Name : Pankaj Umesh Varma

Email ID : [pv43770@gmail.com](mailto:pv43770@gmail.com)

Mob No. : 9372123700

Q.1) Identify the Data type for the Following:

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

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

Nominal, Ordinal, Interval, Ratio.

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

Q.3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Solution:-

Let S be the the Incident that three coins are tossed together, So our S will be ,

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

n(S) =8

Let A be the Incident that ‘Two Heads and One Tail are obtained’, So our A will be ,

A= { HHT , HTH , THH }

n(A)= 3

Probability of A = ( No. Of incidents in A ) / (Total No. Of incidents S)

P(A) = n(A) / n(S)

P(A)= 3/8=0.375=37.5%

Ans:-

If three coins are tossed then the Probability of getting two Heads and one tail is 37.5% .

Q.4) 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 :-

Let S be the Incident that Two dies are rolled together, So S will be,

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) = 36

Let A be the Incident that the sum of dies is Equal to 1, So A will be,

A = { }

n(A) = 0

Probability of A = ( No. Of incidents in A ) / (Total No. Of incidents S)

P(A) = n(A) / n(S)

P(A) = 0 / 36

P(A) = 0

Let B be the Incident that the sum of dies is Less than or equal to 4, So B will be,

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

n(B) = 6

Probability of B = ( No. Of incidents in B ) / (Total No. Of incidents S)

P(B) = n(B) / n(S)

P(B) = 6 / 36

P(B) = 1 / 6=0.166=16.6%

Let C be the Incident that the sum of dies is divisible by 2 and 3, So C will be,

C = { (1,5) , (2,4) , (3,3) , (4,2) , (5,1) , (6,6) }

n(C) = 6

Probability of C = ( No. Of incidents in C ) / (Total No. Of incidents S)

P(C) = n(C) / n(S)

P(C) = 6 / 36

P(C) = 1 / 6=0.166=16.6%

Ans:-

If Two Dice are rolled, then the probability that sum is

1. Equal to 1 is 0

B.Less than or equal to 4 is 16.6%

C.Sum is divisible by 2 and 3 is 16.6%

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 :-

Let S be the incident of drawning 2 random balls from a bag which contains 2 red , 3 green and 2 blue balls , so our S will be ,

Total balls contained by bag = 2 red + 3 green + 2 blue = 7 balls

n(S) = 7! / 5!\*2!

n(S) = 7\*6 / 2 = 21

Let A be the incident of drawing two random balls which are not blue , so our A will be ,

Bag without Blue balls = 2 red + 3 green = 5 balls

We are drawing two balls , so our n(A) will be ,

n(A) = 5! / 3!\*2!

n(A) = 5\*4 / 2 = 10

P(A) = n(A) / n(S)

P(A) = 10 /21

Ans :-

Probability of drawing no blue balls from a bag which contains 2 red, 3 green and 2 blue balls is 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

Solution :-

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count | Probability | Count X Probability |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.80 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.240 |
|  |  |  | E(x) =3.09 |

Ans: The Expected number of candies for randomly slected child is 3.09

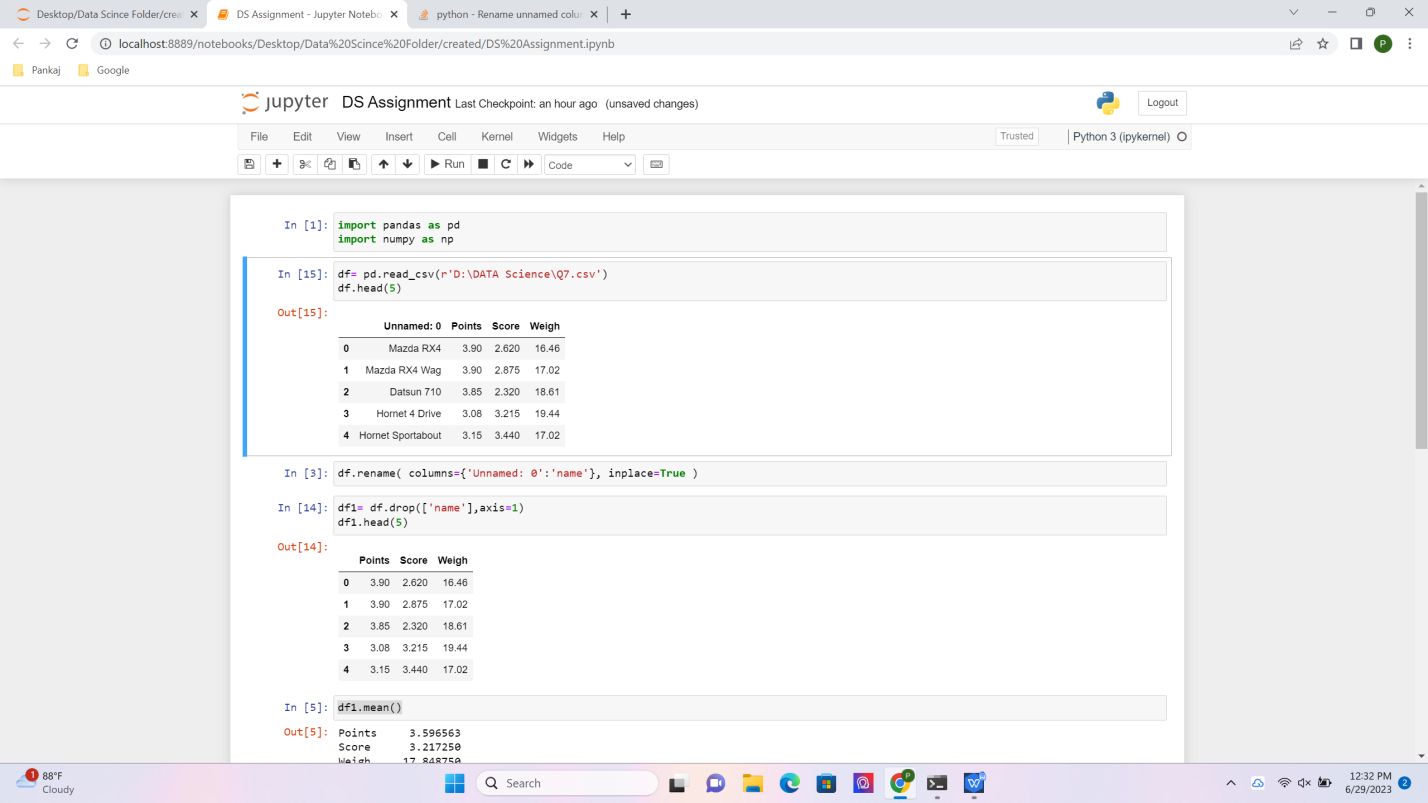
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

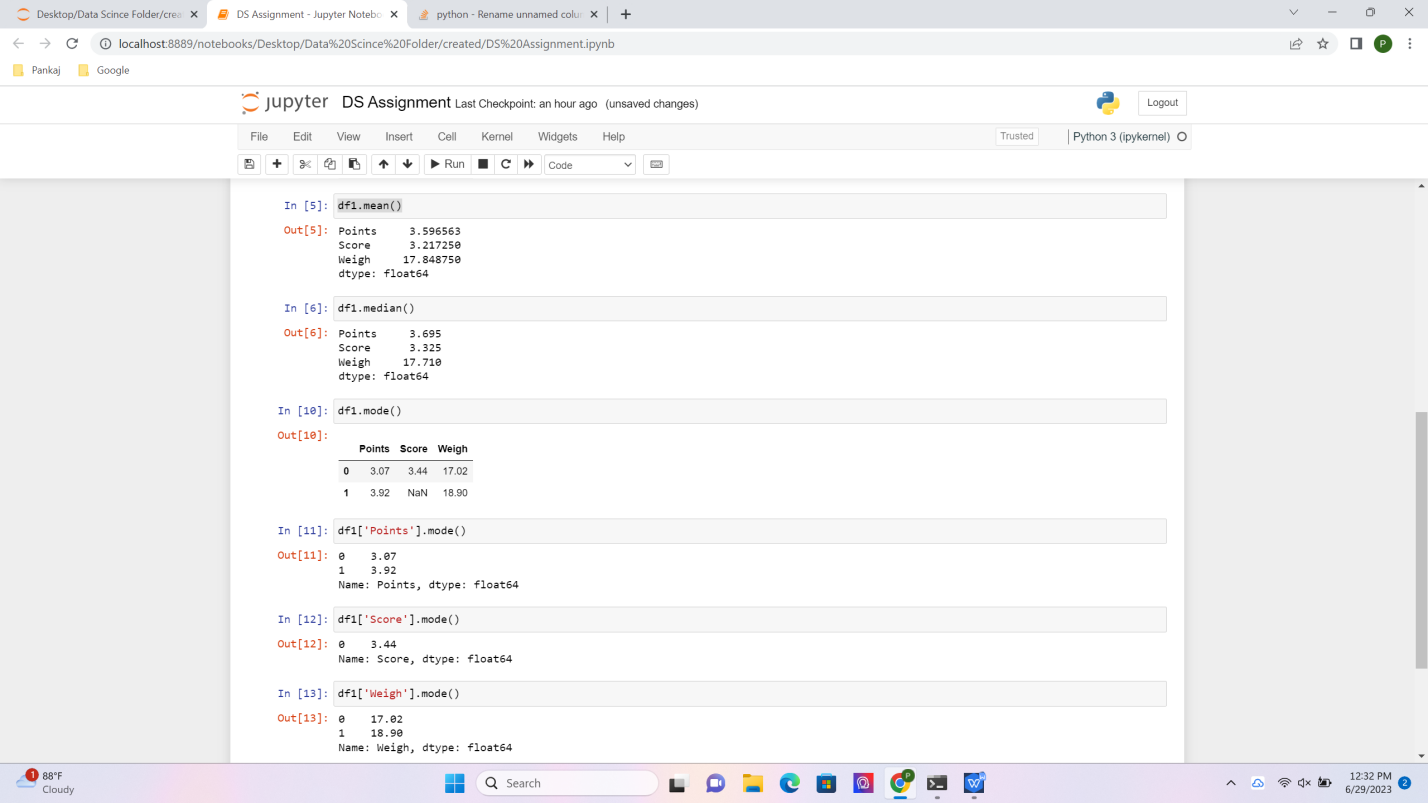
* For Points,Score,Weigh>

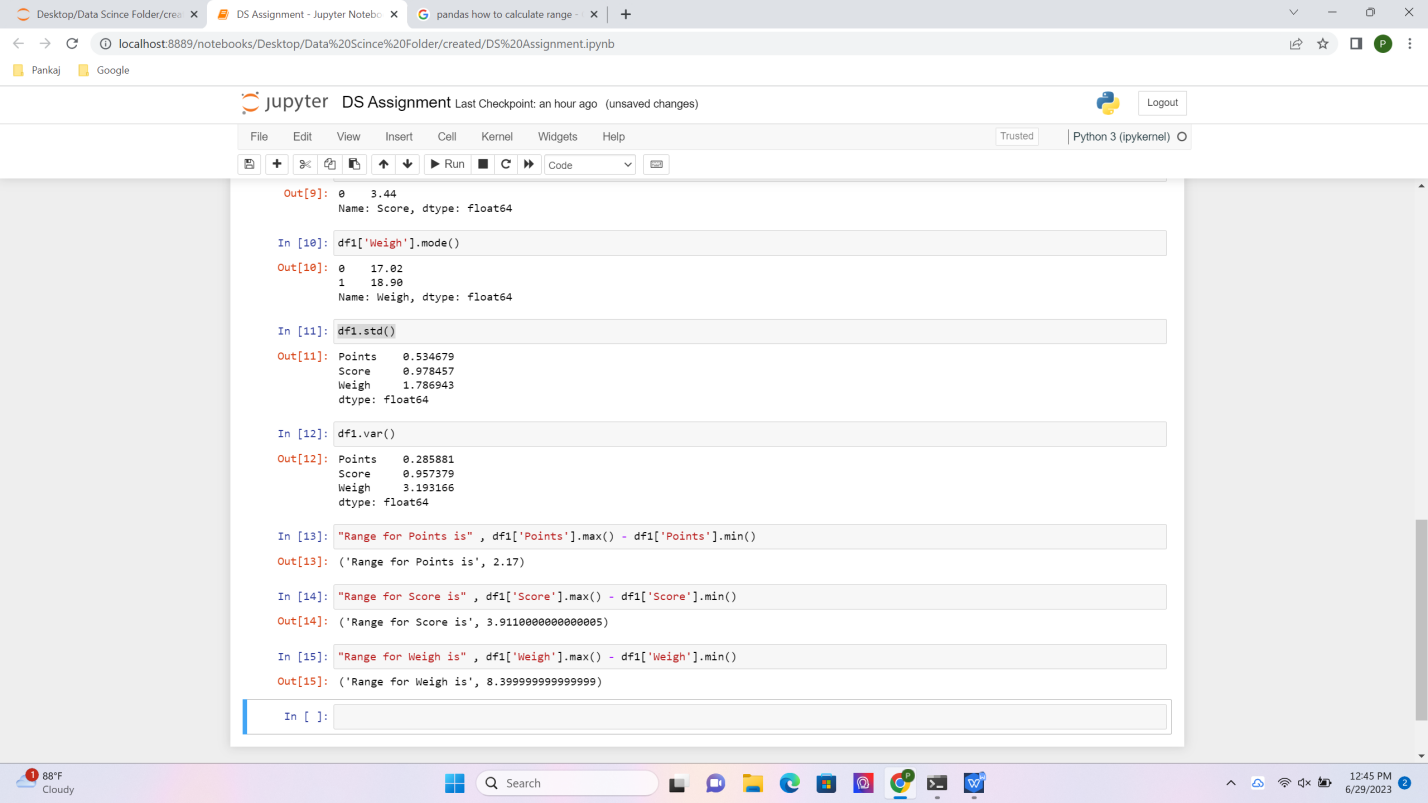
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

