STAT342: Introduction to Statistical Computing and Exploratory Data Analysis

Comprehensive Tutorial Report on Adult, Bank Marketing, and Student Performance Datasets
PROC TABULATE AND PROC SGPLOT

This tutorial is for the students in Dr. Haolun Shi's STAT342 class at Simon Fraser University. In real-life situations, it shows how to use two powerful SAS techniques, **PROC TABULATE** and **PROC SGPLOT**, on three different datasets. Students can easily follow the steps in this report to properly learn information about SAS algorithms. If you are unsure about how to handle the dataset using SAS software, don't worry! This tutorial will enhance your ability to master the task!

- ## *Getting Started*

  We will be focusing on three different datasets from the **UCI Machine Learning Repository** in this tutorial. Each one gives us new information about a different subject. The **Adult Dataset** (https://archive.ics.uci.edu/ml/datasets/adult) has a lot of data about adults, which makes it ideal for studying socioeconomic issues. The **Bank Marketing Dataset** (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) also has a lot of information about how people use banks, which is very useful for business research. The **Student Performance Dataset** (https://archive.ics.uci.edu/ml/datasets/Student+Performance), on the other hand, gives a unique view of how well students are doing in school, which is very important for figuring out what makes students successful. I will use two powerful SAS tools to look at these datasets: **PROC TABULATE** to extract data and show it in detailed tables and **PROC SGPLOT** to use statistical graphics to show results and draw attention to patterns and links.

- ## *Code Explanation of adult.data*

Figure 1 and Figure 2 (Data Loading and Initial Processing)

It starts with **proc contents** to show a lot of information about the work.adult dataset. Then, data work. adult is used to bring in data from a file called adult.data. In this step, you choose the file path, set the delimiter, and decide what to do with missing values and data lines. It says what kind of variable it is and how long it is for each one, like age, education, hours per week, etc.

Figure 3 (Basic Tabulation of Occupation)

The code uses **proc sgplot** to make a scatter plot that shows how Age and HoursPerWeek are related in the dataset. This shows how these two factors are linked.

Figure 4 (Scatter Plot of Age VS. Hours Per week)

After that, **proc tabulate** is used to make a simple table that summarizes the occupation variable in the work.dataset of adults. It shows the number of each job (n), with commas added to make it easier to read.

Figure 5 (Tabulation with Occupation and Race)

It is used again, but this time the data is sorted by both occupation and race. It figures out and shows the number of hours worked and the number of hours worked each week for each job, broken down by race.

Figure 6 (Histogram and Density Plot of Age)

In this part, **proc sgplot** is used to make a density plot and a histogram for the age variable. The histogram has a bin width of 5, which shows how the ages in the information are spread out visually.

- ## *Interpretation of Results*

Figure 7 (Data Summary)

The SAS **proc contents** procedure gives a lot of information about an SAS dataset called work.adult.

Figure 8 (Alphabetical order of variable name)

proc contents, which provides an alphabetical list of variables and their attributes from a SAS dataset.

Figure 9 (Distribution of individuals across various occupational categories)

It is a table listing different occupations along with a column labeled N, which typically stands for Number or Frequency. Each occupation has a corresponding number indicating the count of individuals or entries in each occupational category within a dataset. For example, there are 1,843 entries without an occupation label (indicated by '?'), 3,769 entries for Admin-clerical, and so on, down to 1,597 for Transport-moving.

Figure 10 (relationship between age (on the horizontal x-axis) and hours worked per week (on the vertical y-axis).

When examining the workforce's age distribution in terms of working hours, one can use the scatter plot. This could have effects on job policies, economic studies, planning for retirement, and figuring out how to balance work and life at different stages of life.

Observation

- Data is spread across a wide range of ages, from young adults to those around 80 years old.
- Hours worked per week vary significantly across all ages, with values ranging from very few hours to 100 hours per week, which is unusually high.
- The concentration of data points appears denser in the middle range of hours worked (around 30-60 hours per week), which suggests this is the most common work duration for most individuals.
- There is no clear, strong pattern indicating that age has a direct influence on the number of hours worked per week. However, it does seem that fewer individuals work very long hours (over 60 per week) as age increases.
  1. For younger and middle-aged adults, there is a wider spread in the number of hours worked, including both part-time and full-time hours, as well as some instances of very high work hours (possibly indicating multiple jobs or overtime work).

2. In the higher age ranges (approximately 60 and above), the spread of hours worked is narrower, which might suggest a trend toward reduced work hours or retirement.

