Distribution and Correlation Analysis of a CSV Imported File Using SAS Studio

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

Colorado State University-Global Campus

Dr. Steve Chung

August 14, 2022

Introduction

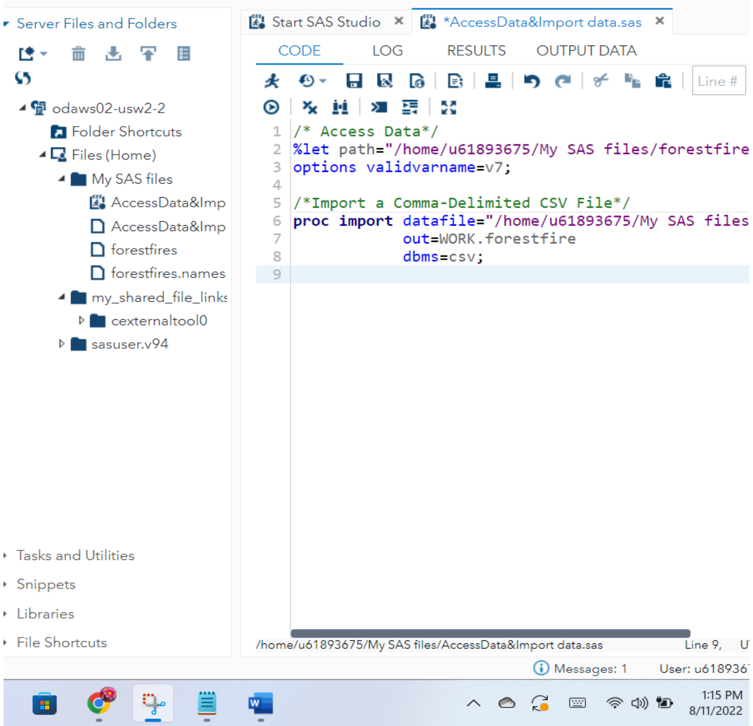
The forest fire dataset contains information that helps the most affected forest fire damage and puts human and animal life at risk. This dataset is publicly available from the UCI Machine Learning Repository. The dataset contains 517 observations with 13 attributes. Before applying and executing the model, we should perform a distribution analysis for each variable and check the correlation relation to individual variables. In this assignment, I will use the forest fire dataset's variables 'temp' and 'RH' to analyze the distribution and correlation with temperature in Celsius degrees from 2.2 to 33.30 and RH is relative to the humidity in percentage from a percentage from 15.0 to 100.

Uploading and Import the forestfire CSV(Comma-delimited) file to SAS Studio

The first step to uploading CSV data from my computer into SAS Studio is right click on the new icon and click Folder. Then, a new page will pop up, giving a new folder name and save. Then select an existing folder and right click new folder to upload a file from my laptop. Now is forestfire.csv online on SAS Studio; change the file name right, click then rename forestfire. Screenshot 2 has details.

Screenshot 1;

SAS code to import forestfire CSV file into Studio on SAS studio



```
/* Access Data*/
%let path="/home/u61893675/My SAS files/forestfire
options validvarname=v7;

/*Import a Comma-Delimited CSV File*/
proc import datafile="/home/u61893675/My SAS files
            out=WORK.forestfire
            dbms=csv;

```

/home/u61893675/My SAS files/AccessData&Import data.sas Line 9, U

Messages: 1 User: u618936

Screenshot 2

That shows the dataset table on OUTPUT DATA on SAS studio.

