**APAN 5205 Project deliverable 2**

Statement of the Problem

Our team would like to understand the data science related job market in the United States. How many data related positions is each company hiring for a given year? Which state or city is in high demand for data science related positions? What are the popular job titles that companies use in their job postings? What are the qualifications that companies have in mind for their ideal candidates? Our project aims at answering these questions by performing analysis on job data retrieved from Indeed in 2018.

The choice of analytical technique.

We have chosen text mining techniques that we learnt in this term because the majority of our data columns are non-numerical. As mentioned in the deliverable 1 report, we started with deleting missing values to ensure there is no blank in any columns because there are a few missing values in the dataset and these missing values are texts. Then we take advantage of Wordcloud to find words with high frequency in the position of data science. The most frequent words are scientist, machine learning, analyst, engineer and manager. And we reclassify positions based on frequency of words in the position data science. In addition, we reprocess data about job descriptions. Initially, we clean job descriptions by deleting extra spaces, excessive punctuation marks, url links, non-english words and new lines. In addition, we remove meaningless words, such as stopwards and job position from job description and conduct lemmatization by grouping the inflected form of word together. Finally, we explore and analyze the differences of job positions in analytics tools, skills and degrees.

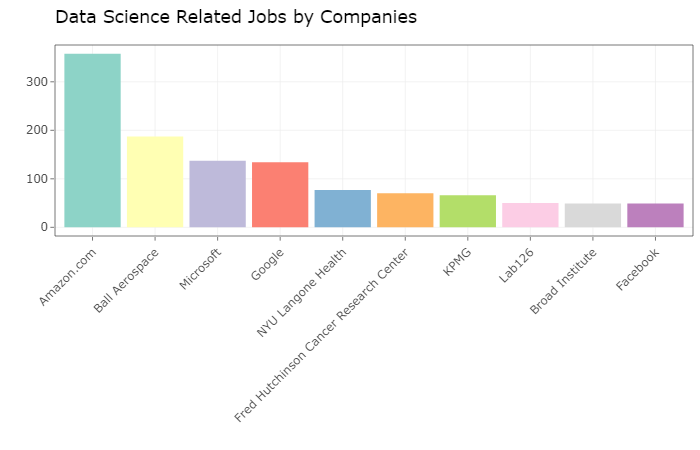
Results

There are many common job titles that data science related positions use. We started the analysis from a preliminary exploration on our dataset to determine which job title should be our main focus.

Question 1 How many data science related positions does each company hire per year (2018)?

Amazon has the highest demand for data related positions among all companies in the United States by posting 358 jobs on Linkedin. Ball Aerospace takes the second place, with approximately half of the number of jobs that Amazon provides. Microsoft and Google each provide about one thirds of the number that Amazon does.

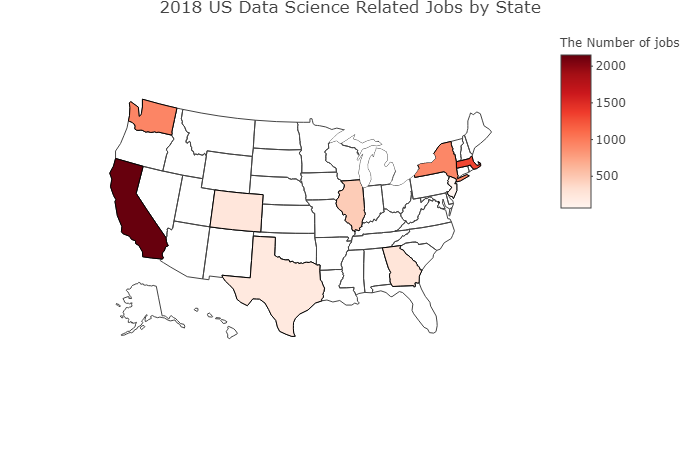
Figure 1 Data Science Related Jobs by Different Companies



Question 2 What is the spatial distribution of data-related jobs(by state) in the United States?

As we can see in the map below, data-related jobs are concentrated on the east and west coasts, with California and Washington state for the west and Massachusetts and New York state for the east.The distribution of data-related jobs is highly correlated with the economy: states with strong economies have more jobs.

