|  |  |
| --- | --- |
| 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 | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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 | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Interval |

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

Solution:

**Total outcomes: 8**

|  |  |  |
| --- | --- | --- |
| C1 | C2 | C3 |
| H | H | H |
| H | H | T |
| H | T | H |
| T | H | H |
| H | T | T |
| T | H | T |
| T | T | H |
| T | T | T |

**Favorable outcome (two heads and one tail): 3**

|  |  |  |
| --- | --- | --- |
| C1 | C2 | C3 |
| H | H | T |
| H | T | H |
| T | H | H |

**Probability of “two heads and one tail” is - Favorable outcome/ Total outcomes** = 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

Solution:

1. Equal to 1

There is no possibility of getting the sum equal to 1 for 2 dies so the favorable outcome is 0 (Zero).

So, the probability of “getting the sum equal to one” is - Favorable outcome/ Total outcomes = 0/36 = 0

1. Less than or equal to 4

Total Outcomes - 36

|  |  |
| --- | --- |
| Die1 | Die2 |
| 6 | 6 |
| 6 | 5 |
| 6 | 4 |
| 6 | 3 |
| 6 | 2 |
| 6 | 1 |
| 5 | 6 |
| 5 | 5 |
| 5 | 4 |
| 5 | 3 |
| 5 | 2 |
| 5 | 1 |
| 4 | 6 |
| 4 | 5 |
| 4 | 4 |
| 4 | 3 |
| 4 | 2 |
| 4 | 1 |
| 3 | 6 |
| 3 | 5 |
| 3 | 4 |
| 3 | 3 |
| 3 | 2 |
| 3 | 1 |
| 2 | 6 |
| 2 | 5 |
| 2 | 4 |
| 2 | 3 |
| 2 | 2 |
| 2 | 1 |
| 1 | 6 |
| 1 | 5 |
| 1 | 4 |
| 1 | 3 |
| 1 | 2 |
| 1 | 1 |

Favorable outcome – 6

Formula in test column is =IF(C25<=4,TRUE,FALSE)

|  |  |  |  |
| --- | --- | --- | --- |
| Die1 | Die2 | SUM | Test |
| 3 | 1 | 4 | TRUE |
| 2 | 2 | 4 | TRUE |
| 2 | 1 | 3 | TRUE |
| 1 | 3 | 4 | TRUE |
| 1 | 2 | 3 | TRUE |
| 1 | 1 | 2 | TRUE |

The probability of getting “Less than or equal to 4” = Favorable outcome/Total Outcome = 6/36 = 1/6 = 1.667 (rounded of till 3rd digit)

1. Sum is divisible by 2 and 3

Total Outcomes - 36

|  |  |
| --- | --- |
| Die1 | Die2 |
| 6 | 6 |
| 6 | 5 |
| 6 | 4 |
| 6 | 3 |
| 6 | 2 |
| 6 | 1 |
| 5 | 6 |
| 5 | 5 |
| 5 | 4 |
| 5 | 3 |
| 5 | 2 |
| 5 | 1 |
| 4 | 6 |
| 4 | 5 |
| 4 | 4 |
| 4 | 3 |
| 4 | 2 |
| 4 | 1 |
| 3 | 6 |
| 3 | 5 |
| 3 | 4 |
| 3 | 3 |
| 3 | 2 |
| 3 | 1 |
| 2 | 6 |
| 2 | 5 |
| 2 | 4 |
| 2 | 3 |
| 2 | 2 |
| 2 | 1 |
| 1 | 6 |
| 1 | 5 |
| 1 | 4 |
| 1 | 3 |
| 1 | 2 |
| 1 | 1 |

Favorable outcome – 6

Formula in test column is =IF(AND(MOD(C2,2) =0, MOD(C2,3) =0), TRUE, FALSE)

|  |  |  |  |
| --- | --- | --- | --- |
| Die1 | Die2 | SUM | Test2 |
| 6 | 6 | 12 | TRUE |
| 5 | 1 | 6 | TRUE |
| 4 | 2 | 6 | TRUE |
| 3 | 3 | 6 | TRUE |
| 2 | 4 | 6 | TRUE |
| 1 | 5 | 6 | TRUE |

The probability of getting “Sum is divisible by 2 and 3” = Favorable outcome/Total Outcome = 6/36 = 1/6 = 1.667 (rounded of till 3rd digit)

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:

Types of Possibilities without Blue are – (2 Red AND 0 Green) OR (1 Red AND 1 Green) OR (0 Red AND 2 Green)

((2C2 \* 3C0) + (2C1 \* 3C1) + (2C0 \* 3C2)) / (7C2)

= ((2\*1) / (2\*1)) + ((2/1) \* (3/1)) + ((3\*2) / (2\*1)) / ((7\*6)/2)

= (1 + 6 + 3)/21

= 10/21

= 0.4761

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

Solution:

The Expected number of candies for a randomly selected child can be obtained as follows –

= (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.

