Introduction to Data Analysis in R

Derek Sollberger

2022-05-31

Contents


```
library("corrplot")
library("forcats")
library("janitor")
library("skimr")
library("tidyverse")
```

Today's Data Set

This survey is the result of a partnership between Axios and Harris Poll to gauge the reputation of the most visible brands in America, based on 20 years of Harris Poll research. From Trader Joe's to Disney, here's how this year's class stacks up.

"Methodology: The Axios Harris Poll 100 is based on a survey of 33,096 Americans in a nationally representative sample conducted March 11-April 3, 2022. The two-step process starts fresh each year by surveying the public's top-of-mind awareness of companies that either excel or falter in society.

"These 100 "most visible companies" are then ranked by a second group of Americans across the seven key dimensions of reputation to arrive at the ranking. If a company is not on the list, it did not reach a critical level of visibility to be measured."

Since we have CSV files (comma-separated values), we can use the `read_csv` function in the `readr` package to load the file into our current programming session.

```
reputation <- readr::read_csv("reputation.csv",
                              name_repair = janitor::make_clean_names)
reputation_wide <- readr::read_csv("reputation_wide.csv",
                                   name_repair = janitor::make_clean_names)
```

Exploring the Data

At this point, you can look at the `environment` pane (upper-right area in RStudio), and click on the name of a data frame (e.g. `reputation`) to open up a viewer to take a look at the data.

Another way to take a quick look at the data is with the `head` command to view the first few rows.

For our purposes, we should look at the structure of each data set. In R, this is processed with the `str` command. In particular, this view clarifies which columns are numerical and which are categorical.

Sample Statistics

To get a sense of the numbers, we can compute sample statistics (such as the mean, median, and standard deviation) for a numerical variable.

The `R` programming language becomes really useful when we want to perform our tedious calculations across several categories. Notice how the inclusion of one line of code below helps us compute the sample statistics for each survey category.

In today's exploration, I am interested in the possible differences in the survey scores across the industries.

Data Visualization

Histogram

A histogram allows us to visualize the distribution of one numerical variable.

Scatterplot

A scatterplot allows us to visualize a pair of numerical variables.

Boxplot

A boxplot allows us to visualize numerical distributions across a categorical variable.

Let us see what happens if we explore the `industry` categorical variable.

To start to ease the complexity, we can (for example)

- focus on two of the categories
- remove the legend