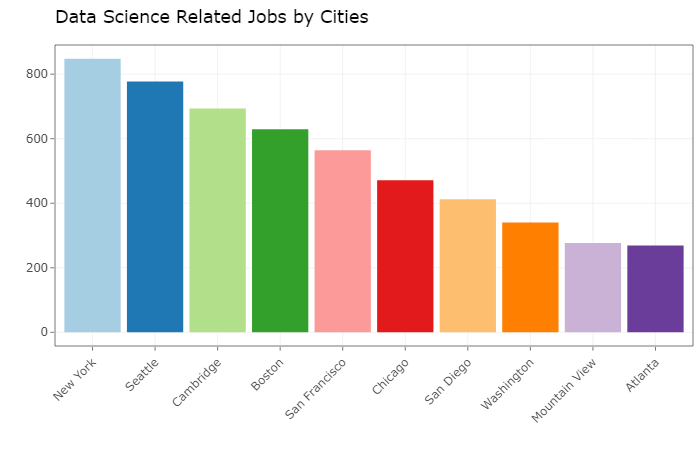
Figure 2 2018 US Data Science Related Jobs by State



Question 3 Which city is in high demand for data related positions?

New York has the highest number of data science related positions across all cities in the United States with over 800 job openings. Seattle, the IT hub on the west coast takes the second place with more than 700 openings. Cambridge, a city known as the home of two top rated universities in Massachusetts, ranks number three on the list.

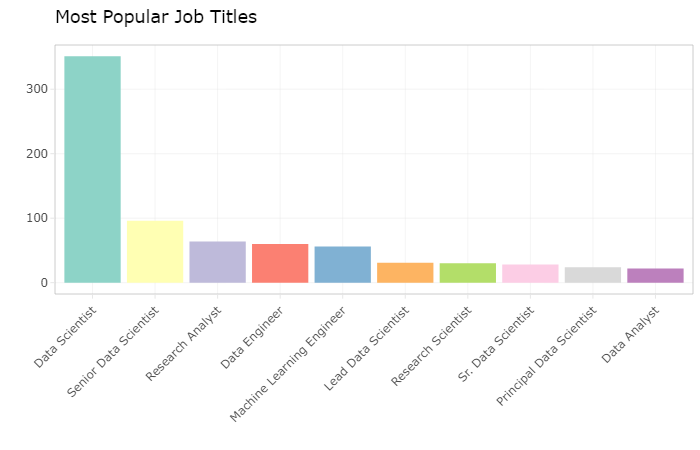
Figure 3 Data Science Related Jobs by Cities



Question 4 What are the popular job titles that companies use in their job postings?

As we can see from the bar chart below, ‘Data Scientist’ is the most popular job title among all data science related positions. Companies also use ‘analyst’ in the title in many of the job postings.

Figure 4 Most Popular Job Titles



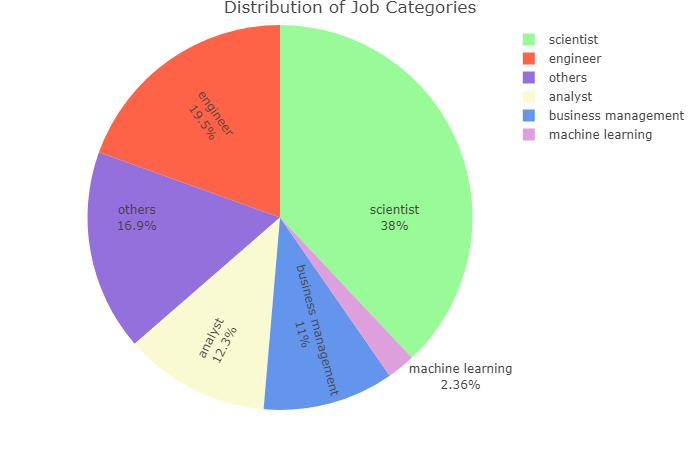
We have decided to categorize data science jobs to understand the qualification required for each job category. We believe that a word cloud is the best way to see frequent words that appear in job positions. We can see the following words stand out from the cloud: data, scientist, analyst, engineer, research, senior and learning.

Figure 5 Word Cloud for Frequent Words in Position column



According to the word cloud, we have categorized all jobs into six categories: Scientist, Engineer, Analyst, Business Management, Machine Learning and Others. We would be analyzing the five categories except “others” for the purpose of this project.