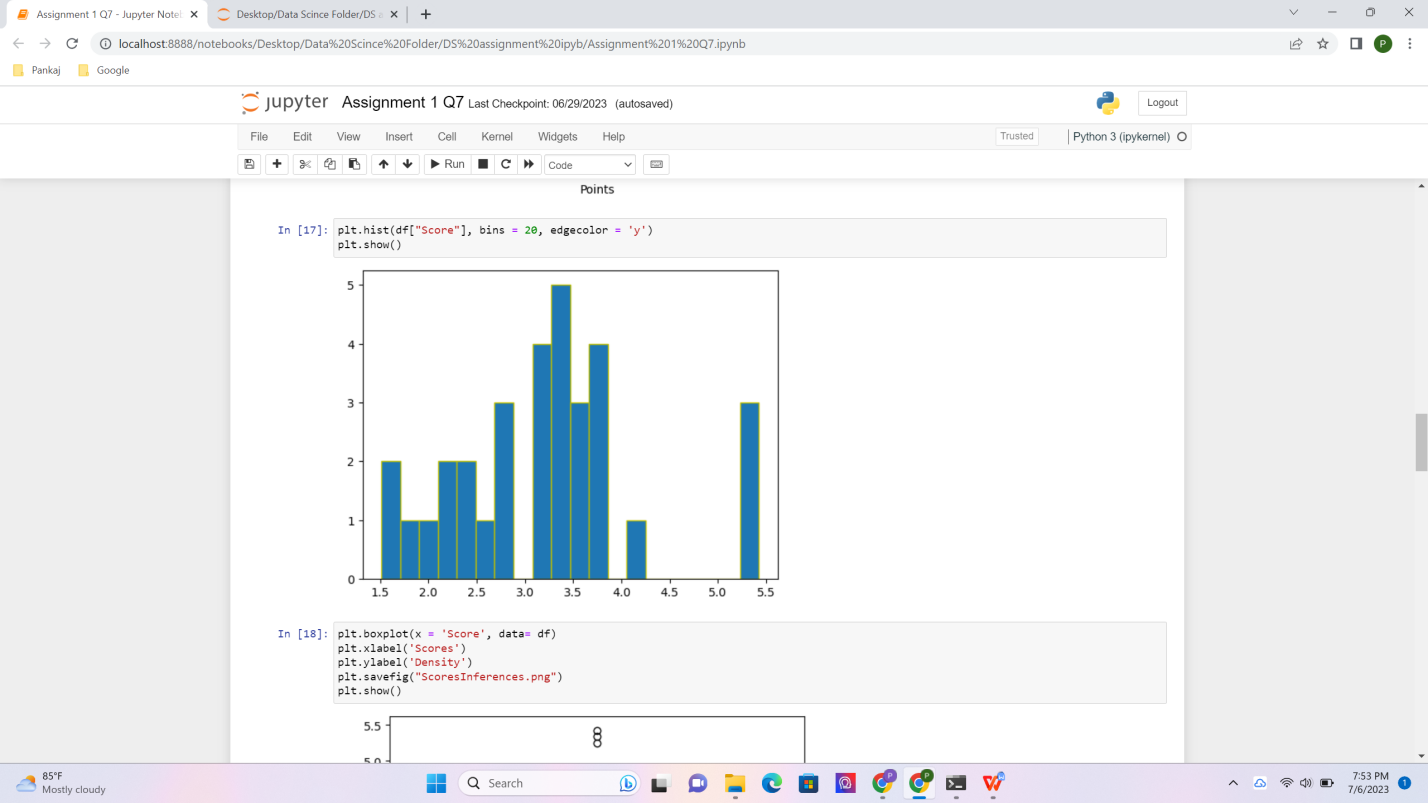
**Ans :-**

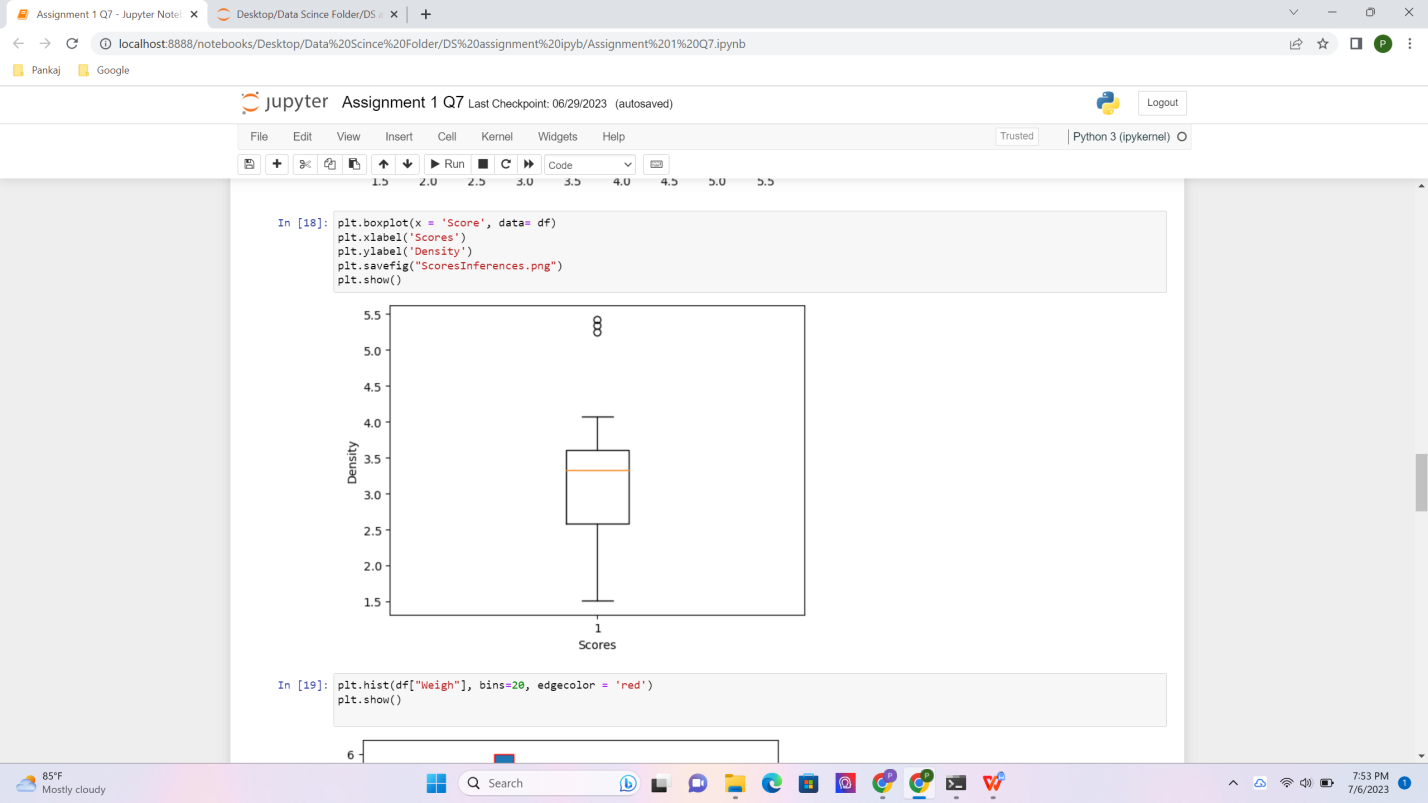




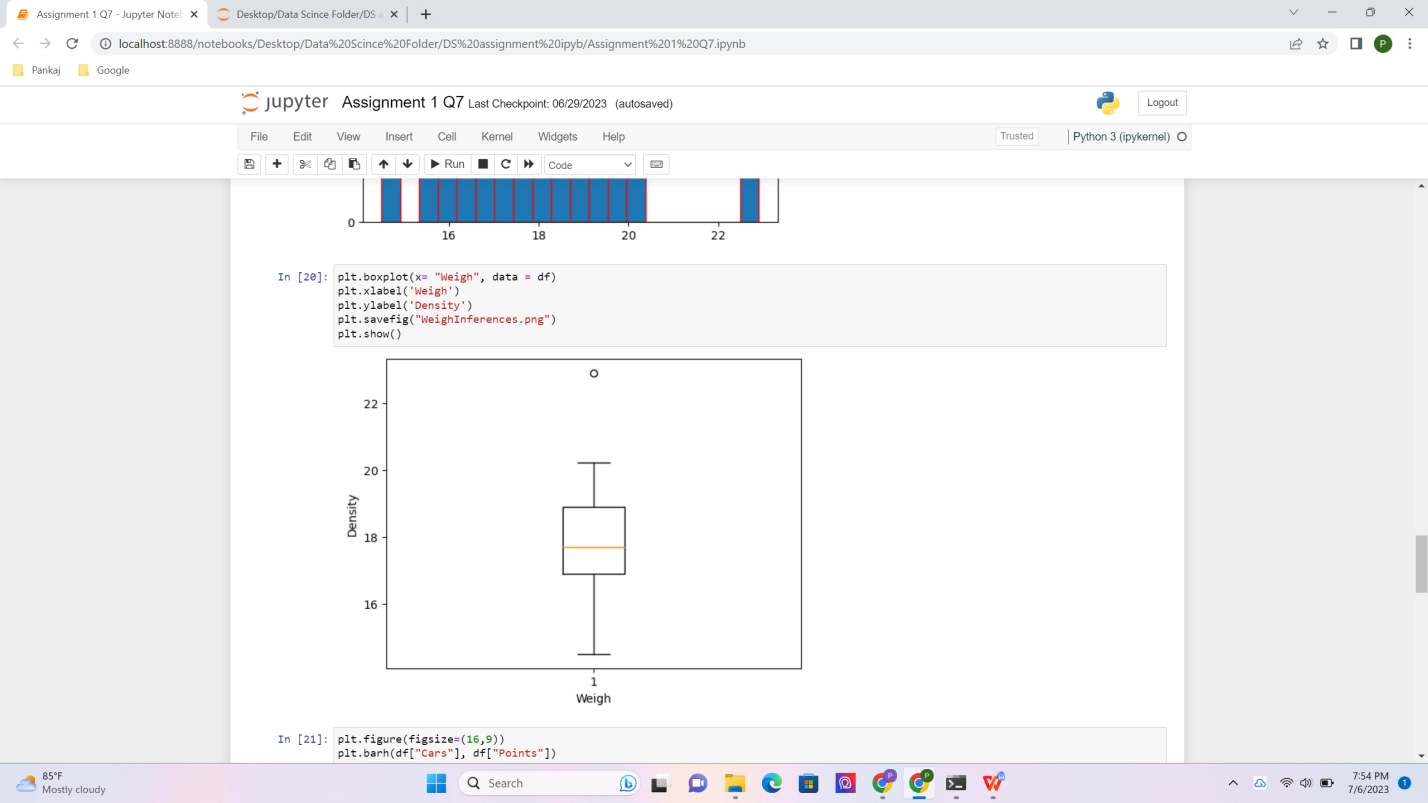


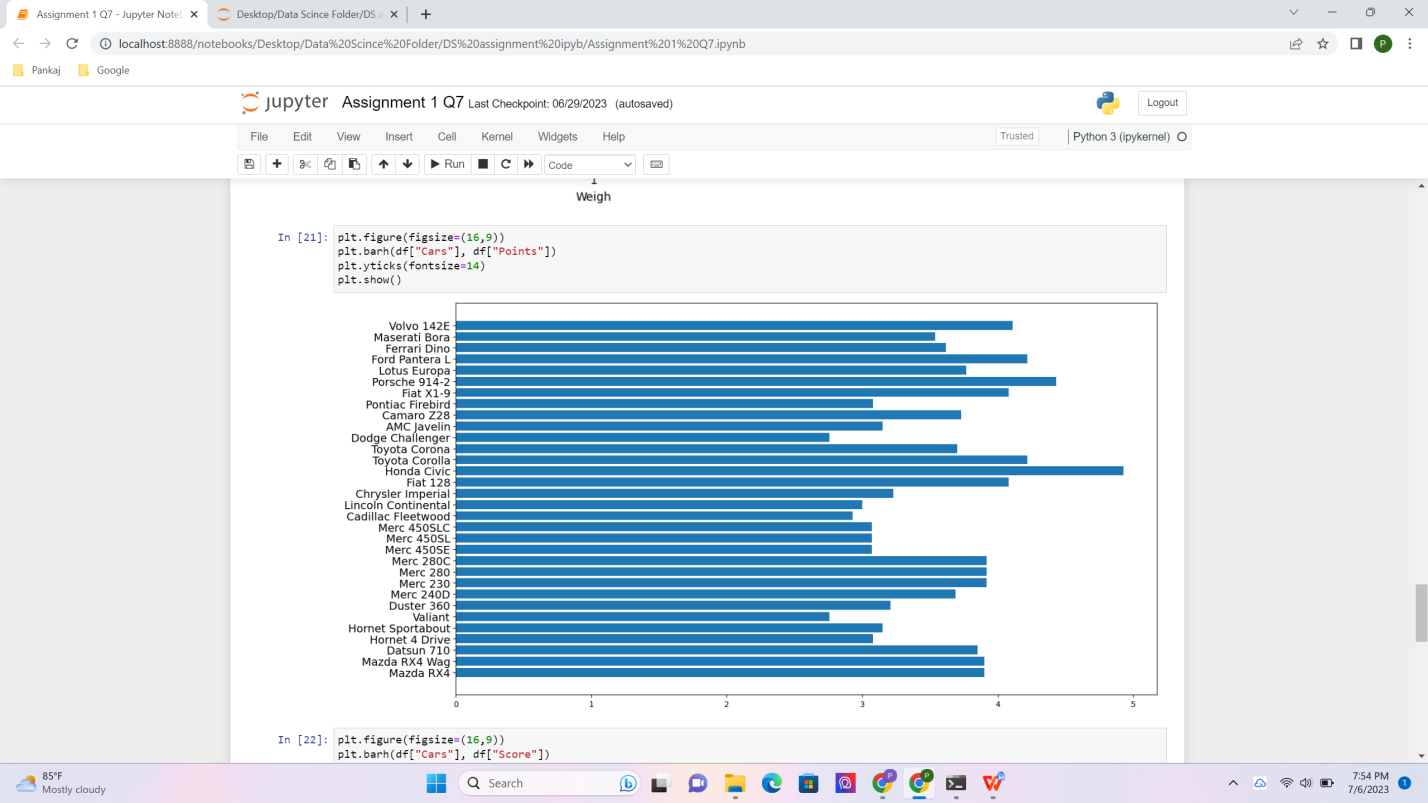


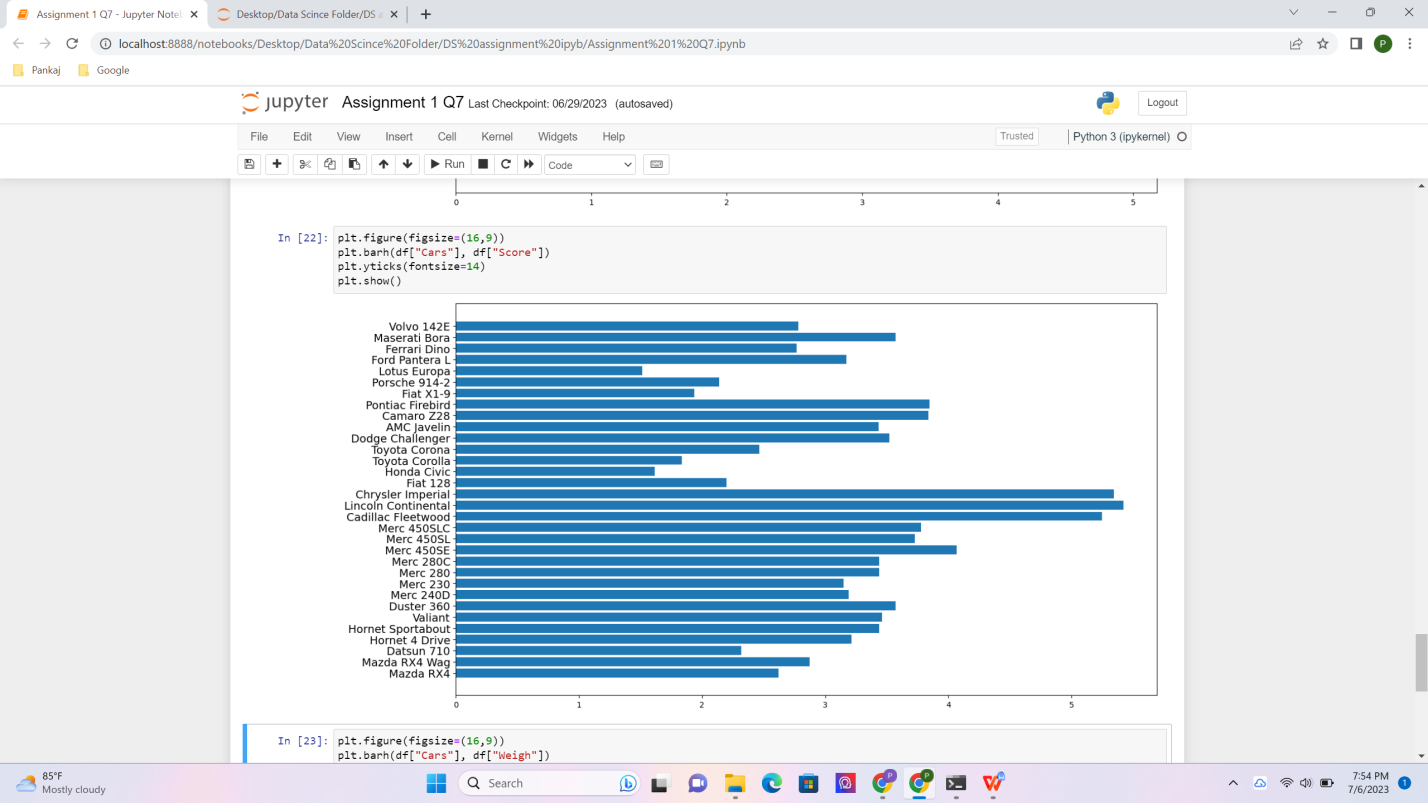


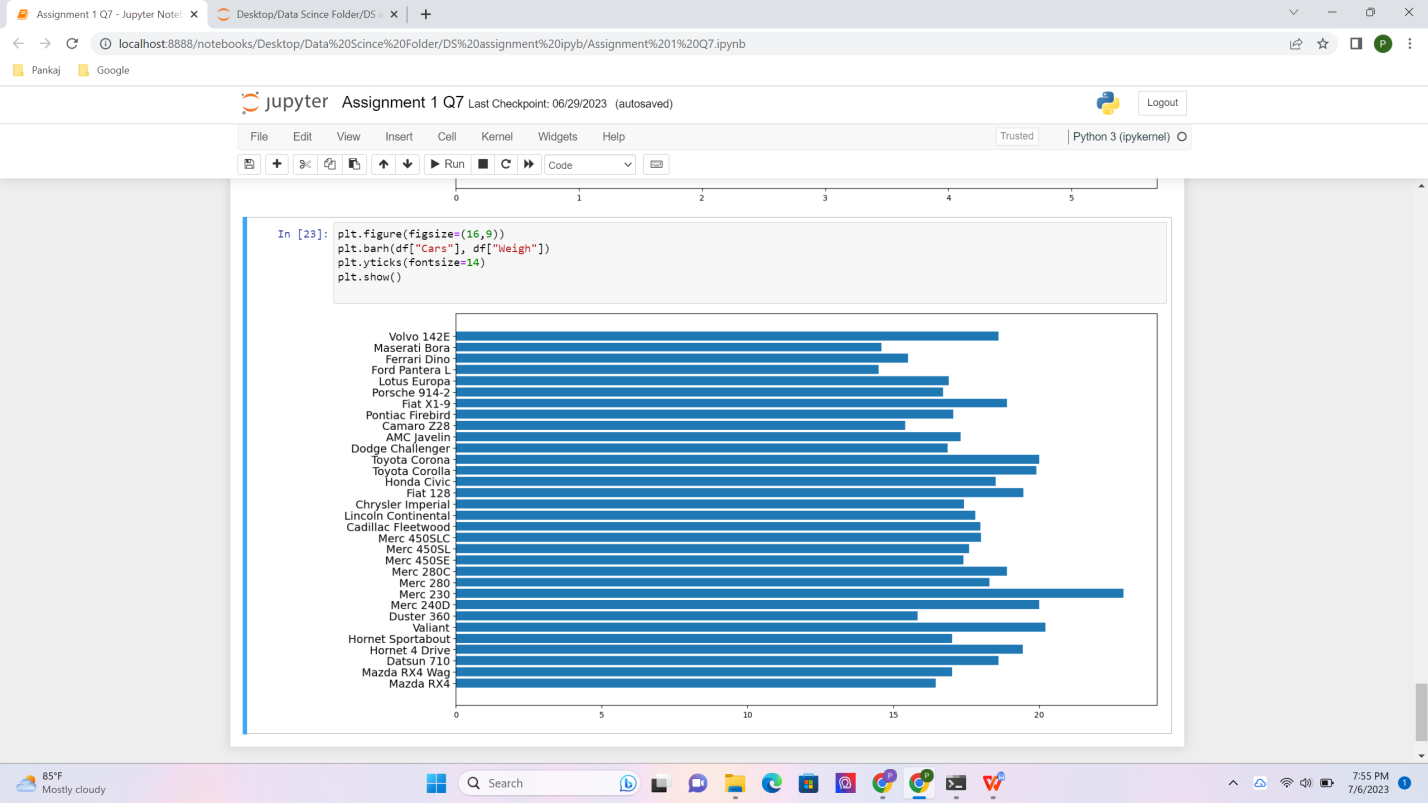












Q8) Calculate Expected Value for the problem below

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 :-

Sum of wieghts = 108 +110+123+134+135+145+167+187+199= 1208

