

Explore the Iris Data Set Using SAS Studio

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

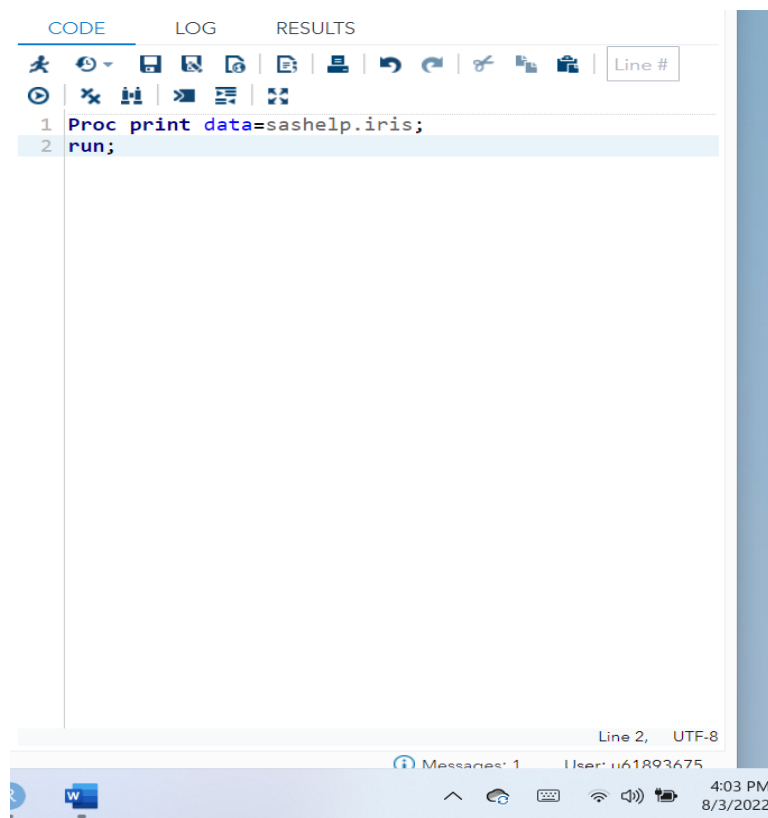
Colorado State University-Global Campus

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.



The screenshot displays the SAS Studio web interface. At the top, there are three tabs: 'CODE', 'LOG', and 'RESULTS'. Below the tabs is a toolbar with various icons for file operations and execution. The 'CODE' tab is active, showing a code editor with two lines of SAS code: '1 Proc print data=sashelp.iris;' and '2 run;'. The code is highlighted in blue. The status bar at the bottom of the code editor indicates 'Line 2, UTF-8'. The Windows taskbar is visible at the bottom of the screen, showing the time as 4:03 PM on 8/3/2022.

```
1 Proc print data=sashelp.iris;
2 run;
```