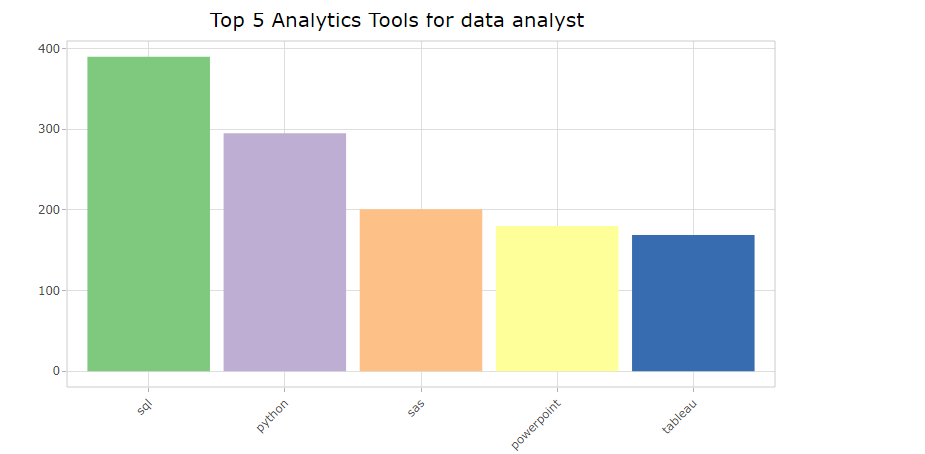
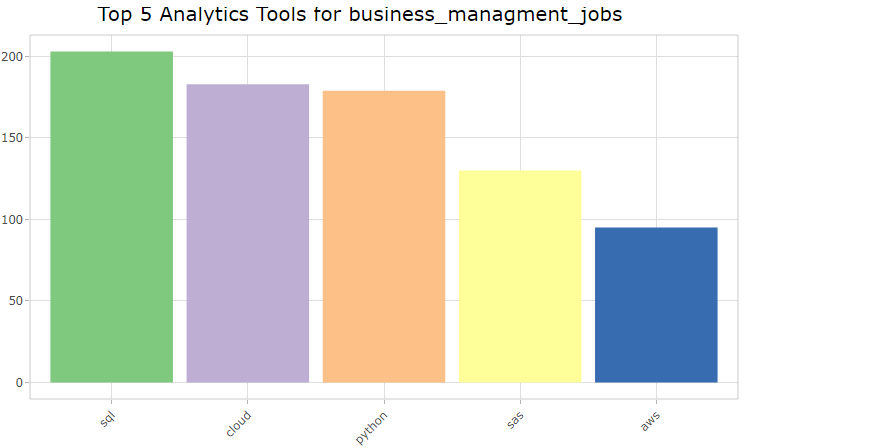
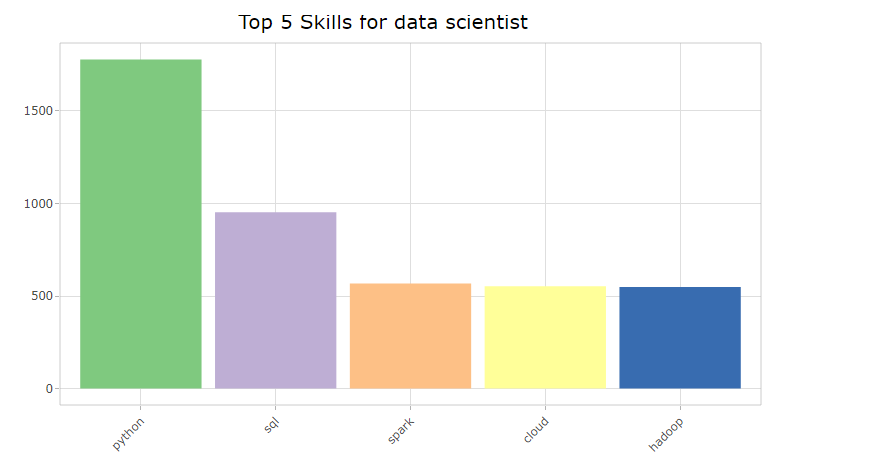
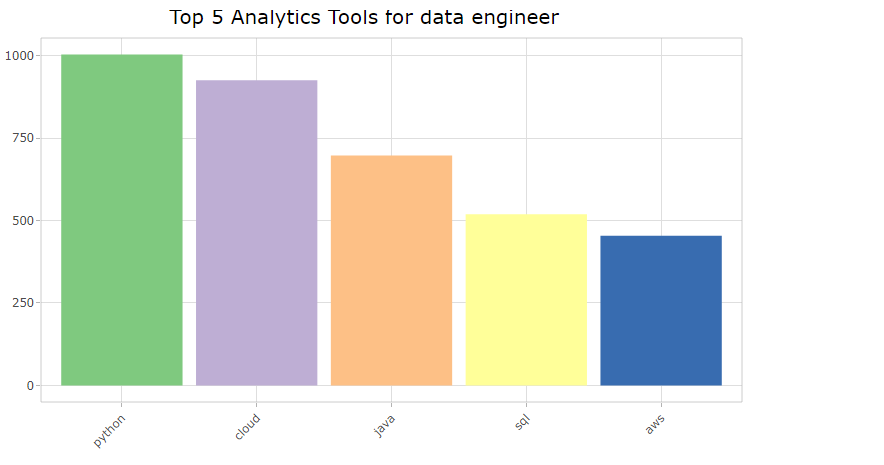
Figure 6 Distribution of Job Categories

