Basic Data Analysis

Using SAS

Dr. Austin Brown

Kennesaw State University

Introduction to Data Analysis

- **Objective**: Learn to generate basic descriptive statistics and visualizations.
- **Importance**: Essential for understanding and summarizing data.

Introduction to Data Analysis

- So far, we have learned how to import, inspect, and perform some basic transformations on data.
- After this is complete, we can now focus on analyzing the data to gain insights and answer questions.
- At a foundational stage, this involves generating descriptive statistics and creating visualizations to summarize and present the data.

Introduction to Data Analysis

- Remember in a previous module on understanding column contents, we learned that we generally have two different types of data:
    - Numeric data: Data that represents quantities or numbers.
    - Categorical data: Data that represents categories or groups.
- This difference is not arbitrary: it has implications for the types of analyses we can perform and methods we have available to us.

Introduction to Data Analysis

- Let's once again use the HEART.csv file to demonstrate some basic data analysis techniques.
- As before, go ahead and upload the file to SAS Studio and then import using the `PROC IMPORT` procedure.
- Now, let's explore some basic analysis of numeric data.

Analyzing Numeric Data

- One of the first steps in analyzing numeric data is to calculate summary statistics.
- In the heart dataset, suppose we want to calculate the sample mean, median, and standard deviation of the `Height` and `Weight` columns.
- While there are several ways to do this, one of the simplest is to use the `PROC MEANS` procedure.
- The `PROC MEANS` procedure provides a concise summary of the data, including the mean, median, standard deviation, and other key statistics.

Analyzing Numeric Data

```
/* Import the HEART dataset */
proc import
  datafile = "HEART.csv"
  out = heart
  dbms = csv
  replace;
  getnames = yes;
run;

/* Calculate summary statistics for Height and Weight */
proc means data=heart mean median std;
  var Height Weight;
run;
```

Analyzing Numeric Data

- Perhaps the most common method for visualizing a numeric variable is to create a histogram.
- A histogram is a graphical representation of the distribution of a numeric variable.
- The widths of the bars represent the intervals into which the data is grouped, while the heights of the bars represent the frequency of observations in each interval.
- It is a quick, visual tool for understanding common and uncommon values in a dataset.
- In SAS, we can use the `PROC SGPLOT` procedure to create a histogram.

Analyzing Numeric Data

```
/* Create a histogram for Height */
proc sgplot data=heart;
  histogram Height;
  title 'Histogram of Height';
run;
```

Analyzing Categorical Data

- For categorical data, one of the most common ways to summarize the data is to create a frequency table.
- A frequency table is a tabular representation of the number of times each category appears in the data.
- In SAS, we can use the `PROC FREQ` procedure to generate a frequency table for a categorical variable.

Analyzing Categorical Data

```
/* Calculate frequency of Weight_Status variable */
proc freq data=heart;
  tables Weight_Status;
run;
```

Analyzing Categorical Data

- A common way to visualize categorical data is to create a bar chart.
- A bar chart is a graphical representation of the frequency of each category in a dataset, similar to a histogram for numeric data.
- In SAS, we can use the `PROC SGPLOT` procedure to create a bar chart for a categorical variable.

Analyzing Categorical Data

```
/* Create a bar chart for Weight_Status variable */
proc sgplot data=heart;
  vbar Weight_Status;
  title 'Bar Chart of Weight Status';
run;
```