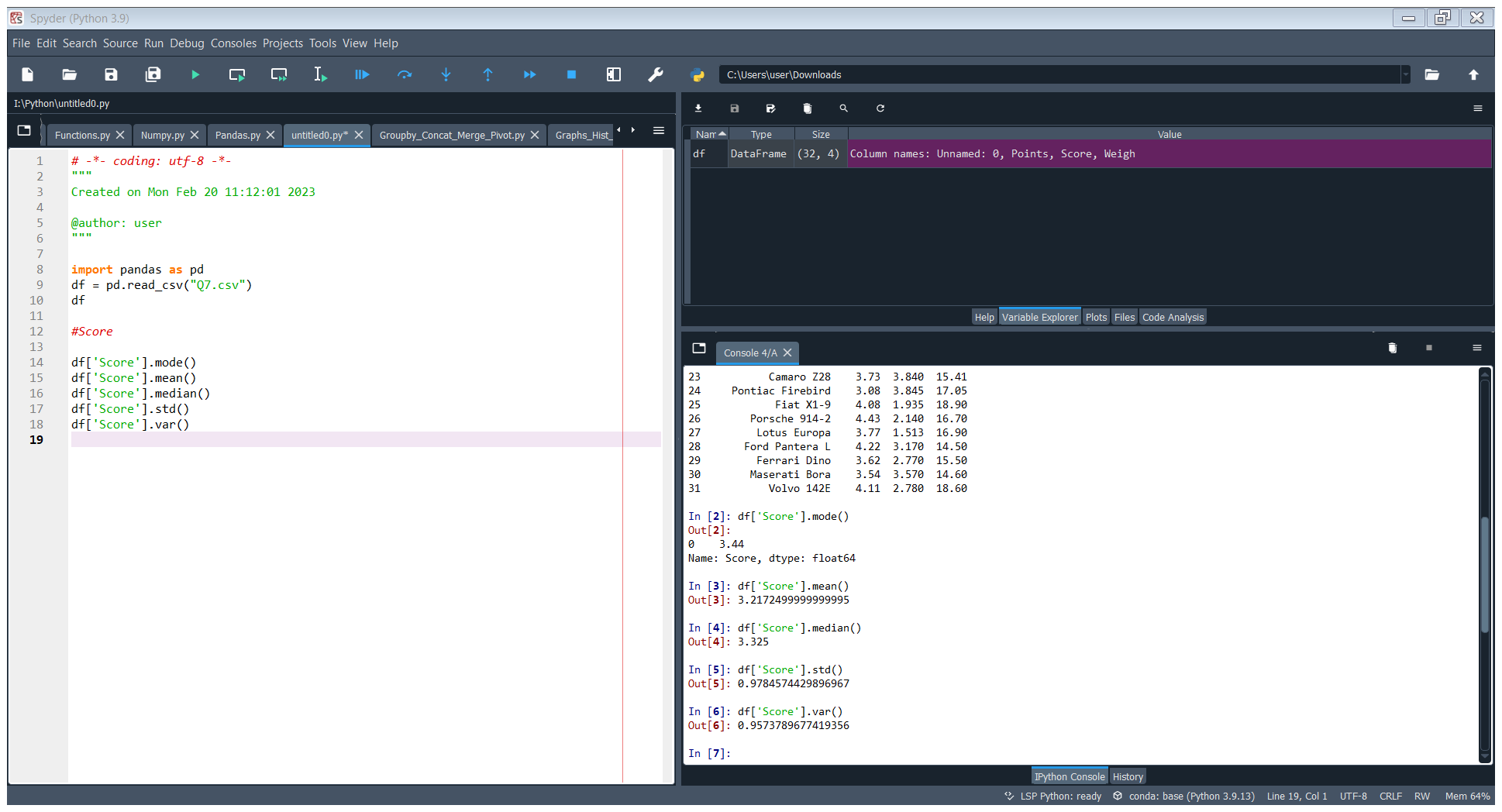
**Use Q7.csv fil**

|  |  |  |
| --- | --- | --- |
| **Points** | **Score** | **Weigh** |
| **3.9** | **2.62** | **16.46** |
| **3.9** | **2.875** | **17.02** |
| **3.85** | **2.32** | **18.61** |
| **3.08** | **3.215** | **19.44** |
| **3.15** | **3.44** | **17.02** |
| **2.76** | **3.46** | **20.22** |
| **3.21** | **3.57** | **15.84** |
| **3.69** | **3.19** | **20** |
| **3.92** | **3.15** | **22.9** |
| **3.92** | **3.44** | **18.3** |
| **3.92** | **3.44** | **18.9** |
| **3.07** | **4.07** | **17.4** |
| **3.07** | **3.73** | **17.6** |
| **3.07** | **3.78** | **18** |
| **2.93** | **5.25** | **17.98** |
| **3** | **5.424** | **17.82** |
| **3.23** | **5.345** | **17.42** |
| **4.08** | **2.2** | **19.47** |
| **4.93** | **1.615** | **18.52** |