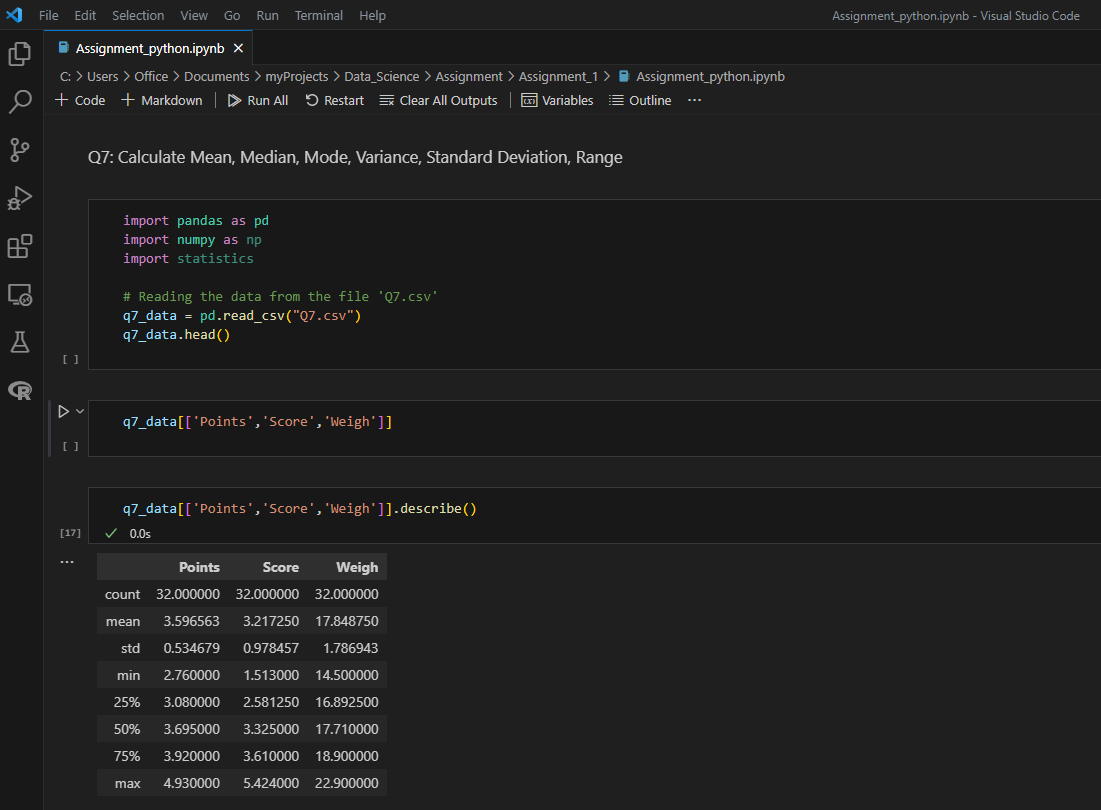
**Use Q7.csv file**

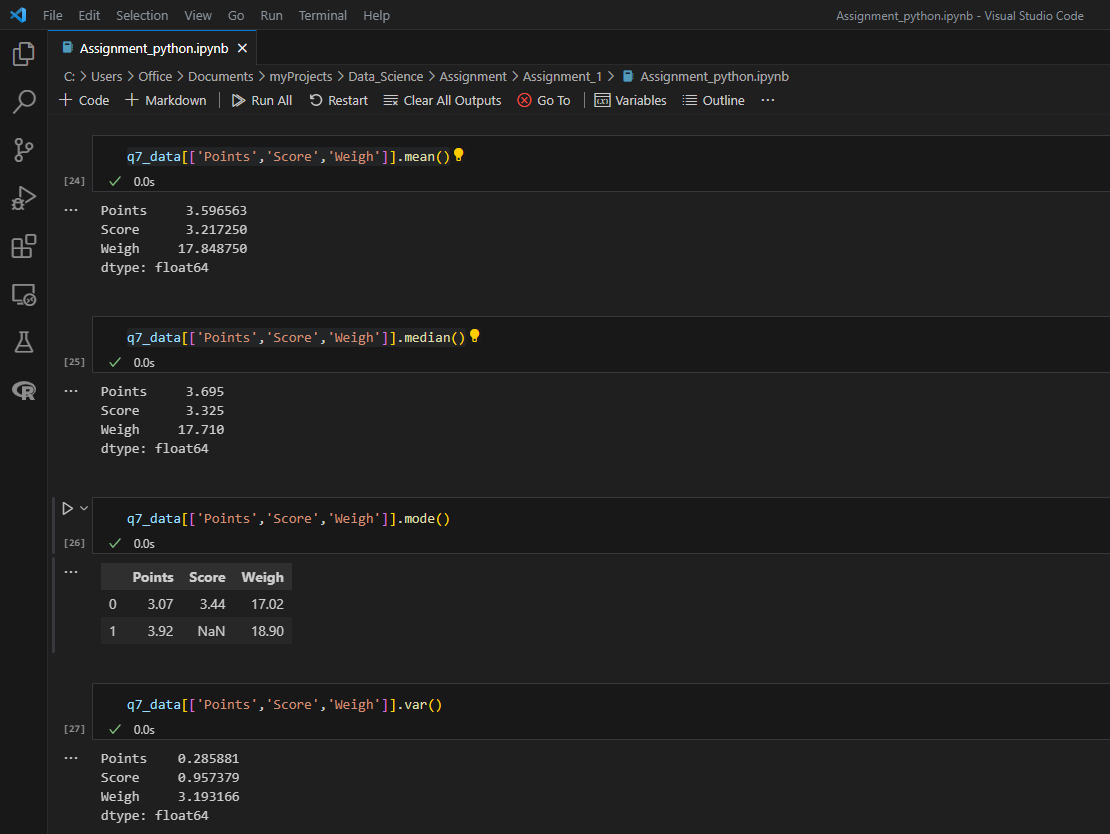
Solution:

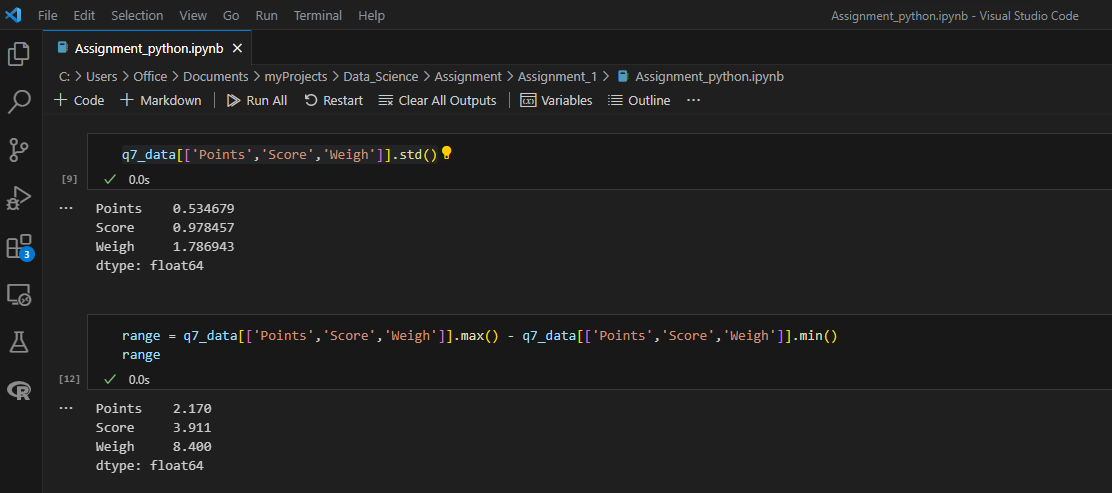
**Based on manual calculation (in excel file Q7\_Manual\_Calculations.csv)**

|  |  |  |  |
| --- | --- | --- | --- |
| Mean | 3.6 | 3.22 | 17.85 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Variance | 0.29 | 0.96 | 3.19 |
| Standard Deviation | 0.53 | 0.98 | 1.79 |
| MAX | 4.93 | 5.424 | 22.9 |
| MIN | 2.76 | 1.513 | 14.5 |
| Range | 2.17 | 3.911 | 8.4 |

**Based on Python:**

****

****

****

**Inference** - The data associated with Points are very close to the mean (~3.597), whereas the data associated with Weight column are the most dispersed and further compared to the mean (~17.85) and high variance, when compared to Points column. The data associated with Score are lesser dispersed from the mean (~3.22) when compared to Weight but more dispersed when compared to Points.

Also, the range between the largest and least value for Score is the least when compared to Score which is the further lesser than that for Weight. This trend is similar to the trend seen for variance and standard deviations for these columns i.e., Points, Score and Weight.

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?

Solution:

The expected value formula-

E(X)= x1​ \* P(x1) + … + xn​ \* P(xn​)

The probability of each event = 1/9 = 0.111111

The expected value is = (108 \* 0.111111) + (110 \* 0.111111) + (123 \* 0.111111) + (134 \* 0.111111) + (135 \* 0.111111) + (145 \* 0.111111) + (167 \* 0.111111) + (187 \* 0.111111) + (199 \* 0.111111)

= 11.999988 + 12.22221 + 13.666653 + 14.888874 + 14.999985 + 16.111095 + 18.555537 + 20.777757 + 22.111089

= 145.333188

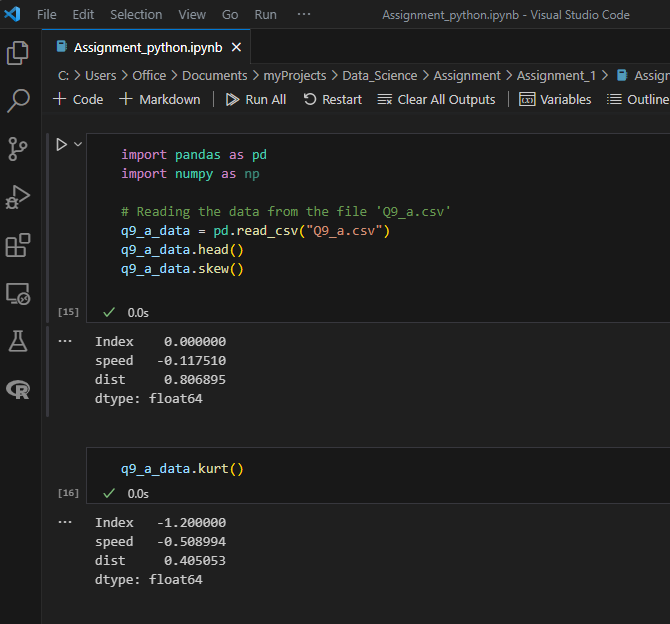
**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

Solution:

Using the following Python commands, I got the result –



**Calculating the skewness using Python:**

**Skewness of speed: -0.117510**

**Inference** – The observations are left skewed (negative skew).

**Skewness of dist: 0.806895**

**Inference** – Majority of observations are near the left with few observations dispersed far towards the right, creating a long tail i.e., highly right skewed (positive skew).

