# Exploring PROC TABULATE and PROC SGPLOT in SAS

Introduction to Statistical Computing and Exploratory Data Analysis - SAS
STAT 342/642 D100
Final Project Assignment

Huy Ngoc Nguyen – 301405437

November 26, 2023

## 1. PREFACE

This report explores two SAS procedures: PROC TABULATE and PROC SGPLOT. We will walk through each procedure and its options to better understand the function of these procedures. We will use the student performance data from the UCI machine learning repositories and import it into SAS using PROC IMPORT. The data contains 30 variables related to students' backgrounds and 3 variables related to their course grades. The data is also divided into two CSV files, so we merge the files into a single dataset using PROC SQL by searching for identical attributes in the student body. In addition, we have created a rating column based on a student's final grade to group similar students together. From there, we will look to produce some examples using PROC TABULATE and PROC SGPLOT.

2. **PROC TABULATE**

PROC TABULATE is a procedure used to create customized tables in tabular format from raw data. It allows one to summarize and display data in simple or highly complex tables, providing a flexible and efficient way for anyone to analyze and present information. The procedure is composed of a wide range of options. In this report, we will discuss the following options from PROC TABULATE: DATA, CLASS, TABLE, VAR, and TITLE.

- The DATA option requires an input data set. In our case, after importing and merging the datasets, and creating a rating column, we can assign a dataset for the TABULATE procedure.
- The CLASS option specifies categorical variables that comprise the rows and columns of the table. Excluding the variable 'absence', which is the number of school absences, as well as course grades, we can use any other variable for our CLASS option.
- The TABLE option specifies the layout of the table and the variables used for its rows and columns. We will explore this option in more detail as it has a few useful sub-options to customize our table.

With just these three options, we can create a simple table using the code below, which also includes the setup of the dataset (Table 1).

```
/* Use PROC IMPORT to import the data */
proc import out = math datafile =
"C:\Users\ADMIN\Desktop\STAT342\Assignments\Project\student-mat.csv"
          dbms = csv replace;
                    getnames = yes;
               delimiter=';';
run;

/* Use PROC IMPORT to import the data */
proc import out = port datafile =
"C:\Users\ADMIN\Desktop\STAT342\Assignments\Project\student-por.csv"
          dbms = csv replace;
                    getnames = yes;
               delimiter=';';
run;

proc sql; /* merging similar students across 2 datasets */
  create table main as
  select *
```

```
  from math as a
  inner join port as b
  on a.school = b.school and a.sex = b.sex and a.age = b.age and a.address = b.address and
a.famsize = b.famsize and a.Pstatus = b.Pstatus
and a.Medu = b.Medu and a.Fedu = b.Fedu and a.Mjob = b.Mjob and a.Fjob = b.Fjob
and a.reason = b.reason and a.nursery = b.nursery
and a.internet = b.internet;
quit;

data final; /* creating a rating */
        set main;
  if G3 < 4 then
    rating = 'Unsatisfactory';
  else if 4 <= G3 <= 7 then
    rating = 'Below Average';
  else if 8 <= G3 <= 11 then
    rating = 'Average';
  else if 12 <= G3 <= 15 then
    rating = 'Good';
  else
    rating = 'Outstanding';
run;

proc tabulate data = final; /* Number of students by rating */
        class rating;
        table rating;
run;
```

Table 1 shows the number of students by their ratings. We can see that the majority of students fall under the 'Average' category, meaning that their final grade is between 8 and 11, followed by the 'Good' category, where their final grade is between 12 and 15. We can also add additional variables to the table to see if there is any relationship.

```
proc tabulate data = final; /* Number of students by gender and rating */
        class sex rating;
        table sex, rating;
run;
```

Table 2 shows the number of students classified by their gender and their ratings. We can see that the highest count in this table is female students whose ratings are average, followed by male students with average ratings. Notice that there is a comma between the variables in the TABLE option. The comma adds an additional dimension to the table, making the gender category on the y-axis and the rating category on the x-axis. The table will be one-dimensional if the comma is removed. We can also change the layout of the table by having a variable nested within another variable.

```
proc tabulate data = final; /* Number of students by gender and rating in one-
dimensional*/
        class sex rating;
        table sex * rating;
run;
```

Table 3 shows the same information in Table 2, but instead of using a comma, we now use a star, which organizes the table in a one-dimensional space where the rating category is nested within the gender category. And so, each gender category now has its own rating category. Until now, we have only displayed frequency counts on our tables. If we want to incorporate numeric statistics such as sum, mean, and frequency percentage, then we will need to use the VAR option.