Figure 11 (average hours worked per week across different occupations and broken down by race).

The table is divided into five columns for different racial groups: American Indian/Eskimo, Asian/Pacific Islander, Black, Other, and White.

Observation

1. **Occupational Distribution**: A different number of people from different racial groups work in each job. For example, there are a lot of people from the White and Asian/Pacific Islander race groups working as 'Admin-clerical'.
2. **Hours Worked**: There is a variation in mean hours worked per week among occupations and among racial groups within those occupations. Some occupations like 'Farming-fishing' and 'Transport-moving' have higher average hours worked across most racial groups compared to others like 'Priv-house serv'.
3. **Racial Representation**: Some occupations have a higher representation from certain racial groups than others. For example, 'Priv-house serv' has a higher percentage of individuals identified as Black compared to other races.
4. **Work Patterns**: The work pattern (hours worked) varies not just by occupation but also by race within the same occupation. For instance, Asian/Pacific Islanders in 'Craft-repair' work on average about 40 hours, whereas those identified as Black in the same occupation work on average around 41 hours.
5. **Diversity in the Workforce**: The table also reflects the diversity within the workforce of different occupations and the racial demographics within those occupations.

Figure 12 (histogram overlaid with a normal distribution curve).

People often use this graph to look at how the ages of people in a community or sample are spread out. It might be useful for studying demographics, planning social services, or figuring out how age affects things in a certain setting.

1. **Age Distribution**: The histogram shows the distribution of age within a population or sample. Each bar represents the frequency or percentage of individuals within certain age ranges.
2. **Skewness**: The distribution appears to be slightly right skewed, as the tail on the right side of the distribution is longer or fatter than the left side. This suggests that there are more individuals in the younger age brackets than the older ones.
3. **Central Tendency**: The peak of the histogram, which corresponds to the mode, appears to be around the 35-45 age range. This indicates that the most common age in this dataset falls within this range.
4. **Fit with Normal Distribution**: The blue line represents a normal distribution curve, which is a theoretical distribution that shows what the data would look like if it were perfectly normal (i.e., symmetric bell-shaped with mean = median = mode). The extent to which the histogram bars fit under this curve indicates how closely the age distribution follows a normal distribution.
5. **Deviation from Normality**: The bars do not perfectly align with the normal distribution curve, particularly in the tails. This discrepancy indicates that the actual distribution of ages is not perfectly normal.

- *Code Explanation of bank data*

Figure 13 (Library Assignment and Data Import and Dataset Contents Display)

The libname line tells the program to put the library reference mydata in a certain place. It then reads a CSV file called bank-additional-full.csv, adds it to the mydata library as an SAS dataset called bank_additional_full, and replaces any other dataset with the same name. A semicolon (;) is used as the separator in the CSV file. The amount of observations, variables, and other attributes for the mydata.bank_additional_full dataset are shown by proc contents.

Figure 14 (Printing First 10 Observations: Tabulation by Marital Status and Education)

proc print is used to display the first 10 observations from the mydata.bank_additional_full dataset to give a quick view of the actual data.

Figure 15 (Tabulation by Job)

proc tabulate creates a table that summarizes the number of clients by marital status and education. The numbers are formatted with commas for better readability, and the box title 'Clients by Marital Status and Education' provides a descriptive header for the table.

Figure 16 (Scatter Plot of Age vs Consumer Confidence Index)

proc sgplot generates a scatter plot of age against the consumer confidence index, grouped by marital status. The plot uses filled circles as markers and includes axis labels and a title to describe the chart.

Figure 17 (Histogram of Employment Variation Rate)

A histogram of the employment variation rate is created using **proc sgplot**, with a bin width set to 0.5. This plot is labeled on both axes and includes a title, providing a visual representation of the distribution of the employment variation rate in the dataset.

- *Interpretation of Results*

Figure 18

Usually, this kind of data is used in predictive modelling to find out what factors affect a certain result, like whether a person signs up for a term deposit. The information could come from a bank or other financial institution. It could be used to help with marketing plans or to figure out how risky a loan is. Indicators such as "emp.var.rate," "cons.price.idx," "cons.conf.idx," "euribor3m," and "nr.employed" point to an economic setting, which could connect a customer's financial situation or economic indicators to their chance of subscribing to a product.

Figure 19 (cross-tabulates clients by marital status against their level of education)

This type of table is useful for understanding the demographics of a client base and can inform targeted marketing strategies, educational program offerings, or other services tailored to specific groups. It also provides insight into the relationship between marital status and education level within a given population.

