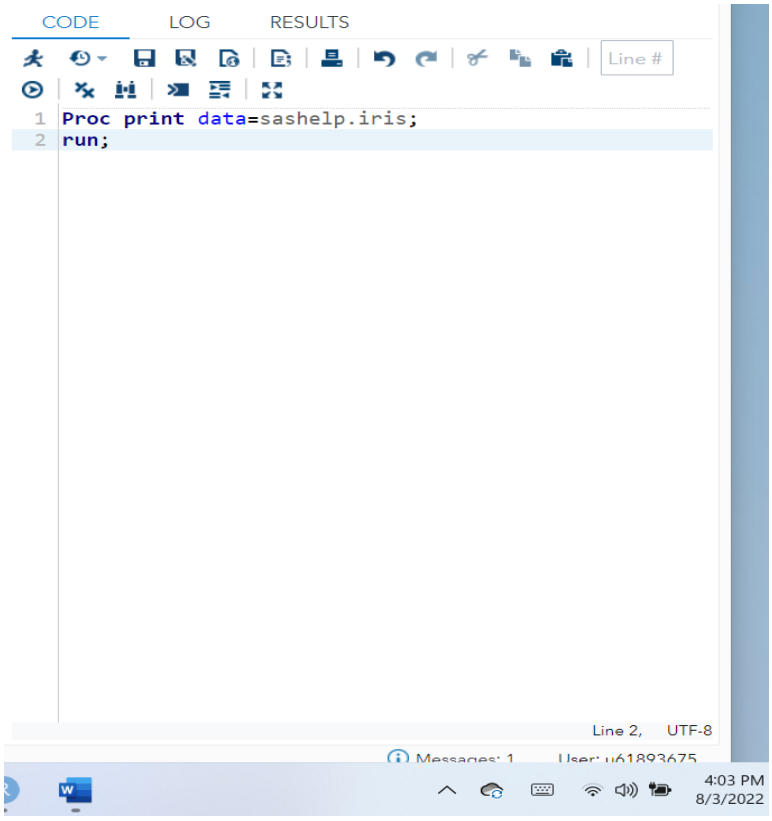
Explore the Iris Data Set Using SAS Studio

Didem Bulut Aykurt
MIS500-1 – Foundations of Data Analytics
Colorado State University-Global Campus