Total no. Of wieghts = 9

Expected value of the weight of that patient = 1208 / 9

=134.22

Ans: Expected value for randomly selected Patient is 134.22

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

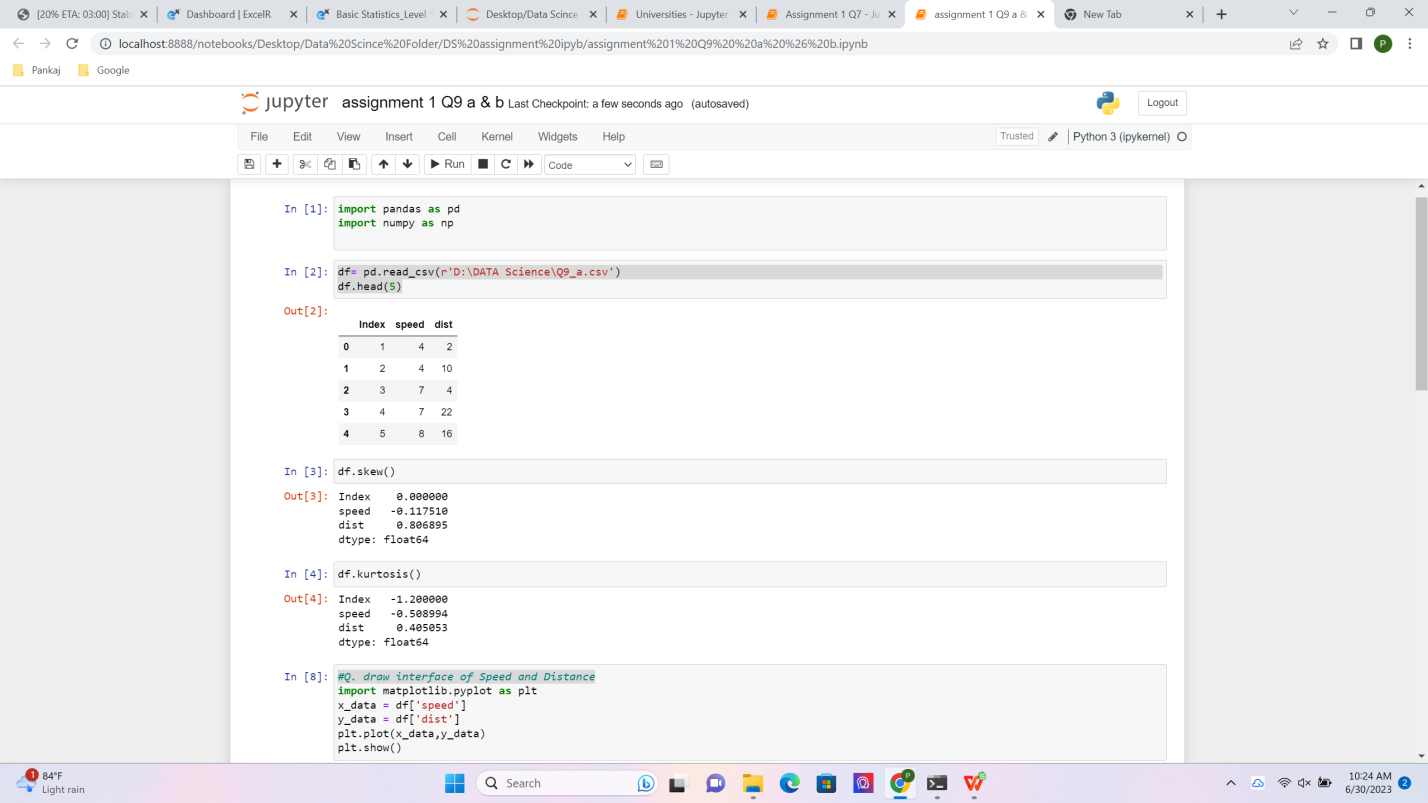
**Cars speed and distance**

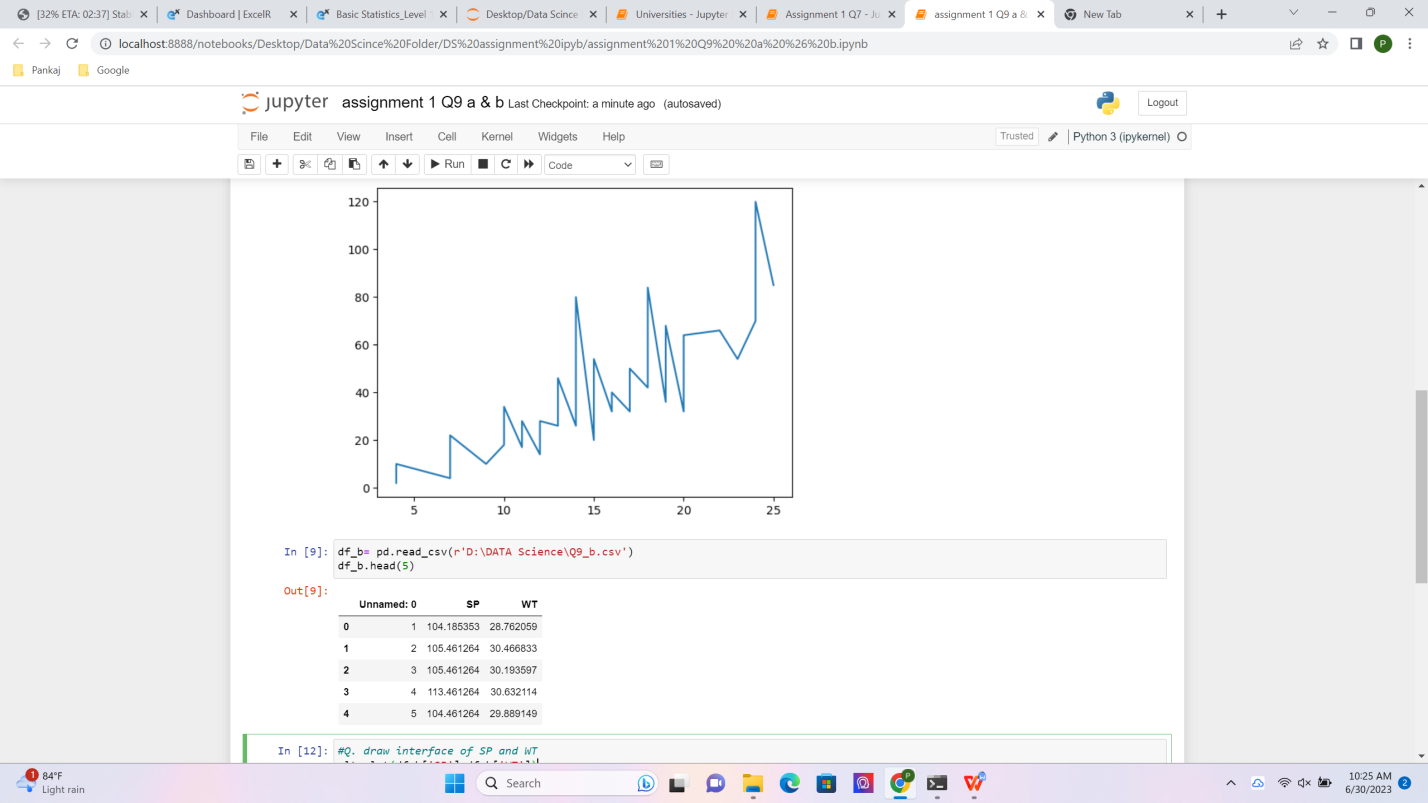
**Use Q9\_a.csv**

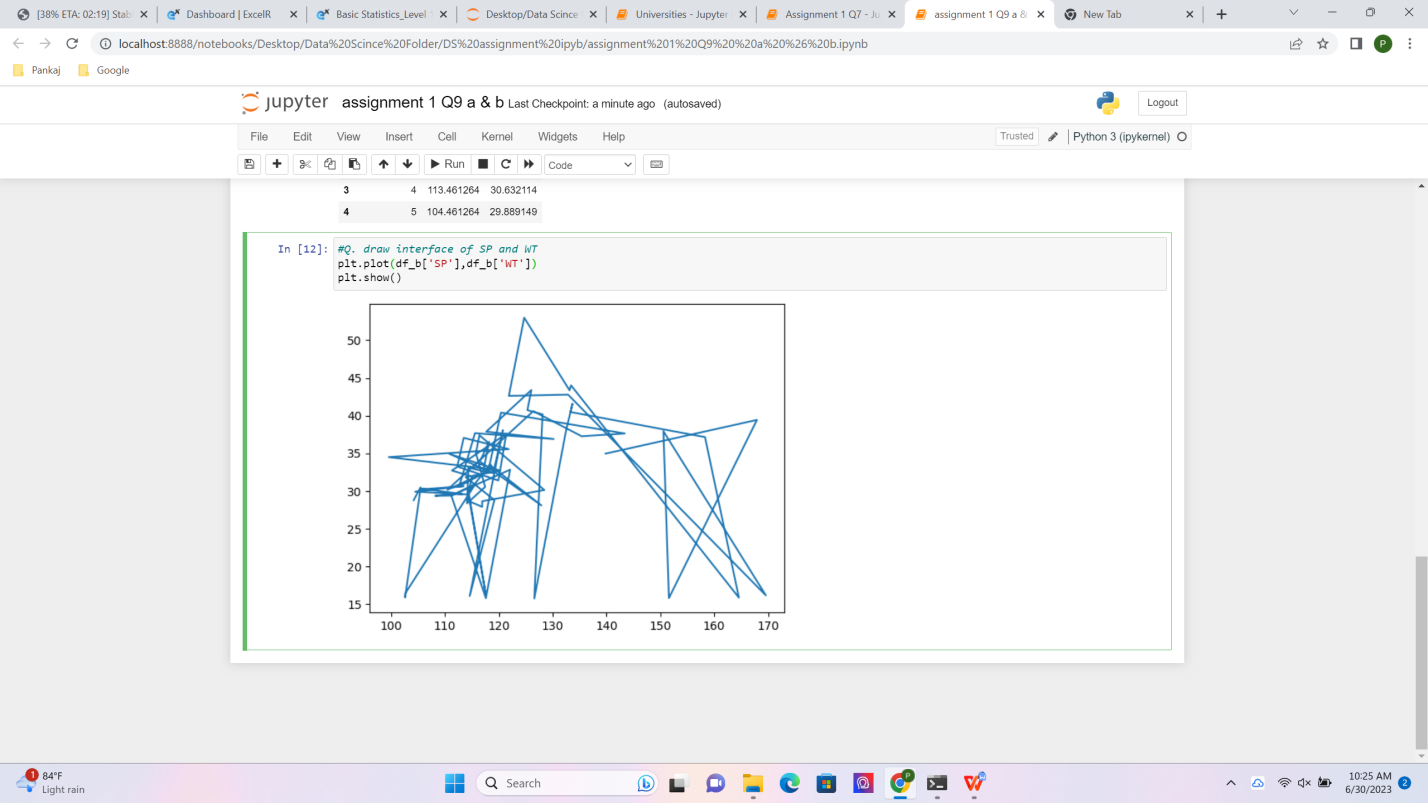
**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS :-**