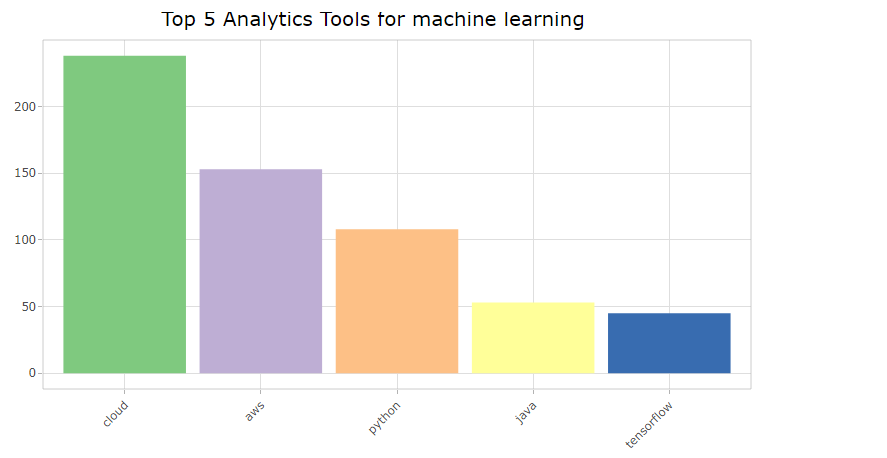


Question 5 What are the top analytic tools that each job category is looking for?

We have compared the top analytical tools that each job category requires by looking at the word frequency. Python is among the top 5 skills for all categories, followed by SQL, Cloud, and SAS. Data Analyst would require proficiency with visualization tools such as Tableau and business communication tools such as PowerPoint. Data Engineer seems to be a more technical position that requires Java proficiency.