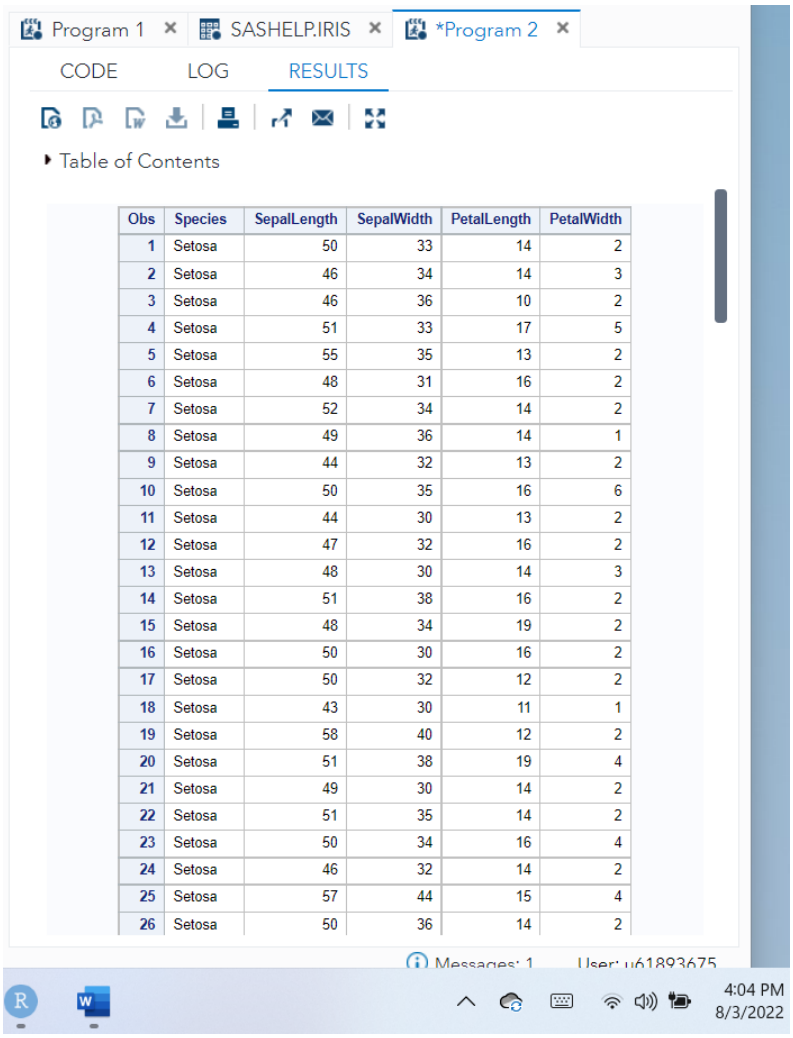
Dr. Steve Chung
August 6, 2022

Mod 3 – Critical Thinking; Option 1

SAS Studio is one of the data analysis tools. What I like about the SAS Studio is the task option. This is a great tool to create an inferential, descriptive, and predictive statistic with just a click without type code. Let's deep dive into SAS Studio with the Iris data set. The Iris data set is a multivariate data set with five variables and 150 rows by the British statistician in 1936. Print the variable with SAS syntax.



