SESUG Paper 220-2023

Visualizing Chronic Lung Disease Incidence in SAS; An Educational Journey to Data Visualization

Natalie A. Jordan, A. Nicole Ferguson, Kevin Gittner; Kennesaw State University; Jessica G. Woo, University of Cincinnati College of Medicine and Cincinnati Children's Hospital Medical Center; Reese H. Clark, Pediatrix Medical Group.

ABSTRACT

Chronic Lung Disease (CLD) in NICU infants is a serious complication of prematurity, and clinicians in the medical field would like to identify infants most at risk based on weight at birth. To address this, birth weight measurements of US NICU infants from the Pediatrix Clinical Data Warehouse born between 2013 and 2018 at 24-33 weeks gestation were classified using percentile cutpoints from published growth curves. Procedures in SAS 9.4 and Microsoft Excel may be used separately or in tandem to identify the incidence of CLD.

To observe differences in CLD incidence rates, the beginner to intermediate level SAS base programmer can generate descriptive statistics and assess the relationship between clinical categorical factors, including gestational age (in weeks), CLD (yes/no), and weight classification. Infants are classified traditionally as small, appropriate, or large for gestational age at birth for weight using available percentile cutpoints. The small and large classifications can both be subdivided into extra and medium small or large for gestational age at birth.

Investigating the relationship between these variables using a PROC FREQ allows users to observe counts and percentages. PROC SGPLOT in SAS generates clustered bar charts with the VBAR option to first group by age. Options in PROC SGPLOT can then be stratified by weight classification using the group option. Clusters are formed per age through the GROUP option. Although PROC SGPLOT shows frequencies by default, percentages can be computed from the frequency table and set as the RESPONSE to show column percentages.

Another way to visualize the relationship for the novice SAS user is to export the three-way frequency table using PROC EXPORT and then create a plot in Excel. This procedure ensures values from a frequency table in SAS export over to Excel. Both programs are used to visualize clinically important relationships for users at different levels of comfort with SAS. The fundamental methods mentioned provide a foundational framework for this clinically relevant issue of identifying which classification of infants has the highest incidence of CLD.

INTRODUCTION

The premise of graph exploration and comparison is to observe how to illustrate data easily and effectively while telling a story to answer research questions. In the medical field, clinically significant questions can be answered through improved data visualizations. Preterm infants are often diagnosed with conditions associated with prematurity like chronic lung disease (CLD). Often, both maternal and neonatal infant factors are related to whether preterm infants develop CLD, such as gestational age and weight at birth. The purpose of this paper is to showcase a visual comparison of SAS and Excel with tables and visuals to answer a clinically meaningful question in a comprehensible way. The beginner to intermediate SAS user as well as the novice SAS user with a strong interest in neonatal and maternal health should walk away with a firm understanding of how relationships between more than two medically relevant categorical factors may be examined descriptively using both SAS and Excel.

DATA DESCRIPTION AND CLEANING PROCESS

For illustrative purposes, we can utilize a subset of birth data from the Pediatrix Clinical Data Warehouse for analysis. Data we collect is on infants admitted to the US neonatal intensive care unit between 2013 and 2018. We can use SAS to import the data using SET statements, apply exclusion criteria, and then classify infant birth weights using published percentile cutpoints. Data can be imported into SAS using a DATA step after we create two separate libraries. First, let's import a copy of the original Pediatrix dataset from its original Excel format into a library. The second library we create stores a personal code file. In addition to the original data set, a second data set - weight - consists of percentile cutpoint values for weight that we can import this dataset using the IMPORT procedure. These are the published percentiles for infant weights in the form of an XLSX file using the following code:

```
PROC IMPORT datafile = "C:\Users\Cutpoint Files\Weight"
     dbms=xlsx
     out = weight;
     getnames = yes;
RUN;
```

Variables we use for analysis include sex (female/male), age (24-33 weeks at birth), weight (a continuous variable measured in grams), and CLD (0/1, where '1' denotes 'Yes' and '0' denotes 'No'). We can categorize weight and age according to clinically important values.