Figure 7-11 Top Analytic Tools for 4 Job Categories

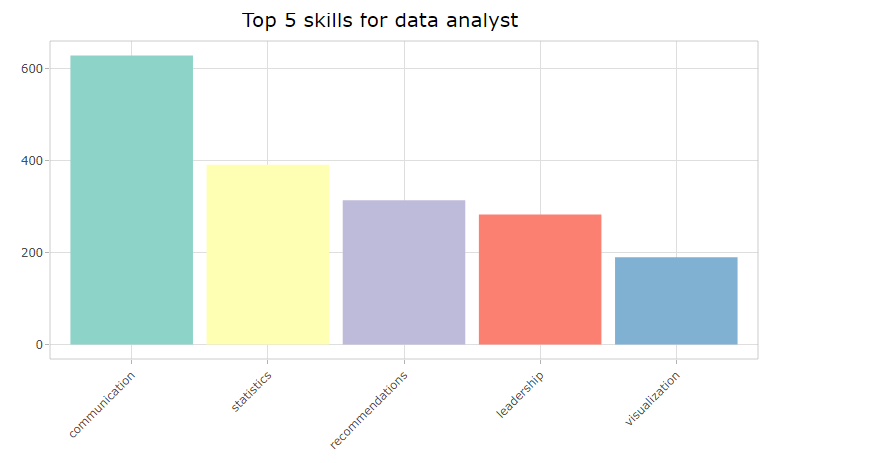
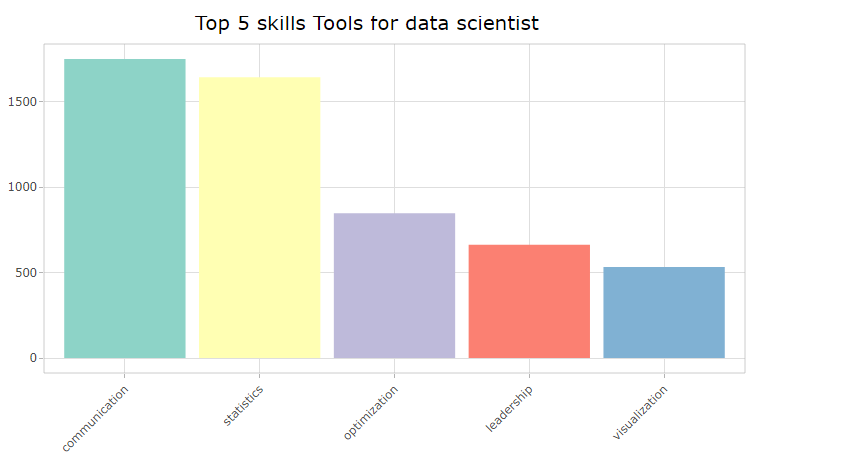


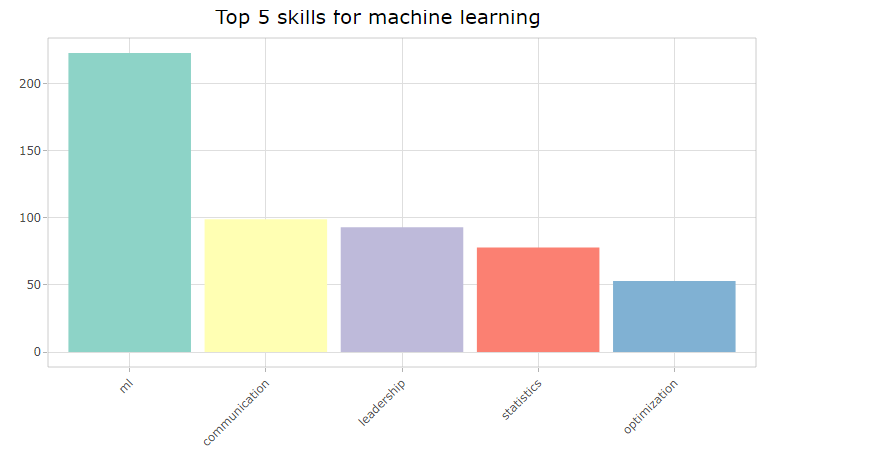
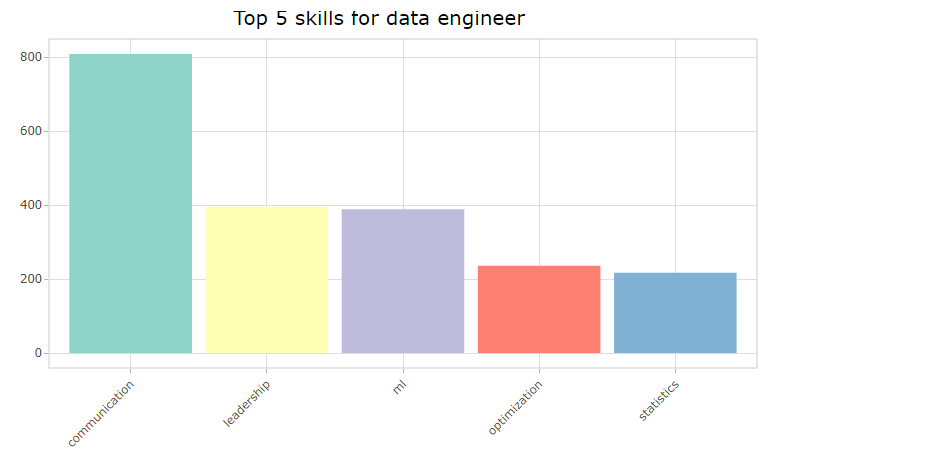