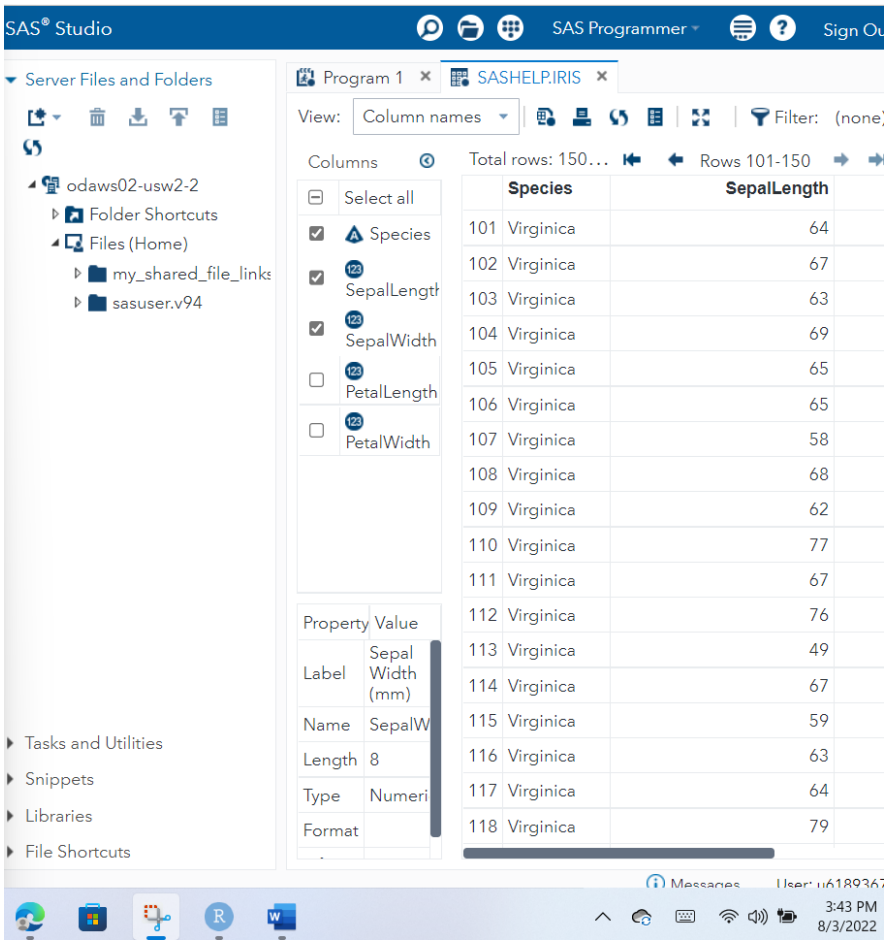
Mod 3 – Critical Thinking; Option 1



Obs	Species	SepalLength	SepalWidth	PetalLength	PetalWidth
1	Setosa	50	33	14	2
2	Setosa	46	34	14	3
3	Setosa	46	36	10	2
4	Setosa	51	33	17	5
5	Setosa	55	35	13	2
6	Setosa	48	31	16	2
7	Setosa	52	34	14	2
8	Setosa	49	36	14	1
9	Setosa	44	32	13	2
10	Setosa	50	35	16	6
11	Setosa	44	30	13	2
12	Setosa	47	32	16	2
13	Setosa	48	30	14	3
14	Setosa	51	38	16	2
15	Setosa	48	34	19	2
16	Setosa	50	30	16	2
17	Setosa	50	32	12	2
18	Setosa	43	30	11	1
19	Setosa	58	40	12	2
20	Setosa	51	38	19	4
21	Setosa	49	30	14	2
22	Setosa	51	35	14	2
23	Setosa	50	34	16	4
24	Setosa	46	32	14	2
25	Setosa	57	44	15	4
26	Setosa	50	36	14	2

SAS also has tools to see tables without complex code, such as dragging the dataset from the left side of My Library to drop the coding side and display the table. This way can filter variables to display.