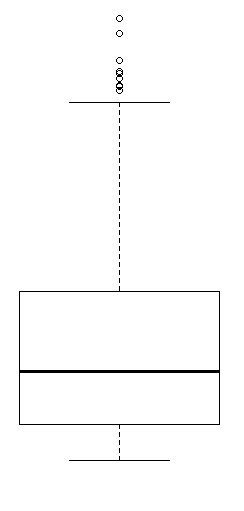




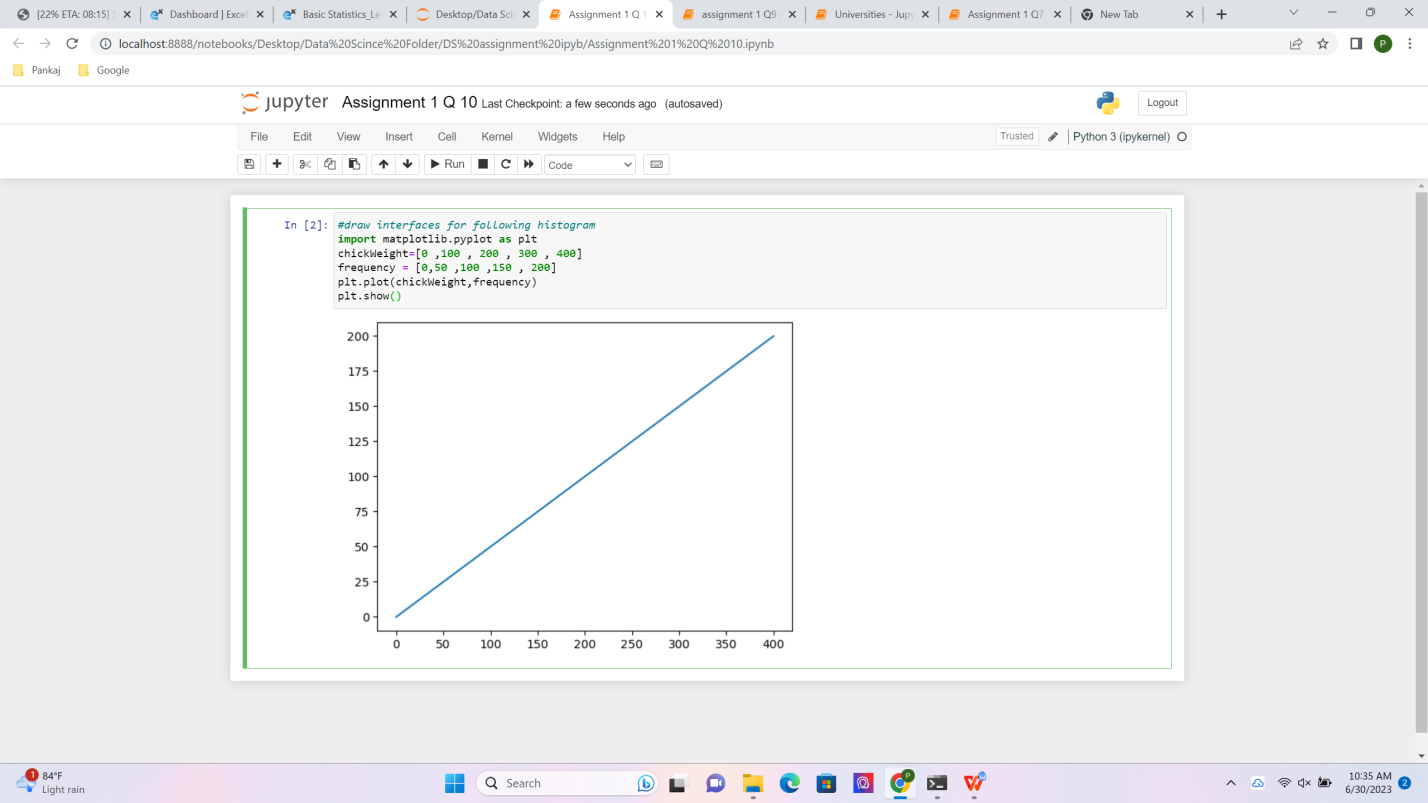


**Q10) Draw inferences about the following boxplot & histogram**

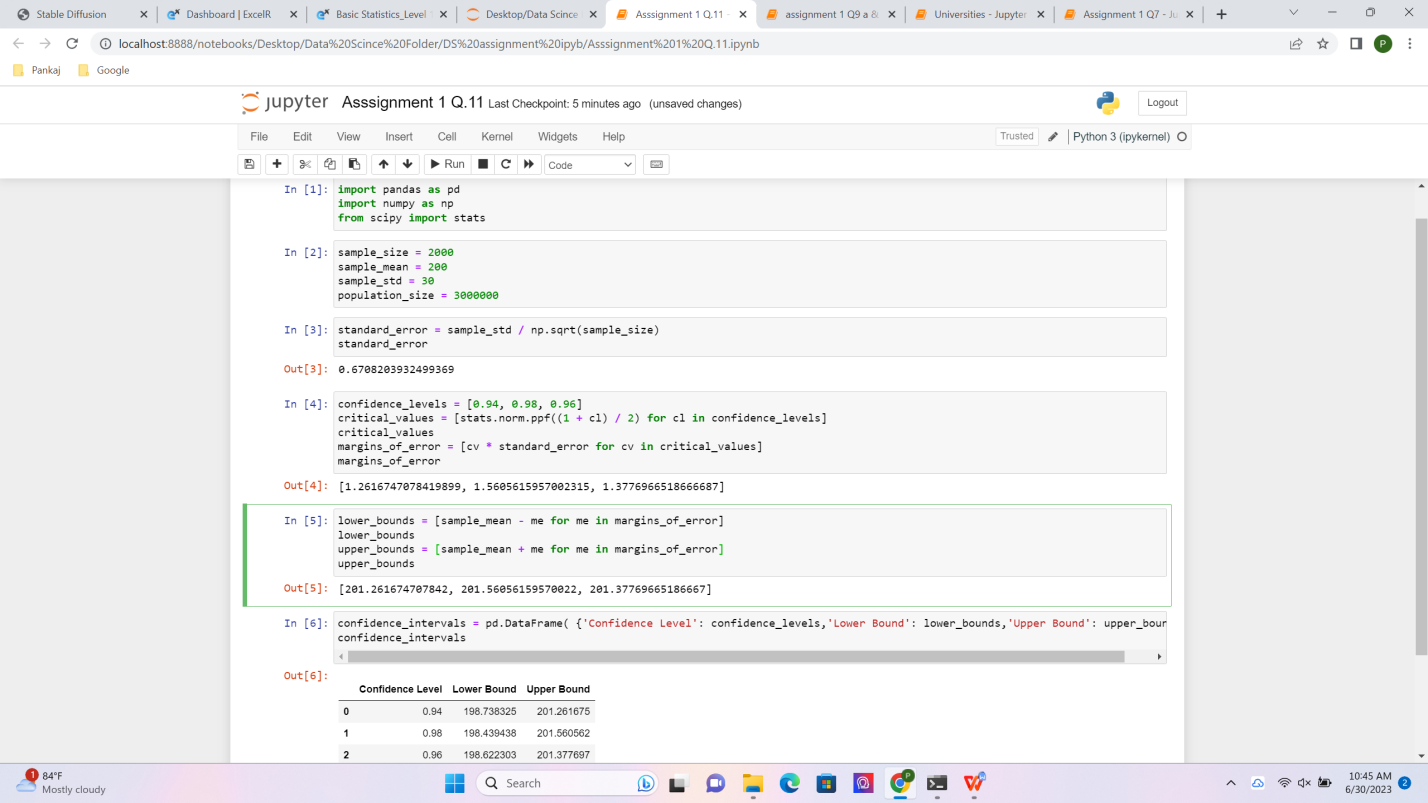




ANS:-



**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?

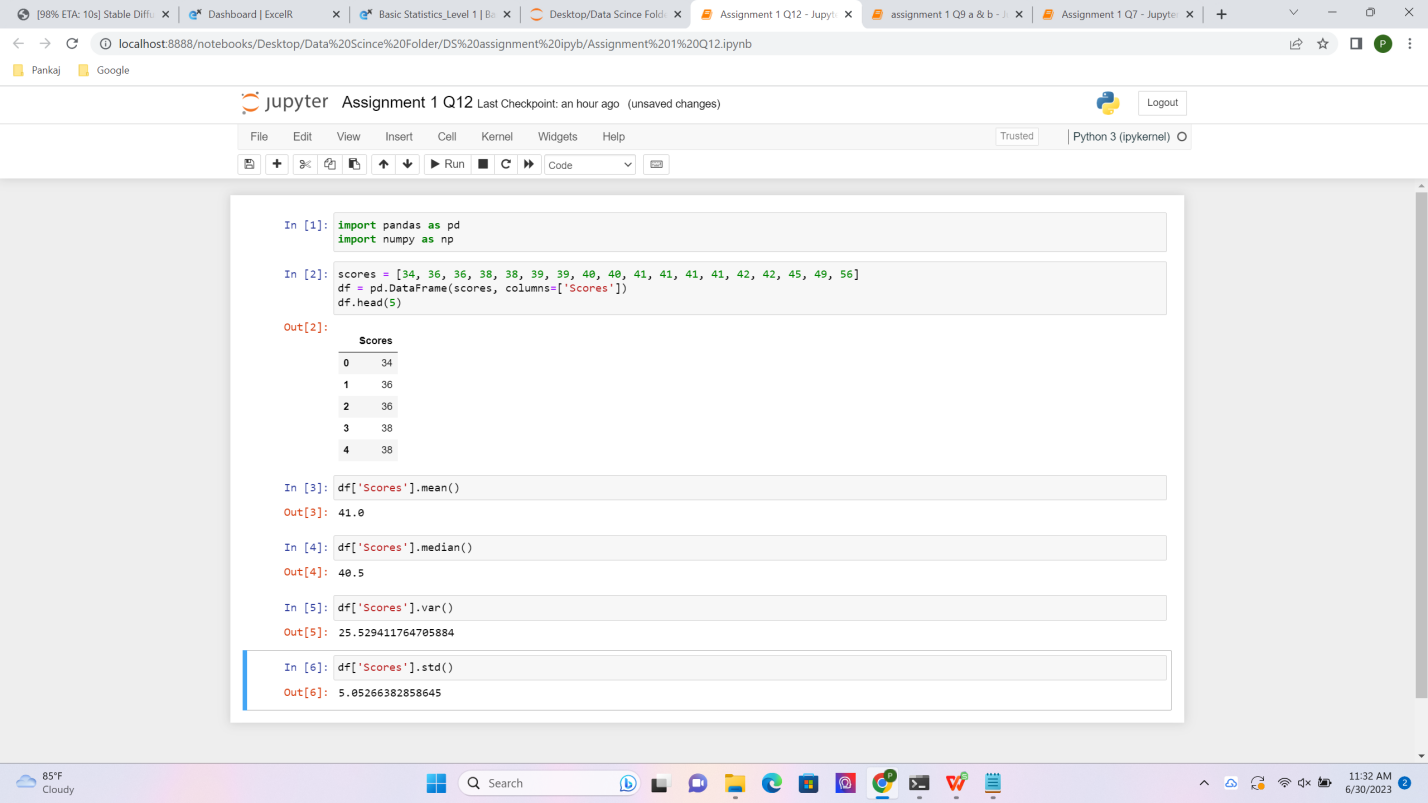


**Q12)** Below are the scores obtained by a student in tests

**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 :-



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

Ans :-

When the mean and median of a dataset are equal, it implies that the data is symmetrically distributed. In a symmetric distribution, the skewness is zero or very close to zero.

Skewness is a measure of the asymmetry of a distribution. It quantifies the extent to which a dataset deviates from being symmetric. If the skewness is positive, it indicates that the tail of the distribution is longer on the right side (right-skewed or positively skewed). Conversely, if the skewness is negative, it means that the tail is longer on the left side (left-skewed or negatively skewed).

When the mean and median are equal, it implies that the dataset has a central tendency in the middle, with no significant skewness in either direction. This suggests that the data is evenly distributed around the central value, resulting in a symmetrical distribution.

In summary, when the mean and median are equal, it indicates a symmetric distribution with zero or very close to zero skewness.

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

ANS :-

When the mean is greater than the median, it indicates that the distribution is positively skewed or right-skewed. In a right-skewed distribution, the tail of the distribution extends towards the higher values, pulling the mean in that direction and causing it to be larger than the median.

In a right-skewed distribution:

- The majority of the data tends to cluster towards the lower values.

- The tail on the right side of the distribution is longer, indicating the presence of some higher values that are relatively far from the central tendency.

- The median is closer to the lower values since it is less affected by extreme values in the tail.

- The mean is pulled towards the higher values due to the influence of the larger values in the tail.

In summary, when the mean is greater than the median, it suggests a right-skewed distribution where the majority of the data is concentrated towards lower values, and there are some higher values that cause the mean to be higher than the median.

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

Ans :-

When the median is greater than the mean, it indicates that the distribution is negatively skewed or left-skewed. In a left-skewed distribution, the tail of the distribution extends towards the lower values, pulling the median in that direction and causing it to be larger than the mean.

In a left-skewed distribution:

The majority of the data tends to cluster towards the higher values.

The tail on the left side of the distribution is longer, indicating the presence of some lower values that are relatively far from the central tendency.

The median is closer to the higher values since it is less affected by extreme values in the tail.

The mean is pulled towards the lower values due to the influence of the smaller values in the tail.

In summary, when the median is greater than the mean, it suggests a left-skewed distribution where the majority of the data is concentrated towards higher values, and there are some lower values that cause the median to be higher than the mean.

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

Ans :-

A positive kurtosis value indicates that a dataset has heavy tails or outliers compared to a normal distribution. Kurtosis is a statistical measure that quantifies the shape of a distribution and describes the concentration of data points in the tails.

When the kurtosis is positive, it means that the distribution has fatter or heavier tails than a normal distribution. This indicates that there is an increased probability of extreme values or outliers in the dataset.

In a distribution with positive kurtosis:

The tails of the distribution are more pronounced, indicating the presence of outliers or extreme values.

The distribution has a sharper peak or a more peaked shape in the center compared to a normal distribution.

The distribution has a higher concentration of data points around the mean.

It's important to note that positive kurtosis does not necessarily indicate a problem or abnormality in the data. It simply suggests that the distribution has heavier tails or a more peaked shape. The interpretation of the data should be based on the context and the specific characteristics of the dataset.

In summary, a positive kurtosis value indicates a distribution with heavy tails or outliers, and a more peaked shape in the center compared to a normal distribution.

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

Ans:-

A negative kurtosis value indicates that a dataset has lighter tails or fewer outliers compared to a normal distribution. Kurtosis is a statistical measure that quantifies the shape of a distribution and describes the concentration of data points in the tails.

When the kurtosis is negative, it means that the distribution has lighter tails than a normal distribution. This suggests that there are fewer extreme values or outliers in the dataset compared to what would be expected in a normal distribution.

In a distribution with negative kurtosis:

- The tails of the distribution are less pronounced, indicating a lack of outliers or extreme values.

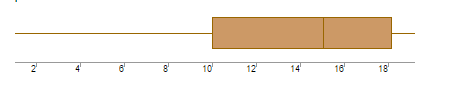
- The distribution has a flatter peak or a more flattened shape in the center compared to a normal distribution.

- The distribution has a lower concentration of data points around the mean.

Similar to positive kurtosis, it's important to note that negative kurtosis does not necessarily indicate a problem or abnormality in the data. It simply suggests that the distribution has lighter tails or a more flattened shape. The interpretation of the data should be based on the context and the specific characteristics of the dataset.