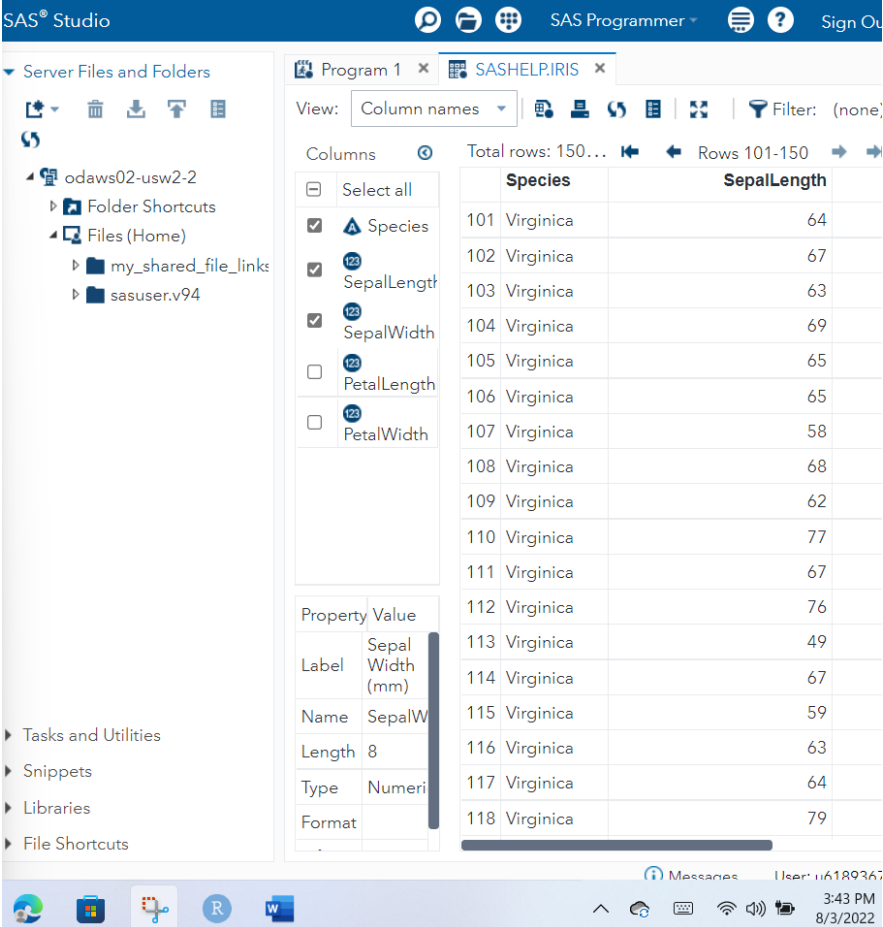
Mod 3 – Critical Thinking; Option 1

The screenshot shows the SAS Studio interface with three tabs: Program 1, SASHELP.IRIS, and *Program 2. The RESULTS tab is active, displaying a table of data. Above the table is a 'Table of Contents' section. The table has six columns: Obs, Species, SepalLength, SepalWidth, PetalLength, and PetalWidth. It contains 26 rows of data for the Setosa species. The bottom of the window shows a taskbar with icons for R, Word, and system status, along with the time 4:04 PM on 8/3/2022.

| 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



The screenshot displays the SAS Studio interface. On the left, the 'Server Files and Folders' pane shows a tree structure with 'odaws02-usw2-2' selected. The main workspace shows the 'SASHELP.IRIS' dataset. The 'Columns' pane lists the variables: Species, SepalLength, SepalWidth, PetalLength, and PetalWidth. The 'Property Value' pane shows the properties for the 'SepalWidth' variable: Label is 'Sepal Width (mm)', Name is 'SepalW', Length is 8, Type is Numerical, and Format is .

| Species | SepalLength |
|-----------|-------------|
| Virginica | 64 |
| Virginica | 67 |
| Virginica | 63 |
| Virginica | 69 |
| Virginica | 65 |
| Virginica | 65 |
| Virginica | 58 |
| Virginica | 68 |
| Virginica | 62 |
| Virginica | 77 |
| Virginica | 67 |
| Virginica | 76 |
| Virginica | 49 |
| Virginica | 67 |
| Virginica | 59 |
| Virginica | 63 |
| Virginica | 64 |
| Virginica | 79 |

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

The screenshot shows the SAS Studio interface. On the left, the 'OPTIONS' pane is expanded to 'STATISTICS', with 'Basic Statistics' checked. The 'CODE' tab is active, displaying the following SAS code:

```

1  /*
2  *
3  * Task code generated by SAS Studio
4  *
5  * Generated on '8/3/22, 4:14 PM'
6  * Generated by 'u61893675'
7  * Generated on server 'ODAWS02-UI'
8  * Generated on SAS platform 'Linux'
9  * Generated on SAS version '9.04'
10 * Generated on browser 'Mozilla/5.0'
11 * Generated on web client 'https://sasstudio.us.sas.com/'
12 *
13 */
14
15 ods noproctitle;
16 ods graphics / imagemap=on;
17
18 proc means data=SASHELP.IRIS char
19           qmethod=os;
20     var SepalLength SepalWidth PetalLength PetalWidth;
21 run;

```

The status bar at the bottom right indicates 'Line 21, UTF-8'.

| 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.

