Week 3 Overview

We have been focusing on the basics of R and Statistics and it is a good point to transition into data analysis. One of the easiest ways to start analyzing data is by visualizing it. So that is where we are going to focus this week. There are some great packages in R that make visualizing data really easy, ggplot2 is the one we will focus on the most in this course, but there are others that can be used. Ggplot2 is likely the most popular package for data viz in R and being comfortable with what you can plot with it will be important as you move into other courses in the program. Try not to get overwhelmed with the visualization options – we have an entire course in the program focused on visualizations. This course is just meant to be a brief introduction with a deeper dive later in the program.

Contents of the Week

Objectives

After completing this week, you should be able to:

Create histograms and probability plots from a dataset to examine the distribution of variables

Produce descriptive statistics and accurately explain the meaning of those results

Weekly Resources

Comprehensive R Archive Network

R Studio

Sage Publications. (2021). Discovering Statistics Using R

RStudio, PBC. (2021). RStudio Documentation

RStudio, PBC. (2021). R Studio Cheatsheets

Week 3 Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:
- *R for Everyone:* Chapter 7
- *Discovering Statistics Using R*: Chapter 4
- Data-visualization-2.1.pdf

Complete the following:
- 3.1 Discussion/Participation
- 3.2 Exercise

Helpful Sources

- RStudio, Tidyverse. ggplot2

3.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is ggplot2?
2. What is a histogram used for?
3. What is a scatterplot used for?
4. What are the benefits of visualizing data?
5. What is the difference between a good chart and a bad chart? What are some examples of each?
6. What is overplotting?
7. What are boxplots used for?
8. What are density plots and how are they used?
9. What are bar charts most used for?

3.2 Exercise

Complete the following 2 assignments. The first is an R Script in GitHub that you will complete. The second will be a script that you create. You will need to import the data using the below dataset and then answer the questions below.

1. Complete assignment03
2. American Community Survey Exercise

    a. For this exercise, you will use the following dataset, 2014 American Community Survey. This data is maintained by the US Census Bureau and are designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment. For this assignment, you will need to load and activate the ggplot2 package. For this deliverable, you should provide the following - make sure you answer each question asked!
        i. List the name of each field and what you believe the data type and intent is of the data included in each field (Example: Id - Data Type: varchar (contains text and numbers) Intent: unique identifier for each row)
        ii. Run the following functions and provide the results: str(); nrow(); ncol()
        iii. Create a Histogram of the HSDegree variable using the ggplot2 package.
            1. Set a bin size for the Histogram that you think best visuals the data (the bin size will determine how many bars display and how wide they are)
            2. Include a Title and appropriate X/Y axis labels on your Histogram Plot.
        iv. Answer the following questions based on the Histogram produced:
            1. Based on what you see in this histogram, is the data distribution unimodal?
            2. Is it approximately symmetrical?
            3. Is it approximately bell-shaped?
            4. Is it approximately normal?
            5. If not normal, is the distribution skewed? If so, in which direction?

6. Include a normal curve to the Histogram that you plotted.
7. Explain whether a normal distribution can accurately be used as a model for this data.

v. Create a Probability Plot of the HSDegree variable.

vi. Answer the following questions based on the Probability Plot:
1. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

vii. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.

viii. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores. In addition, explain how a change in the sample size may change your explanation?