**kurtosis using Python:**

**kurtosis of speed: -0.508994**

**Inference:** The negative kurtosis indicates that the distribution has less peak and is lighter towards the tails when compared to the normal distribution.

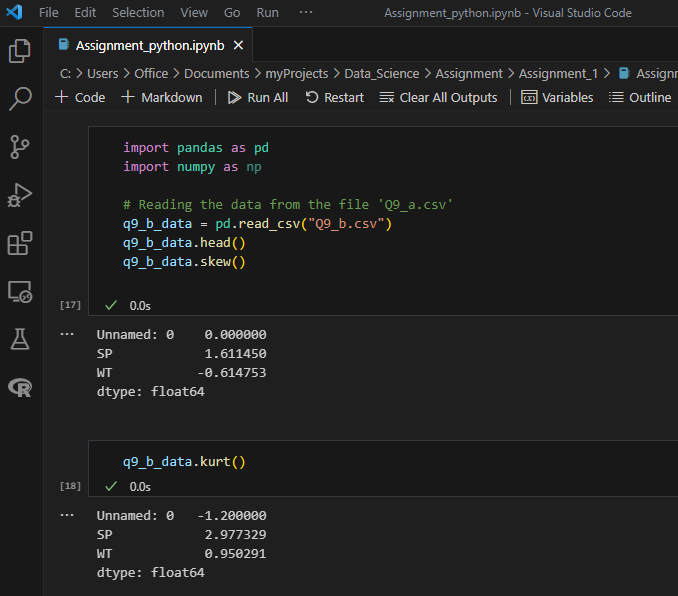
**kurtosis of dist: 0.405053**

**Inference:** Since the kurtosis is positive, it indicates that the distribution is peaked and possesses thick tails. However, when compared to SP, the distance (dist) observations are less peaked and lighter at the tails.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Calculating the skewness using Python:**



**Skewness of SP: 1.611450**

Inference – Majority of observations for SP are near the left with few observations dispersed far towards the right, creating a long tail i.e., highly right skewed (positive skew).

**Skewness of WT: -0.614753**

Inference – The observations of weight are left skewed (negative skew).

**Calculating the kurtosis using Python:**

**kurtosis of SP: 2.977329**

Inference: Since the kurtosis is positive, it indicates that the distribution is peaked and possesses thick tails. Since the value of kurtosis is almost equivalent to 3, it is more towards the normal distribution (having kurtosis of 3).

**kurtosis of WT: 0.950291**

Inference: Since the kurtosis is positive, it indicates that the distribution is peaked and possesses thick tails. However, when compared to SP, the Weight observations are less peaked and lighter at the tails.

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



Solution:

Observations:

In this histogram, we see that the observations are –

1. Skewed to the right side of the peak when compared to the left side.
2. This is a positive skew and the mean will be greater than the median.
3. Here majority of ChickWeight is concentrated between within the range 50 to 200
4. The highest frequency being ~200 between the range 50 and 100.
5. There are very less observations having Weight greater than 200 and extending towards 400 (having the least i.e., between 5-10 frequency), which are the outliers, giving it a long tail.
6. There are few outliers at the Upper Extreme greater than 200 up to 400, causing the long right skew in the curve.



Solution:

Observations:

In this boxplot, we see that –

1. The median is towards the bottom end.
2. We can see some observations towards to the top i.e., Upper Extreme (represented by round circles >9, with some overlaps) that are outliers.
3. We also see that the upper whisker is longer than the lower whisker indicating that the data is more skewed to the top (or right skew, when plotted horizontally) i.e., data is more variable on top.
4. Most of the observations are concentrated towards the lower end.
5. We can find that the median is less than the mean. If a histogram where to be constructed we will see the observations stretched or skewed to the right (positive skew) due to the outliers on the right side and so the mean will be greater than the 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?

Solution:

Manual Calculation –

Since the sample size is >30, it is approximately normally distributed and we will use a z distribution to find the critical values.

Given values –

Sample size (n) = 2000

CI = the confidence interval

X̄ = the population mean = 200

Z\* = the critical value of the z distribution

σ = the population standard deviation = 30

√n = the square root of the population size

CI = X̄± Z× (sigma/√n)

**Z-Score for 94% is 1.881**

Calculation of z-score for 94% -

A 94 % confidence interval has two tails of 6/2 = 3% so it goes from 3% to 97% which leaves 94% in the middle so look up the Z for P(z<Z) = 0.97

The two closest values in the z-table P(z<1.88) = 0.96995 and P(z< 1.89) = 0.97062

Interpolating – 1.88 + (0.97- 0.96995) \* (0.01)/ (0.97062- 0.96995) = 1.880746(approx.)

**Z-Score for 98% is 2.3263**Calculation of z-score for 98% -

A 98% confidence interval has two tails of 2/2 = 1% so it goes from 1% to 99% which leaves 98% in the middle so look up the Z for P(z<Z) = 0.99

The two closest values in the z-table P(z<2.32) = 0.98983 and P(z< 2.33) = 0.99010

Interpolating – 2.32 + (0.99 - 0.98983) \* (0.01)/ (0.99010 - 0.98983) = 2.3263 (approx.)

**Z-Score for 96% is 2.05375**

Calculation of z-score for 96% -

A 96 % confidence interval has two tails of 4/2 = 2% so it goes from 2% to 98% which leaves 96% in the middle so look up the Z for P(z<Z) = 0.98

The two closest values in the z-table P(z<2.05) = 0.97982 and P(z< 2.06) = 0.98030

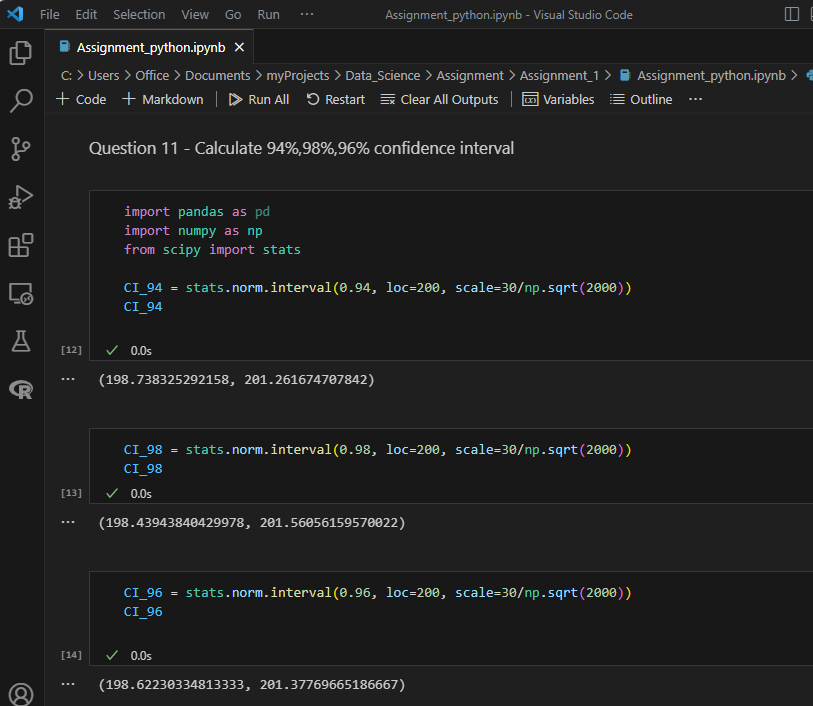
Interpolating – 2.05 + (0.98- 0.97982) \* (0.01)/ (0.98030 - 0.97982) = 2.05375 (approx.)