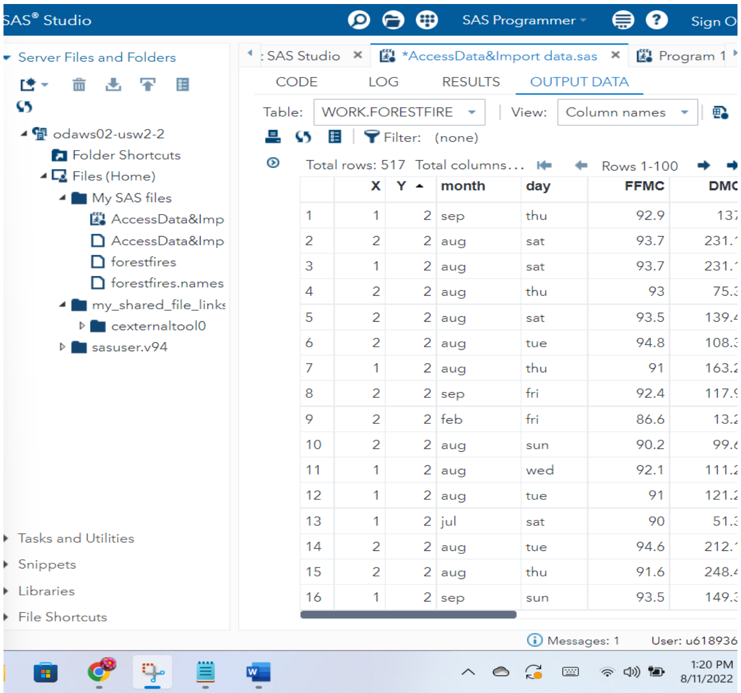


Table: WORK.FORESTFIRE View: Column names

Filter: (none)

Total rows: 517 Total columns... Rows 1-100

	X	Y	month	day	FFMC	DMC
1	1	2	sep	thu	92.9	137
2	2	2	aug	sat	93.7	231.
3	1	2	aug	sat	93.7	231.
4	2	2	aug	thu	93	75.
5	2	2	aug	sat	93.5	139.
6	2	2	aug	tue	94.8	108.
7	1	2	aug	thu	91	163.
8	2	2	sep	fri	92.4	117.
9	2	2	feb	fri	86.6	13.
10	2	2	aug	sun	90.2	99.
11	1	2	aug	wed	92.1	111.
12	1	2	aug	tue	91	121.
13	1	2	jul	sat	90	51.
14	2	2	aug	tue	94.6	212.
15	2	2	aug	thu	91.6	248.
16	1	2	sep	sun	93.5	149.