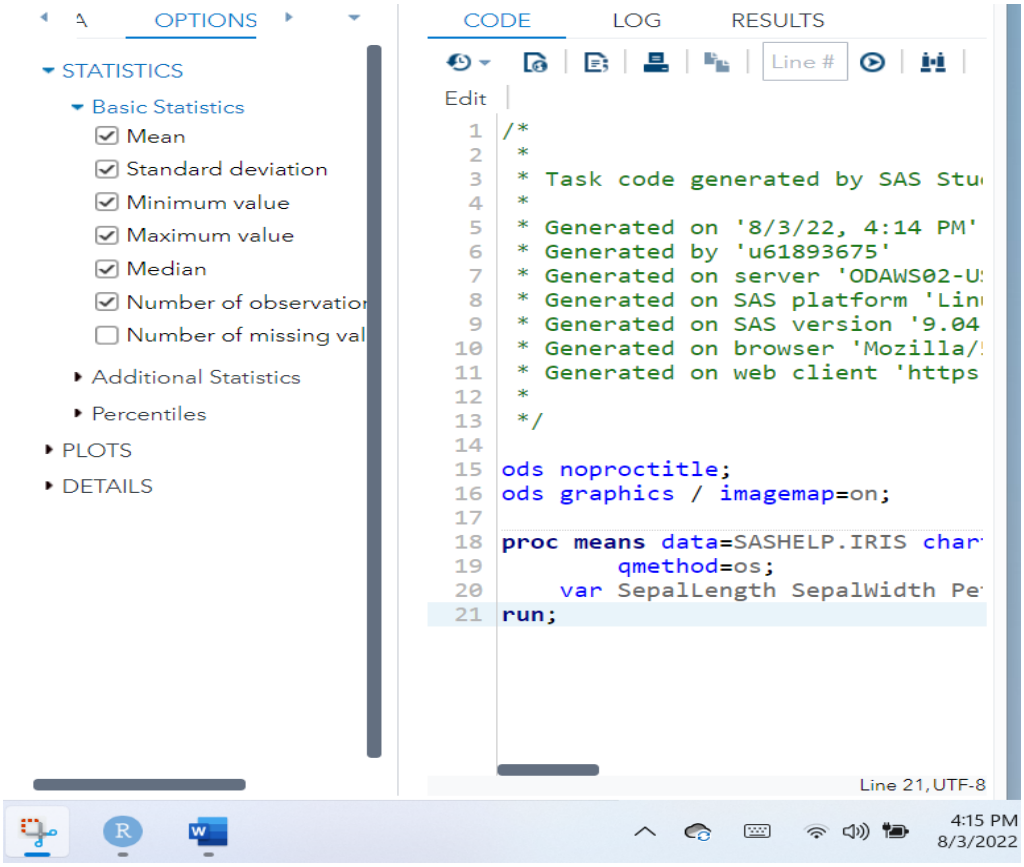
Mod 3 – Critical Thinking; Option 1

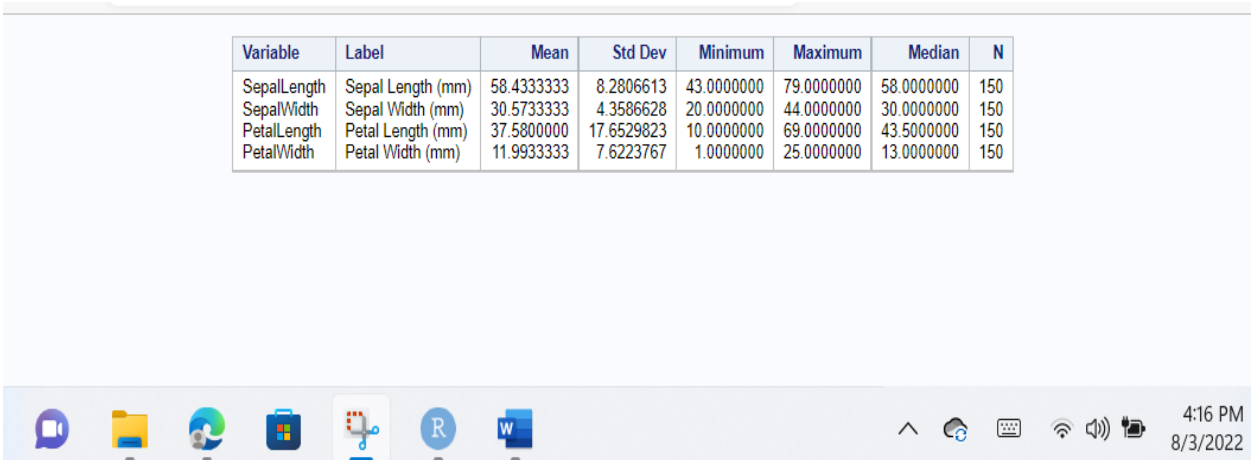


Summary statistics all variables of the iris data set

-Left on the page has the link Task and utilities->Statistic->Summary Statistics. It will open a new Summary Statistic page. Choose the dataset from the first Data section and click the fitting corner table symbol. Then add the variable to the analysis-on-Analysis variable box right corner has + to add variables. Then run the code.