Observation

1. **Marital Status**: Most of the clients are married, followed by single, divorced, and a few unknowns.
2. **Education Levels**: The most common education level among the clients is a university degree, followed by high school education. The least common is illiteracy, with very few clients falling into this category.
3. **Married Clients**: Married clients are the most represented across all education levels except for illiteracy, which has an equal but very low representation across married, single, and unknown marital statuses.
4. **Single Clients**: Single clients have a relatively higher representation among those with a university degree and high school education compared to other education levels.
5. **Divorced Clients**: Divorced clients are most represented in the high school education category.
6. **Educational Distribution Among Marital Status**: There is a diverse distribution of educational attainment within each marital status category, suggesting a varied client base.
7. **Unknown Categories**: Both the marital status and education level have unknown categories, indicating that the data collection process might not have captured all necessary information.

Figure 20 (Job Distribution Among Clients)

This table shows a quick look at the range of professional backgrounds in a client group. It could be useful for looking at employment trends, finding goods or services that are specific to jobs, and learning about the income levels of the customers. For instance, a large number of administrative assistants could mean that most of the clients work in offices, while a large number of blue-collar workers could mean that most of the clients do physical or industrial labour. This information could be very helpful for businesses that lend money, help people find jobs, or study the market.

Figure 21 (relationship between age and the Consumer Confidence Index)

The scatter plot can be useful in determining if there is a relationship between age and economic confidence across different marital statuses. However, based on this visualization, there doesn't seem to be any significant trend or pattern that would suggest a strong relationship between these variables. Further statistical analysis would be needed to draw any definitive conclusions.

Observation

1. **Age Range**: Data points are spread across a wide range of ages, from young adults to those close to 100 years old.
2. **Consumer Confidence Index**: The CCI values span from just above -50 to just below -30. The Consumer Confidence Index is an economic indicator that measures how optimistic or pessimistic consumers are regarding their expected financial situation, with higher values indicating more confidence.
3. **Marital Status Distribution**: Each marital status is represented by a different colour, and it appears that married individuals (blue dots) are the most numerous in the dataset, followed by single (red), divorced (green), and unknown (purple). There's a very small number of dots representing "marital" (possibly an error in categorization, as this group should logically be non-existent or should refer to 'married').
4. **Trends and Correlations**: The plot does not show a clear correlation between age and the Consumer Confidence Index for any marital status group. The distribution of CCI scores is relatively consistent across different ages for each marital status category.
5. **Clustering of Data Points**: There is a noticeable clustering of data points along specific CCI values, which may suggest that the CCI was measured at discrete time points or that the index has common values for large groups of individuals.
6. **Outliers**: There are a few outlier points, particularly in the 'unknown' marital status category, which have much lower CCI values compared to the rest.

Figure 22 (histogram of the Employment Variation Rate)

The result depicts a histogram of the Employment Variation Rate, which is likely an economic indicator that measures changes in employment levels over a period.

Observation

1. **Distribution**: The histogram shows a bimodal distribution with two peaks, one around the -2 to -1 range and another larger peak at the 1 to 2 range.
2. **Most Frequent Values**: The most frequent employment variation rate falls within the 1 to 2 range, suggesting that during the period measured, there was a significant frequency of employment rates increasing by this margin.
3. **Negative Values**: There are instances of negative employment variation rates, particularly between -2 and -1, indicating periods when employment has decreased.
4. **Zero and Near-Zero Occurrences**: There are fewer instances where the employment variation rate was around zero, suggesting that periods of stable employment (no increase or decrease) were less common in this dataset.
5. **Data Range**: The range of the data extends from -3 to 2, indicating that employment variation rates in the dataset have varied widely.

- Interpretation

1. **Economic Trends**: The chart might show economic cycles or trends, with times when employment goes up (growth or recovery phases) and times when employment goes down (recession or contraction phases).
2. **Economic Stability**: The fact that zero variation rates happened less often could mean that the economy was unstable during the time when the data was taken.
3. **Policy Implications**: If used by policymakers or economists, such data can inform decisions about labour markets and

economic policies to either stimulate job growth or prepare for downturns.

- Code Explanation of student-mat data

Figure 23 (Data Import)

You can load a CSV file called student-mat.csv into SAS with the proc import line. This makes a dataset called student_mat. With getnames=yes, the variable names are read from the first row, and the number of possible rows for variable types is set to the maximum. The data is separated by semicolons.