Exclusion criteria are applied using a series of SET statements. Infants missing sex, age, or weight are omitted from our analysis. After we sort by age and sex, to obtain the final analysis data set, we can take the clinical cutpoints and cleaned Pediatrix datasets and merge them. To classify infants, let's use ELSE IF statements to first assign the classifications using clinical cutpoints for weight to create a new character variable, weight_class, which is a birth weight classification variable we create. The LENGTH statement tells SAS this new variable is a character variable with a maximum of 12-character spaces through $12. option, so we can specify the maximum number of character values needed to display length of a variable level name:

```
DATA Merged_Dataset;
   SET Merged_Dataset;
   length weight_class $12;
   if Weight < weight_cut1 then weight_class = "extra small";
   else if Weight < weight_cut2 & Weight >= weight_cut2
      then weight_class = "medium small";
   else if Weight > weight_cut3 & Weight <= weight_cut4
      then weight_class = "medium large";
   else if Weight > weight_cut4 then weight_class = "extra large";
   else if Weight >= weight_cut2 & Weight <= weight_cut3
       then weight_class = "appropriate";
RUN;
```

To assign each character variable to a corresponding character number, we can use IF statements and create a new character variable called Wt_class. Each Wt_class character number is matched to a given character level name:

```
DATA Example_Data;
   SET Example_Data;
   if weight_class = 'extra small' then Wt_class = '1';
   if weight_class = 'medium small' then Wt_class = '2';
   if weight_class = 'appropriate' then Wt_class = '3';
   if weight_class = 'medium large' then Wt_class = '4';
   if weight_class = 'extra large' then Wt_class = '5';
RUN;
```

Then we use the FORMAT procedure and FORMAT statement to apply character names to numbers. This creates a new variable, Wt_class (This is a categorical variable with levels to include extra and medium SGA and LGA cutpoints) which is created by age:

```
PROC FORMAT;
   value $class
   '1' = "extra small"
   '2' = "medium small"
   '3' = "appropriate"
   '4' = "medium large"
   '5' = "extra large";
RUN;


DATA Example_Data;
   SET Example_Data;
   format Wt_class $class.;
RUN;
```

Let's create a new variable, AgeGroup, to divide immature and less immature infants. We can use the following code to categorize age into clinically relevant and similar groups (24-28 and 29-33 weeks at birth). Additionally, note the use of IF statements instead of ELSE IF statements. We use IF statements in this case to have SAS evaluate all IF statements:

```
DATA Example.Data;
   SET Example.Data;
   length AgeGroup $12.;
   if Age < 29 then AgeGroup = "24 - 28";
   if Age >= 29 then AgeGroup = "29 - 33";
RUN;
```

After separating the data into two separate age groups, we can sort the data by the binary outcome of interest, CLD.

PROC SGPLOT

CONTINGENCY TABLES

Contingency tables are one of the most common ways to display relationships of two or more categorical variables. To create visual displays with the SGPLOT procedure in SAS, we may first explore data through contingency tables using the FREQ procedure. In this context, to observe where CLD incidence is the highest within certain age and weight groups, let's create a three-way interaction frequency table. Specifying a WHERE statement limits output to just those with the outcome of interest, which allows the use of the NOROW and NOCOL options to output only the frequency and percent values:

```
PROC FREQ data=Example_Data;
   by AgeGroup;
   tables Agegroup*CLD*Wt_class / NOROW NOCOL;
   where CLD = 1;
RUN;
```

The code above results in a set of frequency tables showing the distribution of CLD incidence across weight class categories stratified by age group with both frequencies and percentages. In each table, the outcome of interest (CLD=1) is in the rows, the five weight classifications (extra small, medium small, appropriate, medium large, and extra large) are in the columns, and a separate table is generated for

each age group. Table 1 shows this distribution for 24-28 week infants and Table 2 shows the distribution for 29-33 week infants.

Frequency Percent	CLD					
	Wt_class					
	Table of CLD by Wt_class					
	Controlling for AgeGroup=24 - 28					
		Wt_class				
CLD	extra small	medium small	appropriate	medium large	extra large	Total
1	331 8.2%	452 11.2%	2,950 72.9%	220 5.4%	92 2.3%	4,045 100.0%
Total	331 8.2%	452 11.2%	2,950 72.9%	220 5.4%	92 2.3%	4,045 100.0%