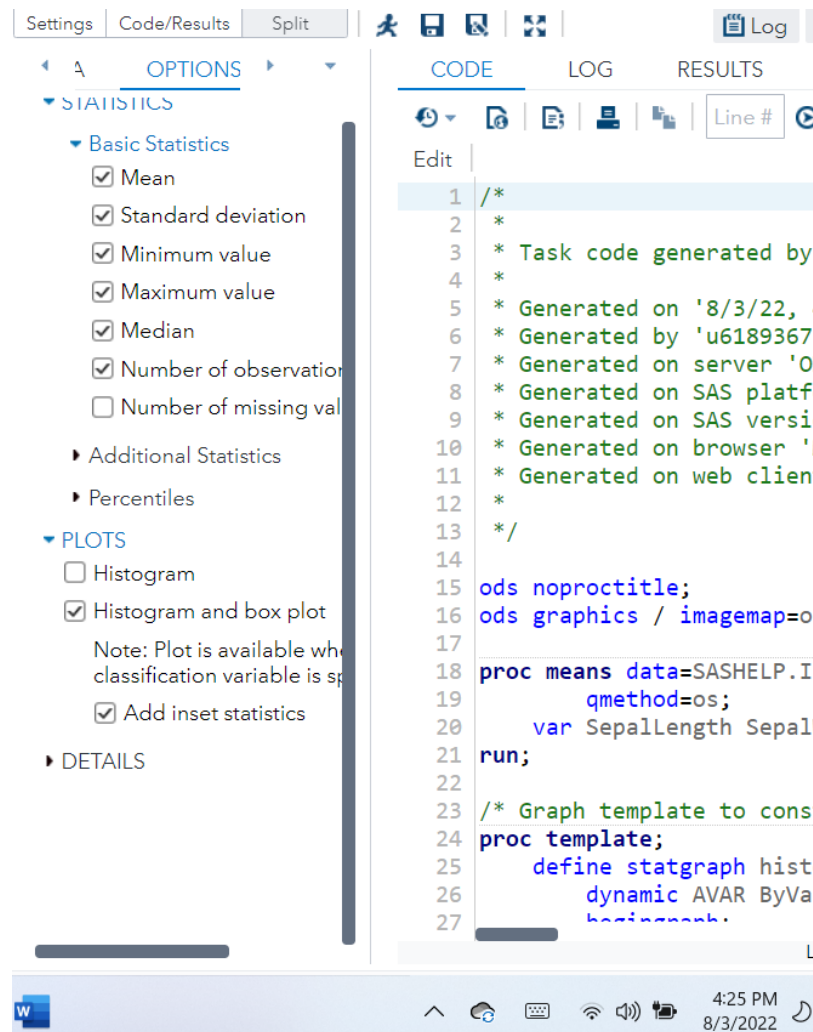
PetalLength has the highest standard deviation is 18, which means