In summary, a negative kurtosis value indicates a distribution with lighter tails or fewer outliers, and a more flattened shape in the center compared to a normal distribution.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

A)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

1. What is nature of skewness of the data?

Ans:

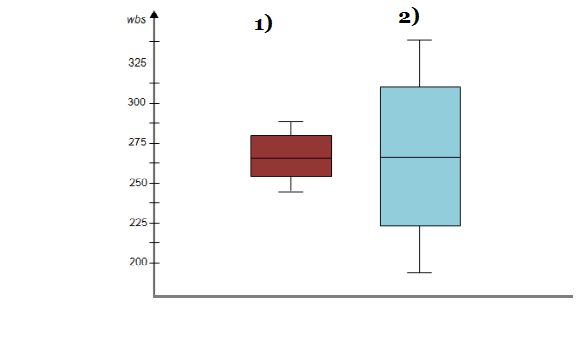
The data is a skewed towards left. The whisker range of minimum value is greater than maximum

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

Ans:

The 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 :

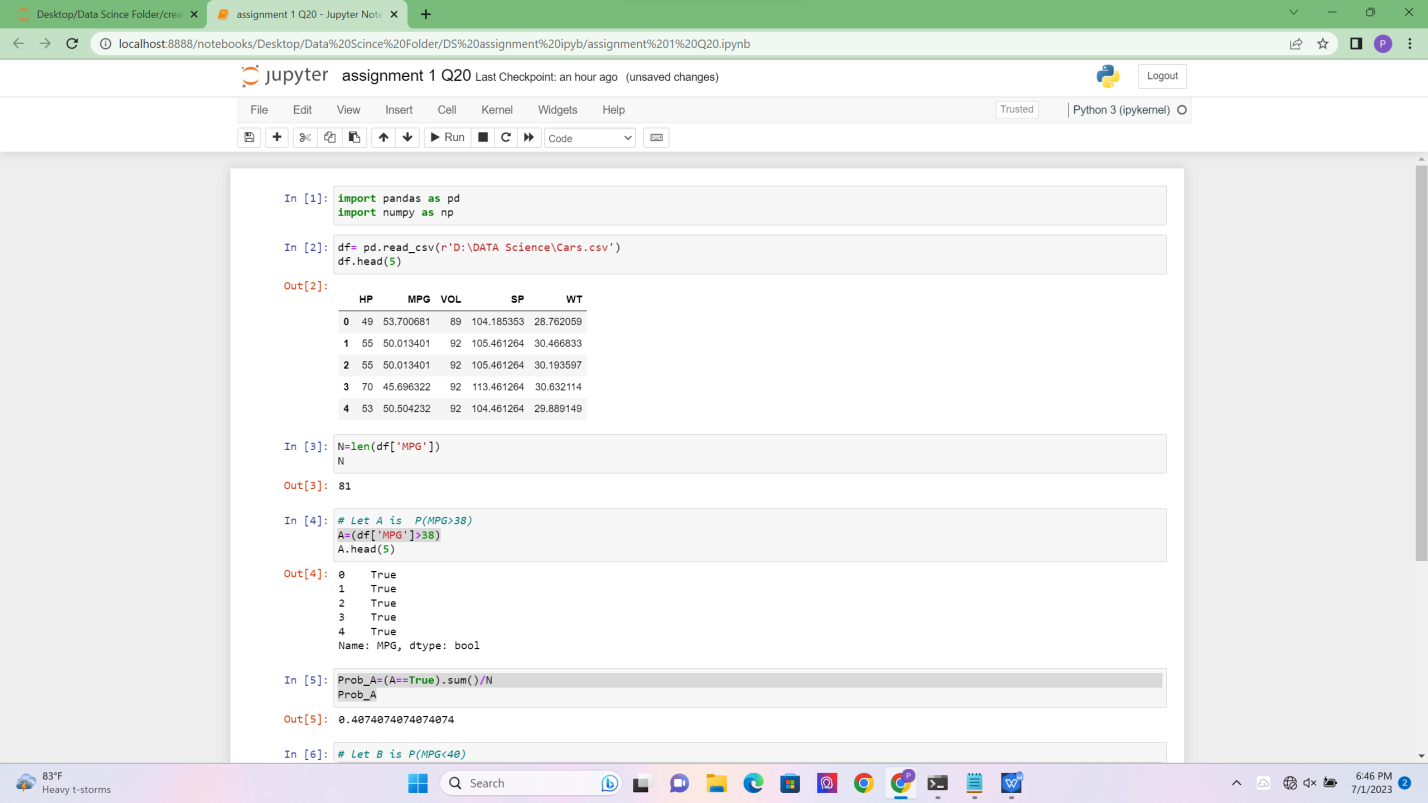
First there are no outliers.

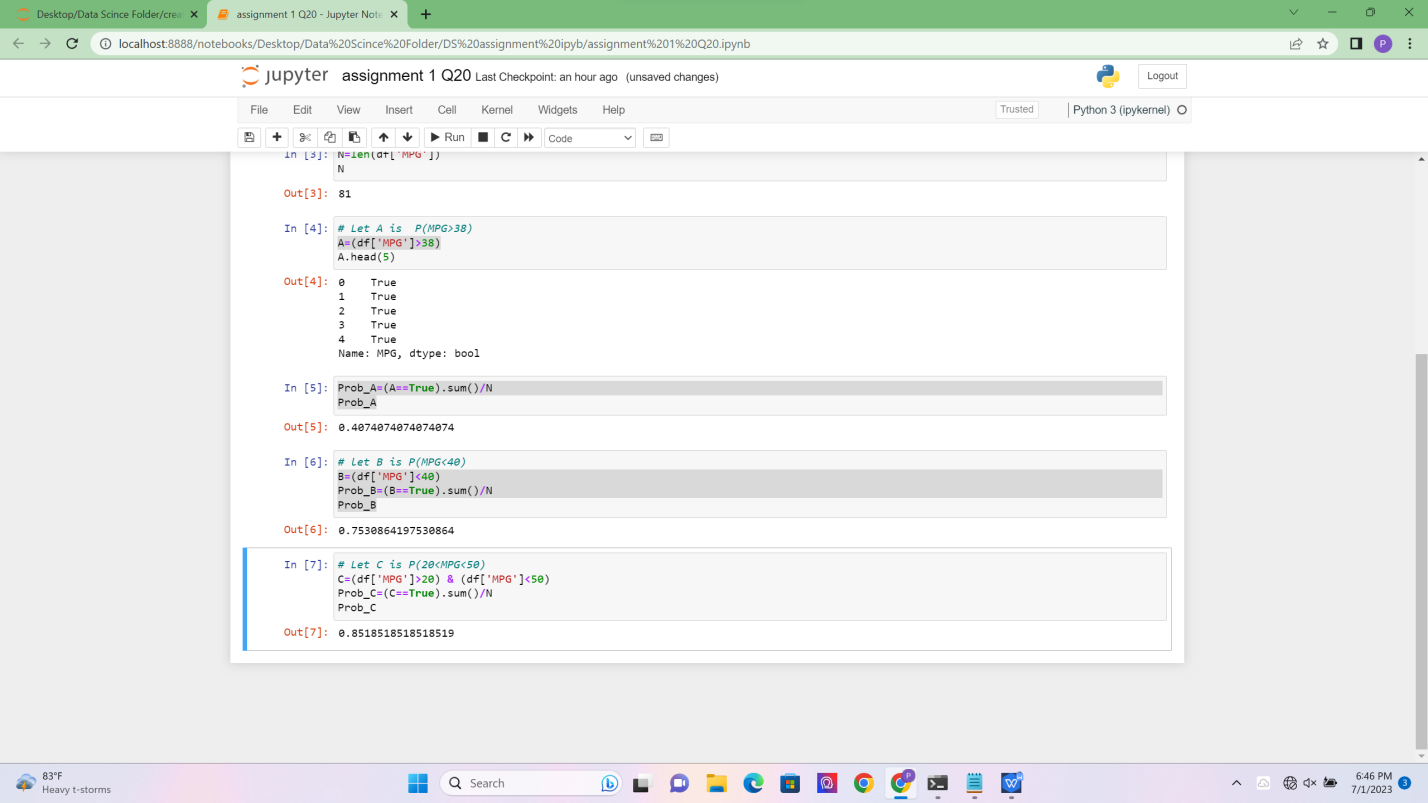
Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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

* 1. Data \_set: Cars.csv .Calculate the probability of MPG of Cars for the below cases. MPG <- Cars$MPG a.P(MPG>38) b.P(MPG<40) c.P (20<MPG<50)

Ans :-





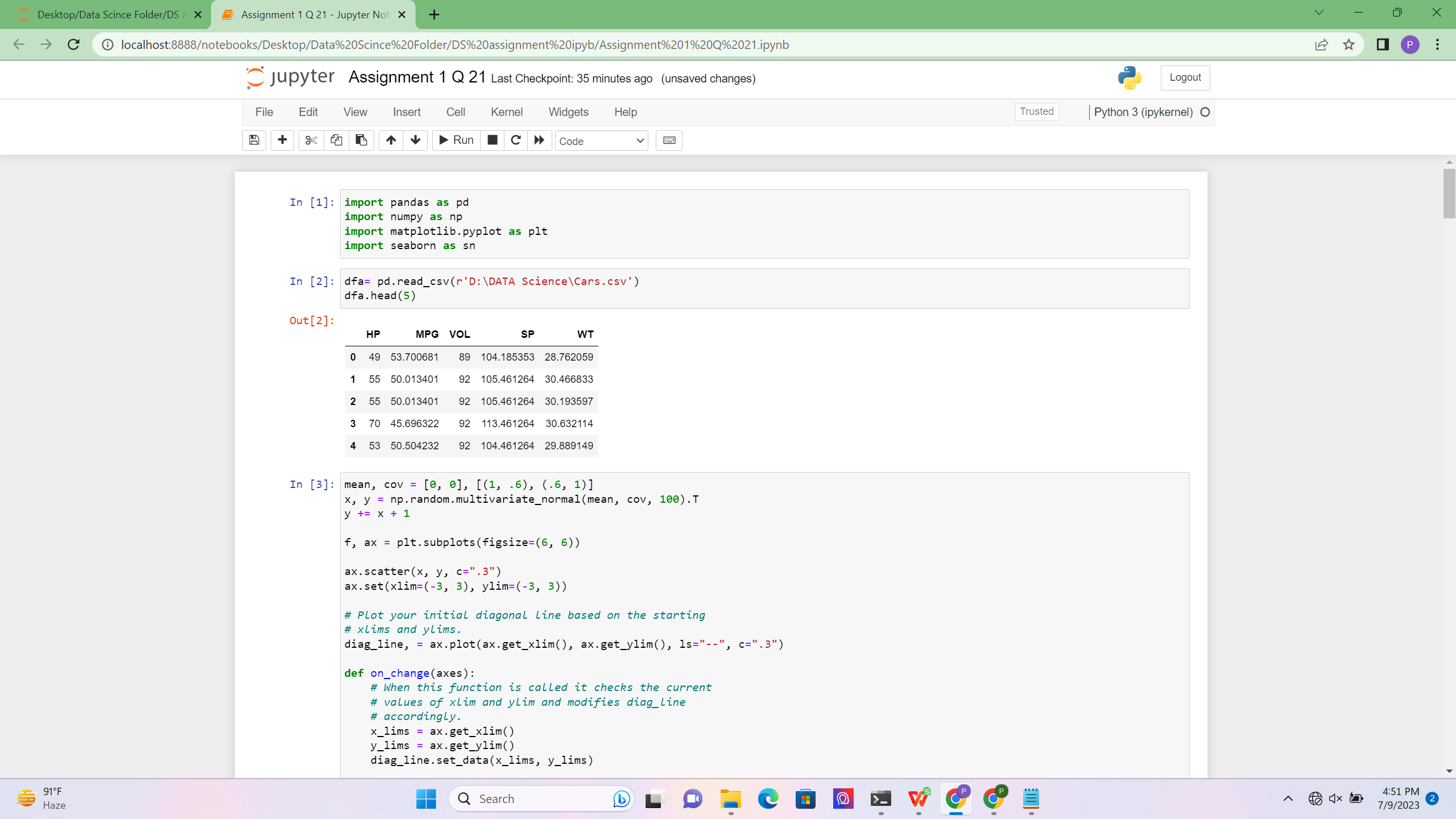
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

