29.05.2020

QMBU450 FINAL PROJECT

MERT ÖZLÜTIRAŞ

53952

Retrieving Data Science Job Postings Insights to Help Job Seekers Break Into the Industry

As a person who would like to transition to data science field, I wanted to analyze job postings in the field to retrieve valuable insights about the path to follow to break into the industry. The main aim of this exploratory analysis is to understand what employers are seeking in the candidates and where/how to find a job in the easiest way in the field.

The data that I used for the analysis can be found here: kaggle

Although the data is from August 2018, as the data science job market is an immature one, I do not think that the nature of the industry has changed drastically ever since.

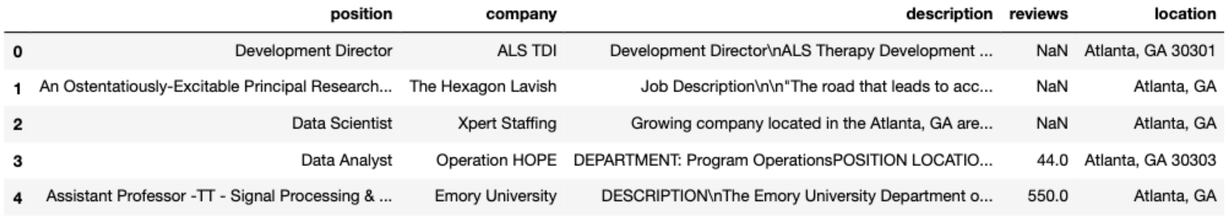
1. Environment & Libraries

Jupyter notebook was selected as I think it is a very suitable tool for making many visualizations represented at the same notebook separately.

Pandas was used for data manipulation and preparation. Seaborn and matplotlib was used extensively for data visualization. Wordcloud was used to generate word clouds to better visualize what is expected from the candidates applying for the data science related positions.

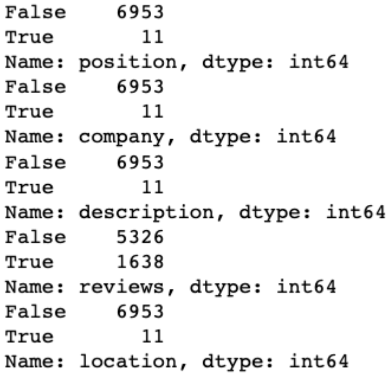
2. Data Preparation

I have looked at the data's head to see the columns and rows we have.



	position	company	description	reviews	location
0	Development Director	ALS TDI	Development Director\nALS Therapy Development ...	NaN	Atlanta, GA 30301
1	An Ostentatiously-Excitable Principal Research...	The Hexagon Lavish	Job Description\n\n"The road that leads to acc...	NaN	Atlanta, GA
2	Data Scientist	Xpert Staffing	Growing company located in the Atlanta, GA are...	NaN	Atlanta, GA
3	Data Analyst	Operation HOPE	DEPARTMENT: Program OperationsPOSITION LOCATIO...	44.0	Atlanta, GA 30303
4	Assistant Professor -TT - Signal Processing & ...	Emory University	DESCRIPTION\nThe Emory University Department o...	550.0	Atlanta, GA

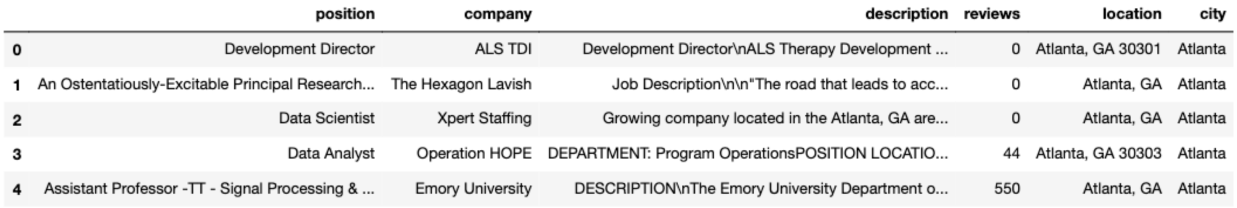
NA values for each column was checked. Position, company, description and location columns all had 11 NA rows, except reviews. Looking at the first rows of data, we can see that it is understandable some new companies to not have reviews. However, old institutions such as Emory University has many reviews.



```
False    6953
True       11
Name: position, dtype: int64
False    6953
True       11
Name: company, dtype: int64
False    6953
True       11
Name: description, dtype: int64
False    5326
True     1638
Name: reviews, dtype: int64
False    6953
True       11
Name: location, dtype: int64
```

NA values in review column were filled with 0s, and the rows where position, company, description or location is NA are dropped.

To generate further insights from the data, I also wanted to have a city column which was generated as the first element of the location column.



	position	company	description	reviews	location	city
0	Development Director	ALS TDI	Development Director\nALS Therapy Development ...	0	Atlanta, GA 30301	Atlanta
1	An Ostentatiously-Excitable Principal Research...	The Hexagon Lavish	Job Description\n\n"The road that leads to acc...	0	Atlanta, GA	Atlanta
2	Data Scientist	Xpert Staffing	Growing company located in the Atlanta, GA are...	0	Atlanta, GA	Atlanta
3	Data Analyst	Operation HOPE	DEPARTMENT: Program OperationsPOSITION LOCATIO...	44	Atlanta, GA 30303	Atlanta
4	Assistant Professor -TT - Signal Processing & ...	Emory University	DESCRIPTION\nThe Emory University Department o...	550	Atlanta, GA	Atlanta

3. Visualizations for Total DS Related Jobs

a. Total Job Postings by Company

A valuable sub-task of this project could be to identify which companies have most job postings in data science field. I grouped the total number of job postings by company and plotted through using seaborn. We see that tech companies such as Amazon, Google, Microsoft are ranked best. However, there are also more traditional firms which are looking for more and more employees in data science field, such as KPMG and McKinsey & Company. This shows us that, tech companies are the early adaptors but the other companies are adapting data science to their business and require employees accordingly.