- The VAR option specifies the numerical variables whose values will be summarized in the table. In this case, we can use the absence category, and the course grades in the VAR statement as they are numerical variables.
- The TITLE option specifies the title of the table. While this is optional, the TITLE statement allows us to give a different title for each of the tables that we are generating.

Using the statements above, we can create a semi-complex table as displayed in Table 4. We can see that the average final grade for students is around 9 or 10, and the average is highest with students who do not receive any additional educational support. In addition, we can also observe that the majority of students in this dataset do not receive extra education support. Instead, they prefer family educational support, which could mean that they spend more time studying with their parents instead of taking extra classes outside of school.

```
proc tabulate data = final; /* displaying the students' final grades based on educational
support, along with the sum, mean, and frequency percentage */
        class schoolsup famsup;
        var G3;
        table (schoolsup)* (famsup),
```

```
                    (G3)* (sum mean pctn);
        title 'final score by educational support';
run;
```

We can also insert an ALL option in the TABLE statement to sum up all levels of the preceding variable. This option acts as a subtotal for each level of the category, which summarizes the information within a level of the category. Table 5 illustrates this by showcasing the final grade of the students based on their ratings and weekly study time. The subtotal row sums up the final grade information in each rating, which allows us to compare between ratings.

```
proc tabulate data = final; /* Understanding if educational support have an effect on
rating */
        class rating studytime;
        var G3;
        table (rating)* (studytime all = 'Subtotal'),
                    (G3 = "Final Grade")* (sum min max pctn);
        title 'Final score by rating and study time';
run;
```

From Table 5, we can see that unsatisfactory students, whose final grades are all zero, comprise 10% of the student body. Regardless of rating, the majority of students spend 2 to 5 hours studying. Looking at the outstanding students, we observe that there is at least one student who received the highest score possible on their final grade by spending more than 10 hours studying weekly.

## 3. PROC SGPLOT

PROC SGPLOT is a procedure used to create various types of statistical plots and graphs on a single set of axes. This includes line plots, scatter plots, histograms, regression plots, and many more. Similar to the PROC TABULATE, this procedure requires an input dataset in the DATA statement. However, depending on the type of graph one would like to use on their program, the syntax and the options can vary. We will walk through some of these graphs to see how our data is represented.

```
proc sgplot data = final;
        SCATTER x = absences y = age;
        TITLE "Scatter plot of age and absences";
run;
```

The code above outputs a scatterplot by using the SCATTER statement, as shown in Figure 1. The x-axis is the 'absences' variable, which counts the number of school absences of a student, and the y-axis is the student's age. The TITLE statement adds a title to the plot. From the scatterplot, we can see the majority of absences are between 15 and 18-year-old students, with a few outliers who have more than 50 school absences.

```
proc sgplot data = final;
        HISTOGRAM studytime;
        density studytime;
        TITLE "Histogram of students by rating";
run;
```

To graph a histogram using SGPLOT, we set the HISTOGRAM option to the desired variable. We can also add a normal density curve on the histogram by setting the DENSITY option to the same variable. Figure 2 shows the distribution of students by their weekly study time, accompanied by a normal density curve. We can observe that almost 50% of students study between 2 to 5 hours.

```
proc sgplot data= final;
        xaxis label = "Rating category";
        yaxis label = "Total Hours";
        vbar rating / response = studytime;
        vbar rating / response = traveltime
                barwidth = 0.5
                transparency = 0.2;
        TITLE "Bar charts of the sum of study time and travel time of students by rating";
run;
```

To graph vertical bar charts, we can use the VBAR statement and specify the explanatory and response variables. The code chunk above shows two VBAR statements, meaning that we are graphing two vertical bar charts on the same chart using two different response variables (Figure 3). Notice that the second VBAR statement has additional options for “bar width” and “transparency”, which customizes the layout of the second vertical bar chart for easier viewing. The x-axis and y-axis label statements rename the axes of the graph. From the figure, we can see that average students spend the most time studying weekly as well as traveling to school.