Figure 24 (Conversion of Grades to Numeric)

A data step is used to convert the grade variables G1 and G2 from character to numeric format within the student_mat dataset. The original variables are dropped, and the new numeric variables are renamed to G1 and G2.

Figure 25 (Tabulation of Study Time by Sex and Age)

proc tabulate is utilized to summarize the studytime variable, classifying by sex and age. It provides the count (n) and mean study time for each age within each sex category.

Figure 26 (Scatter Plot of Age and Study Time)

A scatter plot is generated with **proc sgplot**, plotting studytime against age, grouped by sex. The plot uses filled circle markers and includes axis labels.

Figure 27 (General Linear Model Tabulation)

proc tabulate creates a more complex table, classifying by sex and Pstatus (parent's cohabitation status), and summarizing the variables absences, G1, G2, and G3 (grades) with the number of observations, mean, maximum, and minimum values.

Figure 28 (Scatter Plot with Grades G2 and G3)

This section uses **proc sgplot** to plot a scatter plot of G2 (second period grade) against G3 (final grade), grouped by sex. It includes a regression line for each sex category and is labeled appropriately.

Figure 29 (Histogram for Final Grade G3)

The final code block creates a histogram for the variable G3 using **proc sgplot**, with a specified bin width of 2. It also overlays a density plot on top of the histogram and labels the x-axis with the variable name Final Grade (G3).

- *Interpretation of Results*

Figure 30 (study time by age and sex)

Observations

1. Girls (F) study for an average of 2.28 hours a week, while boys (M) study for an average of 2.07 hours a week.
2. The average amount of time people spends studying goes down with age, though this varies between men and women. For example, 16-year-old girls learn for an average of 2.47 hours more than 15-year-old girls.
3. The average amount of time spent studying is always low for people ages 21 and 22 (one hour for both sexes). This could mean that fewer people this age are studying, or the data could be affected because the sample size was so small (N=1 for both ages and sexes).
4. At age 16, girls' study for an average of 2.47 hours, and at age 15, boys study for an average of 2.07 hours.
5. The sample size (N) goes down with age, which happens a lot in educational statistics because students drop out or get full-time jobs.

Figure 31 (study time as it relates to age and sex)

In short, the figure shows that study time decreases with age for both men and women. Some reasons for this could be more duties, having a job, or not needing as much structured study time after high school. To give more nuanced views, more information about the dataset and the subjects would be needed.

1. **Axes**: The horizontal axis (x-axis) represents age, ranging from 15 to 22 years. The vertical axis (y-axis) represents study time in hours.
2. **Data Points**: There are two sets of data points, one for females (F) shown in red, and one for males (M) shown in blue.
3. **Trends**: For both sexes, the study time seems to decrease with age. This is indicated by the concentration of higher data points (more study hours) at younger ages and lower data points at older ages.
4. **Sex Comparison**: At certain ages, such as 16 and 18, females appear to have higher study times than males, whereas at age 17, males have a slightly higher study time. However, the differences are not pronounced.
5. **Variability**: There is variability in study time within the same age group, suggesting individual differences among the subjects.
6. **Discrepancies at Older Ages**: For ages 20-22, study time is consistently low (1 hour) for both sexes. This could suggest a lack of data, an error in data collection, or that individuals of this age group generally study less.
7. **Sample Size**: The plot does not provide information on the sample size for each age group, which is important for determining the reliability of the mean study time.

Figure 32 (compares two groups identified by sex (female and male) across various variables)

This table could be used to analyze the impact of gender and Pstatus (possibly parental status or a similar demographic factor) on school attendance and performance.

- **Absences**
  1. Females have a higher mean number of absences (9.70) compared to males (7.63).
  2. The maximum number of absences recorded for females is significantly higher (75) compared to males (30).
- **G1, G2, G3 (Scores in three different assessments)**
  1. The mean scores for females are slightly higher in G1 and G3, but lower in G2 compared to males.
  2. Females have a higher minimum score in G1 and G3, suggesting fewer low scores, but a lower minimum score in G2 compared to males.
  3. The maximum scores for both sexes are relatively close across G1, G2, and G3.