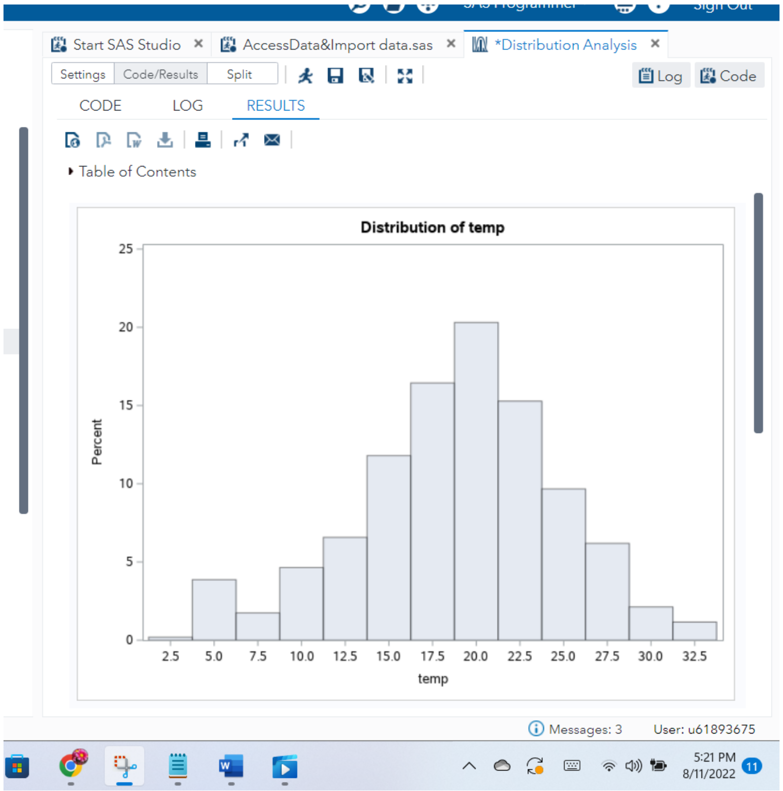
Messages: 1 User: u618936

Analyzing the distribution of temp ‘temp’ variable

Analyze the distribution of the temp variable with SAS Studio; click Tasks and Utilities from in the navigation menu, then Distribution Analyst under the Statistic as it opens a new code page. The upper ribbon has the list of the Data and Options menu. Add forest fire dataset and analysis variable is ‘temp’ following select ‘Histogram’ and ‘Normal quantile-quantile plot’ adds inset statistics—the result on screenshots 3 and 4.

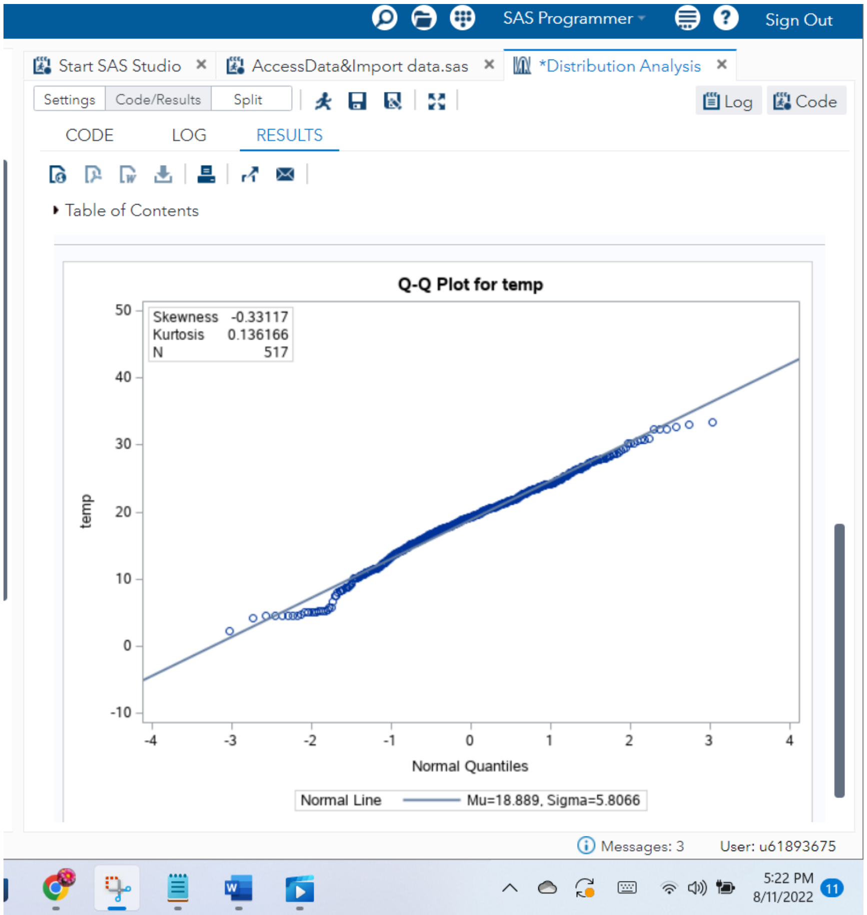
Screenshot 3

The resulting distribution analysis is a histogram of the ‘temp’ variable on SAS studio.



Screenshot 4

The resulting distribution analysis is the Norma quantile-quantile plot of the ‘temp’ variable on SAS studio.



Skewness	-0.33117
Kurtosis	0.136166
N	517

Normal Line Mu=18.889, Sigma=5.8066

Skewness tells dataset has an asymmetric distribution or not that measures three different distributions. One zero skew means the distribution is symmetrical. Others negative skew when the number is negative and positive skew when the number is positive. The skewness of temp shows a negative number of -0.33, which tells the left skew.