Mod 3 – Critical Thinking; Option 1





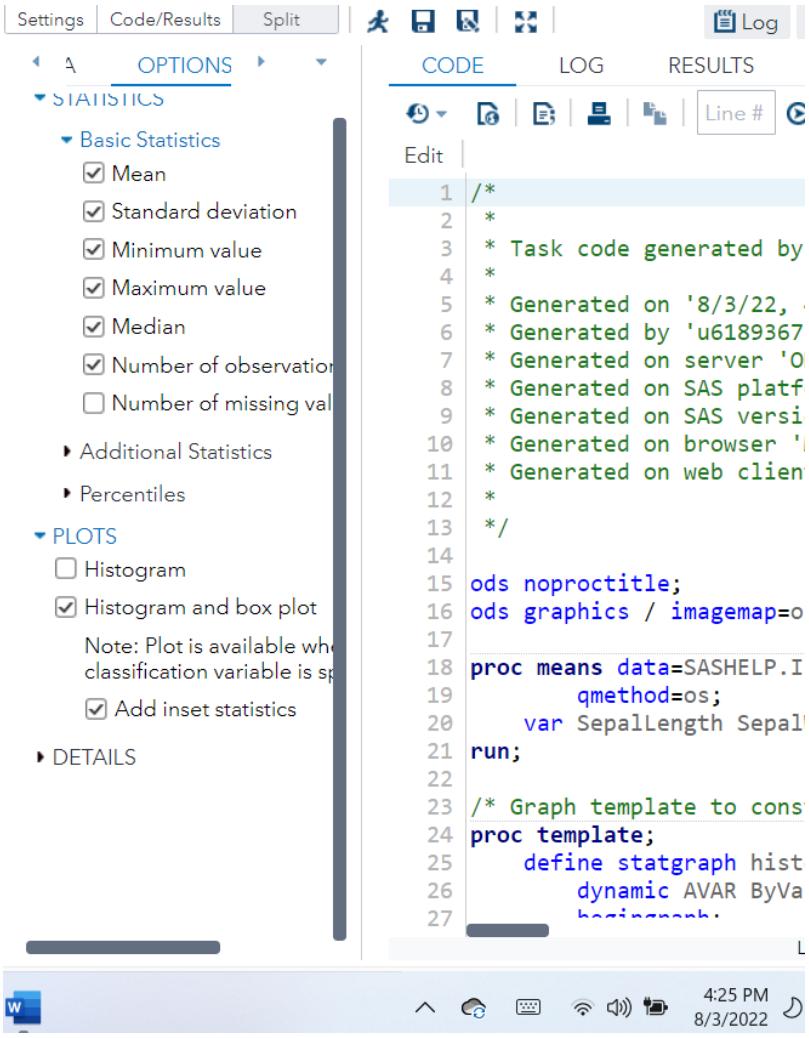
Variable	Label	Mean	Std Dev	Minimum	Maximum	Median	N
SepalLength	Sepal Length (mm)	58.4333333	8.2806613	43.0000000	79.0000000	58.0000000	150
SepalWidth	Sepal Width (mm)	30.5733333	4.3586628	20.0000000	44.0000000	30.0000000	150
PetalLength	Petal Length (mm)	37.5800000	17.6529823	10.0000000	69.0000000	43.5000000	150
PetalWidth	Petal Width (mm)	11.9933333	7.6223767	1.0000000	25.0000000	13.0000000	150

SepalLength and SepalWidth"s mean close to their median.

PetalLength has the highest standard deviation is 18, which means the mean is far from the data point. SepalWidth's standard

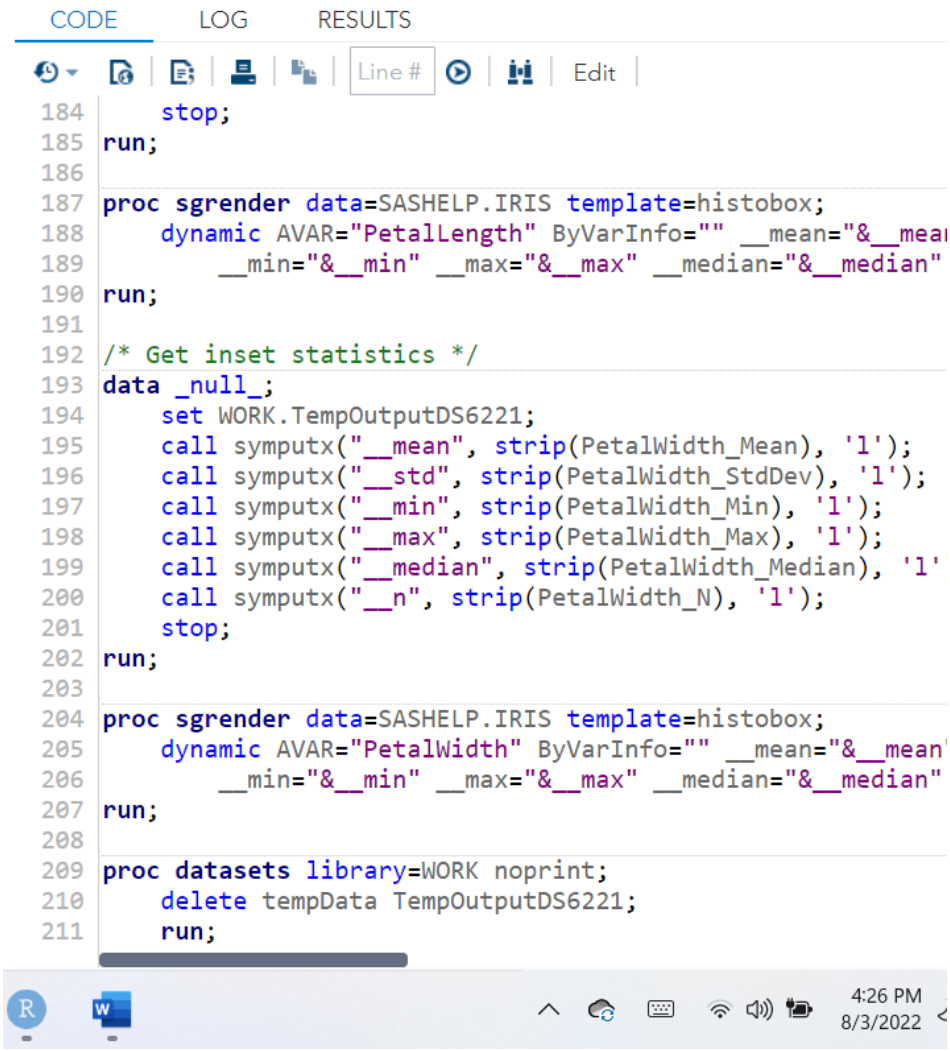
deviation is the lowest (4.3), representing the data point close to the mean.