## APPENDIX

| rating | | | | |
|---|---|---|---|---|
| Average | Below Average | Good | Outstanding | Unsatisfactory |
| N | N | N | N | N |
| 157 | 30 | 114 | 42 | 39 |

Table 1: Student frequency count by rating

| | rating | | | | |
|---|---|---|---|---|---|
| | Average | Below Average | Good | Outstanding | Unsatisfactory |
| | N | N | N | N | N |
| sex | | | | | |
| F | 87 | 21 | 51 | 16 | 23 |
| M | 70 | 9 | 63 | 26 | 16 |

Table 2: Student frequency count by rating and gender in a two-dimensional table

| sex | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| F | | | | | M | | | | |
| rating | | | | | rating | | | | |
| Average | Below Average | Good | Outstanding | Unsatisfactory | Average | Below Average | Good | Outstanding | Unsatisfactory |
| N | N | N | N | N | N | N | N | N | N |
| 87 | 21 | 51 | 16 | 23 | 70 | 9 | 63 | 26 | 16 |

Table 3: Student frequency count by rating and gender in a one-dimensional table

| | | G3 | | |
|---|---|---|---|---|
| | | Sum | Mean | PctN |
| schoolsup | famsup | | | |
| no | no | 1392.00 | 10.71 | 34.03 |
| | yes | 2092.00 | 10.41 | 52.62 |
| yes | no | 147.00 | 10.50 | 3.66 |
| | yes | 337.00 | 9.11 | 9.69 |

Table 4: Student final grades (sum, mean, frequency percentage) by extra/family education support

**Final score by rating and study time**

| | | Final Grade | | | |
|---|---|---|---|---|---|
| | | Sum | Min | Max | PctN |
| rating | studytime | | | | |
| Average | 1 | 363.00 | 8.00 | 11.00 | 9.95 |
| | 2 | 841.00 | 8.00 | 11.00 | 22.51 |
| | 3 | 233.00 | 8.00 | 11.00 | 6.28 |
| | 4 | 87.00 | 8.00 | 11.00 | 2.36 |
| | Subtotal | 1524.00 | 8.00 | 11.00 | 41.10 |
| Below Average | studytime | | | | |
| | 1 | 45.00 | 5.00 | 7.00 | 2.09 |
| | 2 | 120.00 | 4.00 | 7.00 | 5.24 |
| | 3 | 7.00 | 7.00 | 7.00 | 0.26 |
| | 4 | 6.00 | 6.00 | 6.00 | 0.26 |
| | Subtotal | 178.00 | 4.00 | 7.00 | 7.85 |
| Good | studytime | | | | |
| | 1 | 440.00 | 12.00 | 15.00 | 8.64 |
| | 2 | 689.00 | 12.00 | 15.00 | 13.35 |
| | 3 | 292.00 | 12.00 | 15.00 | 5.50 |
| | 4 | 122.00 | 12.00 | 15.00 | 2.36 |
| | Subtotal | 1543.00 | 12.00 | 15.00 | 29.84 |
| Outstanding | studytime | | | | |
| | 1 | 192.00 | 16.00 | 19.00 | 2.88 |
| | 2 | 269.00 | 16.00 | 19.00 | 4.19 |
| | 3 | 173.00 | 16.00 | 19.00 | 2.62 |
| | 4 | 89.00 | 16.00 | 20.00 | 1.31 |
| | Subtotal | 723.00 | 16.00 | 20.00 | 10.99 |
| Unsatisfactory | studytime | | | | |
| | 1 | 0.00 | 0.00 | 0.00 | 3.40 |
| | 2 | 0.00 | 0.00 | 0.00 | 4.45 |
| | 3 | 0.00 | 0.00 | 0.00 | 1.57 |
| | 4 | 0.00 | 0.00 | 0.00 | 0.79 |
| | Subtotal | 0.00 | 0.00 | 0.00 | 10.21 |