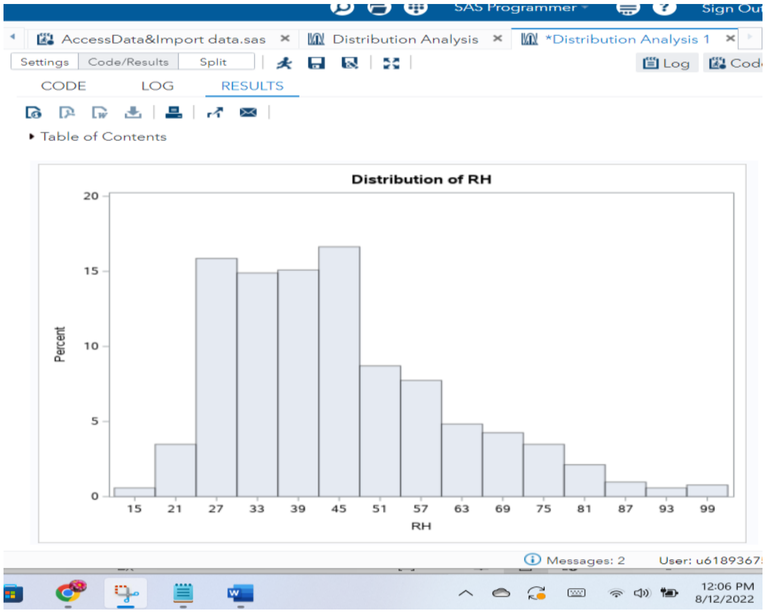
Kurtosis shows the variable's probability or frequency, which also helps to compare which variable has a heavy distribution tail with three kurtosis types. I found so many different ranges people use, and I would like to use several zero for the normal distribution of kurtosis because I check outliers and best explanate for the case outliers with the number of zero for kurtosis. Medium tails are **mesokurtic(kurtosis=0)**, low kurtosis is **platykurtic(kurtosis<0),** and high kurtosis is **leptokurtic(kurtosis>0).** The temp variable has leptokurtic because the temp's kurtosis is 0.13 as positive kurtosis more than zero as a normal distribution kurtosis's number is zero or a close zero. Thus, the temp distribution should have outliers; the skewness result is close to zero, which can be accepted for normal distribution if outliers exceed what we expect.

Analyzing the distribution of temp 'RH' variable

First, I change the variable from the Data section on the top ribbon Option has a checklist as I use Histogram under the exploring data, a Normal quantile-quantile plot with inset statistic under checking for Normality, and a number of observations, skewness, and kurtosis. Screenshot 5 has a result for histogram distribution analysis, and 6 shows a Q-Q plot for the 'RH' variable.