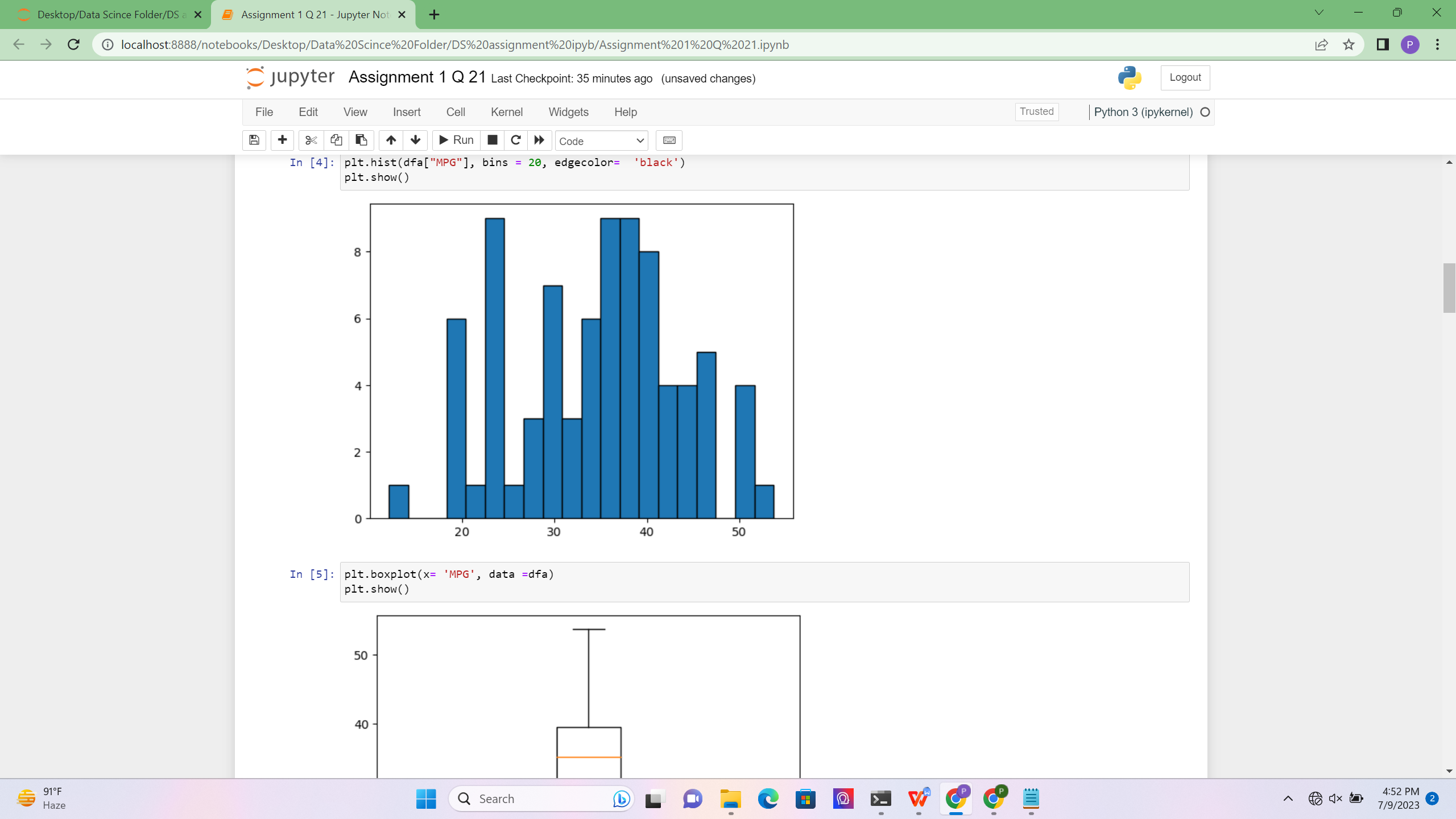
Dataset: Cars.csv

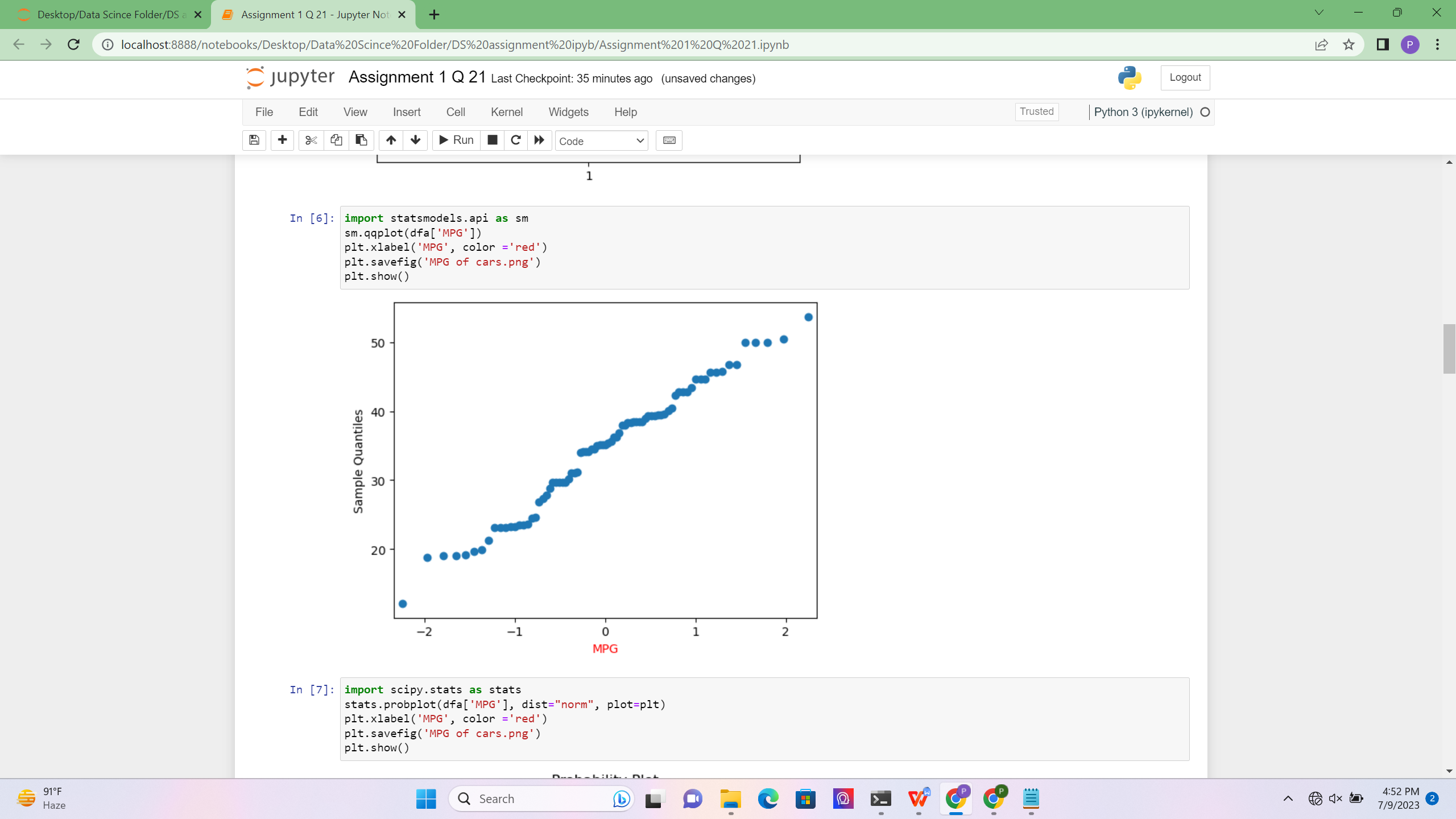
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