the mean is far from the data point. SepalWidth's standard

Mod 3 – Critical Thinking; Option 1

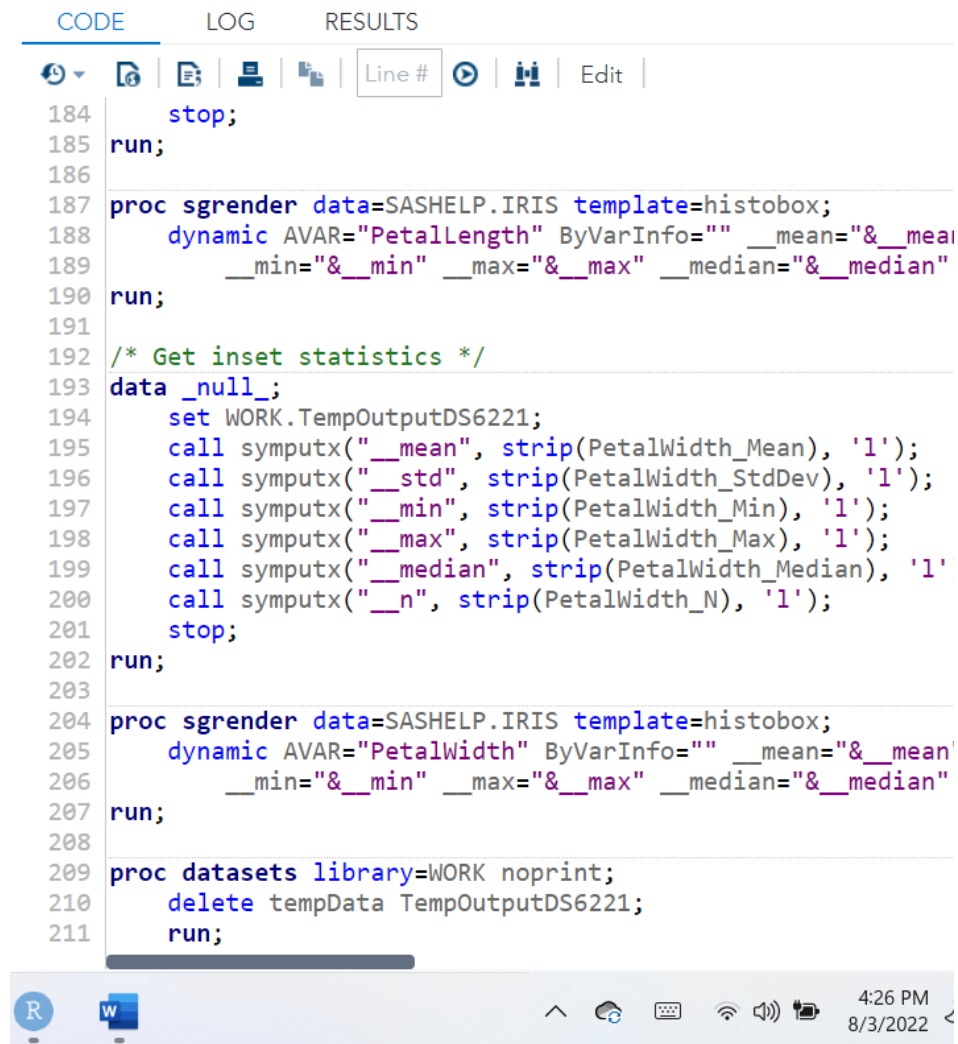
deviation is the lowest (4.3), representing the data point close to the mean.

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.