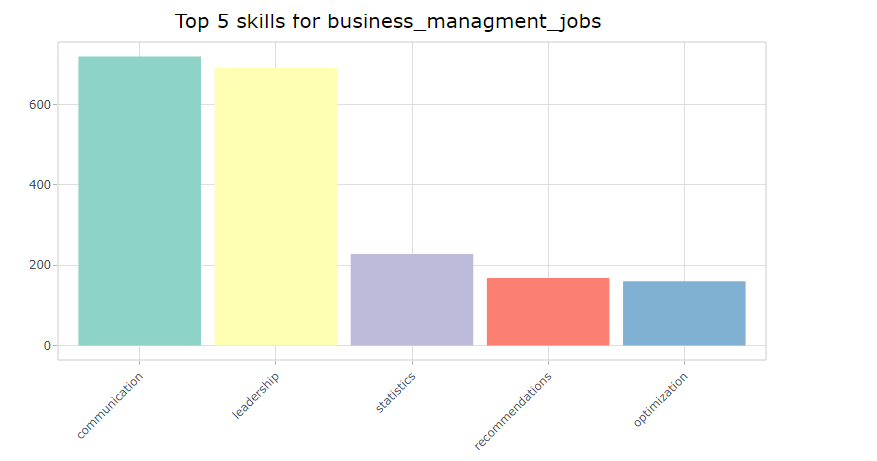
Question 6 What are the top skills that each job category is looking for?

As we can see from charts below, communication is the key skill for all job categories since all these positions need to present results to others. Leadership and statistics are also common skills that all categories require. Data engineers and Machine Learning related positions will require additional hard skills such as understanding of machine learning concepts and model optimization.