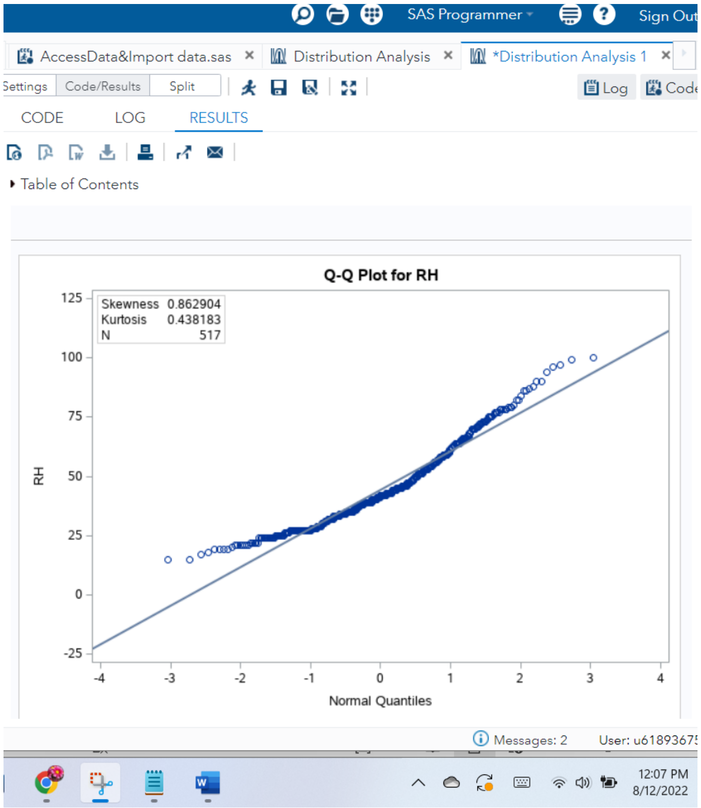
Screenshot 5

The resulting distribution analysis is a histogram of the 'RH' variable on SAS studio on SAS studio.



Screenshot 6

The resulting distribution analysis is the Norma quantile-quantile plot of the ‘temp’ variable on SAS studio.



Skewness	0.862904
Kurtosis	0.438183
N	517

RH’s histogram(screenshot5) shows more variables on the right side which means the distribution has a strong positive skew in the dataset. That also indicates that the' RH' distribution is not normal.

Q-Q plot for ‘RH’s results skewness is 0.86 that also say same what we found histogram strongly positive skewness.

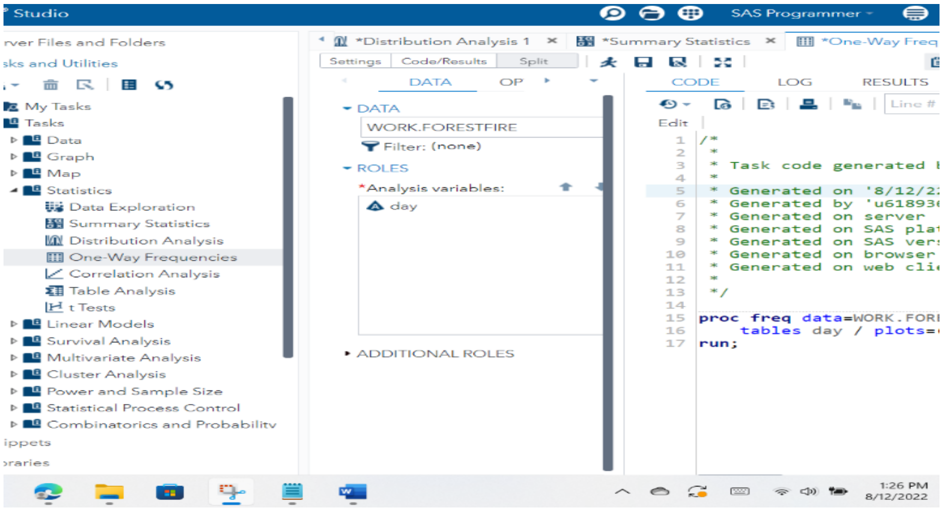
Kurtosis’s result is 0.43, as positive kurtosis would have more outliers. Thus, ‘RH’ has a positive skew and leptokurtic distribution.

One-Way Frequencies analysis by ‘day.’

What tools are ideally used for nominal variables to count frequency by each category? I worked on SAS studio’s main menu, Tasks and Utilities, then One-way Frequencies. Add variable ‘day’ to show each day’s frequency on the histogram. Screenshot 7 shows the code detail, and screenshot 8 has the result page.

Screenshot 7

The code of the one-way frequencies by ‘day’ on SAS studio.