- **Pstatus 'A'**
  1. For the group with Pstatus 'A', females have a lower mean number of absences (5.78) than the overall female group and lower than males in the same Pstatus 'A' group (4.68).
  2. In Pstatus 'A', the mean scores for G1 are higher for females than males, while for G2 and G3, males have higher mean scores.
- **Pstatus 'T'**
  1. This group has more data points (N=185 for females and males), suggesting it is a larger group or more common status.
  2. Absences mean is lower for both sexes in Pstatus 'T' compared to the overall and Pstatus 'A' groups.
  3. The mean scores in G1, G2, and G3 are relatively similar between sexes in Pstatus 'T', with males scoring slightly higher in G1 and G3, and females scoring higher in G2.

Figure 33 (relationship between the grades students received in the second period (G2) and their final grades (G3)

Second period grades are a good predictor of final grades.

There is no immediate visible trend that suggests a difference in this relationship based on sex.

The consistency in performance from G2 to G3 might suggest that the grading is stable and that students maintain their performance level.

Observation

1. **Positive Correlation**: There is a positive correlation between G2 and G3 grades, indicated by the upward trend of the data points and the line of best fit. This suggests that students who scored higher in the second period also tended to score higher in their final grades.
2. **Gender Distribution**: Both female and male students are represented across the range of grades. There does not appear to be a significant difference in performance between the sexes based on this graph alone, as both red and blue dots are interspersed throughout.
3. **Data Spread**: The spread of data points around the line of best fit is relatively tight, indicating a strong relationship between G2 and G3 grades.
4. **Range of Grades**: Both second period and final grades range from low to high (approximately from 0 to 20), showing a wide distribution of student performance.
5. **Line of Best Fit**: The slope of the line is positive and close to being linear, which implies a steady increase in final grade with an increase in the second period grade.

Figure 34 (distribution of final grades (G3) for a group of students)

Professors could use this information to figure out how hard the course is or how well the teachers are doing. It could also point to grade inflation or a problem with the test that makes it hard for students to get the best score possible.

Observation

1. **Skewness**: The distribution of final grades is left-skewed, meaning that there is a tail with lower frequency on the lower grade side, and most students have scored above the midpoint of the possible grades.
2. **Most Common Grades**: The histogram shows that the most common final grades are around the 10 to 15 mark. There is a significant drop in frequency as grades approach 20.
3. **Range of Grades**: Final grades range from 0 to 20, which is likely the full range of possible scores.
4. **Comparison to Normal Distribution**: The normal distribution curve, which represents what the distribution of grades would look like if it were normal (bell-shaped with the mean equal to the median), doesn't fit the actual distribution perfectly. This indicates that the grades are not normally distributed.

- *Code Explanation of student-por data*

Figure 35 (Data Import and Variable Definition and Dataset Structure Display)

The data step reads a CSV file named student-por.csv into SAS, creating a dataset called student_por. Delimiters are set to semicolons, and reading starts from the second row. The length statement defines the variables with their types and lengths, and the input statement specifies how SAS reads each variable from the file. **proc contents** is executed to show the details of the student_por dataset structure, such as the variables it contains and their attributes.

Figure 36 (Average Grades Tabulation by Sex and Age)

proc tabulate is used to summarize the variables G1, G2, and G3 (grades) by Sex and Age, calculating the mean (average) for each combination of sex and age group.

Figure 37 (Scatter Plot of Age and Grade 1)

A scatter plot is created with **proc sgplot**, plotting Grade 1 (G1) against Age, differentiated by Sex. The axes are labeled "Age" and "Grade 1" to indicate what the plot represents.

Figure 38 (Box Plot of Final Grade by Sex)

The final block uses **proc sgplot** to create a box plot of the Final Grade (G3) by Sex. This visualizes the distribution of final grades for different sexes, with axes labeled "Sex" and "Final Grade (G3)".

- *Interpretation of Results*

Figure 40 (SAS dataset, displayed as part of the output from the contents procedure)

This table helps you understand the SAS dataset's structure, size, and properties, which is important for handling data and getting ready for analyses. It can also help find problems with file compatibility or access rights, and it can take a picture of the dataset's contents for documentation reasons.

Figure 41 (list of variables from a dataset, detailing the type and length of each variable)

This kind of variable list is very important for data analysis because it shows what the data in a dataset is about and can help guide the analysis process. For instance, it's important to know the types of variables when picking statistical methods or changing the data to make a graph or model.

Figure 42 (average grades for students across three different grading periods (G1, G2, and G3) by age and sex)

Grade Trends: For both females and males, average grades tend to decrease with age. This is consistent across all three grading periods. The decline in grades appears to be more pronounced for males, especially in the transition from age 18 to 19. For age 22, the average grades for both G2 and G3 show a significant drop, which could be due to a smaller sample size or other factors not specified in the data.