Table 1 Frequency Table of Distribution for 24-28 Week Infants

Frequency Percent	CLD					
	Wt_class					
	Table of CLD by Wt_class					
	Controlling for AgeGroup=29 - 33					
		Wt_class				
CLD	extra small	medium small	appropriate	medium large	extra large	Total
1	154 7.7%	262 13.1%	1,472 73.4%	86 4.3%	32 1.6%	2,006 100.0%
Total	154 7.7%	262 13.1%	1,472 73.4%	86 4.3%	32 1.6%	2,006 100.0%

Table 2 Frequency Table of Distribution for 29-33 Week Infants

CLUSTERED BAR CHARTS TO SHOW FREQUENCIES

To achieve a desired color scheme to distinguish each weight classification, we can use the REGISTRY procedure to look up available color names the SAS system may provide:

```
PROC REGISTRY list startat = "COLORNAMES";
RUN;
```

Clustered bar charts help with visualizing the relationships between the three variables with more clarity. Let's take a look at aesthetics when making a clustered bar chart. The TRANSPARENCY and DATASKIN options allow us to change the color and transparency of the bar colors. Altering both of these with the DATACOLORS options, we choose BLACK for extra small, DODGERBLUE for small, LIGHTGRAY for appropriate, INDIGO for large, and PLUM for extra large. Note that this will be the color scheme for all visual charts moving forward. The VALUEATTRS option enables us to change the color, font family, font weight, font style, and size for the axis tick-value labels or legend value labels. The LABELATTRS option does the same, but for reference label lines:

```
PROC SGPLOT data = Example_Data;
  styleattrs datacolors = (Black DodgerBlue LightGray Indigo Plum);
   by CLD;
   vbar AgeGroup / group = Wt_class  groupdisplay = cluster
   transparency=0.3
   dataskin=matte name='bar';
         xaxis valueattrs = (size = 10pt color = navy);
```

```
            yaxis valueattrs = (size = 10pt color = navy);
            xaxis labelattrs = (size = 12pt weight = bold);
            yaxis labelattrs = (size = 12pt weight = bold);
RUN;
```

Figure 1 shows the resulting clustered bar charts we obtain from the previous code, with frequency values of CLD for infant weights. These are the same frequencies seen in Table 1 and Table 2.

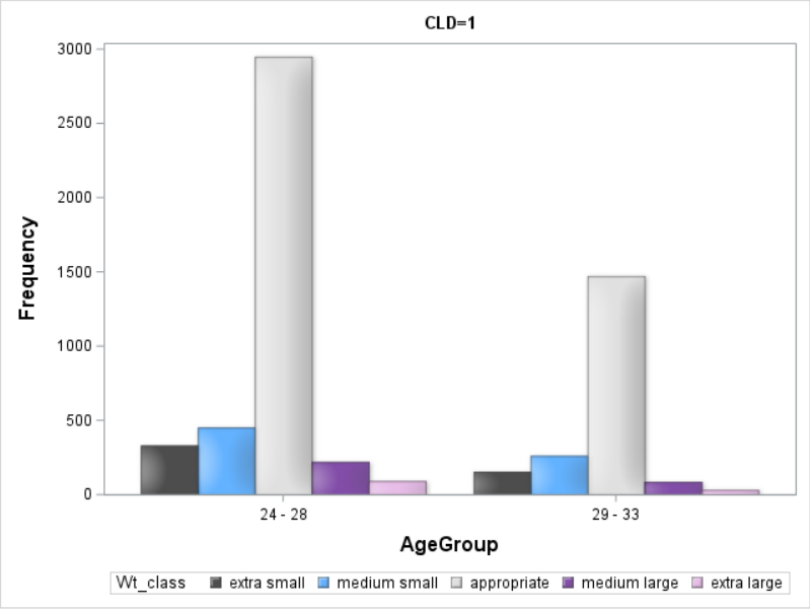

Figure 1 Clustered Bar Chart of CLD Frequencies Stratified by AgeGroup

CLUSTERED BAR CHARTS TO SHOW COLUMN PERCENTAGES

The first analysis above focuses on the distribution across weight classifications, which we observe in Table 1 and Table 2 and Figure 1. Let's now calculate column percentages to compare the percentage of infants within a certain weight classification who have CLD (e.g., the incidence rate). This is a separate analysis which focuses on the incidence of CLD per weight classification (this is seen in Table 4, Figure 2, and Figure 6). To conduct this analysis, let's first run the ODS OUTPUT option within a PROC FREQ. This outputs a readable output dataset which automatically calculates and creates a column percentage variable for each weight classification for infants with and without the condition for each age group:

```
PROC FREQ data=Example_Data;
   tables AgeGroup*CLD*Wt_class / norow nopercent;
   ods output CrossTabFreqs = CrossTabFreqs;
RUN;
```

Table 3 is the SAS view table of the first five observations of the new dataset we create from the ODS OUTPUT option which shows the calculated column percentages as a new variable. This ultimately gives us a glimpse of how incidence of CLD in each of the weight classifications for 24-28 week infants (AgeGroup = 24-28) would look like for CLD = 0 (or no CLD).