The next step is to create histogram and box plots for all variables. The exact page on Summary Statistics does the same thing add dataset, then add variable. The top on the ribbon side has DATA – OPTION, so on click OPTION, the middle side has a list choose the Plots - select histogram or histogram box plot.



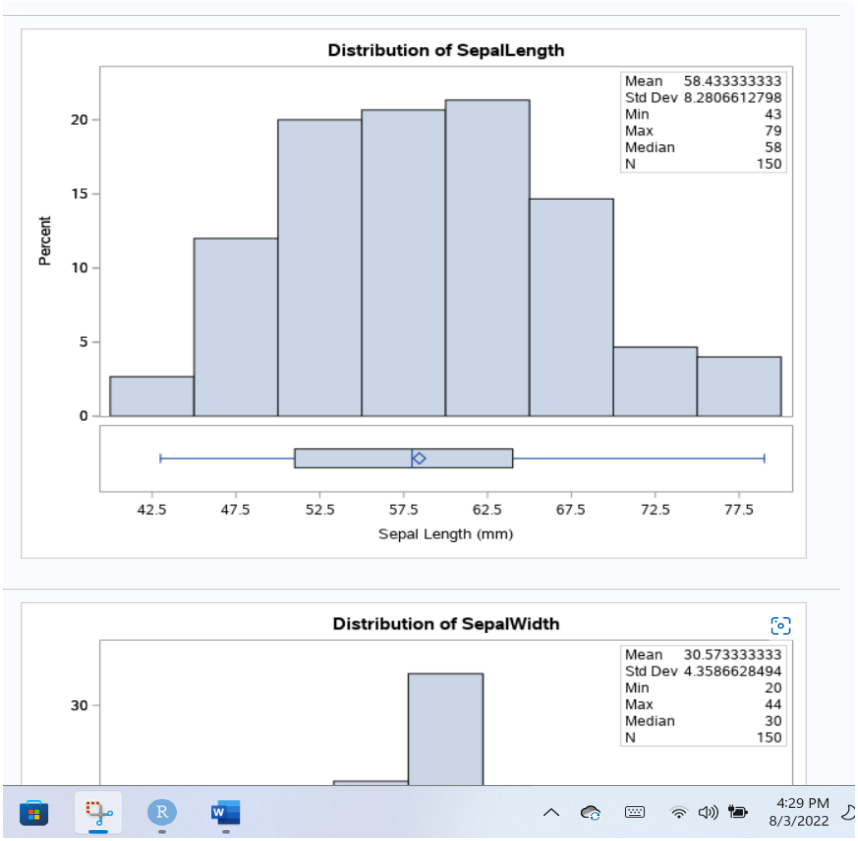
Mod 3 – Critical Thinking; Option 1

This is the code page SAS's TASK function automatically creates all code. It's so cool.