Gender Comparison: At age 16, females and males have similar average grades in G1 and G3, but females have a slightly higher average in G2. In the 17-18 age group, females have higher averages than males in all grading periods. At age 19, the average grades for females are higher in G1 and G3, but males have a slightly higher average in G2. At ages 20 and 21, females have higher averages in all grading periods except for G2 at age 21, where the average is equal for both sexes. By age 22, the sample size is likely very small, as indicated by the absence of data for females in G2 and G3, and a significant drop in males' average grades.

Potential Implications: The overall trend suggests that academic performance may decline with age, which could be related to a variety of factors such as increased academic difficulty, changes in study habits, or external responsibilities The differences in performance between females and males could indicate different learning experiences or external factors affecting educational outcomes.

Figure 43 (students' grades (Grade 1) against their age, with data points differentiated by sex (female in red, male in blue)

1 **Age Range**: The students' ages range from about 15 to 22 years.

2 **Grade Variation**: There is a wide variation in grades across all ages for both sexes. Grades range from the lowest (0) to the highest possible score (20).

3 **Gender Representation**: Both female and male students' grades are represented across the age spectrum. There is no clear pattern that suggests a difference in grades based on sex; both females and males have both high and low grades.

4 **Grade Distribution by Age:** Many students who get grades close to 20 are younger (around 15 to 16 years old). The grades of students get less even as they get older, with some older students (ages 20–22) getting lower scores. There are clear signs of lower grades among students ages 19 to 22. This could be due to a rise in the difficulty of their coursework, changes in their personal lives, or other issues that affect their academic success at this age.

5 **Data Points Distribution**: There are more data points for younger ages, which suggests that the dataset may have more kids who are younger. There are fewer data points for older ages, which could mean that there are fewer students in the group or that more students drop out as age rises.

Figure 44 (comparing the final grades (G3) of students by sex (F for females and M for males)

Both sexes have a similar average performance, as indicated by the mean and median final grades. Male students have a wider range of final grades, indicating more variability in their performance. The presence of outliers suggests that there are both female and male students who have grades that deviate significantly from the norm.

1. **Central Tendency**: The diamonds inside the boxes represent the mean final grades for each sex. It appears that the mean grades are very close for both sexes, possibly indicating no significant difference in average performance between females and males.
2. **Spread of Data**: The boxes, which represent the interquartile range (IQR), show the middle 50% of the data. The IQR for males seems slightly larger than for females, suggesting that male students' final grades have a broader spread.
3. **Median**: The line within each box represents the median (the 50th percentile) of the final grades. The medians are also similar for both groups, which further suggests that the central tendency of final grades is comparable between sexes.
4. **Variability**: The "whiskers" of the box plot (the lines extending from the top and bottom of the boxes) show the range of the data excluding outliers. The whiskers extend further for males, indicating that male students have a wider range of final grades.
5. **Outliers**: The circles represent outliers, which are data points that fall outside the typical range (1.5 times the IQR above the third quartile and below the first quartile). There are outliers on both the high and low ends for both sexes, with males having one more outlier than females.

● *Conclusion of SAS Tutorial for STAT342 Students*

- As we come to the end of this helpful lesson for Dr. Haolun Shi's STAT342 class, it's clear that the trip through **PROC TABULATE** and **PROC SGPLOT** has been both educational and useful! This lesson was carefully made for you! It has made it easy to connect what you've learned in the classroom with what you can use in real life.
- *Adult*, *bank marketing*, and *student performance* datasets have all been great places to learn and find new things. You now understand more about the complicated world of social factors thanks to the *Adult Dataset*. With the help of the *Bank Marketing Dataset*, we can now better understand how bank customers act. In the meantime, the *Student Performance Dataset* has given us a look into how academic success changes over time.
- It's not just the information. This lesson has helped you learn how to use SAS software better. If you've ever been unsure about how to work with datasets in SAS, this lesson has helped you not only get over your doubts but also get better at working with and analyzing data.
- Remember that everything you've learned in this lesson, from loading data to making plots and tables that are easy to read, is a skill that you can use outside of school. These are tools that will help you in your future work, whether it's in business, academia, or study.
- Now that you've learned something, take it with you and feel confident as you go forward. You're not just students learning how to use software; you're also new researchers who are ready to take on data problems in the real world. The field of data is very big and always changing. You are now ready to join this exciting journey. Do more research, think more deeply, and keep growing!