Problem Set 5

Samuel Hunt

March 4, 2025

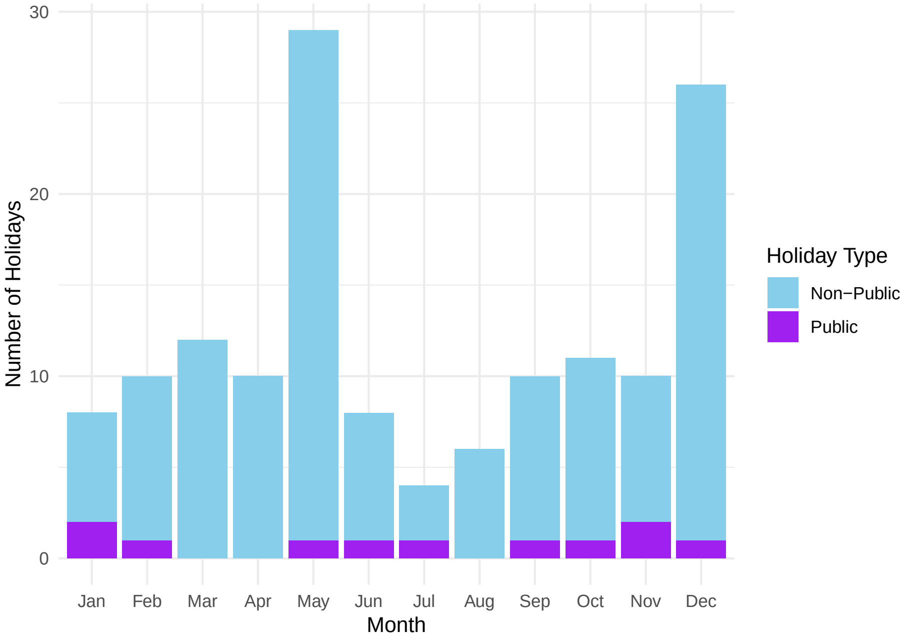
1 Webpage Scraping

The data I chose to scrape for this Problem Set is a ranking of the top Economics Ph.D programs. My intended final project for this class will involve American tariff datasets, however, I found that the type of data that will end up being useful for that project is not typically of the form that scraping would make sense for. Instead, I decided that I would scrape a rankings for Economics Ph.D programs as that is going to be useful to me in the future. When I apply for these programs sometimes in the next few years I will likely use some spreadsheet to keep track of my progress, and so this tabular ranking that updates as the website changes will be useful to me. I originally intended to use the USA today programs rankings, but I had difficulty in my program processing the website, possibly because the website is too big or because it has some process to prevent web scraping. Instead, I used RePEc's Economics departments rankings, which ranks economics departments on the quality of their publications. This table is slightly less useful for my purposes than the USA today rankings, however an initial look shows that the two rankings are highly correlated. This data was very simple to scrape and convert into a neat data frame object, and I did not use any tutorials besides ChatGPT to help create the code.

2 API Data

For the API Data collection, I decided that I wanted to obtain a dataset that contained information on every holiday in the United States. I was interested in this data because I felt that the early year months of January, February, and March must have the least number of holidays in the year and I wanted to see if that was true. The website Holiday API had easily accessible API keys for its holiday data sets, so I used it to collect this data. When I created a graph that showed the number of holidays per month, I was surprised to see the summer months were actually those with the least. I was also very interested to see that May had the most holidays. To accomplish this I used the httr, jsonlite, and xml2 packages to obtain the data from the API, and tidyverse, lubridate, and ggplot2 to create the graph. I again used ChatGPT and Claude to generate the code. On the next page I have attached the graph.

Number of Public and Non−Public Holidays in the U.S. by Month


Number of Holidays
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Holiday Type
Non−Public
Public