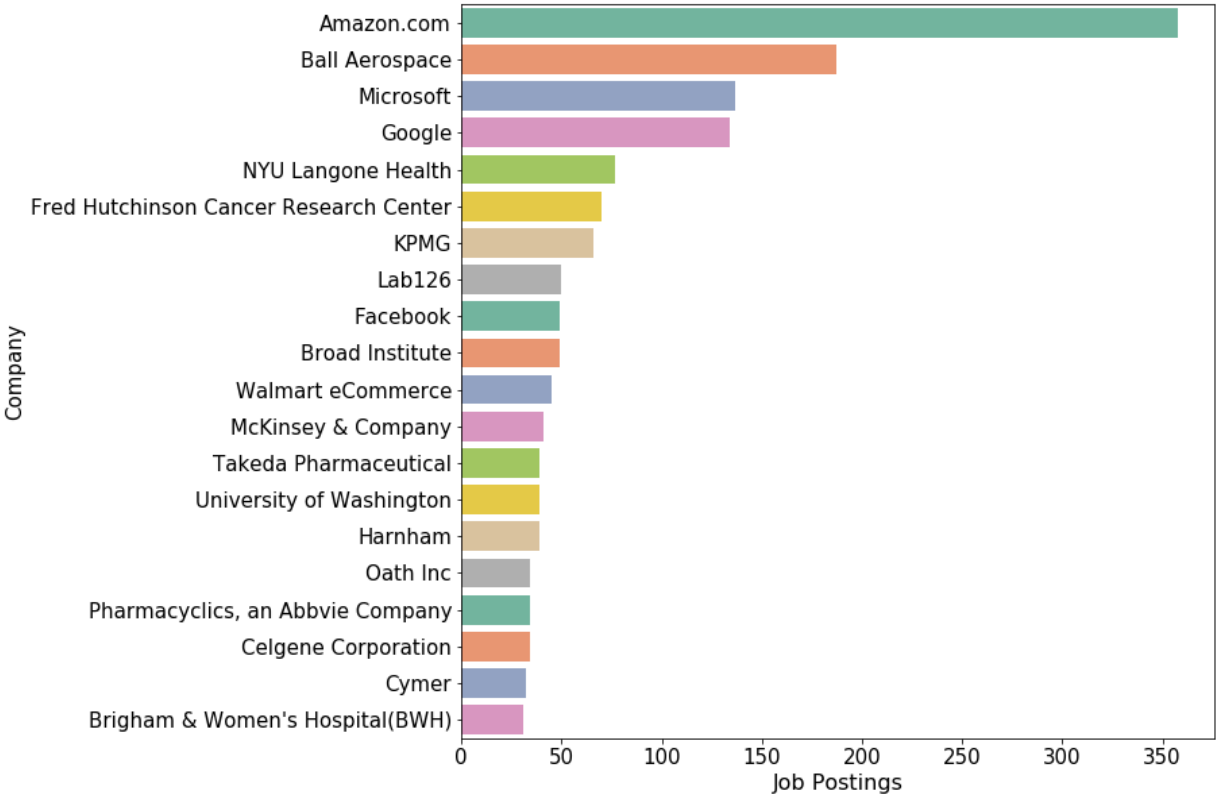


Figure 1: Top 20 Companies by Number of Job Postings

b. <u>Total Job Postings by Position</u>

To see what positions are posted most in the job opportunities, I grouped job postings by the position name. Visualized top 10 positions. We see that 70% of job postings for top 10 positions is specifically data scientist jobs (46% data scientist + 13% senior data scientist + 4% lead data scientist + 4% sr. data scientist + 3% principal data scientist). It shows us how heavily the industry relies on data scientists.