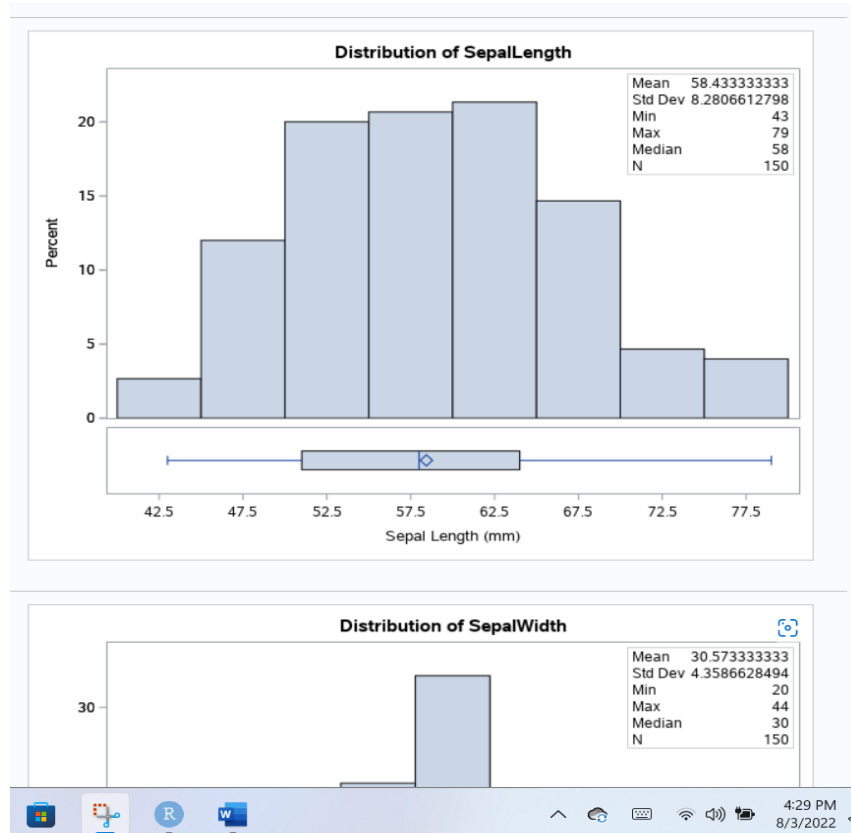
The screenshot shows the SAS Studio interface with three tabs: CODE, LOG, and RESULTS. The CODE tab is active, displaying a SAS program. The program includes a `stop;` statement at line 184, followed by a `run;` statement at line 185. At line 187, the `proc sgrender` statement is used to create a histogram for the `PetalLength` variable, with the `data=SASHELP.IRIS` and `template=histobox` options. The `dynamic` option is used to dynamically generate the `AVAR` variable, and the `ByVarInfo` option is used to generate the `__mean`, `__min`, `__max`, and `__median` variables. The `run;` statement is at line 190. At line 192, a comment `/* Get inset statistics */` is present. The `data _null_;` statement is at line 193, followed by a `set WORK.TempOutputDS6221;` statement at line 194. The `call symputx` macro is used to store the values of `__mean`, `__std`, `__min`, `__max`, `__median`, and `__n` into the `WORK.TempOutputDS6221` dataset. The `run;` statement is at line 202. At line 204, the `proc sgrender` statement is used to create a histogram for the `PetalWidth` variable, with the `data=SASHELP.IRIS` and `template=histobox` options. The `dynamic` option is used to dynamically generate the `AVAR` variable, and the `ByVarInfo` option is used to generate the `__mean`, `__min`, `__max`, and `__median` variables. The `run;` statement is at line 207. At line 209, the `proc datasets` statement is used to delete the `tempData` dataset, with the `library=WORK` and `noprint` options. The `delete` statement is at line 210, and the `run;` statement is at line 211. The bottom of the screenshot shows the Windows taskbar with the time 4:26 PM and date 8/3/2022.

```
184 stop;
185 run;
186
187 proc sgrender data=SASHELP.IRIS template=histobox;
188     dynamic AVAR="PetalLength" ByVarInfo="" __mean="__mean"
189         __min="__min" __max="__max" __median="__median"
190 run;
191
192 /* Get inset statistics */
193 data _null_;
194     set WORK.TempOutputDS6221;
195     call symputx("__mean", strip(PetalWidth_Mean), '1');
196     call symputx("__std", strip(PetalWidth_StdDev), '1');
197     call symputx("__min", strip(PetalWidth_Min), '1');
198     call symputx("__max", strip(PetalWidth_Max), '1');
199     call symputx("__median", strip(PetalWidth_Median), '1');
200     call symputx("__n", strip(PetalWidth_N), '1');
201 stop;
202 run;
203
204 proc sgrender data=SASHELP.IRIS template=histobox;
205     dynamic AVAR="PetalWidth" ByVarInfo="" __mean="__mean"
206         __min="__min" __max="__max" __median="__median"
207 run;
208
209 proc datasets library=WORK noprint;
210     delete tempData TempOutputDS6221;
211 run;
```

The RESULT page shows all graphs and charts. The first chart is the distribution of `SepalLength`. $\mu=58.4$, median=58, and $\sigma=8.2$ 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