Table 5: Student final grades (sum, mean, frequency percentage) by rating and weekly study time

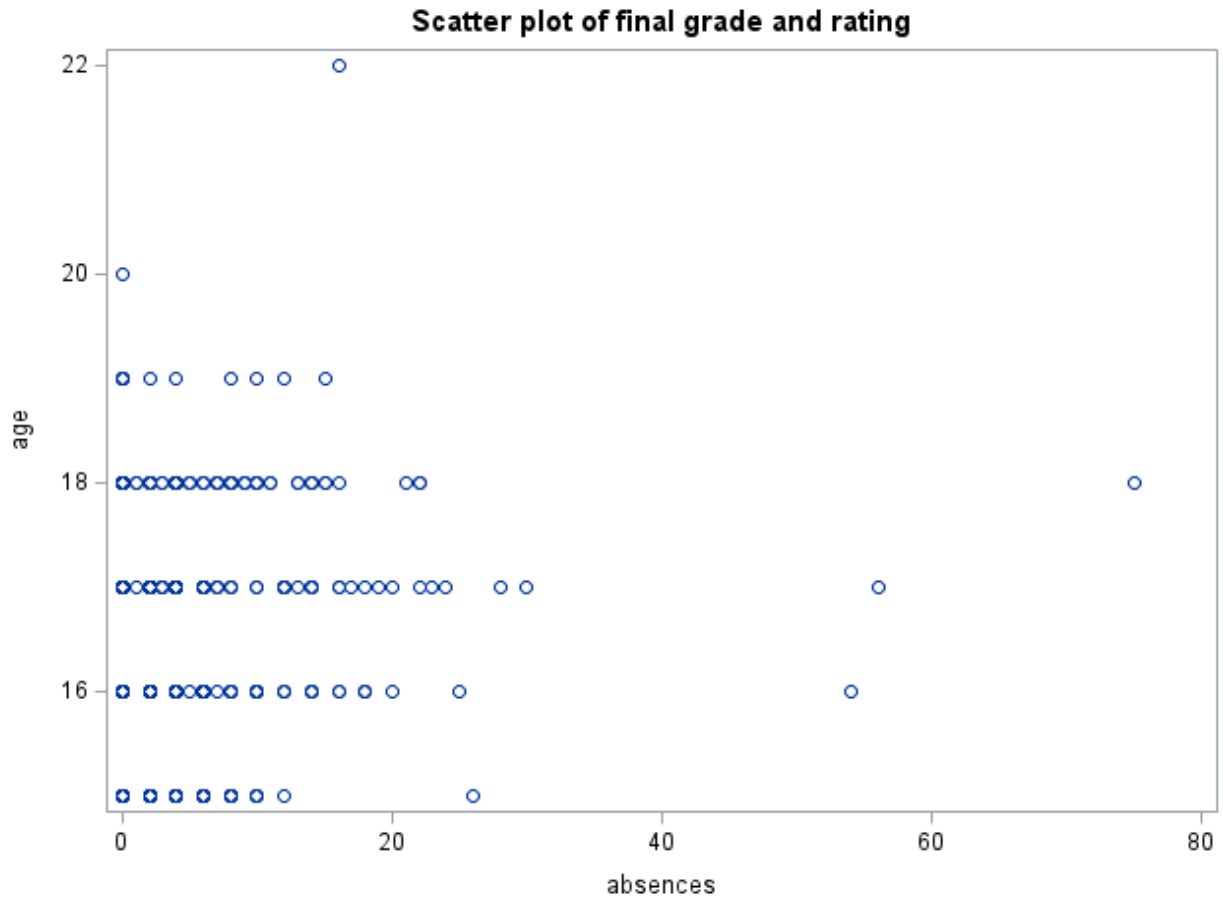