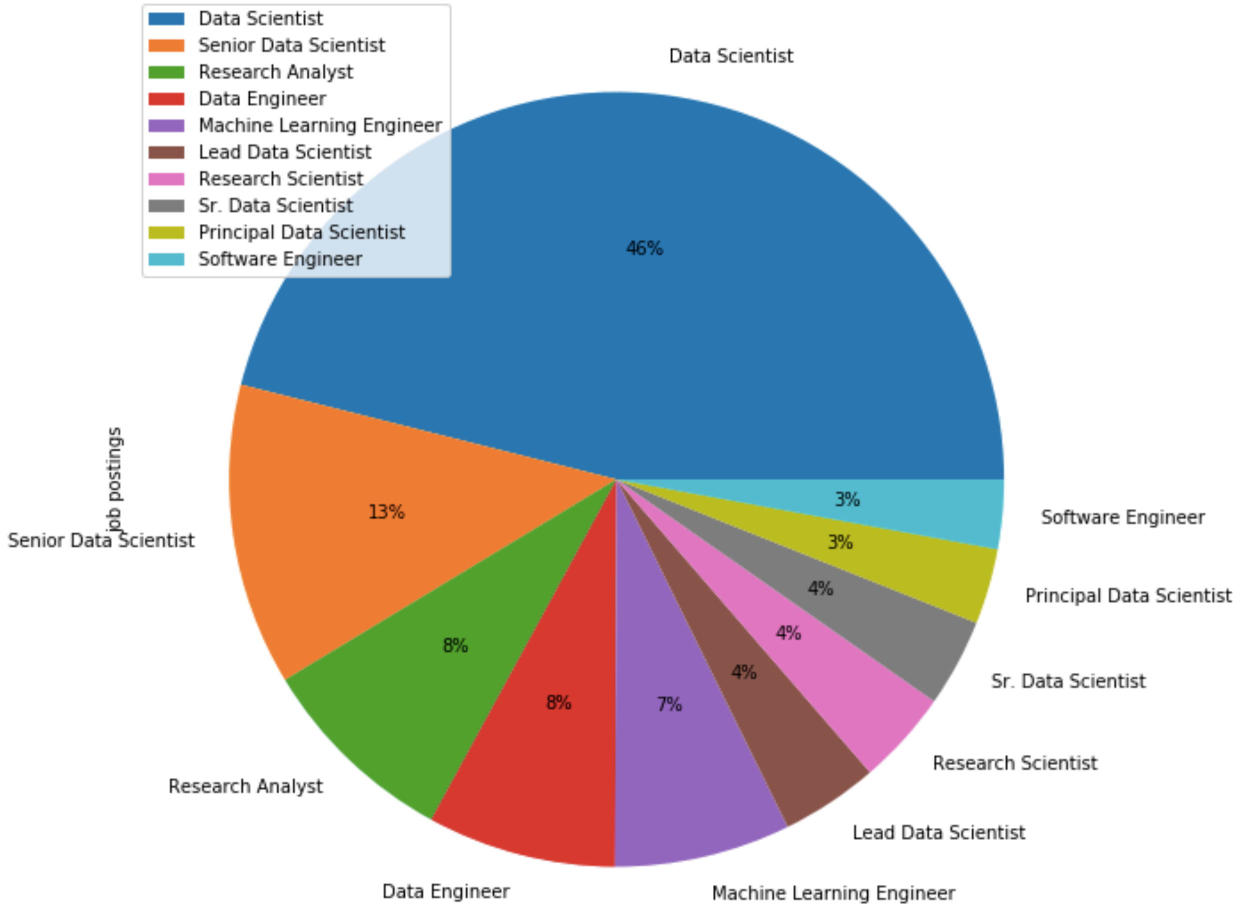


Figure 2 Pie Chart for Top 10 DS Related Positions' Job Postings

To see the absolute job posting volumes for Top 10 positions, I have also included a bar plot visualized using seaborn. We see the same ranking here, this time based on absolute volumes.

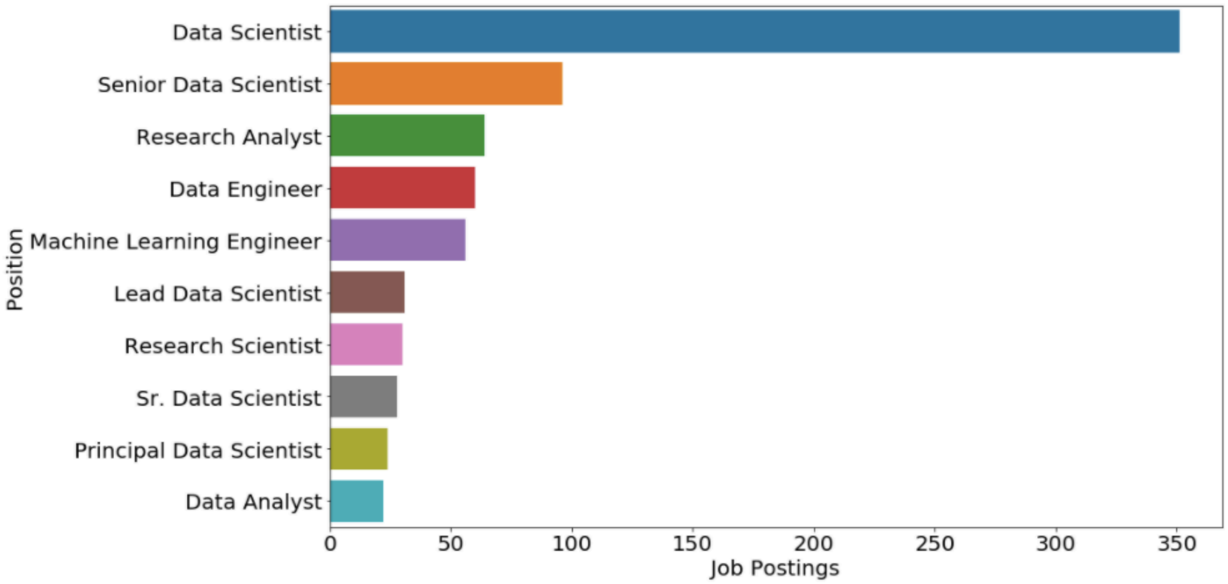


Figure 3 Bar Plot for Top 10 DS Related Positions' Job Postings (top10)

c. Total Job Postings by City

To better understand to where to look at for a job hunter looking for a job in United States, I also looked at total data science related job postings by city. In addition, I thought it would create an important insight on whether data science positions are only existent in metropolitan cities, or also in rural areas. We see that most DS related jobs are in larger cities such as New York, Seattle, Boston, San Francisco etc. It shows us that to maximize odds for getting a DS related job, one should look at larger cities.