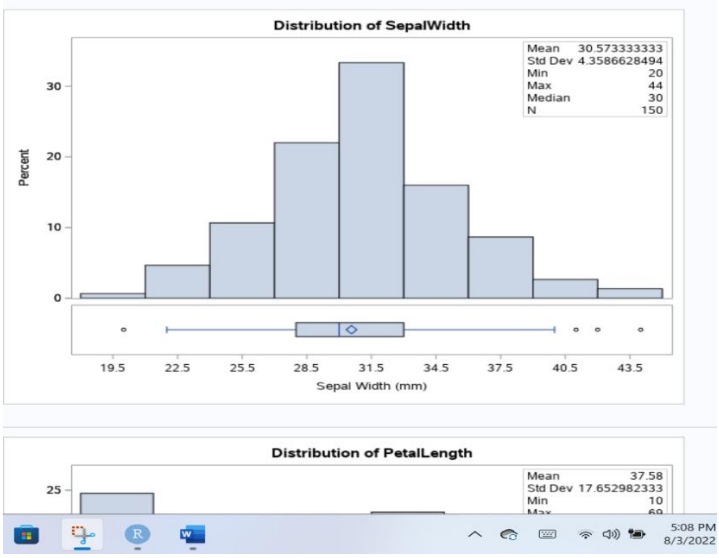


The RESULT page shows all graphs and charts. The first chart is the distribution of SepalLength. μ=58.4, median=58, and σ=8.2 thus, data points are close to each other, and the SepalLength data set is qualified for the conviction of statistical tests. The boxplot of the graft shows the mean and median so close to each other that Q2

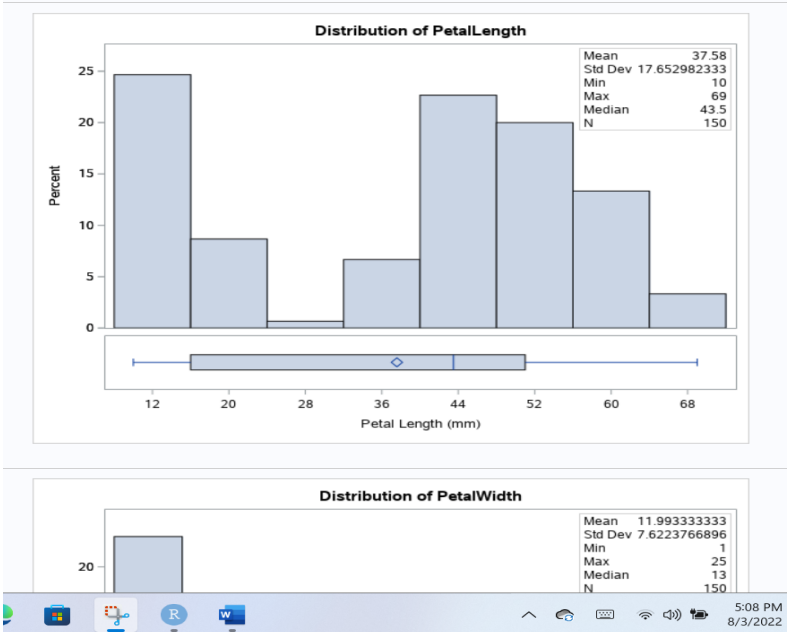
and Q3 are close in size. The first quartile is smaller than the fourth, which tells us the most variable from the right side.



Text histogram is the distribution of SepalWidth; this means it is lower than the sepal length's mean and standard deviation. Thus, the sepal length probability is higher than the sepal width. And a histogram is like a normal distribution. The boxplot of SepalWidth shows nonsymmetric because the median is 30 and the mean is 30.59 as Q2's size is smaller than close to Q3, so the distribution is negatively skewed.



PetalLength is one of the bimodal distributions of the histogram with a high standard deviation compared to the other variables, which is why the cure is wider than the others and shorter. It has two peaks. The data should be apart and analyzed. The boxplot tells the median is higher than the mean as the distribution is left-skewed. Q2 size is more extended than the Q3 size.



The last variable is PetalWidth has a lower mean and also lower probability. Thus, variables Mu =12 and sigma = 8 have a distribution. The boxplot shows the data set has been negatively skewed because the mean is lower than the median.