| **4.22** | **1.835** | **19.9** |
| **3.7** | **2.465** | **20.01** |
| **2.76** | **3.52** | **16.87** |
| **3.15** | **3.435** | **17.3** |
| **3.73** | **3.84** | **15.41** |
| **3.08** | **3.845** | **17.05** |
| **4.08** | **1.935** | **18.9** |
| **4.43** | **2.14** | **16.7** |
| **3.77** | **1.513** | **16.9** |
| **4.22** | **3.17** | **14.5** |
| **3.62** | **2.77** | **15.5** |
| **3.54** | **3.57** | **14.6** |
| **4.11** | **2.78** | **18.6** |

**Points**: Mean =3.596563, Median= 3.695, Mode= 3.07,3.92 , Variance= 0.2858814, Standard deviation= 0.5346787, Range (Min-Max) – [3.59 – 4.93]

****

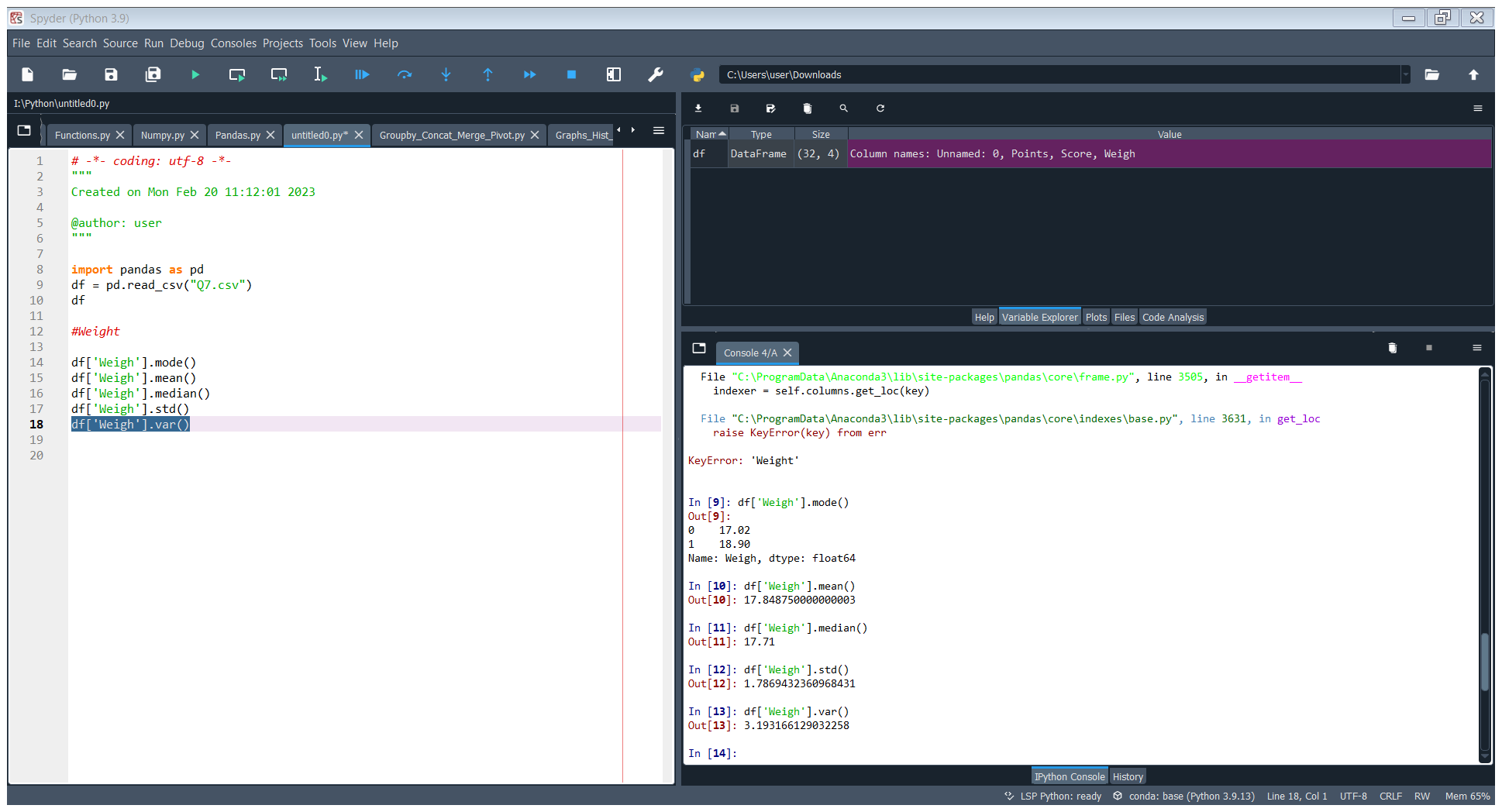
**Score**: Mean= 3.21725, Median= 3.325, Mode= “numeric”, Variance= 0.957379, Standard deviation= 0.9784574, Range (Min – Max) – [3.21-5.42]