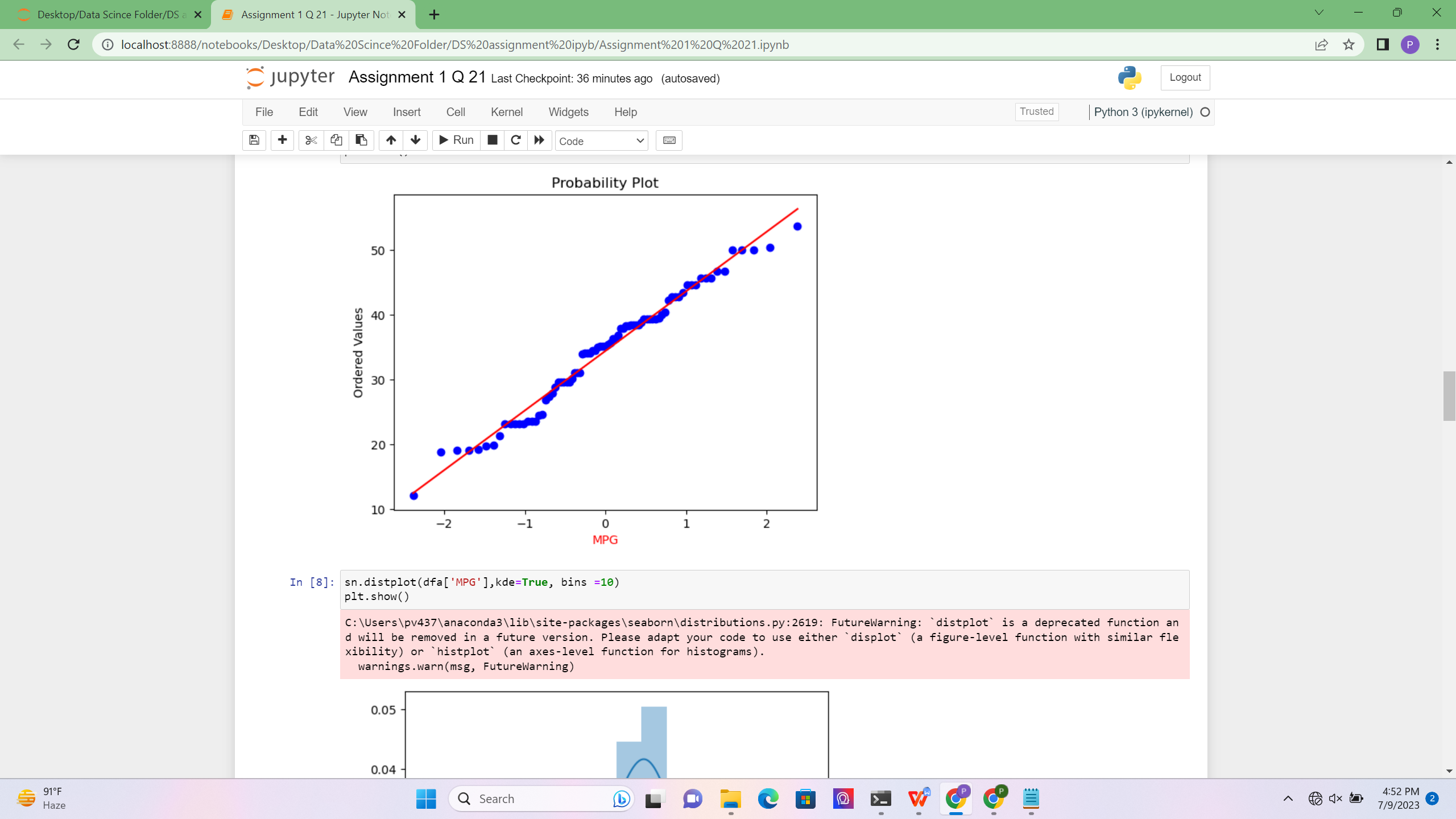
Dataset: wc-at.csv

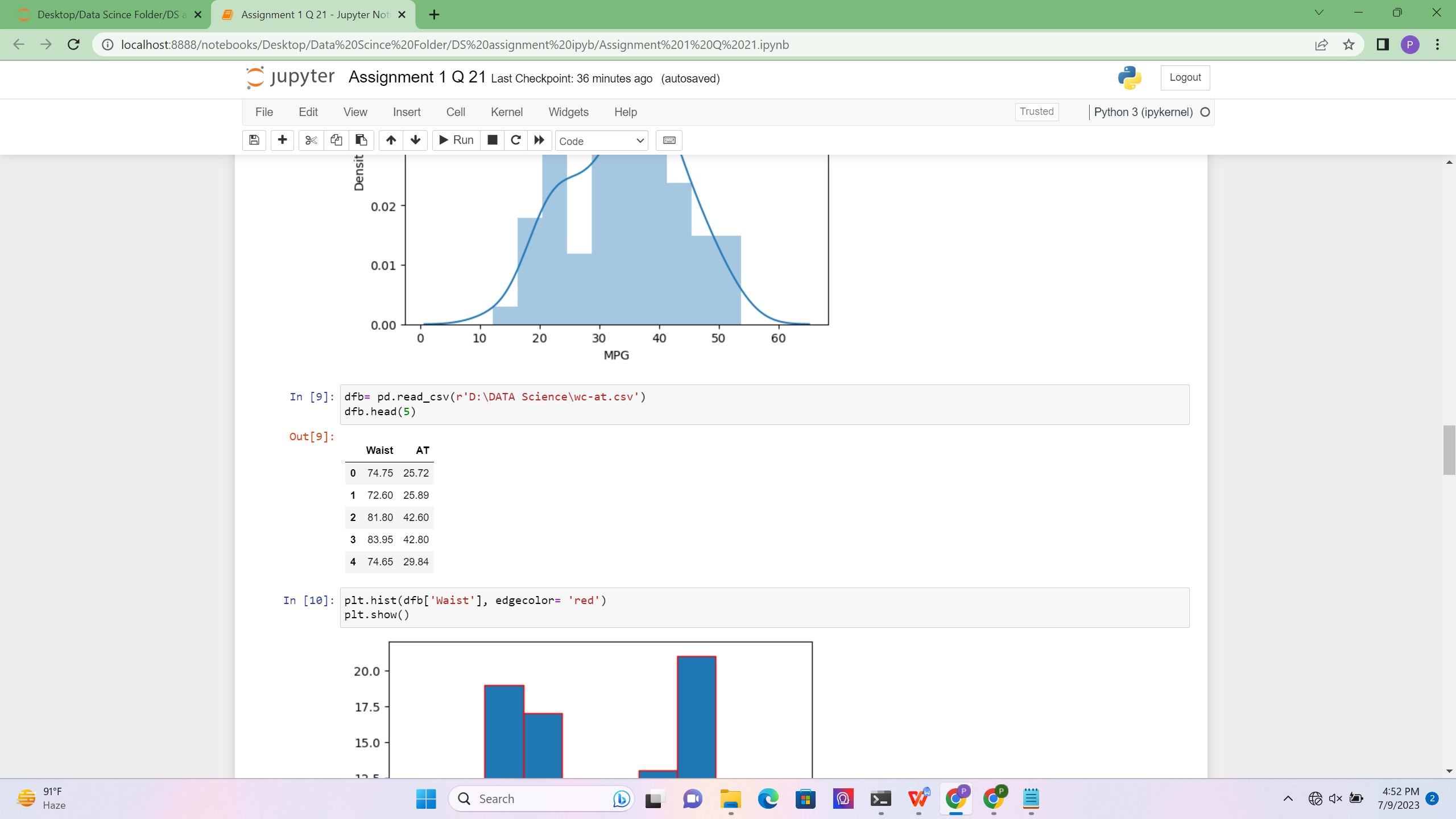


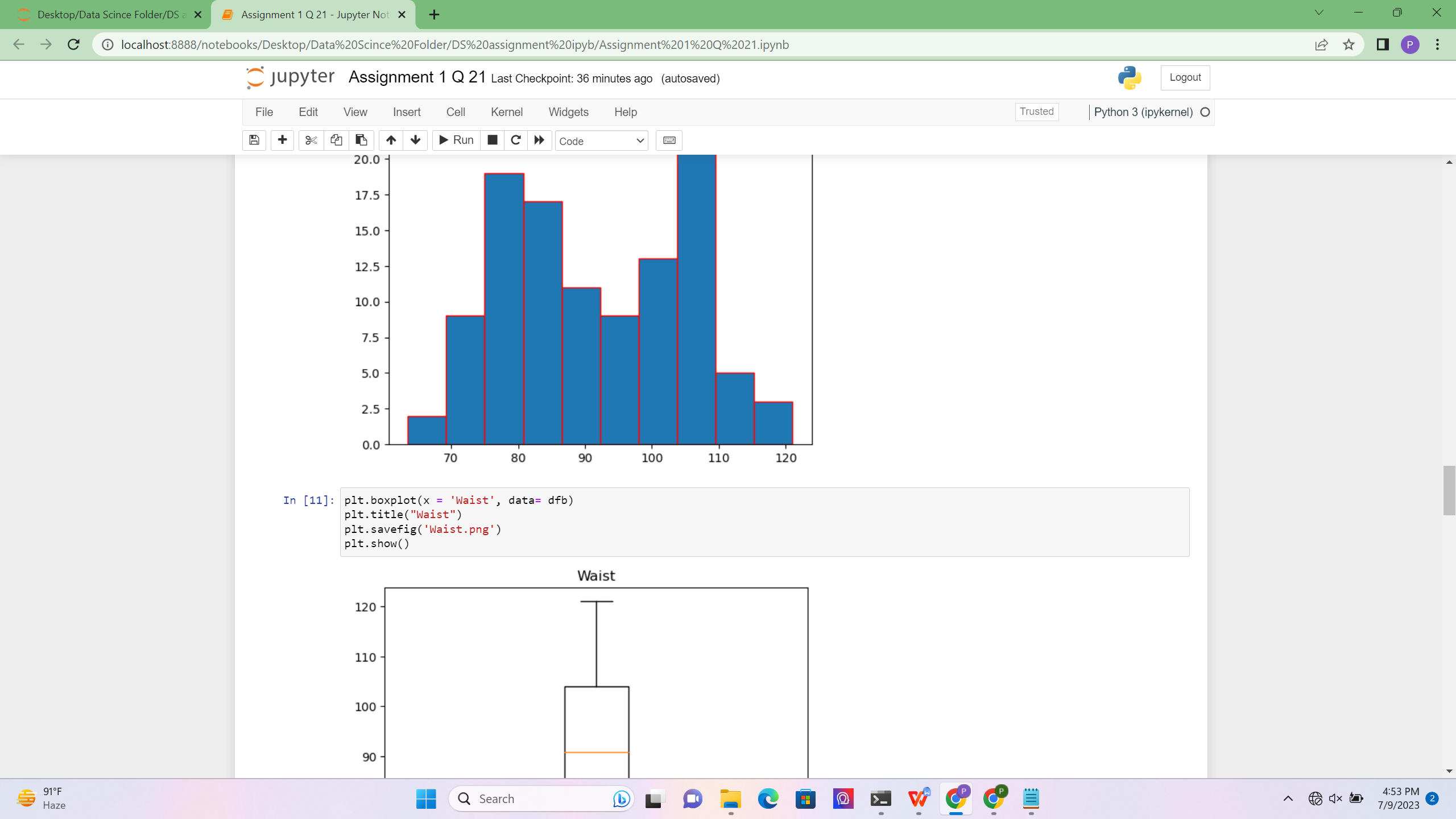


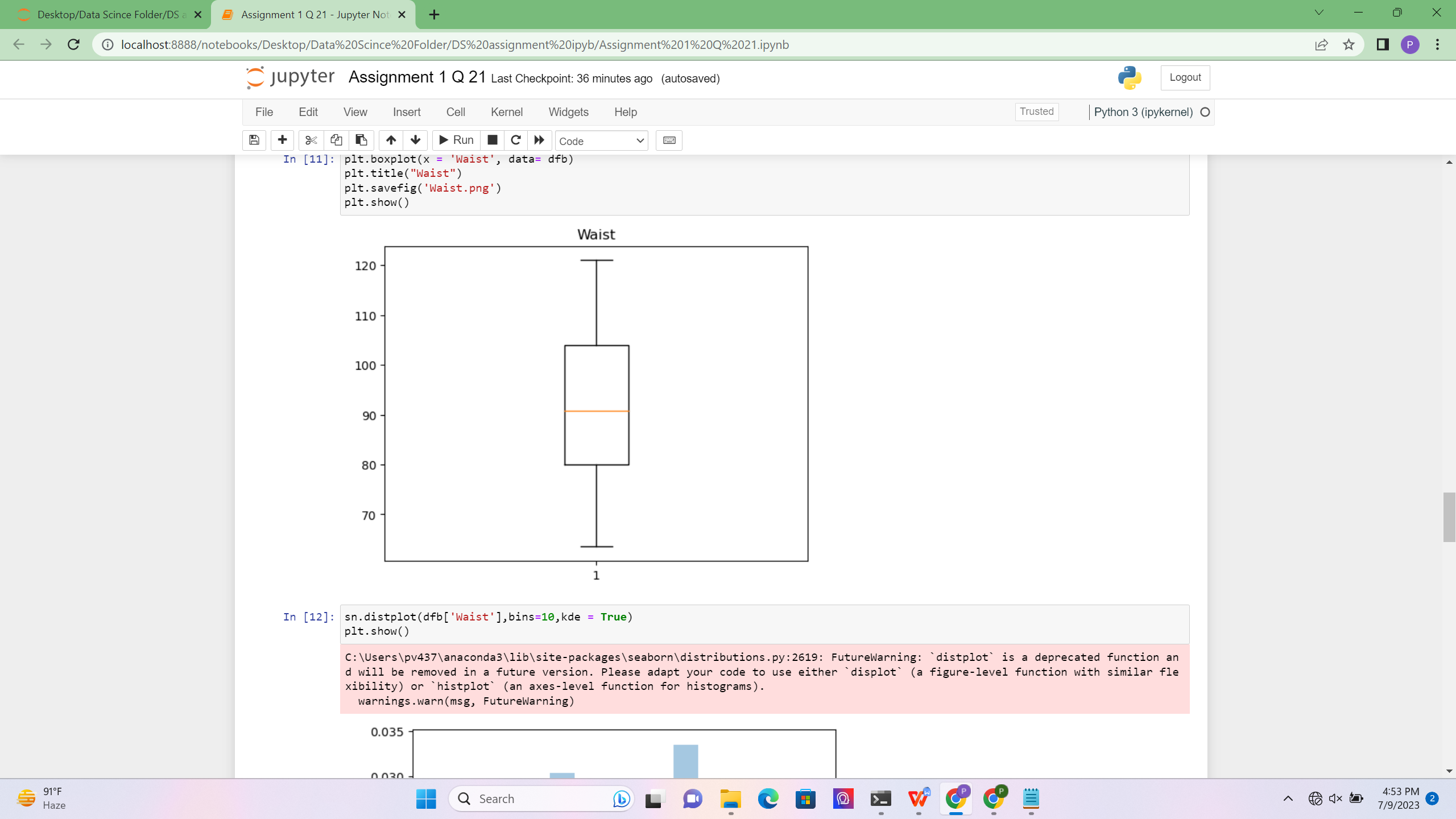


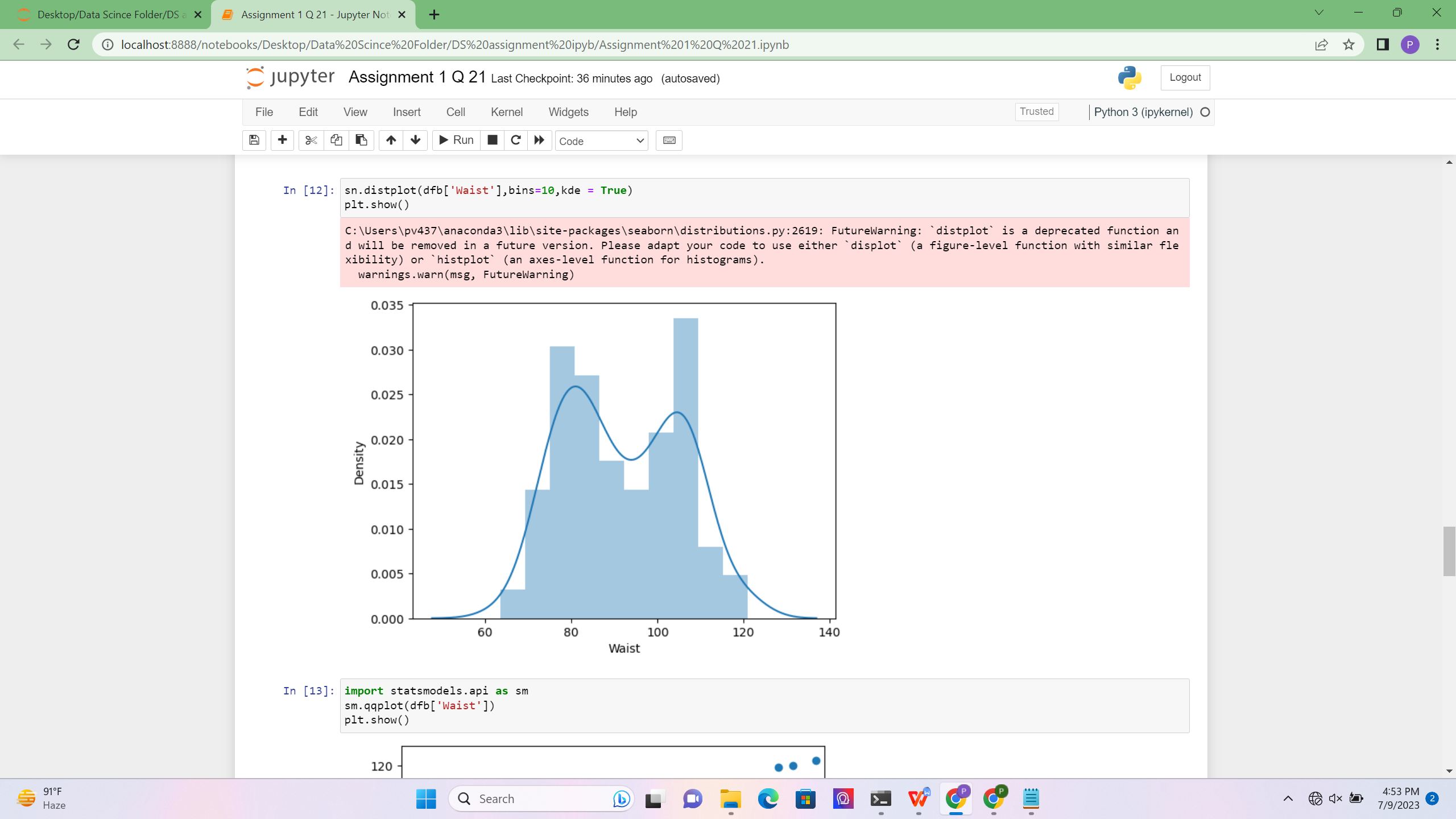


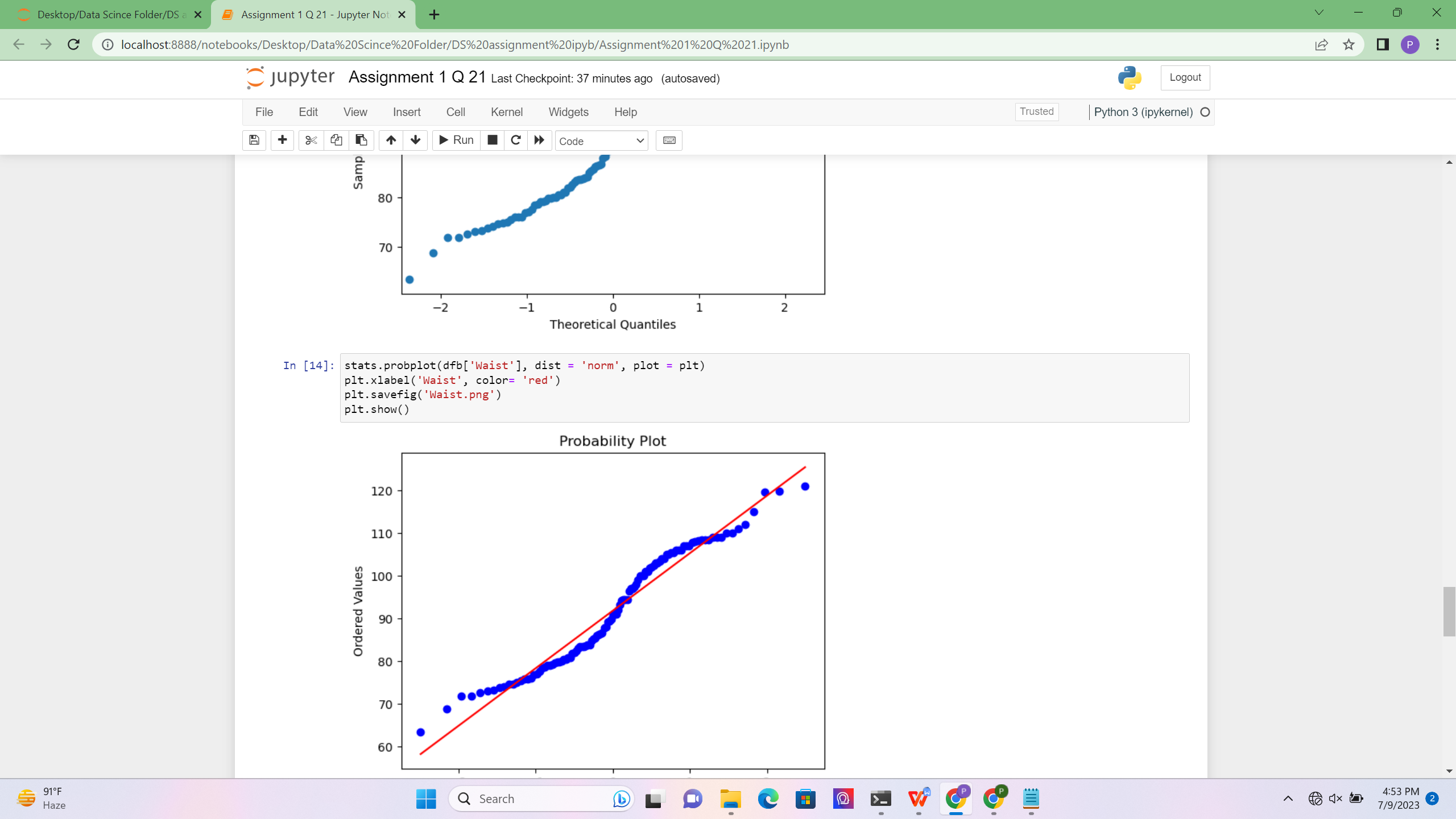


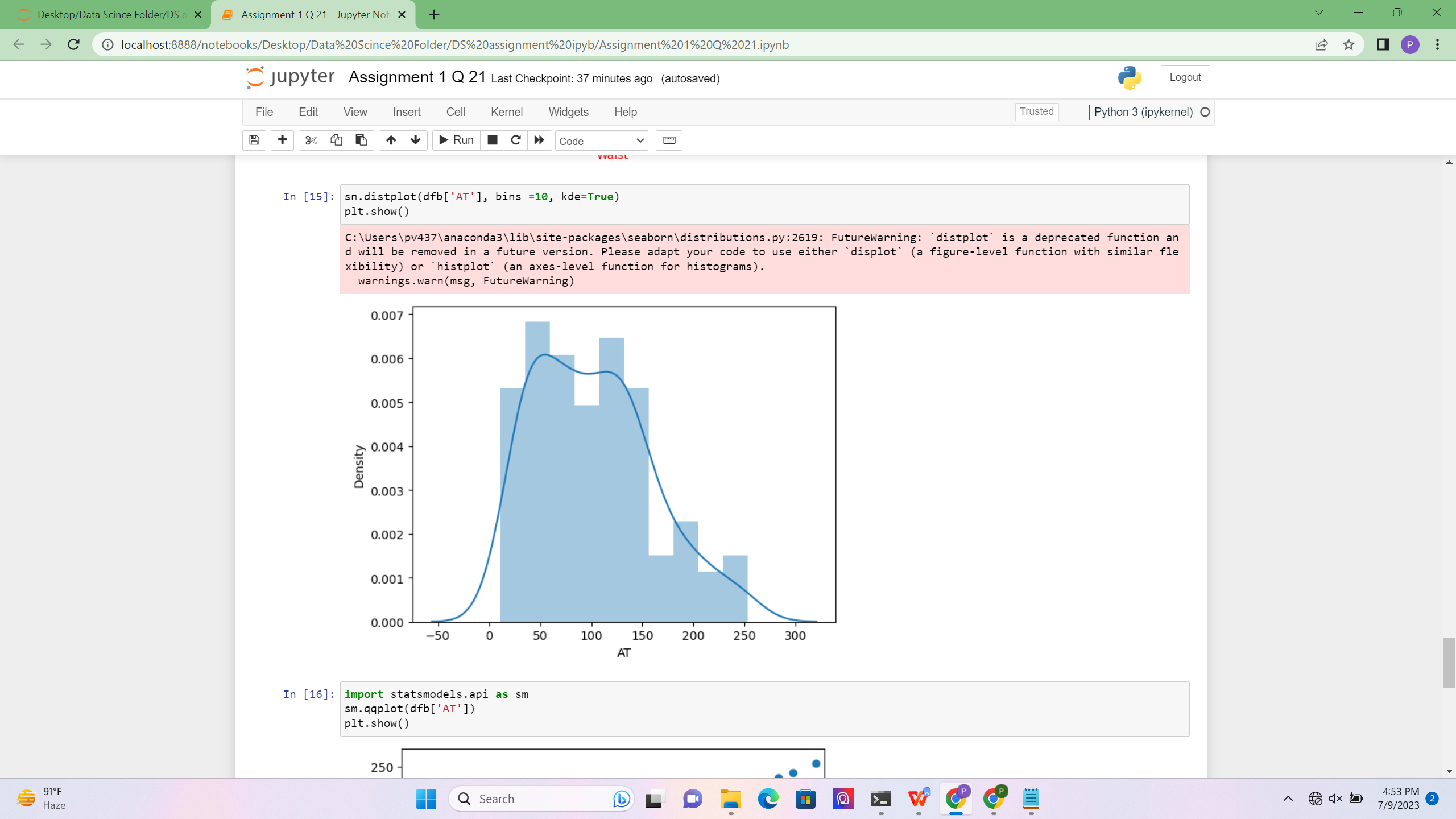


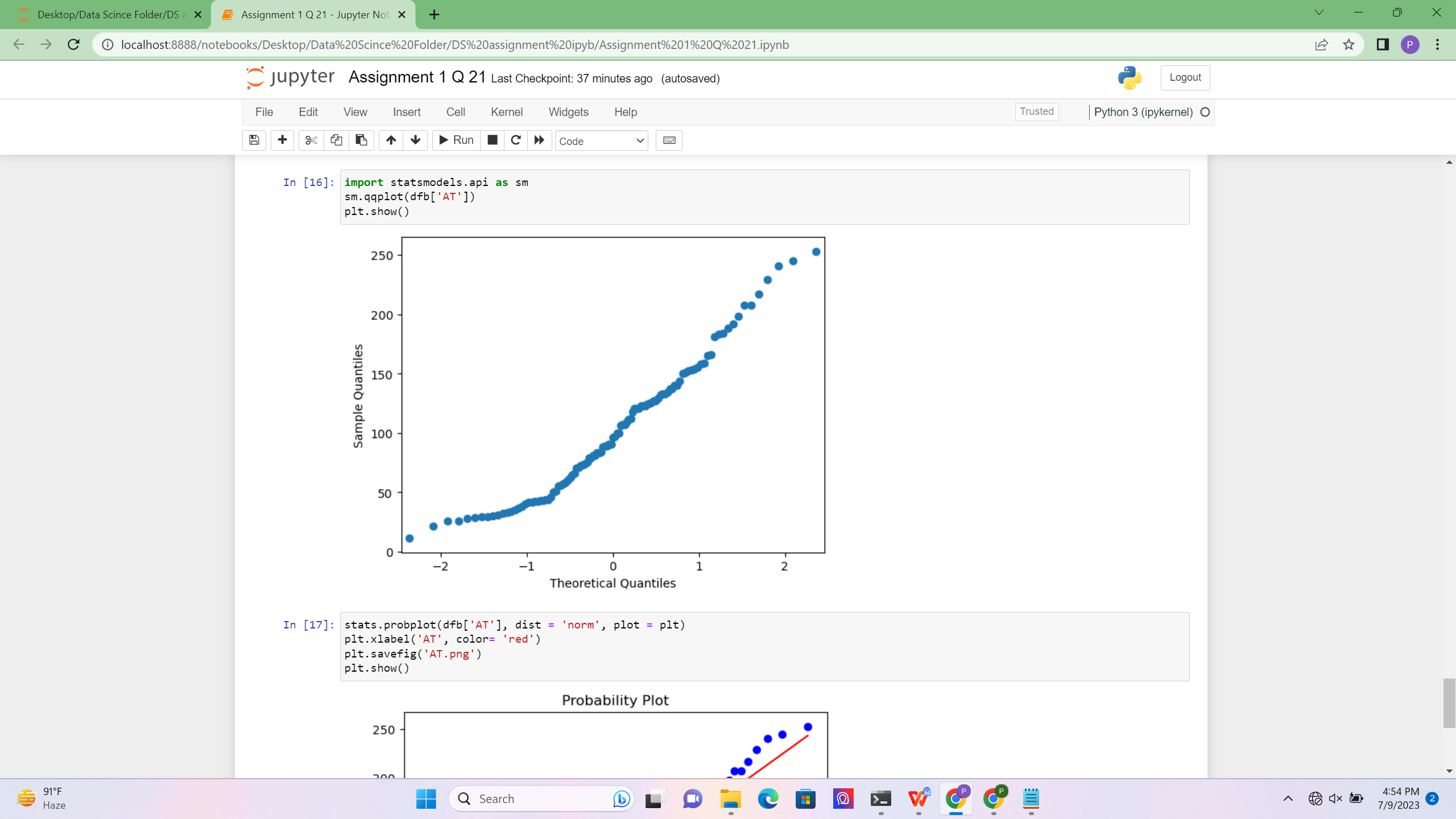


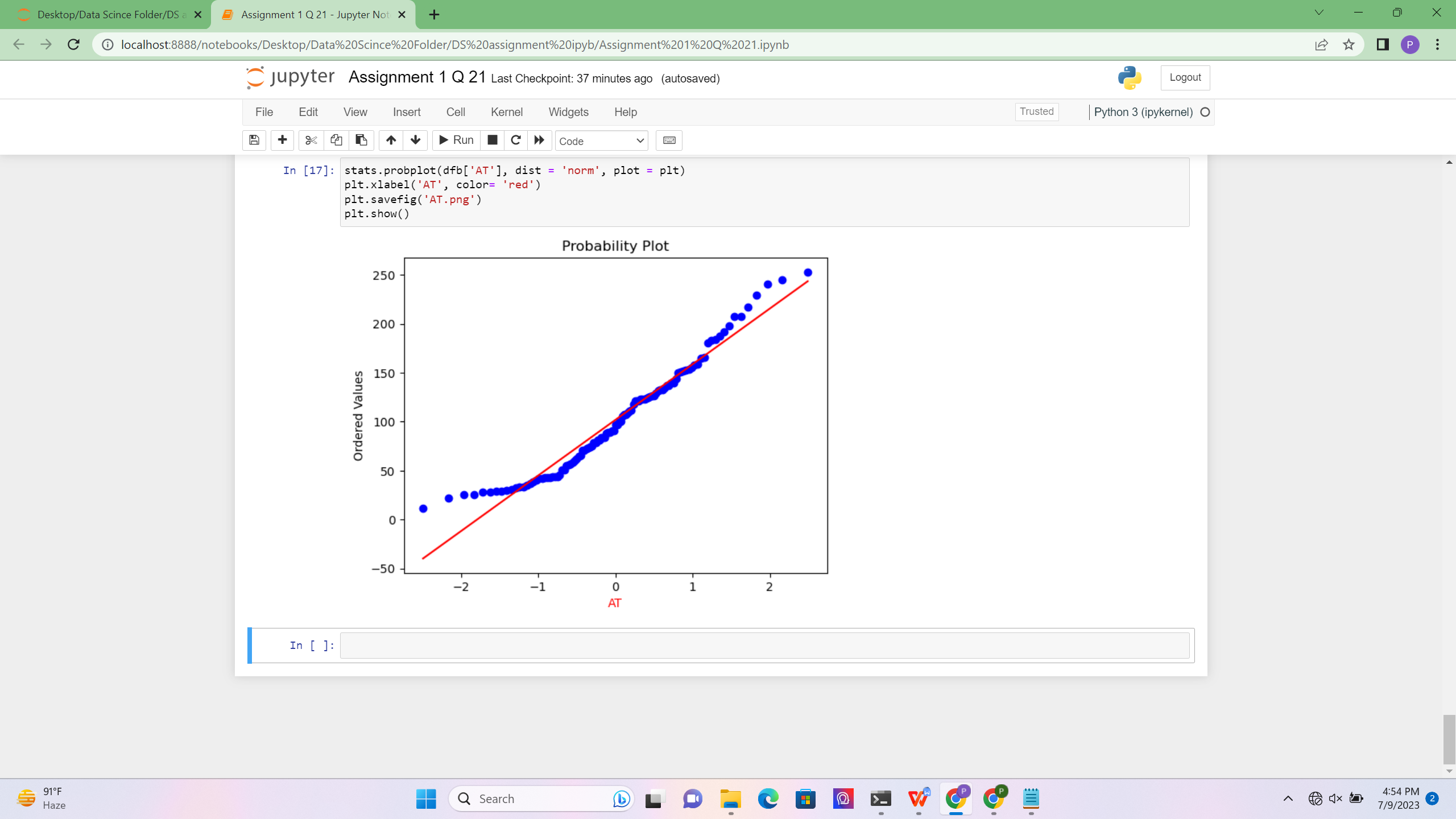












Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

ANS :-

To calculate the Z scores for different confidence intervals, we need to refer to the standard normal distribution table. The Z score corresponds to the number of standard deviations away from the mean. Here are the Z scores for the given confidence intervals:

90% Confidence Interval:

For a 90% confidence interval, the area under the standard normal distribution curve is divided equally into two tails, leaving 5% in each tail. Thus, we need to find the Z score that corresponds to an area of 0.05 in the tails.

Looking up the Z score for a cumulative area of 0.05 in the standard normal distribution table, we find that the Z score is approximately -1.645.

94% Confidence Interval:

For a 94% confidence interval, the area under the standard normal distribution curve is divided equally into two tails, leaving 3% in each tail. Thus, we need to find the Z score that corresponds to an area of 0.03 in the tails.

Looking up the Z score for a cumulative area of 0.03 in the standard normal distribution table, we find that the Z score is approximately -1.881.

60% Confidence Interval:

For a 60% confidence interval, the area under the standard normal distribution curve is divided equally into two tails, leaving 20% in each tail. Thus, we need to find the Z score that corresponds to an area of 0.20 in the tails.

Looking up the Z score for a cumulative area of 0.20 in the standard normal distribution table, we find that the Z score is approximately -0.8416.

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

ANS :-