DATA

WORK.FORESTFIRE

Filter: (none)

ROLES

*Analysis variables:

day

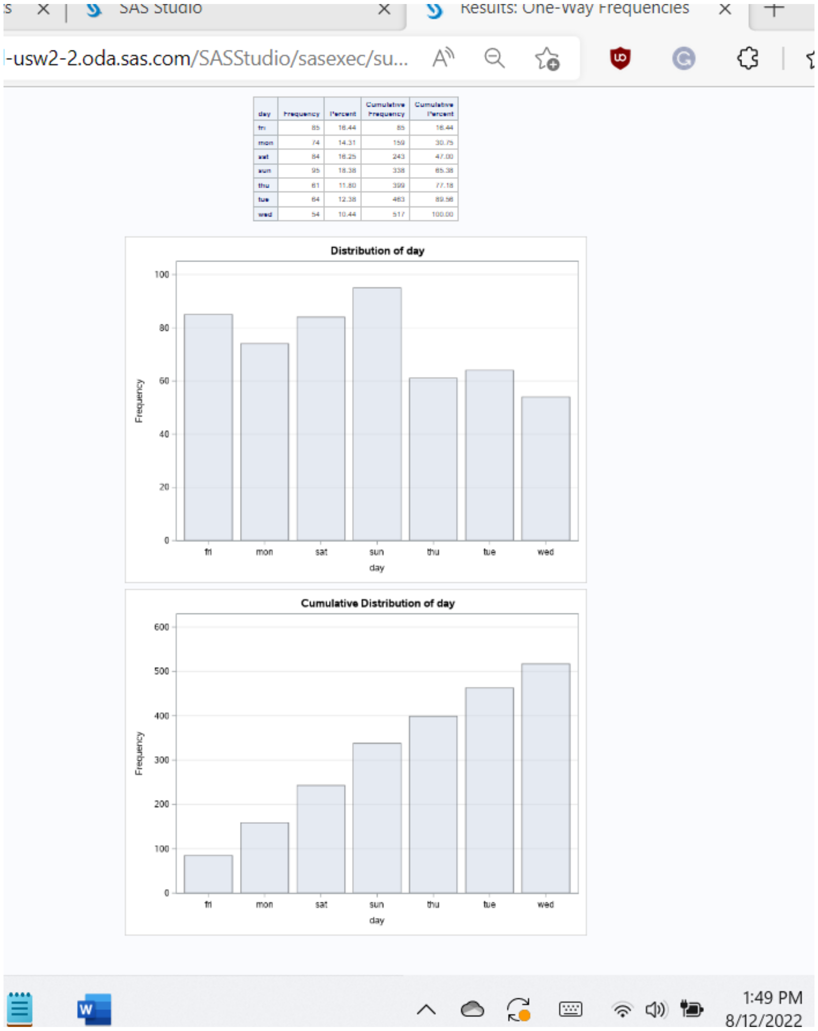
ADDITIONAL ROLES

```
/*
 *
 *  Task code generated I
 *
 *  Generated on '8/12/2
 *  Generated by 'u61893
 *  Generated on server
 *  Generated on SAS pla
 *  Generated on SAS ver
 *  Generated on browser
 *  Generated on web cli
 *
 */

proc freq data=WORK.FOR
    tables day / plots=
run;
```

Screenshot 8

The result of the distribution of 'day' by One-way Frequencies on SAS studio.


SAS Studio
Results: One-Way Frequencies
-usw2-2.oda.sas.com/SASStudio/sasexec/su...
1:49 PM
8/12/2022

day	Frequency	Percent	Cumulative Frequency	Cumulative Percent
fri	85	16.44	85	16.44
mon	74	14.31	159	30.75
sat	84	16.25	243	47.00
sun	95	18.38	338	65.38
thu	61	11.80	399	77.18
tue	64	12.38	463	89.56
wed	54	10.44	517	100.00