Note: Mean value are closer for both ‘Point’ and ‘Score’.

**Weight**: Mean= 17.84875, Median= 17.71, Mode= “17.02, 18.90”, Variance= 3.193166, Standard deviation= 1.786943, Range (Min – Max) – [17.84 – 22.9]

Weigh has two modes. So, it is bi modal in nature.



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?

Sol: Expected value = Sum (X \* Probability of X) = (1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199) = 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**

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



Sol : Most of the data points are concentrated in the range 50-100 with frequency 200. And least range of weight is 400 somewhere around 0-10. So the expected value the above distribution is 75.

We can infer from the figure that the Median is less than mean.

**Skewness**- We can notice a long tail towards right so it is heavily right skewed.



Sol:

We have outliers on the upper side of box plot and there are less data points between Q1 and bottom point.

**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?

Sol:

Identify a sample statistic. Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (200) as the sample statistic

We are working with a 94 % confidence level.

Find standard error.

The standard error (SE) of the mean is:

SE = s √ n = 30 √ 2000 = 0.670

Compute alpha (α): α = 1 - (confidence level / 100) = 0.9933

Critical probability (p\*): p\* = 1 - α/2 = 1 – 0.9933/2 = 0.4966

Degree of freedom (df): df = n-1 = 2000 – 1 = 1999

The critical value is the t score having 1999 degrees of freedom and a probability equal to 0.4966.

Critical value is -0.009

Critical value \* standard error = - 0.009 \* 0.94 = -0.00846

= - 0.009 \* 0.98 = -0.00882

= -0.009 \* 0.96 = - 0.00846

**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.

Sol: Mean= 41, Median= 40, variance= 24.111, Standard deviation= 4.910

1. What can we say about the student marks?

Sol : 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?

Sol: Symmetrical

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

Sol: Right skewed

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

Sol: Left Skewed

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

Sol : The data is normally distributed and kurtosis value is 0 and the curve is more peaked.

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

Negative Kurtosis means the curve will be flatter and broader

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



What can we say about the distribution of the data?

Ans: The above Boxplot is not normally distributed the median is towards the higher value.

What is nature of skewness of the data?

Ans: The data is a skewed towards left. Median is greater than mean

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

Ans : Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile = 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.

Ans : Firstly, there are no outliers. Secondly, both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range. The distribution is symmetrical.

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:

Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3) print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

* 1. P(MPG<40)

Ans: prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

* 1. P (20<MPG<50)

Ans: prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20)) p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))

P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20) print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

P(20<MPG<50)= 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: a.) 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: Adipose Tissue (AT) and Waist does not follow Normal Distribution

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

Ans:

print('Z score for 90% Confidence Interval =',np.round(stats.norm.ppf(.05),4))

Z score for 90% Confidence Interval = -1.6449

print('Z score for 94% Confidence Interval =',np.round(stats.norm.ppf(.03),4))

Z score for 94% Confidence Interval = -1.8808

print('Z score for 60% Confidence Interval =',np.round(stats.norm.ppf(.2),4))

Z score for 60% Confidence Interval = -0.8416

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

Ans:

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))

T score for 95% Confidence Interval = -2.0639

print('T score for 94% Confidence Inteval =',np.round(stats.t.ppf(0.03,df=24),4))

T score for 94% Confidence Inteval = -1.974

print('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

T score for 95% Confidence Interval = -2.7969

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

Sol:

Python Code:

from scipy import stats

from scipy.stats import norm

#Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

#Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

#find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

t**=**(260**-**270)**/**(90**/**18**\*\***0.5)

t

t = -0.4714045207910317

# Find P(X>=260) for null hypothesis

#p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

p\_value

p\_value = 0.32167411684460556

Here p-value is greater than alpha(0.05), So, Ho is accepted and Ha is rejected.

Hence, Average life of Bulb >= 260 days.