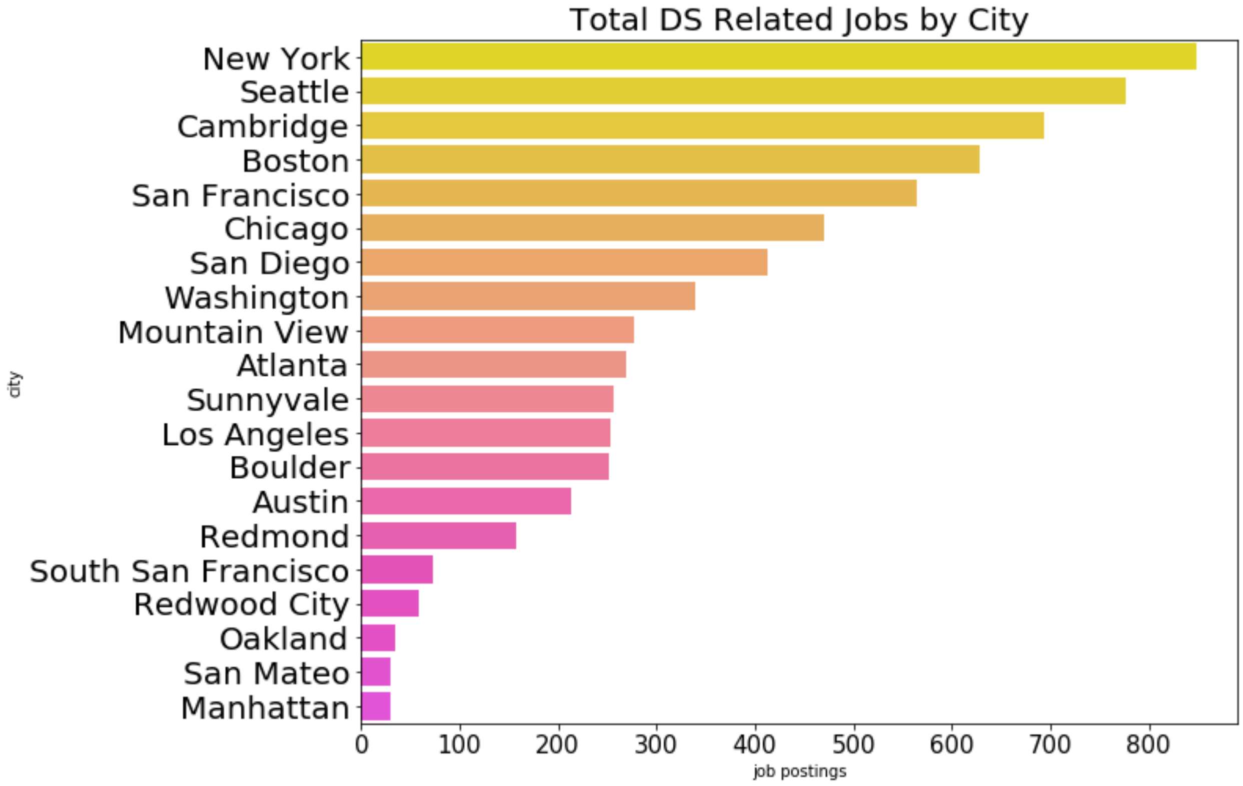


Figure 4 Total DS Related Jobs by City (top 20)

4. Visualizations for Specific Positions

I have analyzed 4 most common group of DS related jobs in depth. These are data scientist, machine learning engineer, data engineer and data analyst. I have looked again most popular cities and required skills for the specified positions.

a. Data Scientist Jobs

I grouped “data scientist”, “senior data scientist”, “lead data scientist” and “sr. data scientist” positions as “data scientist”; since they are of the very same nature. Similar to DS related jobs, Data Scientist jobs are also mostly in the larger cities. Here, the difference between large and small cities is even more significant.