VIEWTABLE: Natalie.Colpercent_table

	Table	AgeGroup	CLD	Wt_class	_TYPE_	_TABLE_	frequency	ColPercent
1	Table 1 of CLD * Wt_class	24 - 28	0	eSGA	111	1	6,660	43.63
2	Table 1 of CLD * Wt_class	24 - 28	0	SGA	111	1	14,479	54.98
3	Table 1 of CLD * Wt_class	24 - 28	0	AGA	111	1	179,109	70.07
4	Table 1 of CLD * Wt_class	24 - 28	0	LGA	111	1	18,938	76.78
5	Table 1 of CLD * Wt_class	24 - 28	0	eLGA	111	1	8,882	78.82

Table 3 SAS ViewTable with New "ColPercent" Variable Added

With data summarization complete, let's run PROC SGPLOT from the table we output. Include a WHERE statement to have CLD = 1, the VBAR option to focus on age as the first category (VBAR = AgeGroup), and the weight classifications as the second category (GROUP = Wt_class). Remember that this Wt_class variable contains five levels. To ensure we get a cluster effect, set GROUPDISPLAY = CLUSTER. This allows for weight classifications within each age group to be "clustered" together. The column percentages which we calculate from PROC FREQ may be set as the RESPONSE to show the variable ColPercent from Table 3 which provides column percentages:

```
PROC SGPLOT data=Colpercent_table pctlevel = graph;
   where CLD = 1;
     VBAR AgeGroup/ RESPONSE = ColPercent GROUP = Wt_class
     GROUPDISPLAY = cluster;
     styleattrs datacolors = (Black DodgerBlue LightGray Indigo Plum);
     xaxis label = "Age (Weeks)";
     yaxis label = "Percent Incidence CLD" values=(0 to 100 by 10);
RUN;
```

Here, we now see a trend that infants at earlier ages (24-28 Weeks) have a higher incidence (% occurrence) of CLD (Figure 2). Those classified as small, regardless of age group, have a higher incidence, as well.

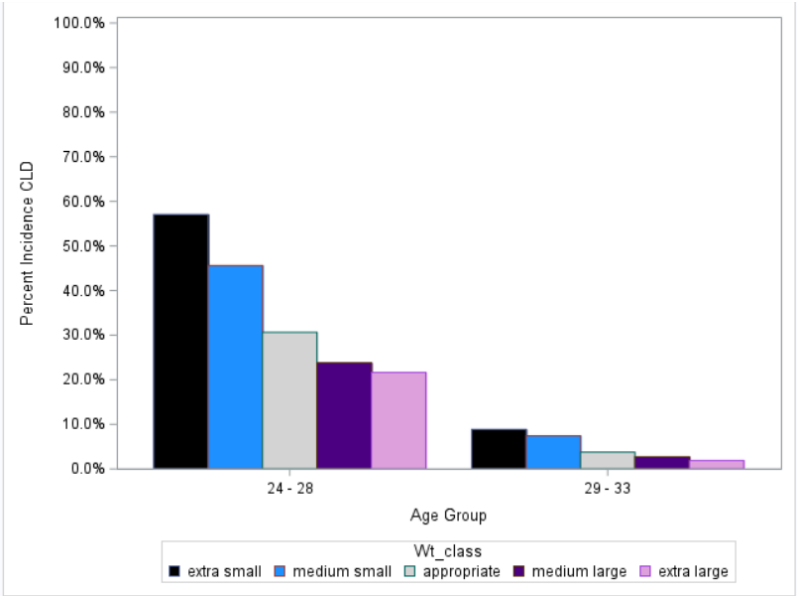

Figure 2 SAS Clustered Bar Charts with CLD Incidence Rate Stratified by AgeGroup

MICROSOFT EXCEL

IMPORTING SAS DATA INTO EXCEL AND CREATING A TABLE

Clustered Bar Charts can also be created in Excel using the frequency tables created in SAS. Using the EXPORT procedure in SAS, let's export the CrossTabFreqs_table as an Excel spreadsheet with the new sheet named "%CLD" using the SHEET option:

```
PROC EXPORT Data = CrossTabFreqs_table
   File = "C:\Users\col%table.xlsx"
   DBMS = XLSX REPLACE;
   SHEET = "%CLD ";
RUN;
```