To calculate the t-scores for different confidence intervals with a sample size of 25, we need to consider the degrees of freedom (df) associated with the sample. For a sample size of 25, the degrees of freedom would be 25 - 1 = 24. Here are the t-scores for the given confidence intervals:

95% Confidence Interval:

For a 95% confidence interval with 24 degrees of freedom, we can look up the t-score in the t-distribution table. The t-score corresponds to the critical value at which the cumulative probability is 0.025 (2.5% in each tail).

Looking up the t-score for 0.025 cumulative probability and 24 degrees of freedom, we find that the t-score is approximately 2.064.

96% Confidence Interval:

For a 96% confidence interval with 24 degrees of freedom, we need to find the t-score at which the cumulative probability is 0.02 (2% in each tail).

Looking up the t-score for 0.02 cumulative probability and 24 degrees of freedom, we find that the t-score is approximately 2.171.

99% Confidence Interval:

For a 99% confidence interval with 24 degrees of freedom, we need to find the t-score at which the cumulative probability is 0.005 (0.5% in each tail).

Looking up the t-score for 0.005 cumulative probability and 24 degrees of freedom, we find that the t-score is approximately 2.797.

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 :-

Population mean (CEO's claim): μ = 270 days

Sample mean: x̄ = 260 days

Sample standard deviation: s = 90 days

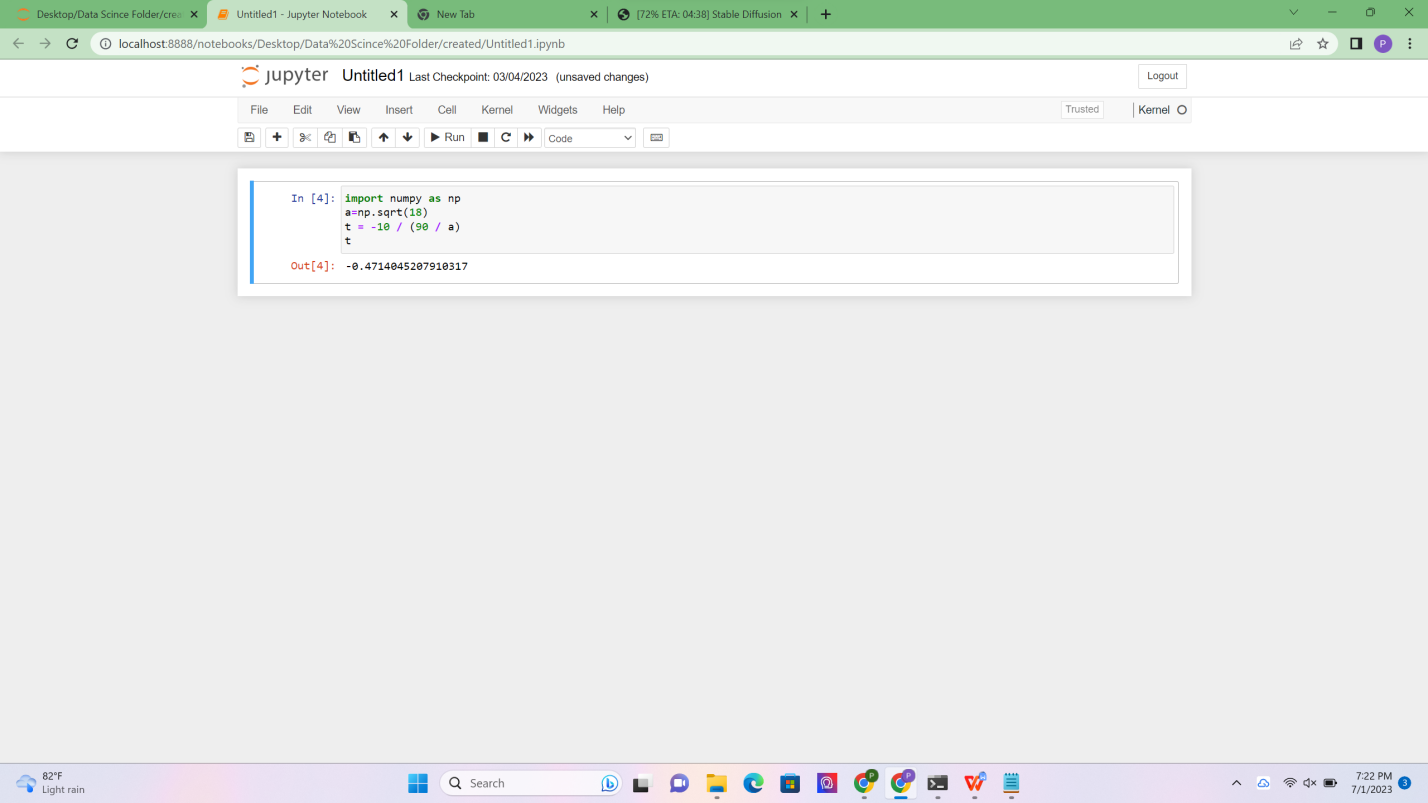
Sample size: n = 18 bulbs

First, we calculate the t-score using the formula:

t = (x̄ - μ) / (s / √n)

t = (260 - 270) / (90 / √18)

t = -10 / (90 / √18)



t=-0.4714045207910317

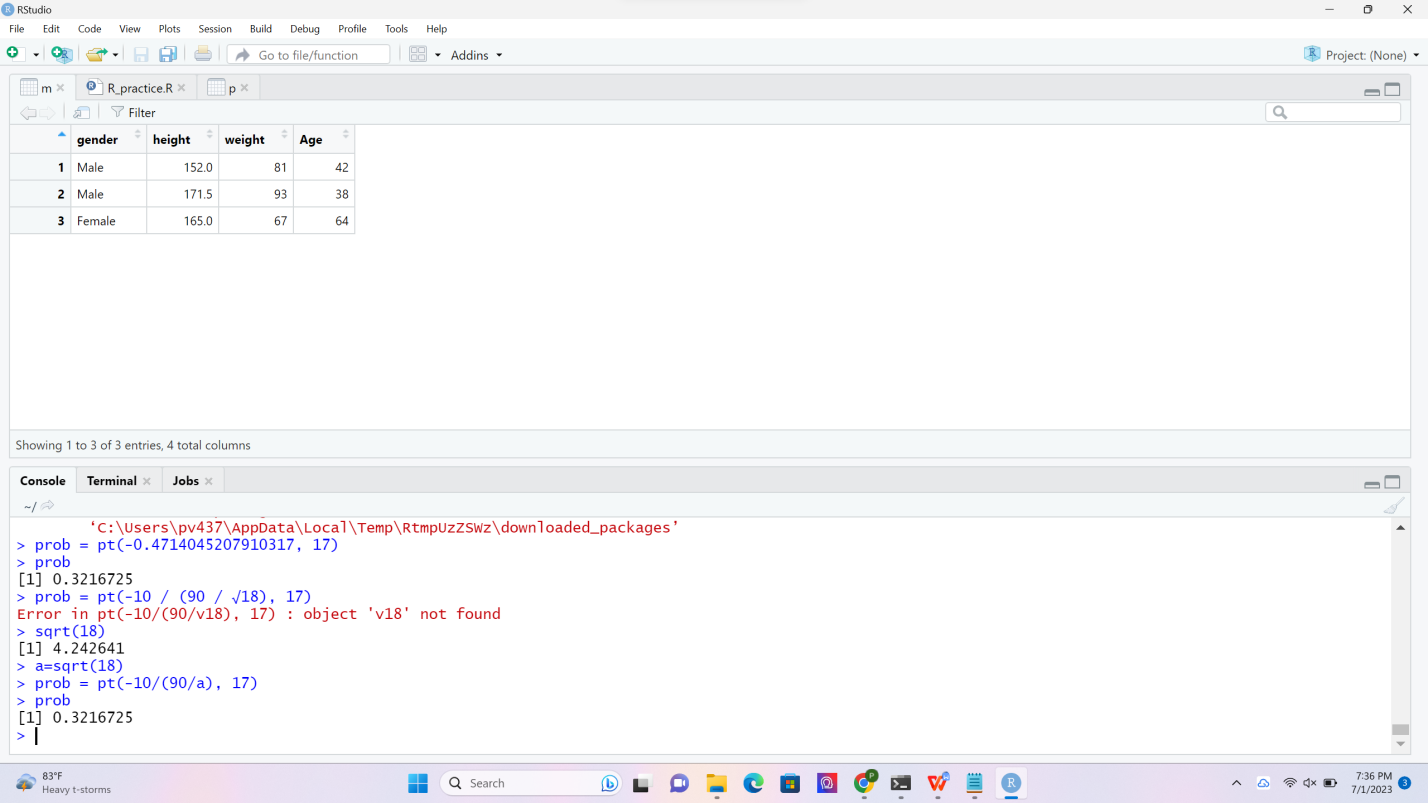
Next, we need to determine the degrees of freedom for the t-distribution. For a sample size of 18, the degrees of freedom is given by

df = n - 1 = 18 - 1 = 17.

Using the R code pt(tscore, df) to calculate the probability, we have:

prob = pt(t, df)

prob = pt(-0.4714045207910317, 17)



Evaluating this expression using R we find that the probability is approximately 0.3216725, or 3.27%.

Therefore, if the CEO's claim were true, there is a 3.27% probability that 18 randomly selected bulbs would have an average life of no more than 260 days.