**Manual Calculation to get Confidence interval of 94%** = 200 ± 1.8808\*(30/√2000) = 200 ± 1.881\*(0.671) = 200 ± 1.262 = (198.738 - 201.262)

**Manual Calculation to get Confidence interval of 98%** = 200 ± 2.3263\*(30/√2000) = 200 ± 2.3263\*(0.671) = 200 ± 1.5609473 = (198.439 - 201.561)

**Manual Calculation to get Confidence interval of 96%** = 200 ± 2.05375\*(30/√2000) = 200 ± 2.05375\*(0.671) = 200 ± 1.378 = (198.622 - 201.377)

**Using Python:**



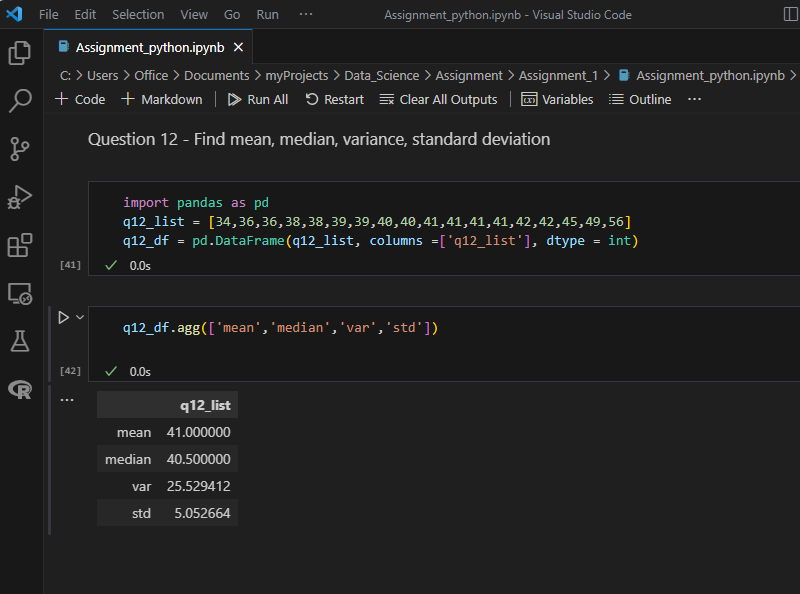
**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

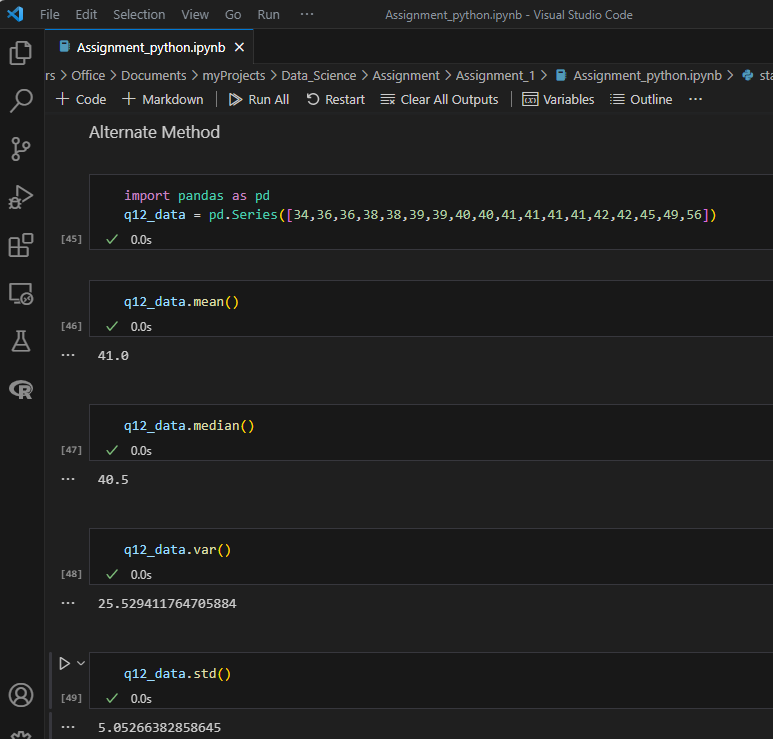
1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

Solution:

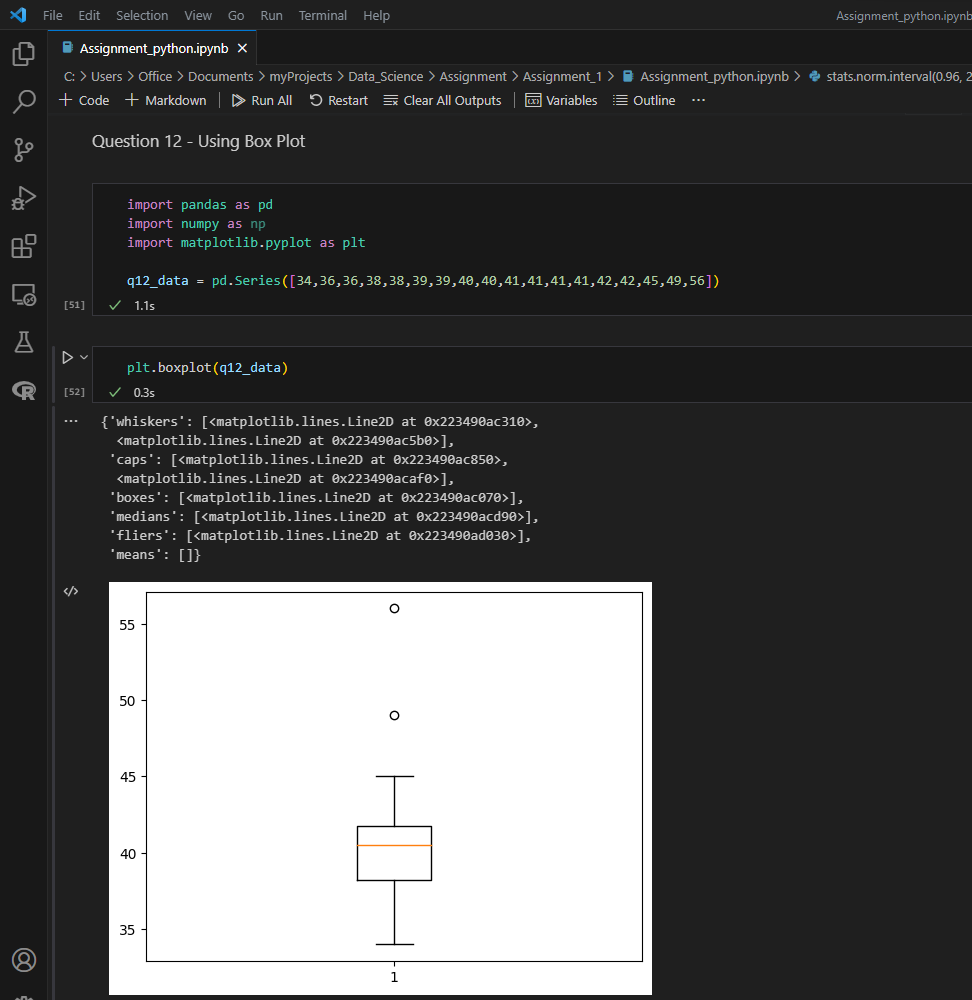
1. **Finding mean, median, variance and standard deviation using Python:**



1. **Alternate Method to find mean, median, variance and standard deviation using Python:**



**Using Box Plot in Python to visualize the data:**



**Observations:** In this boxplot, we see that –

1. The median is towards the bottom end.
2. We can see some observations towards to the top i.e., Upper Extreme (represented by 2 round circles) that are outliers.
3. Based on the box plot we see that the median is closer to the lower or bottom quartile.
4. A distribution is right skewed (in other words "Positively Skewed") when mean > median. It means the data constitute higher frequency of high valued scores.