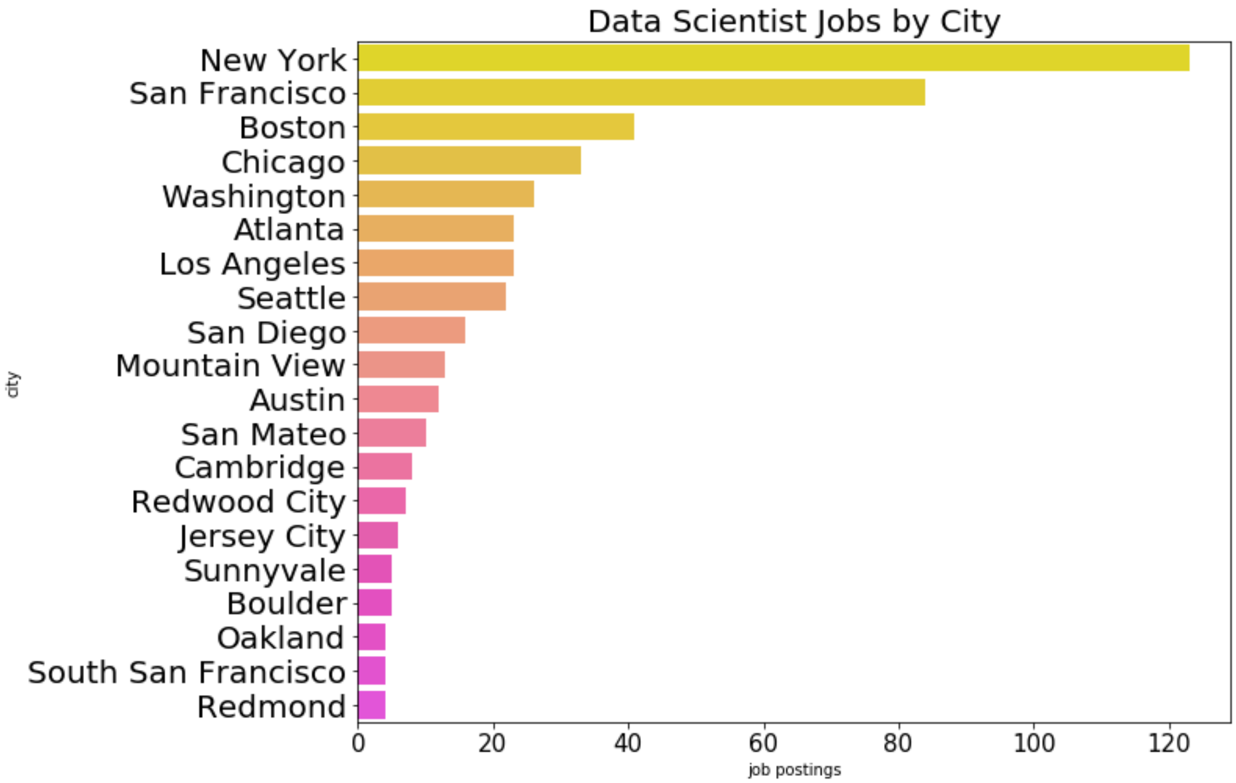


Figure 5 Data Scientist Job Postings by City

To further analyze the skills required for each job position (data scientist, data analyst, data engineer, ml engineer) I have defined a function named "desc_to_text". This function is used to extract words used in the descriptions of all data scientist job postings. Then, this list is given to wordcloud as an argument to create a word cloud showing most important skills. We see that "experience", "machine learning", "data science", "years of" , "ability to", "product", "project" are the most outstanding keywords for data scientists.

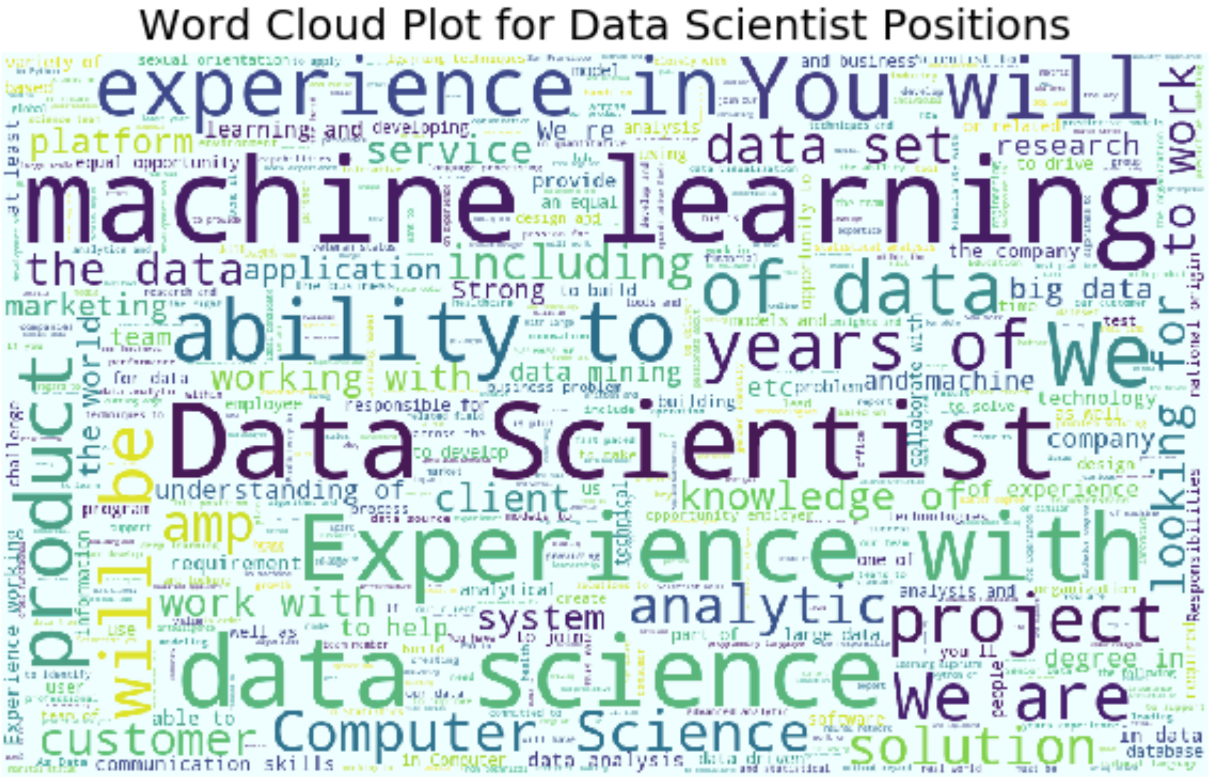


Figure 6 Word Cloud of Skills Required for Data Scientists

Lastly, what percentages of data scientist jobs required the most common knowledge skills were explored. These skills are Python, SQL, Java and Machine Learning. The percentage was calculated as the division of the number of job postings containing the specified keywords (e.g. “python”, ”Python”, “PYTHON”) by total number of job postings. Surprisingly, it’s seen that Machine Learning as at least as important as Python for data scientists. Expectedly, Java is not very essential for a data scientist, existing only in 25% of job postings.