Figure 12-16



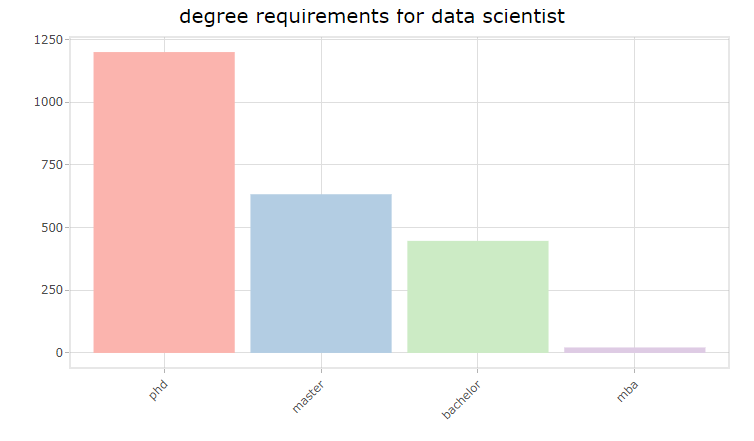


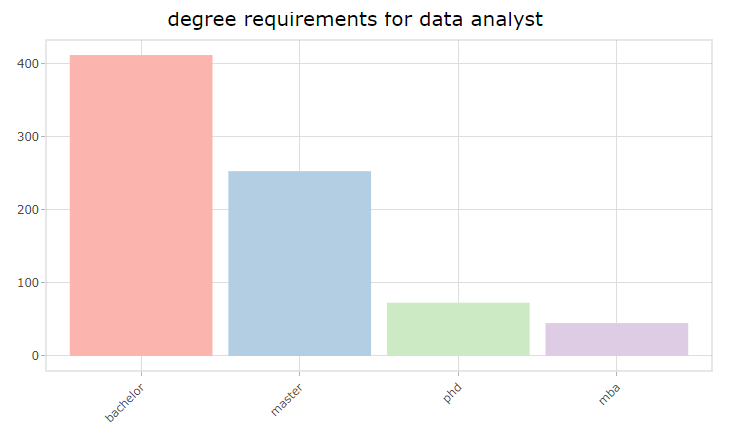


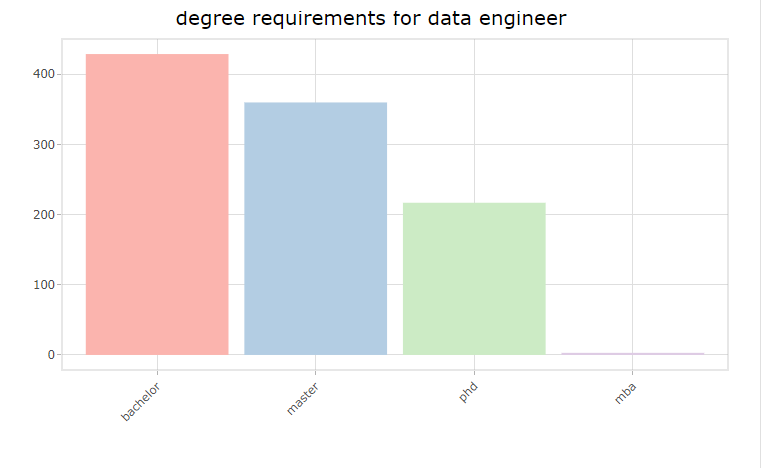
Question 7 What are the ideal educational background that each job category is looking for?

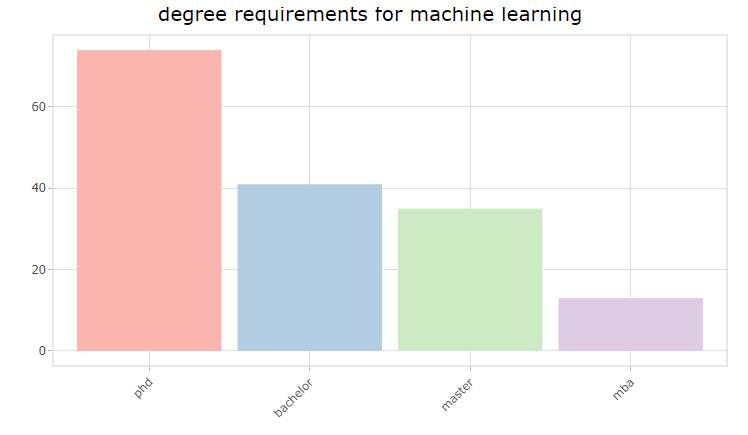
Majority of technical positions such as Data Scientist and Machine Learning jobs require a PhD or master degree. On the other hand, Data Analyst, Business Management Jobs and Data Engineer only require a Bachelor’s degree.

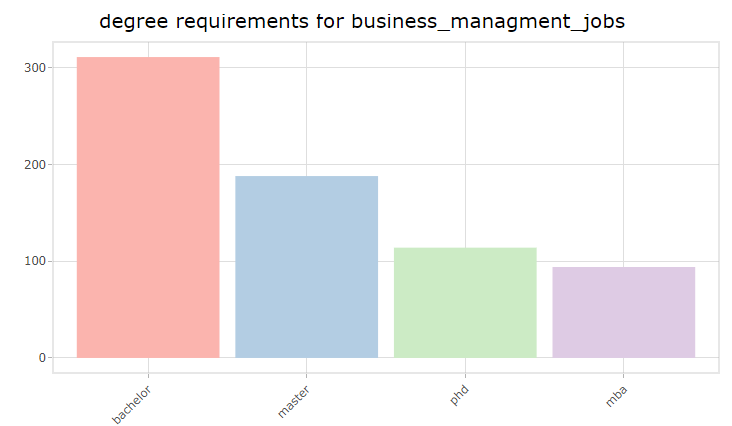
Figure 17-21











Conclusion

Our project will provide insights for people on job seeking. Geographically speaking, if you are interested in getting a data science related job in the United States, then consider the following states: Massachusetts, New York, California and Washington. There are more positions in these areas so better chances for you to find a suitable job. Big technology companies have more open positions so job seekers should pay attention to the job alert from those companies. No matter what type of data science related job you are seeking for, you would need to have proficiency in Python and SQL, plus excellent communication skill. Additional skill requirements will vary based on the detail of the job posting. You would need a PhD or at least a masters degree to get a job as a data scientist or machine learning professional, but for other positions a bachelor's degree may be enough.