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

Solution:

When the mean and median of the data are equal, the distribution has zero skew and it is symmetrical i.e., it is normal distribution. In such cases, the observations are distributed equally on left and right side if seen in a histogram i.e., the left and right side are mirror images. In a perfectly symmetrical distribution, the mean and median are the same. The thumb rule suggests that in such cases the skewness is between -0.5 and 0.5

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

Solution:

When the mean is greater than the median, then it implies –

1. The distribution is skewed to the right, with the distribution being longer/stretched to the right side of the peak than on its left side. This distribution is called right skewed or positive skew.
2. The thumb rule suggests that in such cases the skewness is between 0.5 & 1 (positive skewed). If the skewness is greater than 1, the data is extremely skewed.

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

Solution:

When the median is greater than the mean, then it implies that the distribution is skewed to the left, with the distribution being longer/stretched to the left side of the peak than on its right side. This distribution is called left skewed or negative skew. The thumb rule suggests that in such cases the skewness is between -1 & -0.5 (negative skewed). If the skewness is lesser than -1, the data is extremely skewed.

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

Solution:

Kurtosis is a measure of the tailedness of a distribution. Tailedness is how often outliers occur. Kurtosis is indicated as the measure of peakedness of the distribution. Peakedness in a data distribution is the degree to which data values are concentrated around the mean.

Datasets with positive kurtosis indicate that the distribution has a distinct peak near the mean and tend to decline rapidly, and has heavy/thick tails.

An extremely Positive values of kurtosis i.e., kurtosis greater than 3 indicates sharply peaked with greater number of observations located in the tails of the distribution instead of around the mean. Also termed a leptokurtic distribution, it has a higher peak (thin bell) and taller (i.e., fatter and heavy) tails than a normal distribution (kurtosis = 3).

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

Solution:

Datasets with a negative kurtosis indicates that the distribution has lighter tails than the normal distribution. Also called a platykurtic distribution it is flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter (i.e., lighter and thinner) tails.

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



What can we say about the distribution of the data?

Solution: The boxplot is not normally distributed as the median is to the higher value. The data is left skewed (negatively skewed) with a long lower whisker in the Lower Extreme (LE).

What is nature of skewness of the data?

Solution: The data is left skewed (negatively skewed) with a large tail on the left.

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

IQR = UQ – LQ = 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.

Solution:

**Inferences –**

1. Both Box Plots i.e., “Box-Plot-1” and “Box-Plot-2” have the same median i.e., ~262.5
2. The Box Plot “Box-Plot-2” appears to be symmetrical and follows a normal distribution as the upper and lower whiskers are of the same size and Median is exactly in between the Upper and Lower Quartile. However, “Box-Plot-1” appears to be very slightly skewed to the right (positively skewed) as the median appears to be more towards the lower end.
3. The variance for “Box-Plot-2” is greater than for “Box-Plot-1”
4. There are no outliers in both box Plots.
5. The Inter Quartile Range (IQR) for “Box-Plot-1” is approximately 25 (UQ – LQ = 281.25-256.25). The Inter Quartile Range (IQR) for “Box-Plot-2” is approximately 87.5 (UQ – LQ = 312.5-225)

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

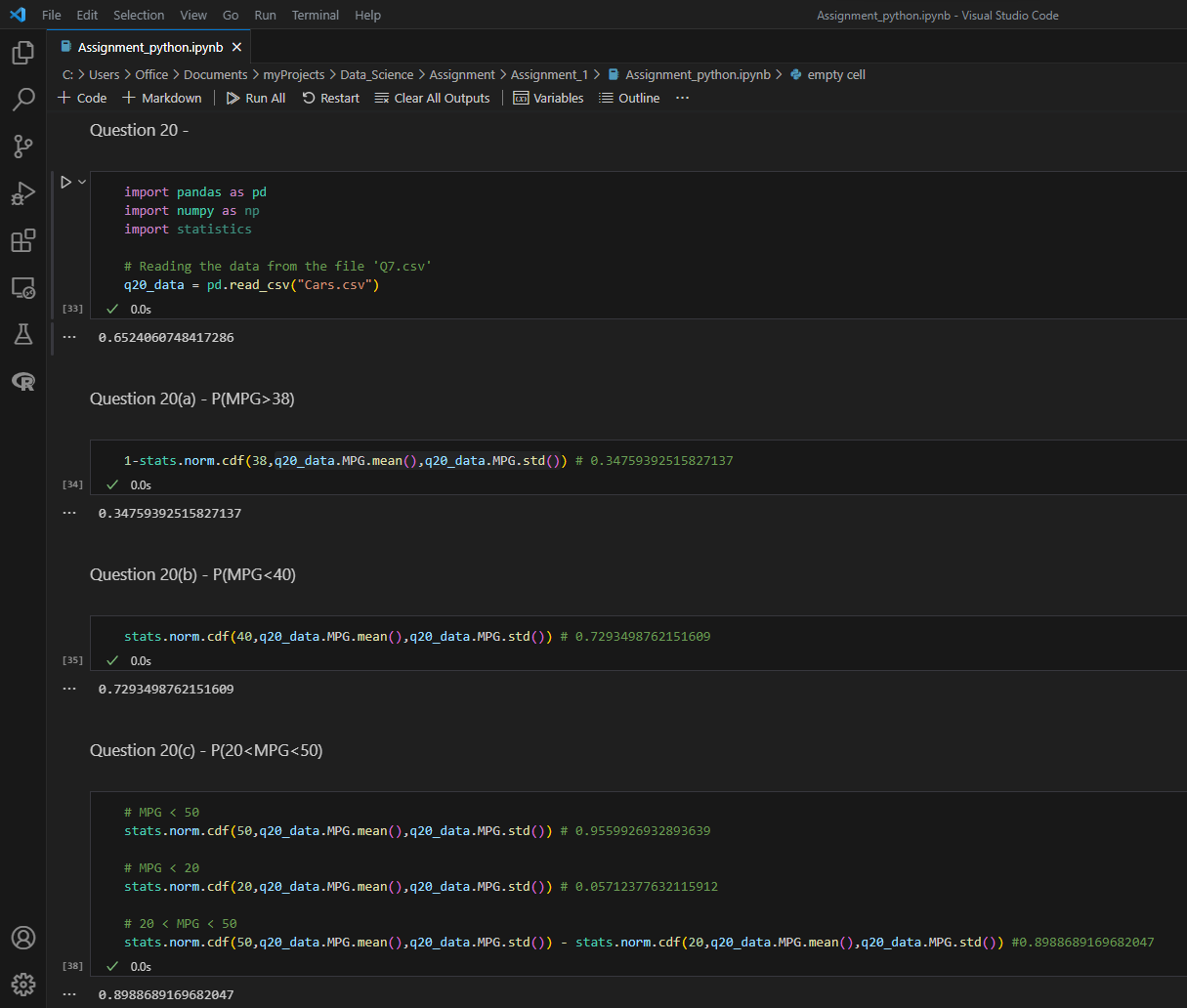
Data \_set: Cars.csv

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

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

Solution:



1. P(MPG>38) - 0.34759392515827137 i.e., 34.75
2. P(MPG<40) - 0.7293498762151609
3. P (20<MPG<50) - 0.8988689169682047

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

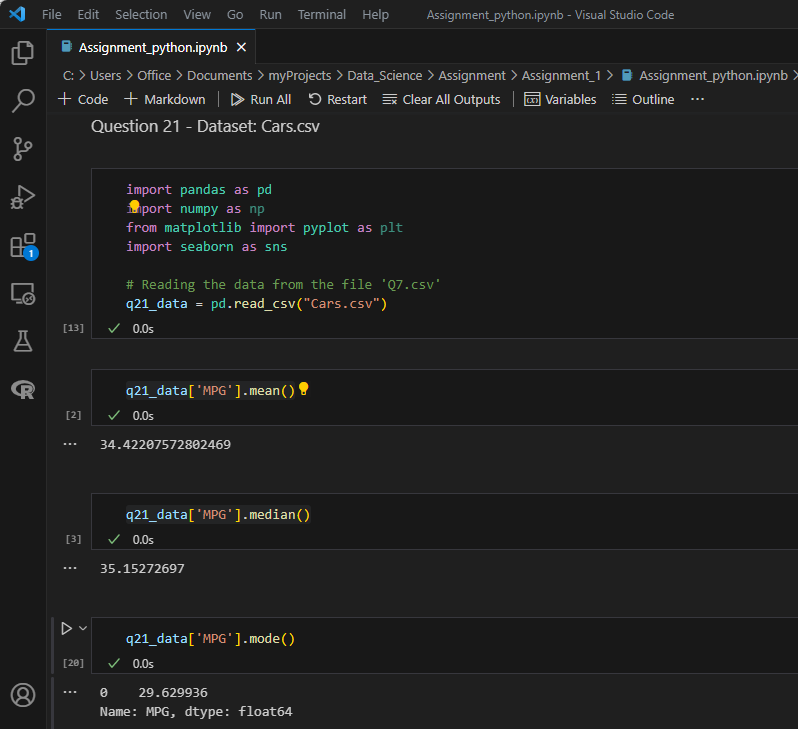
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

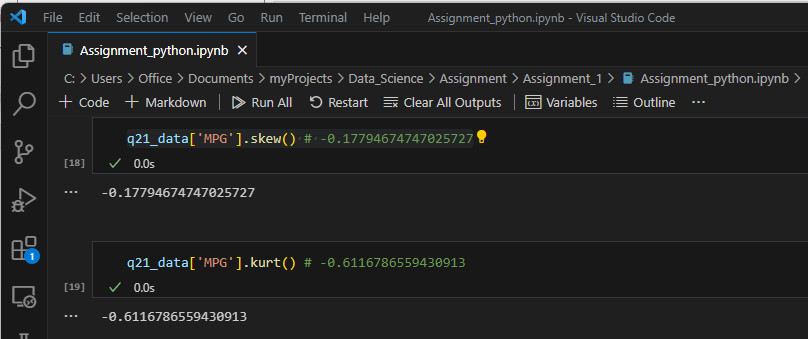
Solution:

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

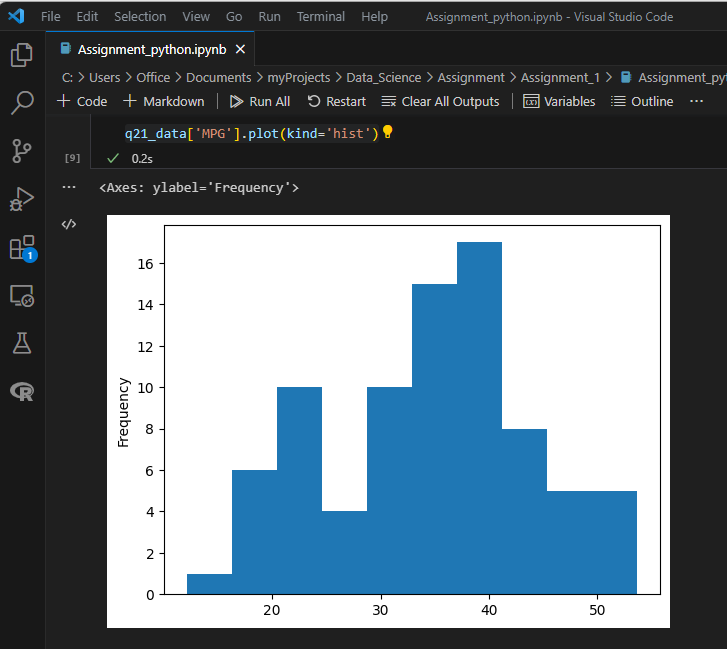


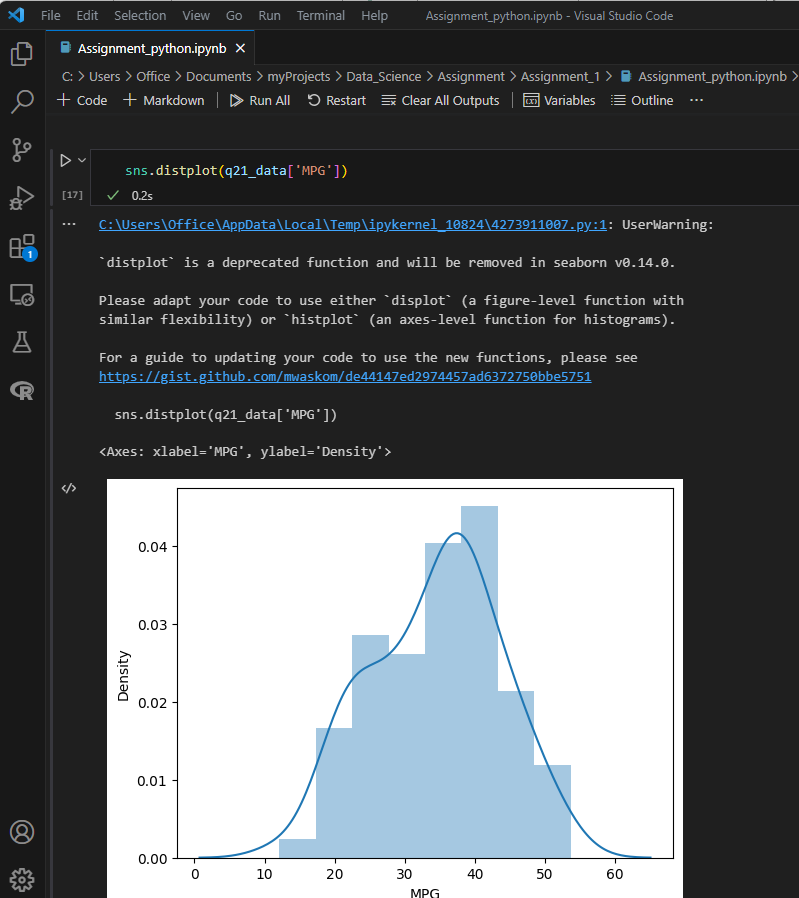
Since the mean, median and mode are close to each other, we can say that the data is fairly symmetric and follows a normal distribution.



Skewness and kurtosis are close to zero, so it is almost a normal distribution.