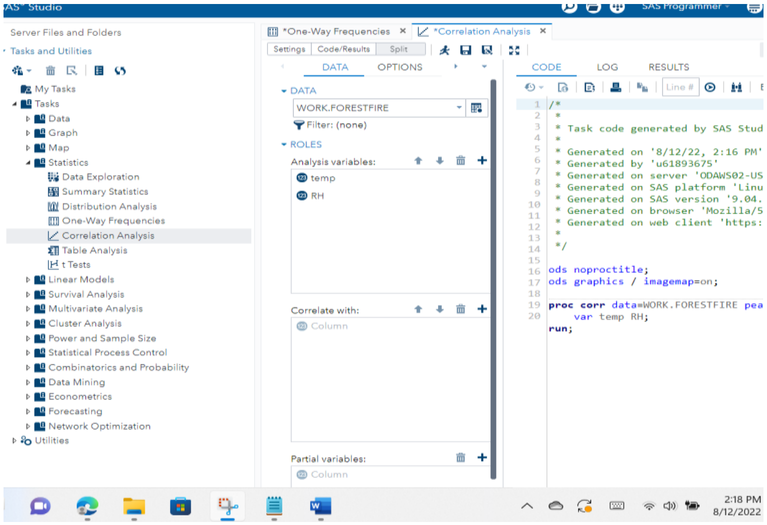
As we can easily compare, which day has more frequency than other days with frequency distribution is an outline of all prominent values. Screenshot 8, the top on the graph, shows the most frequent day on Sunday. The bottom graph is a cumulative frequency distribution that helps if there is a specific value and to see how probability is.

Correlation Analysis of the Variables of ‘temp’ and ‘RH.’

Correlation analysis helps us to know how to relation between two variables. I use the Tasks and Utilities main menu, double-click Correlation Analysis, add variables ‘temp’ and ‘RH’ on the DATA page, and move to the OPTIONS page select Matrix of scatter plots. All steps are on screenshots 9 and 10. The result page is in screenshot 11.

Screenshot 9

The code of the correlation analysis of ‘temp’ and ‘RH’ on SAS Studio.



DATA

WORK.FORESTFIRE

Filter: (none)

ROLES

Analysis variables:
- temp
- RH

Correlate with:
- Column

Partial variables:
- Column

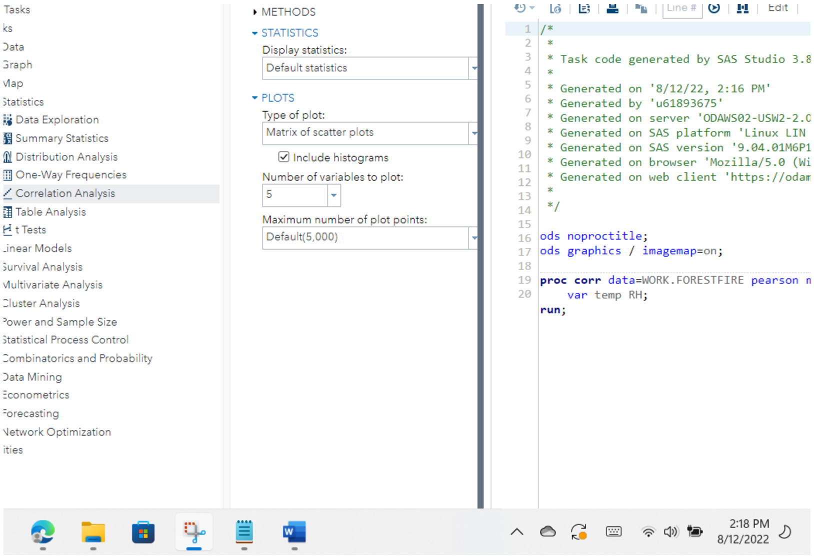
```
/*
 *
 * Task code generated by SAS Stud
 *
 * Generated on '8/12/22, 2:16 PM'
 * Generated by 'u61893675'
 * Generated on server 'ODAWS02-US
 * Generated on SAS platform 'Linu
 * Generated on SAS version '9.04.
 * Generated on browser 'Mozilla/5
 * Generated on web client 'https:
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.FORESTFIRE pea
    var temp RH;
run;
```

Screenshot 10

The code of correlation analysis Option page for ‘temp’ and “RH’ variables on SAS Studio.



