Tanner Sax

Capstone Project – Data Wrangling

The data for this project was collected from a web scrape on Glassdoor.com. A search for ‘Data Analyst’ was performed for each of four major tech cities in the northeast – including New York, Boston, Washington D.C., and Philadelphia. The data consists of the job title, company name, location, and a full job description from the Glassdoor job postings. The web scrape collected the data for each city separately and stored it in a data frame. There were several job description links that were unable to be scraped due to ads. These values were set to return NA in the data frames. The rows containing NA in the job description column were then removed using the complete.case() function.

*New\_York <- New\_York[complete.cases(New\_York), ]*

Since there were only 10 NA’s total in 3,555 results, removing them did not have a large impact on the sample.

The analysis requires keyword extraction to identify the most common programming language skills and degree fields. Separate columns were created that that identify keyword matches through binary variables. This was done through the mutate function for each of the keywords.

*New\_York <- New\_York %>%*

*mutate(python = ifelse(grepl("python", job\_description, ignore.case = TRUE), 1, 0))*

The grepl function searches for a match and returns a logical vector. If it returns TRUE, then according to the ifelse statement, a ‘1’ will be returned. Otherwise, a ‘0’ will be returned.

The means of these columns then represent the percentage of job postings that mention the keyword. Then two separate tidy data frames were created for the language skills and degree fields with these values. To have a tidy data frame, vectors were created for the cities and languages/fields that placed them in the correct order with the means. The vectors were then combined into a tibble data frame.

*language\_df <- tibble(cities1, language, Percent\_of\_jobs\_language)*

*degree\_field\_df <- tibble(cities2, degree\_field, Percent\_of\_jobs\_degree)*

The first data frame includes the city, programming language, and percent of job postings that mention the programming language. The second data frame contains the city, degree field, and percent of job postings that mention the degree field.