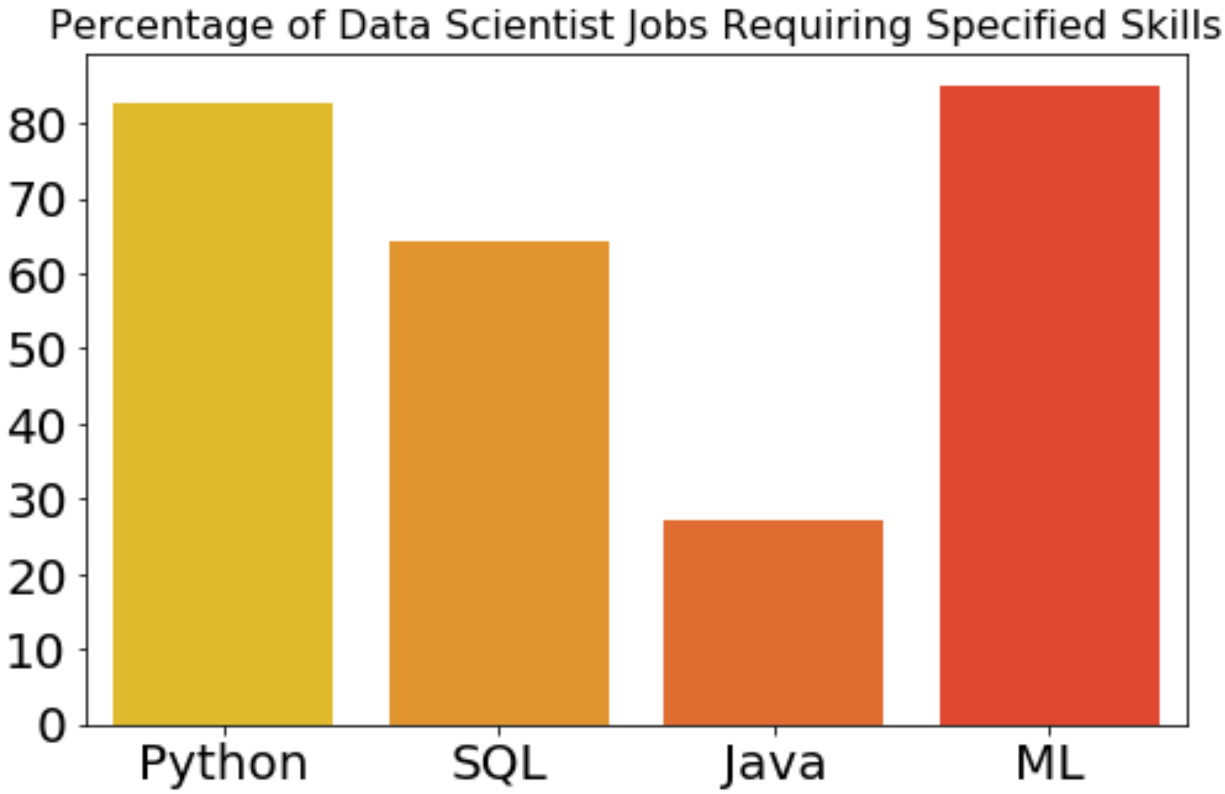


Figure 7 Percentage of Data Scientist Jobs Requiring Python, SQL, Java & ML

b. Machine Learning Engineer Jobs

Similar methods were used to analyze ML Engineer jobs as in Data Scientist jobs. Only the position named “Machine Learning Engineer” was considered for this job position. The volumes of ML Engineer jobs were plotted by city. As we have seen in the beginning of our analysis, this dataset doesn’t contain a large number of ML engineer jobs. This could be a limitation for this part of the analysis. I wanted to put top 20 cities with most ML engineer job positions. However, there are only 14 cities in our dataset with ML engineer job postings. Still, we see that large cities offer most of the job opportunities in this field too. San Francisco offers more job postings than New York for the first time in our analysis.