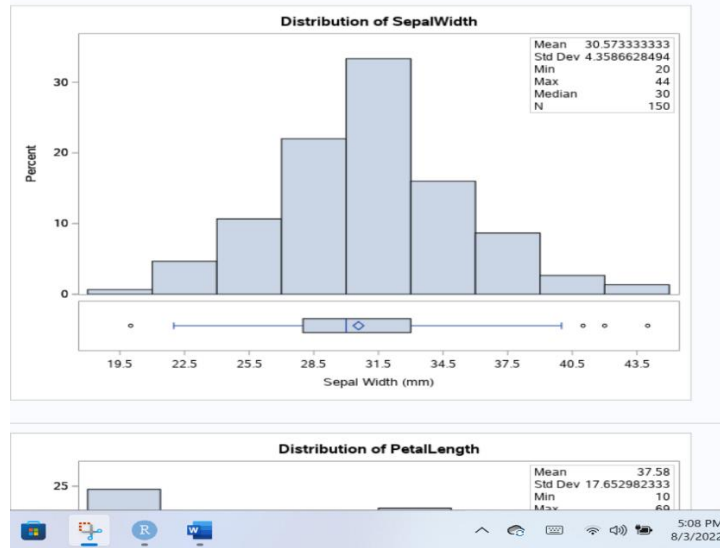
Mod 3 – Critical Thinking; Option 1

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

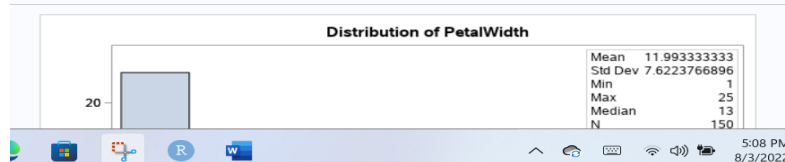
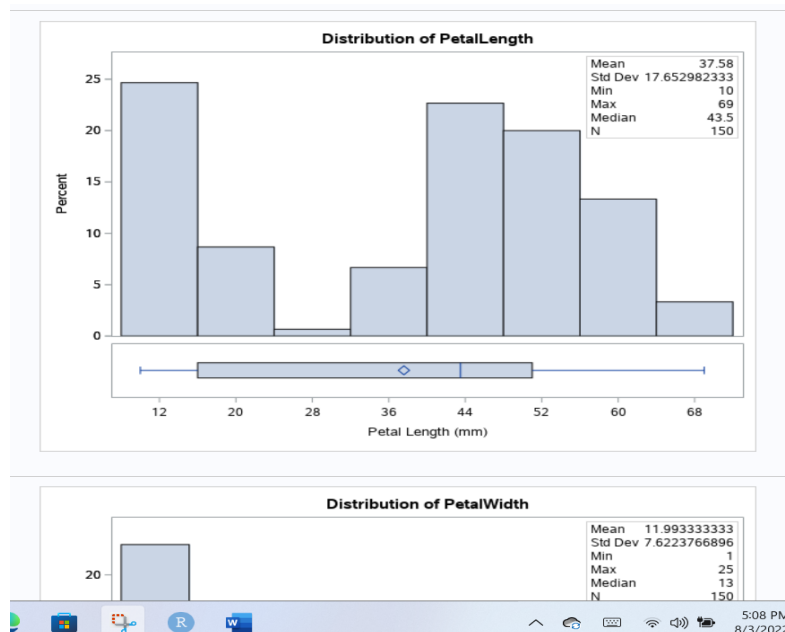


Text histogram is the distribution of SepalWidth; this means it is lower than the sepal length's mean and standard deviation. Thus, the sepal length probability is higher than the sepal width. And a histogram is like a normal distribution. The boxplot of SepalWidth shows nonsymmetric because the median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

Mod 3 – Critical Thinking; Option 1



PetalLength is one of the bimodal distributions of the histogram with a high standard deviation compared to the other variables, which is why the curve 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.



Mod 3 – Critical Thinking; Option 1

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.