Once in Excel, we can select insert from the ribbon, and select insert column or bar chart in the charts section. Click on more column charts and select the leftmost clustered column chart that appears. Let's copy and paste column percent values to a new table in Excel which includes two separate columns as the age groups where each row corresponds to a weight classification. We can use Table 4 to create clustered bar charts in Excel.

	AgeGroup	
Wt_class	24 - 28	29 - 33
extra small	56%	9%
medium small	45%	7%
appropriate	30%	4%
medium large	23%	3%
extra large	21%	2%

Table 4 Excel Table of CLD Incidence Rate by Weight Classification Stratified by AgeGroup

CREATING A CLUSTERED BAR CHART IN EXCEL

To create a clustered bar chart in Excel, highlight the content in Table 4 and click insert chart. Then we can select the clustered column option in the Column Section and select appropriate customize chart elements from there (as shown in Figure 3). We can then choose the rightmost figure, as we would like two separate clusters (one cluster is for AgeGroup 24-28 and the second cluster is for AgeGroup 29-33), with each cluster having five separate weight classifications.

Figure 3 Inserting Chart Screen from Excel

To customize the color scheme in Excel, you right click each bar in the clustered bar chart, then select the shape fill option just beneath the format data point option. This process of customizing color is shown in Figure 4 and Figure 5, where the plots pictured have the custom colors after selecting More Fill Colors.

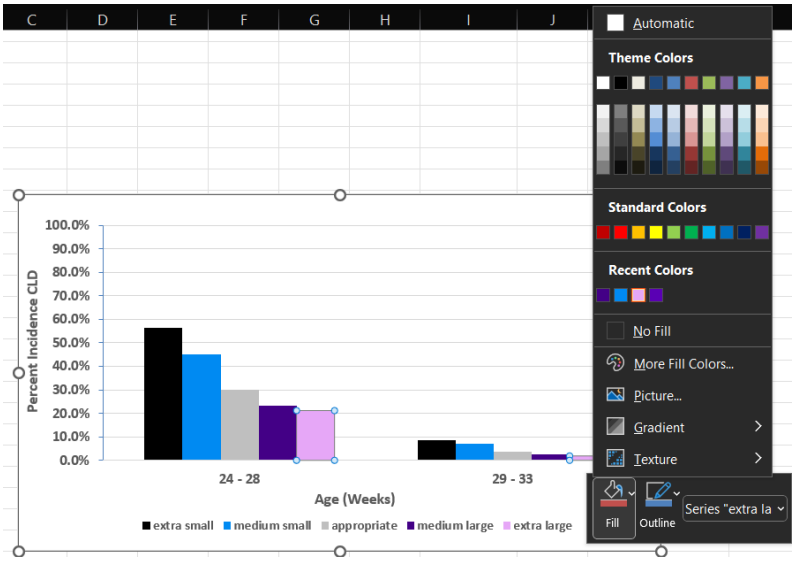

Figure 4 Customizing Bar Colors in Excel

Let's select one of the standard colors to choose from for the extra small weight classification.

