Introduction

The purpose of this project is to research and analyze specific skills companies are looking for when hiring a Data Scientist. While some of these skills can be assumed, programming, analysis, statistics, which skills do we need to make sure we stand out from the pool of applicants when seeking a Data Science job.

Research Question

What are the most valued data science skills?

Group Members & Roles/Contribution

We are a group of 4. Below is a list of all group members, their role in the project, what they contributed, and their Rpub link to show their individual work that helped to complete this project. All of our work is stored in a group GitHub Repo

- Salma Elshahawy - Project Manager of the team. Responsible for scraping Data Scientist job postings from ai-job.net. Managed group GitHub repo. Rpubs link
- Maryluz Cruz - Responsible for combining all Data Scientist job posting data from SimplyHired, Monster, and ai-job.net into one dataframe for analysis. Created visualization using ggplot, tidywords, kable, and WordCloud. Rpubs link
- Suwarman Sufian - Responsible for scraping Data Scientist job postings from Monster. Cleaned data using REGEX and created visualizations using ggplot. Rpubs link
- LeTicia Cancel - Responsible for scraping Data Scientist job postings from Monster and for the group documentation. Rpubs link

Implementation

In order to make sure we had a large sample size of data from multiple sources, most of the team was responsible for scraping data from a different website. This data was then combined into one R dataframe and analyzed for the frequency of major hard and soft skills keywords. This project is made of 3 major sections - Data Collection, Cleaning the Data, and Text Analysis and Visualizations.

Data Collection

1. First, we went to our assigned website and searched Data Scientist positions. The main URL on the first page of the search results was used as our base URL in our R code.
2. We found the html_nodes for each piece of data needed to pull and added this to the R code. The html_nodes allowed R to locate the job title, company, location, and job description for each job listed.
3. A For loop was used for a set number of web pages using the URL from step 1 and the html_nodes in step 2. The data found in each loop was then dumped into a data frame.
4. A csv file was created for each website and all files were uploaded to GitHub.

Cleaning the Data

1. After all of the Data Scientist job data was collected and posted to GitHub, we could then write R code that pulled each CSV file in the individual GitHub repos.
2. Once the data from each file was collected, the column names had to be modified. Although the data in each file was the same, we needed to make sure there was uniformity in the column header labeling.
3. The job descriptions were then cleaned by removing unwanted words and characters such as punctuation marks, beginnings of URLs (HTTP text), and stop words[1].
4. The words remaining in each description became the main words needed for our analysis.
5. We decided to take the tf-idf approach for our word analysis. Tf-idf measures how frequently a word occurs in a document. Since we are looking for soft skills we want to measure the frequency of words such as collaborative, creative, organized, leadership, and teamwork.

Text Analysis and Visualizations

After stop words, punctuation marks, and other unnecessary text was removed, we began our text analysis by checking the frequency of each remaining word.

1. The first round of analysis checked the frequency of each word. This allowed us to see the words that were used the most, and likely to be the most important, and also the words used the least.
2. From this initial analysis, we then created visualizations to show to top hard skills and the top soft skills based on the frequencies.
3. Using ggplot we created a bar chart to see the top 8 hard skills terms mentioned in all job posters were Modeling, Machine Learning, Statistical, Statistics, Programming, Quantitative, Regression, and Debugging.
4. A ggplot bar chart was also used to see the top soft skills. The top 8 soft skills terms mentioned in all job postings were Research, Communication, Passion, Visualization, Critical, Collaborative, Creative, and Leadership.

Conclusion

Based on the text analysis from Data Scientist job postings from 4 different websites, we were able to see the top skills companies are looking for when hiring a Data Scientist. Companies are looking for Data Scientist candidates with skills in modeling, statistics, programming, debugging, regression, and quantitative skills. Companies are also looking for Data Scientist candidates with research skills, ability to communicate, passion, are collaborative, creative, and also have leadership skills. As students seeking a Data Science degree, I think the assignment was a

good way to get a broad look at what is needed beyond what we are learning in class in order to successfully find work in this field.

Further Expectations

I believe a deeper analysis can be conducted with the job posting data we collected. It would be interesting to see this information grouped by industry. I would also like to see if the location makes any difference, both locally in New York City, and across all states. Some websites also offer salary ranges. For the data that includes salary, it would be interesting to see the difference in salary based on the top skills soat after by companies. This salary data could also be included in the location analysis.

Notes:
[1] Stop Words are words such as “can”, “the”, “is”, “etc”. These words are important within the job descriptions but are not needed in our word frequency analysis.