Figure 1: Scatterplot of students by absences and age

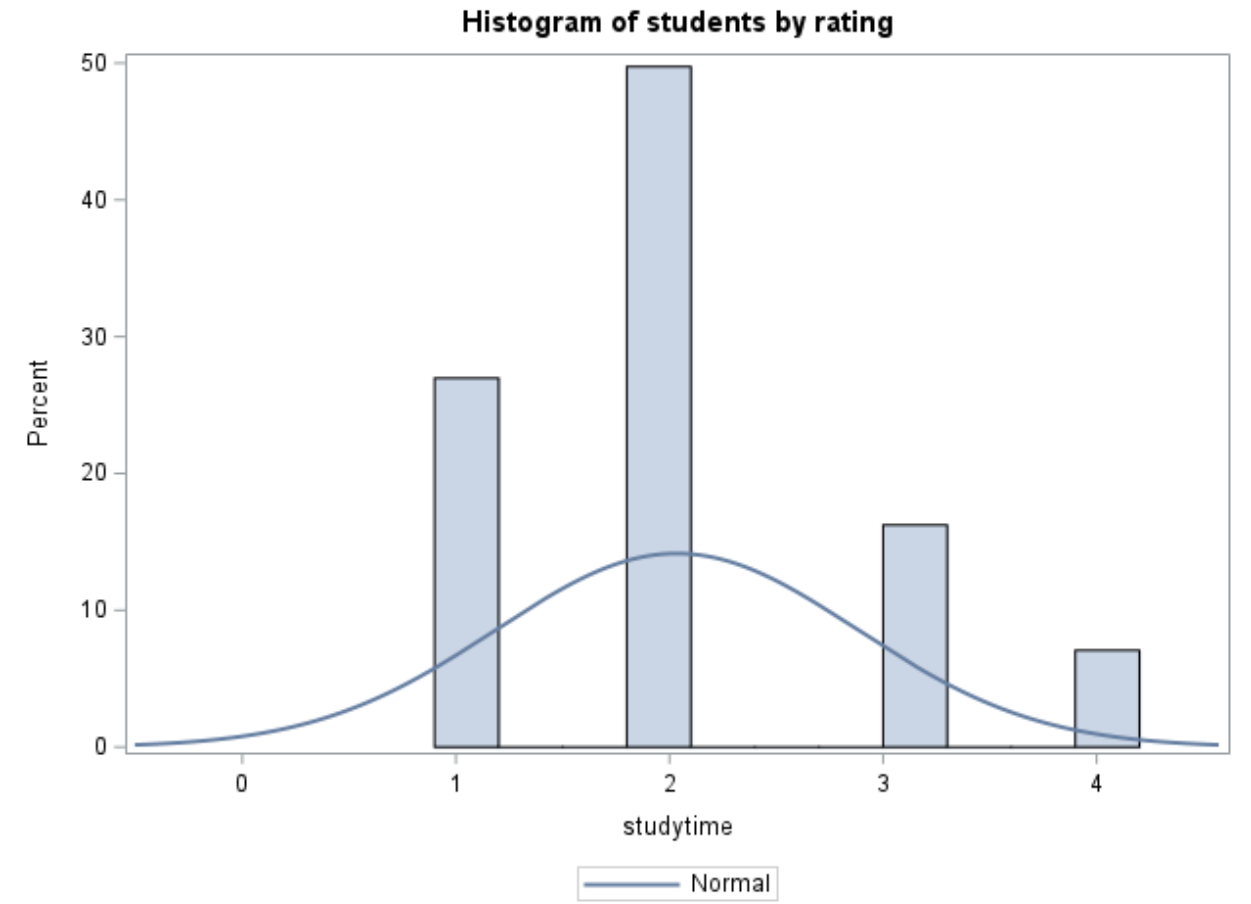


Figure 2: Histogram of students with different weekly study time