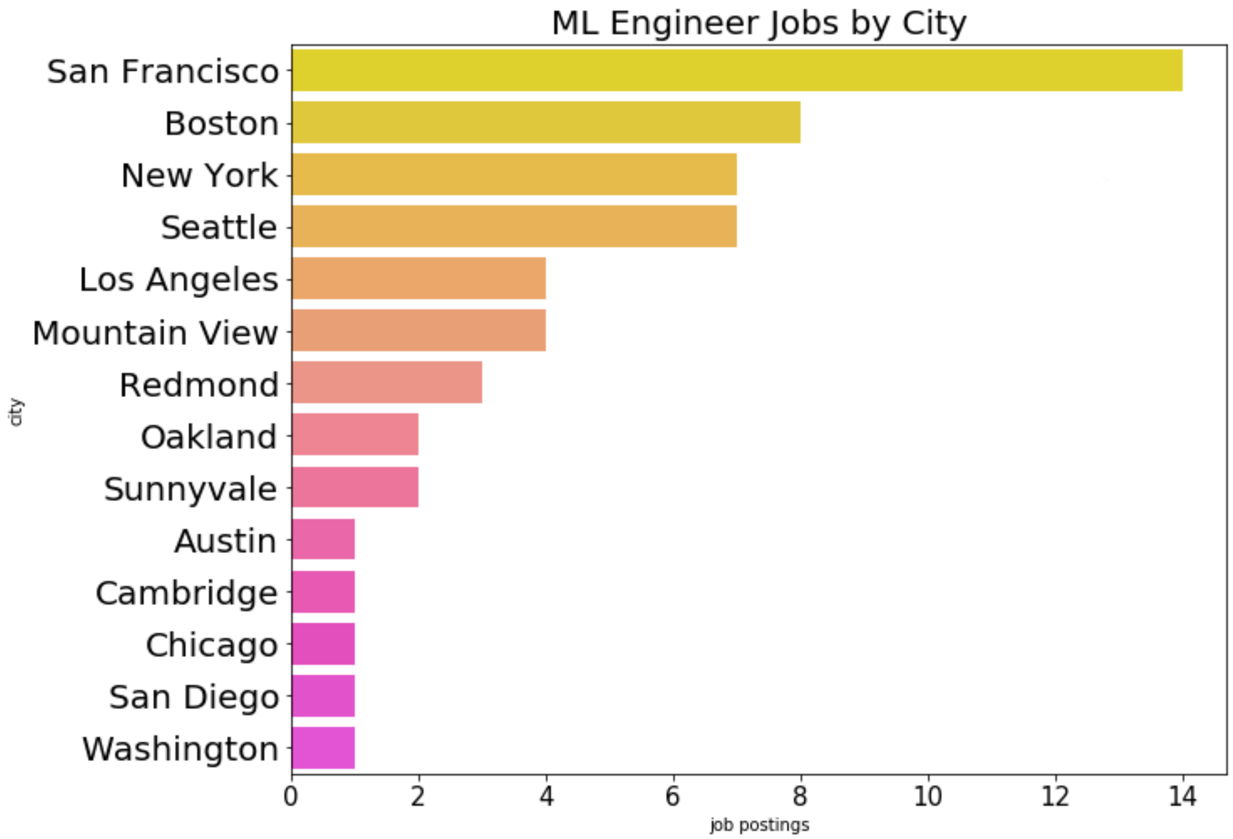


Figure 8 ML Engineer Job Postings by City

A word cloud is generated to better visualize what skills is sought from a machine learning engineer, similar to what we have done with data scientist job postings. “Machine learning”, “data”, “team”, “experience”, “product”, “Python”, “AI” are seen as the outstanding keywords. Gaining experience in these skills would be the wisest decision for a ML engineer candidate to break into the industry.

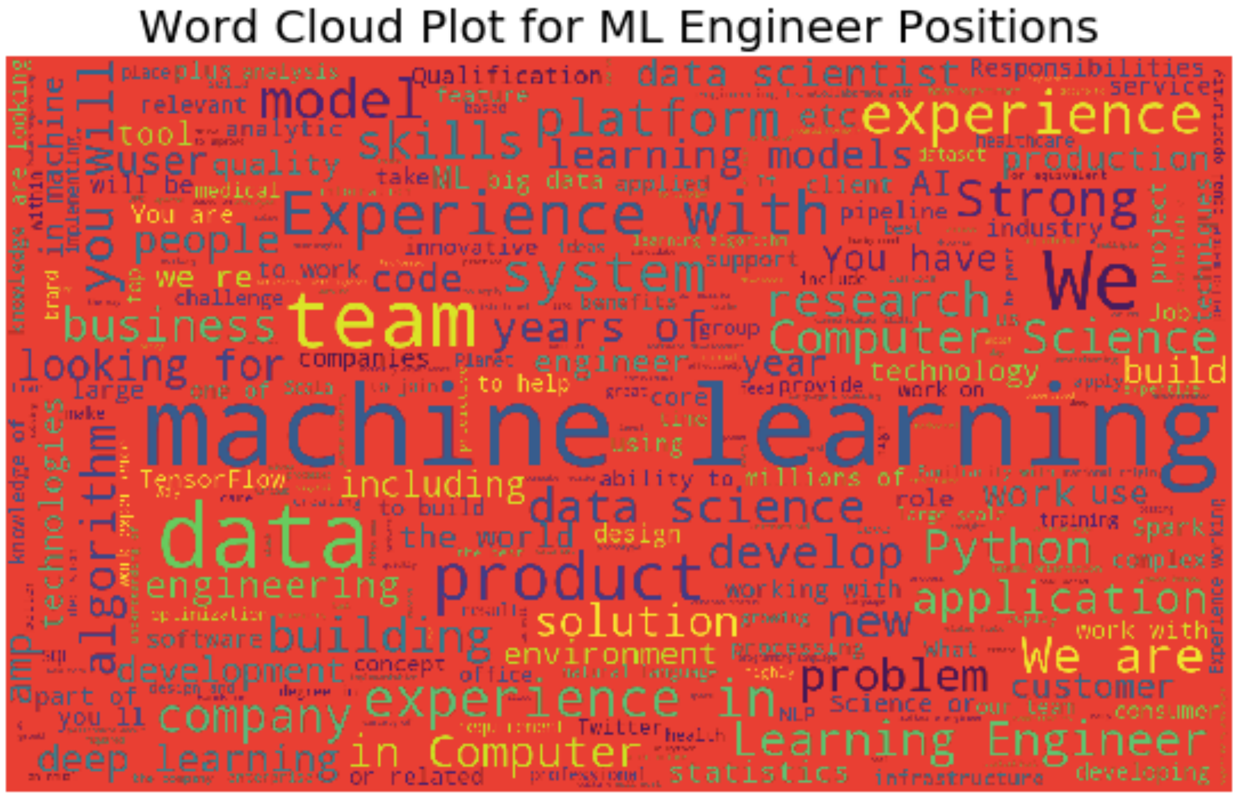


Figure 9 Word Cloud of Skills Required for ML Engineers

Again, what percentages of ML engineer jobs required the most common knowledge skills were explored. Obviously, ML is the most important skill for ML engineers. Second skill is Python, then Java and lastly, SQL. Java and SQL don't look like very essential, especially compared to ML and Python.