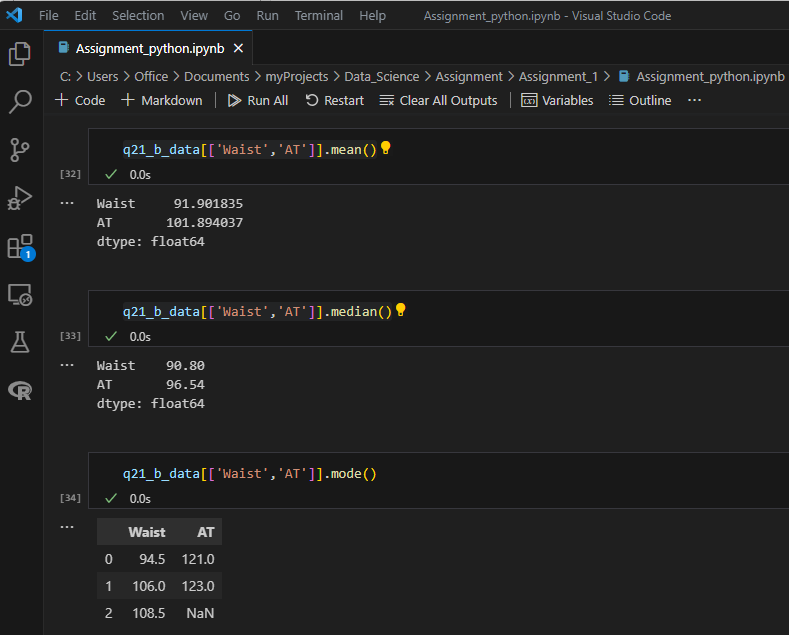
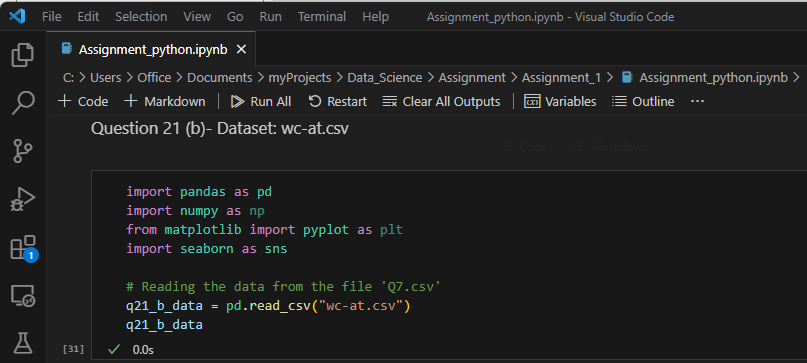
Based so the plots, we see that that the density plot and histogram are fairly symmetrical and follows a normal distribution.

