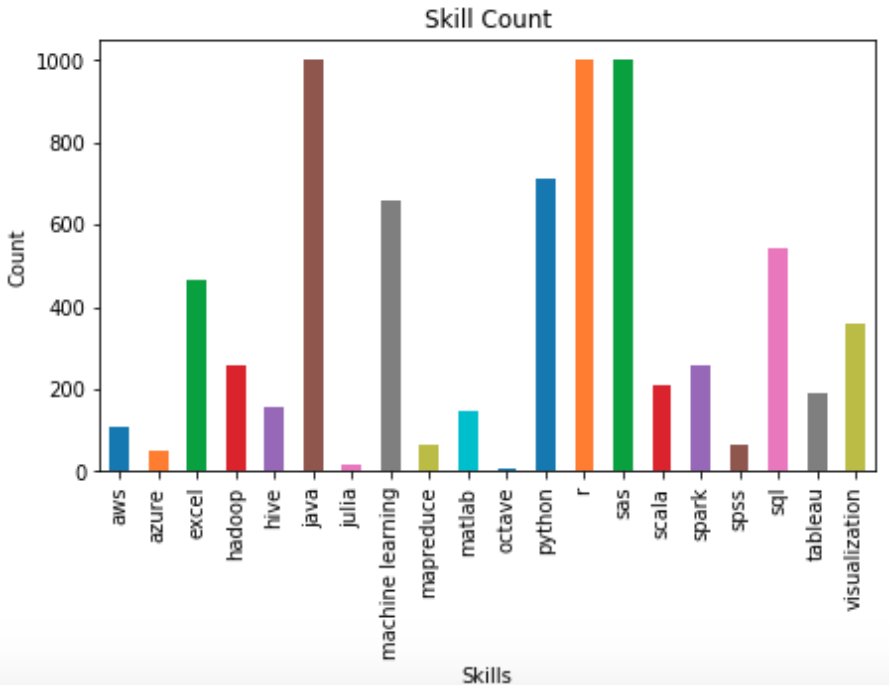
Pratik Patel
Lab4

Exploratory Data Analysis on Data Scientists

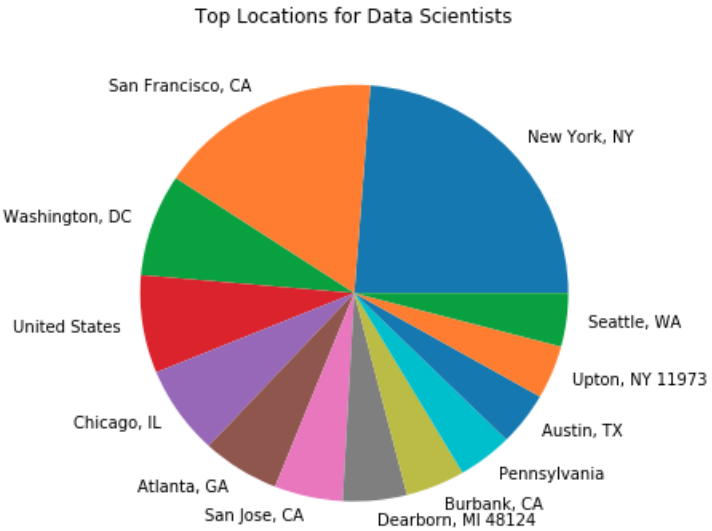
With the invention of the Internet in the late 1960's, a great change occurred in the world which not only affected people's daily lives but affected business too which is the reason the current era is called the Information age. Nearly all devices are connected to the internet nowadays and those devices produce massive amounts of data so that business' can learn more about the users to better themselves. According to IBM roughly 2.5 quintillion bytes of data is produced every day. With so much data being produced the need of data scientists who can analyze and extract meaningful insights from enormous amounts of data has vastly grown.