▸ METHODS

▾ STATISTICS

Display statistics: Default statistics

▾ PLOTS

Type of plot: Matrix of scatter plots

☑ Include histograms

Number of variables to plot: 5

Maximum number of plot points: Default(5,000)

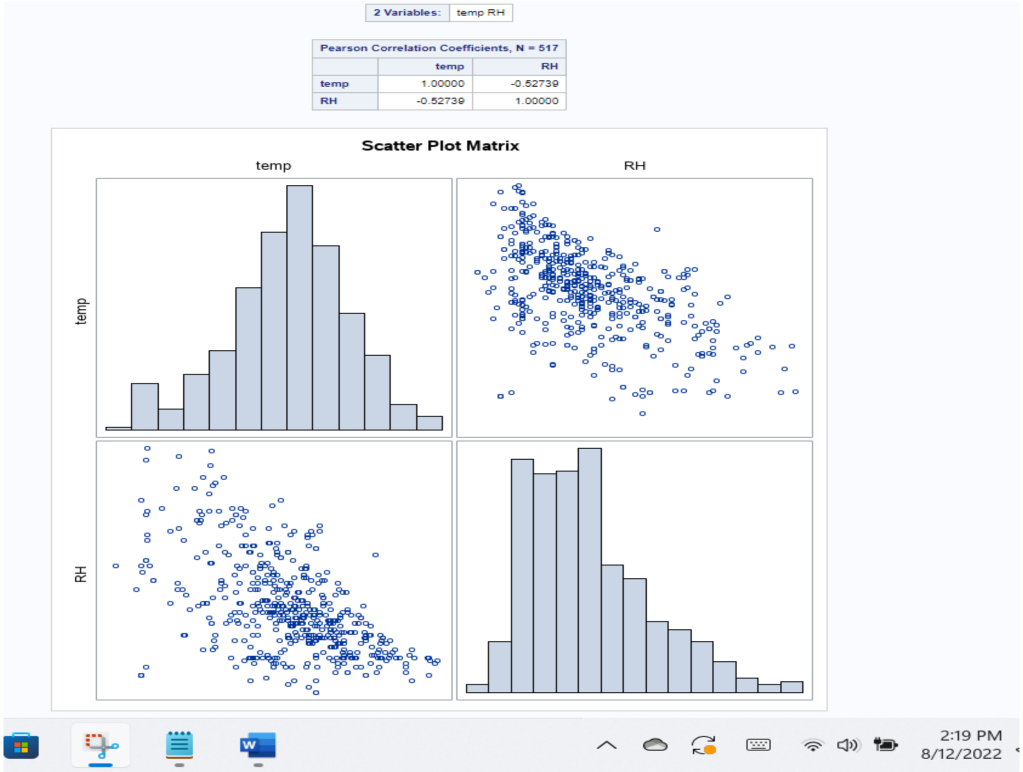
```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '8/12/22, 2:16 PM'
 * Generated by 'u61893675'
 * Generated on server 'ODAWS02-USW2-2.0
 * Generated on SAS platform 'Linux LIN
 * Generated on SAS version '9.04.01M6P1
 * Generated on browser 'Mozilla/5.0 (Wi
 * Generated on web client 'https://odam
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.FORESTFIRE pearson n
    var temp RH;
run;
```

Screenshot 11

The result of correlation analysis for ‘temp’ and ‘RH’ variables on SAS Studio.



2 Variables:	temp RH

Pearson Correlation Coefficients, N = 517		
	temp	RH
temp	1.00000	-0.52739
RH	-0.52739	1.00000

Thus, ‘temp’ and ‘RH’ have a negative correlation of -0.52, which means when the temperature increases, the humidity decreases (see screenshot 11).

Conclusion

That great to learn how to upload complex files online on SAS Studio and Import the file. I use Distribution Analysis for each variable to see how the data behave. The one-way frequencies excellent tools that I use for nominal variables; in this case, I used the ‘day’ variable. The correlation analysis with the matrix plot menu shows the ‘temp’ and ‘RH’ variable relation.

References

UCI Machine Learning Repository: Forest Fires Data Set