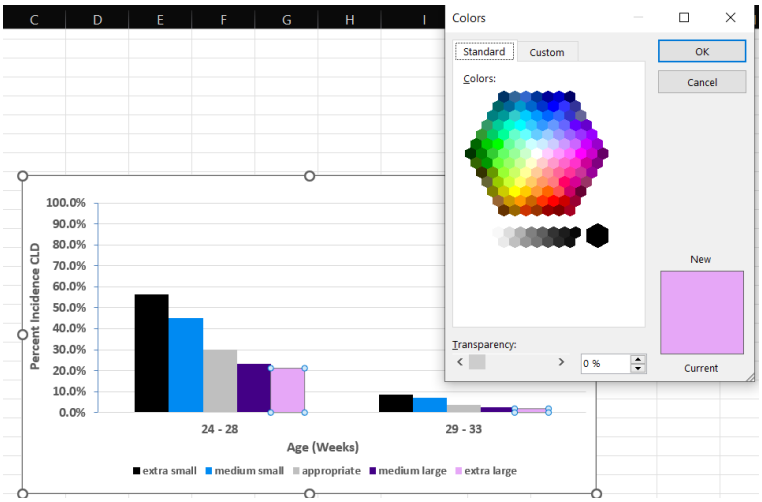

Figure 5 Selecting from the Standard Colors Available

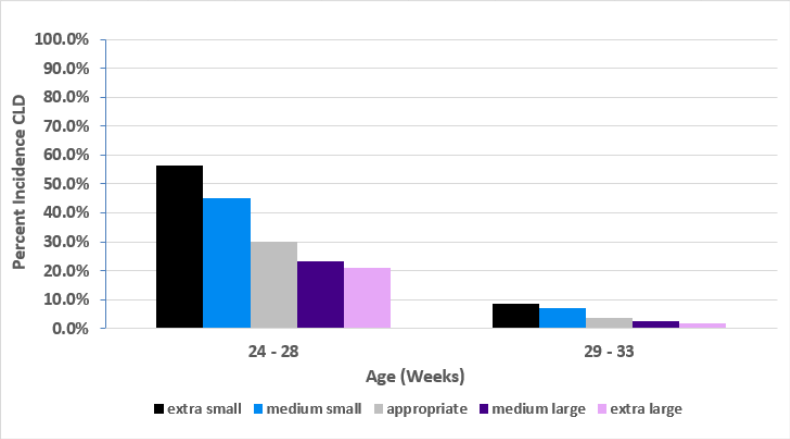

Figure 6 Excel Clustered Bar Chart of CLD Incidence Rate Stratified by AgeGroup

The plot from Excel is similar to the plot created in SAS, but the point-and-click interface may make plot customization easier for novice SAS users.

CONCLUSION

Visualization methods for exploring categorical variable relationships are essential in formulating initial predictions in outcomes research, and specifically for tracking CLD incidence in infants in this case scenario. This is especially true during the exploratory data analysis phase. We can deduce that SAS and Excel plots are similar visually while methods to achieve each with the appropriate column percentages differ. Generating the SAS clustered bar charts involves outputting a table that calculates column percentages. We then use the data from the table to run the clustered bar charts. In juxtaposition, exporting the column percent data to Excel to create the charts may be used with more comfort for the novice SAS user. Both plots identify infants most at risk of CLD since they are the same, and creating these visuals in SAS and Excel both utilize contingency table output of the PROC FREQ. PROC SGPLOT provides a customizable plot of the table for the moderate SAS users, while Excel may be a better option for the novice SAS user. New SAS users who have prior experience with Excel may feel more at ease making quick visual changes on Excel, especially when working on a project collaboratively in a team setting. SAS is useful in that it can process and analyze large data and is a procedural based programming language, which means it’s an ideal interface to apply specific data cleaning methodology.

REFERENCES

Leibel SL, Ye XY, Shah P, Shah V; Canadian Neonatal Network. Chronic lung disease in preterm infants receiving various modes of noninvasive ventilation at ≤30 weeks' postmenstrual age. J Matern Fetal Neonatal Med. 2020 May;33(9):1466-1472. doi: 10.1080/14767058.2018.1519798. Epub 2018 Sep 26. PMID: 30176762.

Olsen IE, Groveman SA, Lawson ML, Clark RH, Zemel BS. New intrauterine growth curves based on United States data. Pediatrics. 2010 Feb;125(2):e214-24. doi: 10.1542/peds.2009-0913. Epub 2010 Jan 25. PMID: 20100760.

ACKNOWLEDGMENTS

I would like to say thanks to my professors and mentors, Dr. Ferguson, Dr. Gittner, and Marion Granger for supporting and being a part of seizing this opportunity and embarking on this journey that includes research and conference prep. I sincerely could not have made it this far without them.

This work was supported by funding from The Gerber Foundation.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Natalie Jordan
Kennesaw State University
470-578-2865
College of Computing and Software Engineering
School of Data Science and Analytics
Njorda19@students.kennesaw.edu