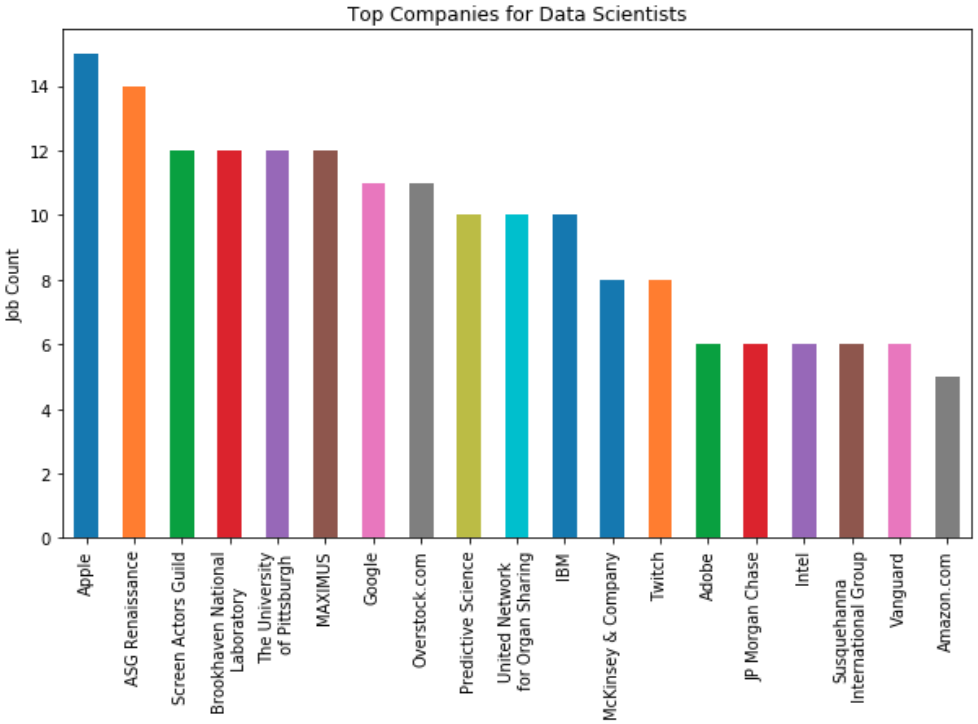
Using the indeed website, a popular site to find jobs, we can analyze all the different data scientist job offers out there and come up with some meaningful insights just like how data scientists do. The indeed website was used to scrape data for 1000 job offerings looking for the title, company, location, and 20 different skills needed for a data scientist. One of the trends seen is that most popular skills required for a data scientist is to know how to use R, Python, Java, and SAS as well as known machine learning algorithms seen from basically every job offering containing those keywords, which makes sense as those are currently the most popular languages used by data scientists today. The



graph also shows how Julia is not really a skill required yet with a count of less than 20 as it is a newer language compared to the others, with it first appearing in 2012 and only recently releasing a stable version in 2017, so there are still some bugs.

Another insight drawn from the data which is an important factor to finding a job is the location of where the jobs are being offered. The location of a job is typically one of the important decision-making factors when accepting or rejecting a job and most of the time job seekers are willing to move to get better job offers. The following pie chart shows the top locations with job offers for data scientists. New York city has the highest job offerings with San Francisco, Washington, D.C , Chicago, and Atlanta being the next locations with highest data scientist positons which is believable as all places are populous cities where there are more companies. Looking closer to the pie chart it can be concluded that most of the positions are offered in California due to San Jose and Burbank being in the chart which is reasonable as that is where Silicon Valley is.





The reason why those locations are where most data scientist jobs are also has to due to the fact that the big technology companies are located there like Google (CA), Apple (CA), IBM(NY) and etc. which all are looking for data scientists seen from the bar graph.

Overall, it looks like an indiviudla pursing a career focused just on data scientist needs to learn the top languages/softwares data scientist currently use like R and Python and need to be willing to live in populous citis in the United States like New York and San Fransciso.

Websites

https://medium.freecodecamp.org/which-languages-should-you-learn-for-data-science-e806ba55a81f

https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/