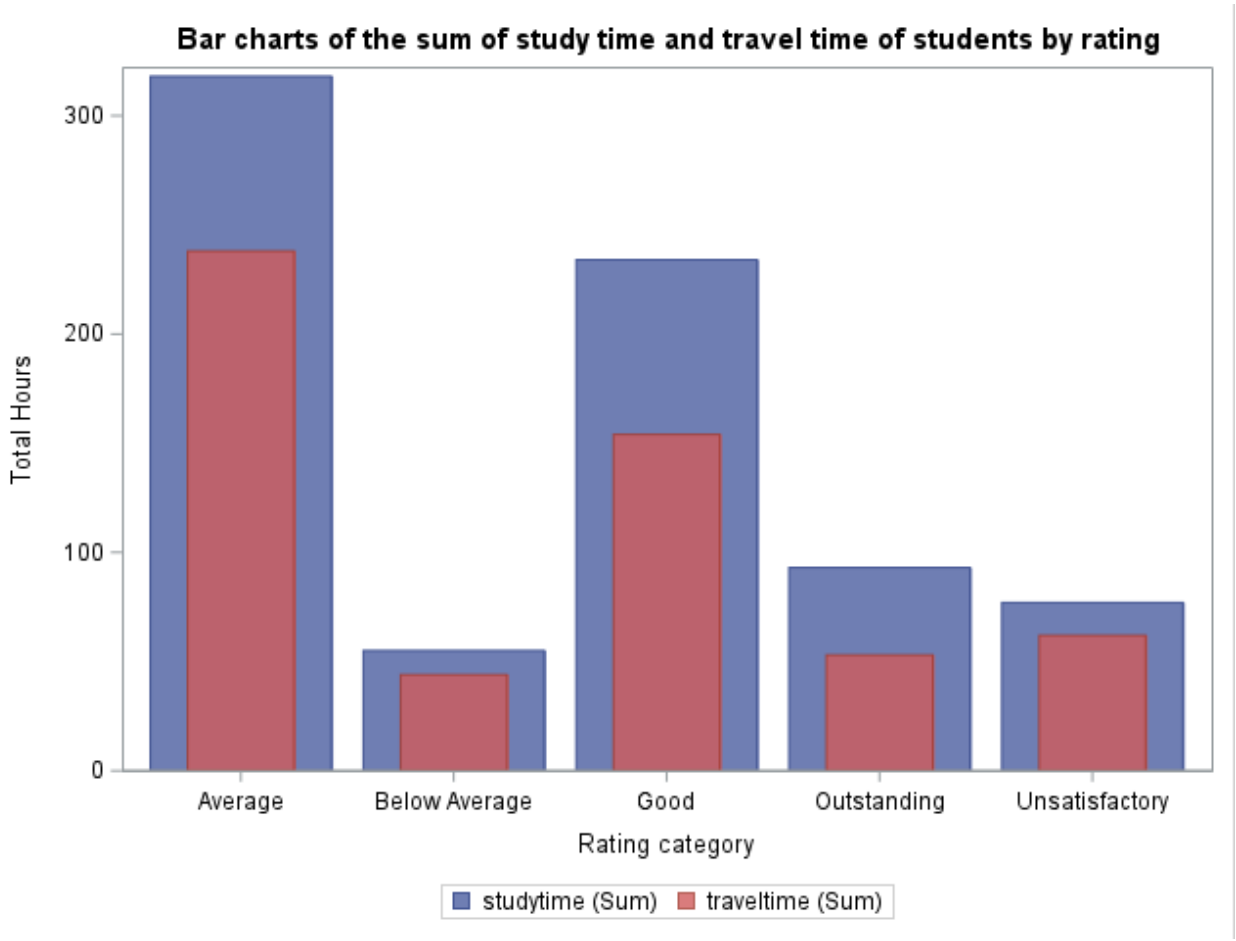


Figure 3: Bar charts of the sum of study time and travel time of students by rating