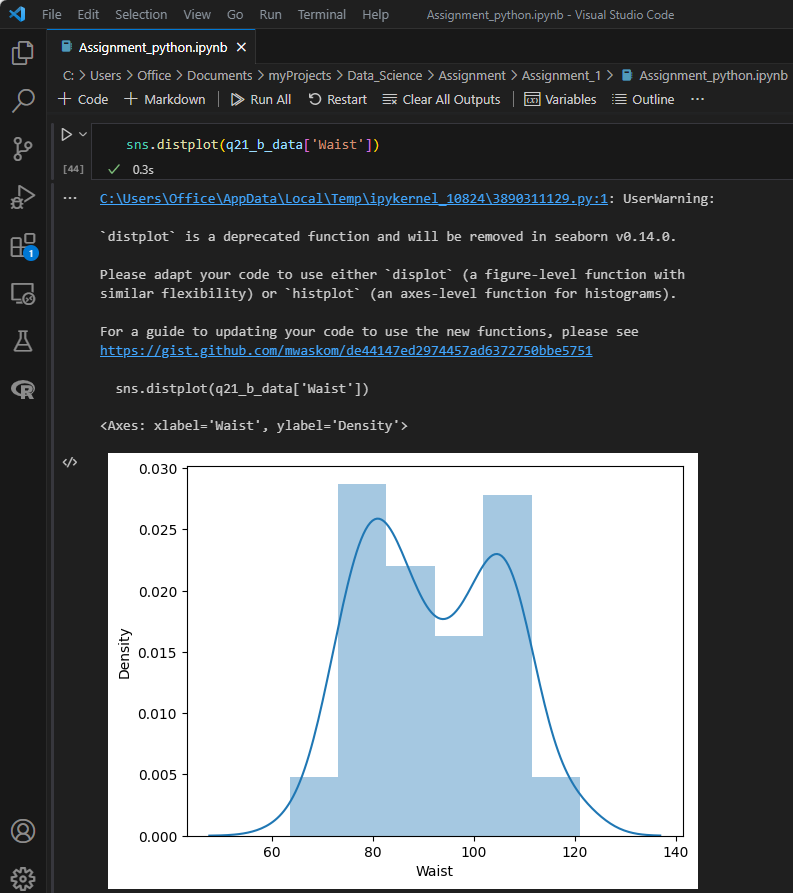


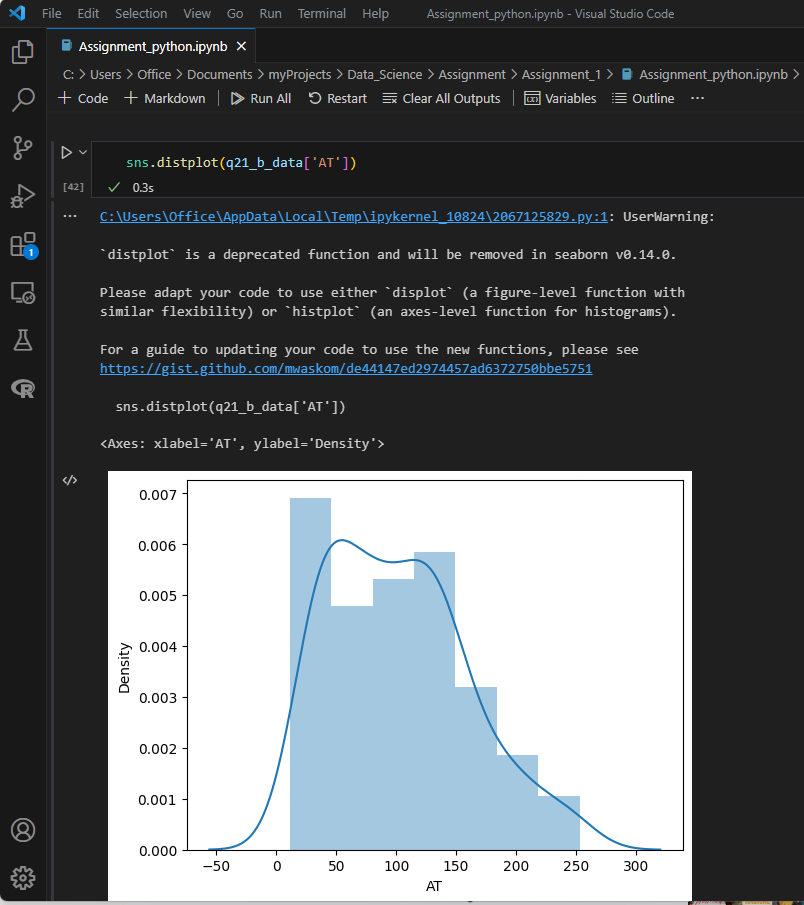


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv



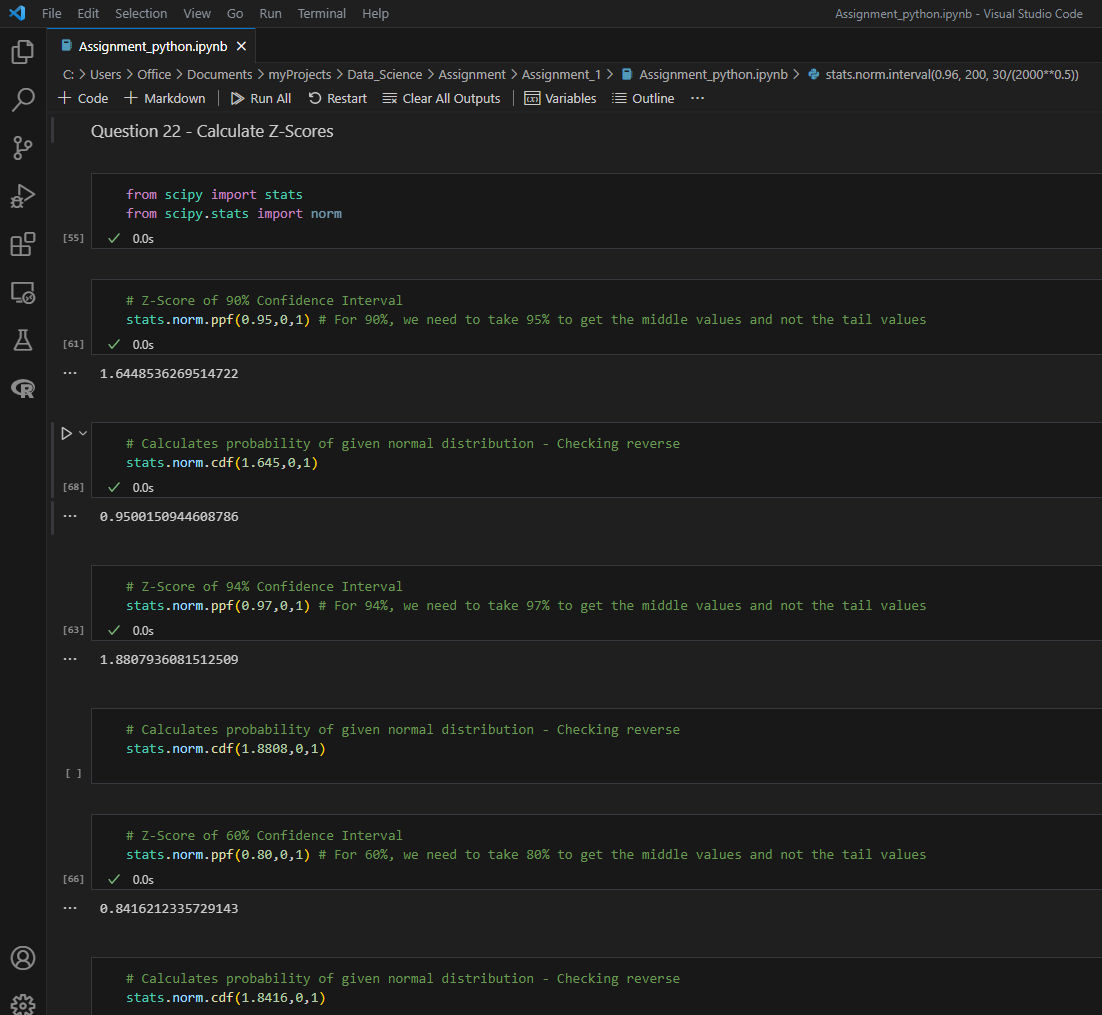




Based on the plots, we see that they are not normally distributed.

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

**Finding Z-Scores using Python:**



**Z-Score for 90% is**

Calculation of z-score for 90% -

A 90% confidence interval has two tails of 10/2 = 5% so it goes from 5% to 95% which leaves 90% in the middle so look up the Z for P(z<Z) = 0.95

The two closest values in the z-table P(z<1.64) = 0.94950 and P(z< 1.65) = 0.95053

Interpolating –

1.64 + (0.95 - 0.94950) \* (0.01) / (0.95053 - 0.94950) = 1.645(approx.)

**Z-Score for 94% is 1.881**

Calculation of z-score for 94% -

A 94 % confidence interval has two tails of 6/2 = 3% so it goes from 3% to 97% which leaves 94% in the middle so look up the Z for P(z<Z) = 0.97

The two closest values in the z-table P(z<1.88) = 0.96995 and P(z< 1.89) = 0.97062

Interpolating –

1.88 + (0.97- 0.96995) \* (0.01) / (0.97062- 0.96995) = 1.880746(approx.)

**Z-Score for 60% is**

Calculation of z-score for 60% -

A 60% confidence interval has two tails of 40/2 = 20% so it goes from 20% to 80% which leaves 80% in the middle so look up the Z for P(z<Z) = 0.80

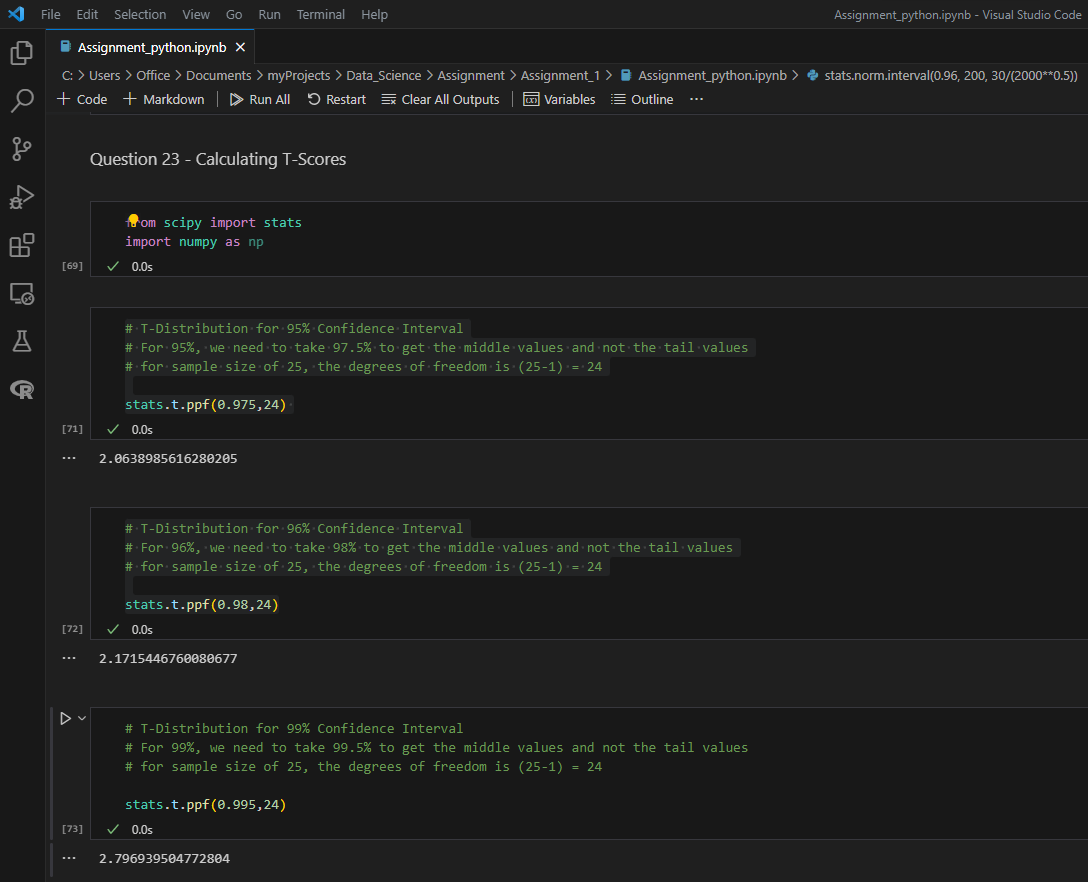
The two closest values in the z-table P(z<0.84) = 0.79955 and P(z<0.85) = 0.80234

Interpolating –

0.84 + (0.80 - 0.79955) \* (0.01) / (0.80234 - 0.79955) = 0.8416 (approx.)

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

Solution:

**Finding T-Distribution using Python:**

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

Solution:

μ = 270, n = 18, X̄ = 260, s = 90

|  |  |  |
| --- | --- | --- |
| **Parameters Provided** | | **Values** |
| **Sample Mean (X̄)** | | 260 |
| **Standard Deviation (s)** | | 90 |
| **Population Mean (μ)** | | 270 |
| **Sample Size (n)** | | 18 |
| **t-value** | Formula - (X̄ - μ)/(s/√n) | -0.471404521 |

**t-value** = -0.471404521

**Calculation of t-value and p-value using Python**