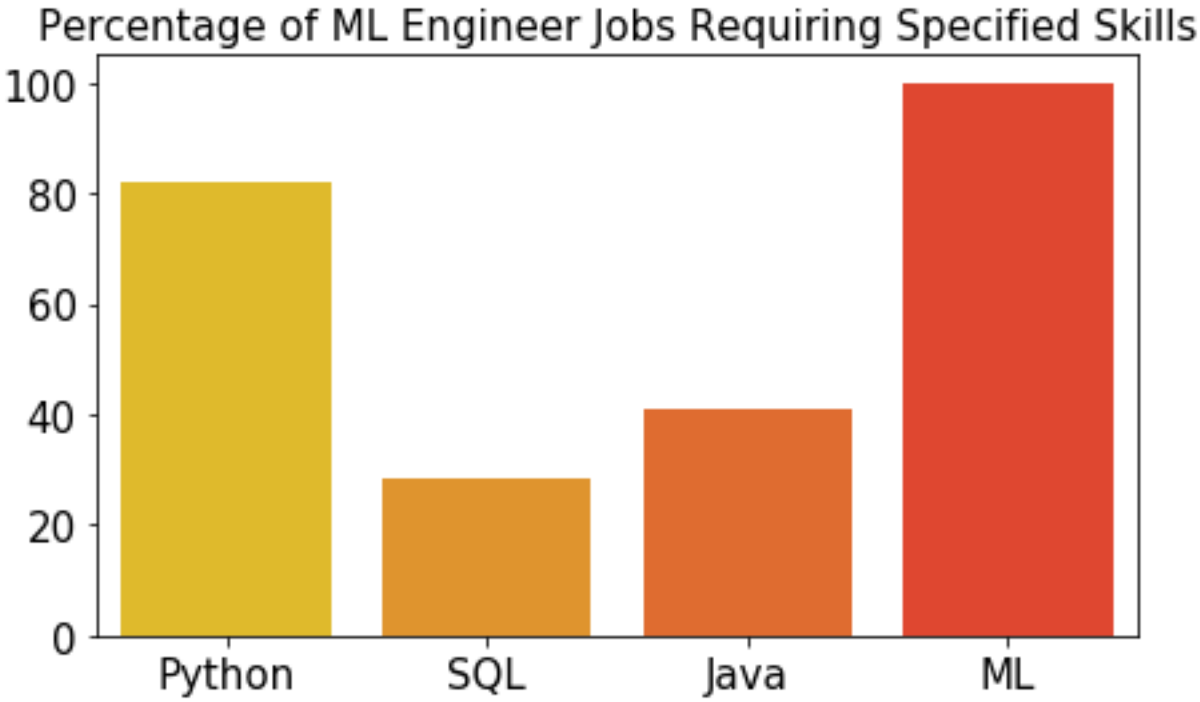


Figure 10 Percentage of ML Engineer Jobs Requiring Python, SQL, Java & ML

c. <u>Data Engineer Jobs</u>

Again, same methods were used to analyze Data Engineer jobs. Only the position named "Data Engineer" was considered for this job position. The volumes of Data Engineer jobs were plotted by city. A similar picture is acquired, largest cities offer the most job opportunities. As in the most part of the analysis, the top 2 cities are New York and San Francisco.

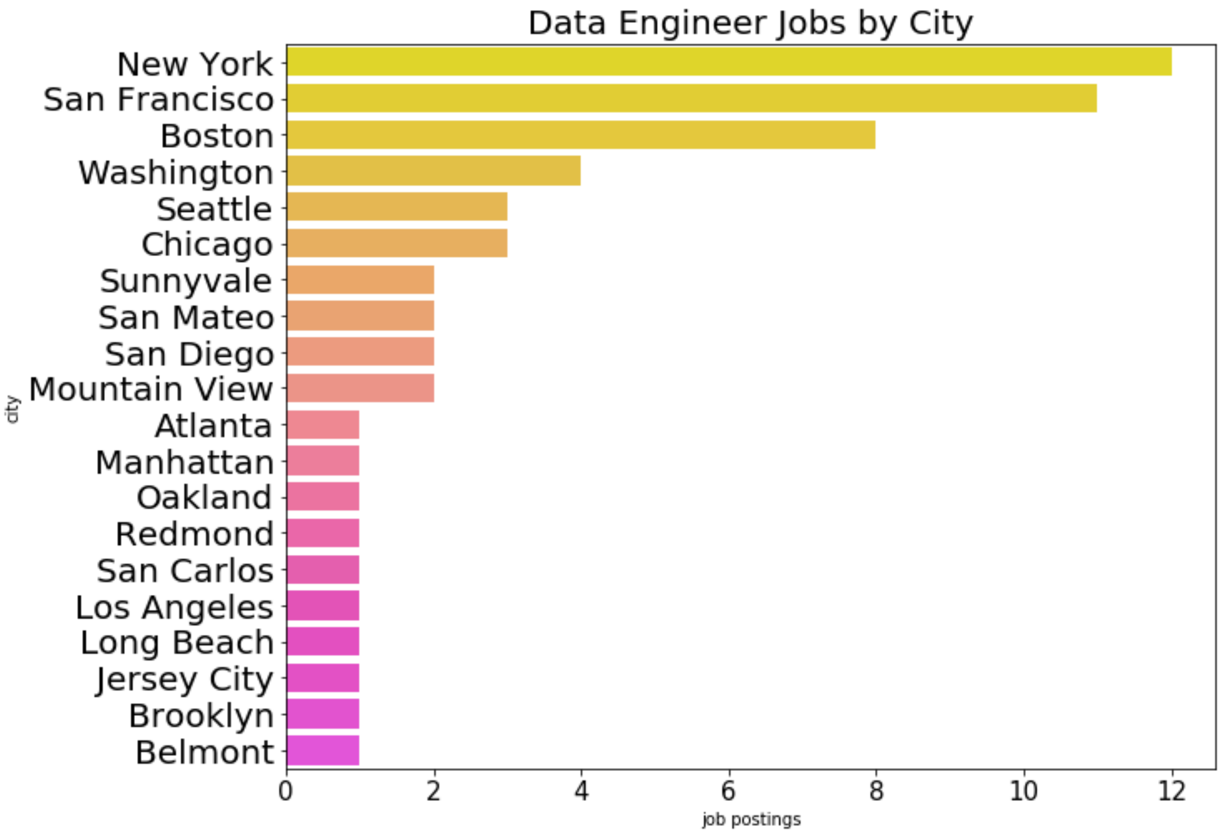


Figure 11 Data Engineer Job Postings by City

Another word cloud is generated for the skills asked by employers to data engineer candidates. The size of "data" is interesting in this word cloud. In none of the previous word clouds for other positions, we have faced with such size for any word. It shows us how much it is important for a data engineer to be able to manage data. "We", "team", "system", "ability", "analytic", "data scientist", "experience with" are the other keywords that catch the most attention.

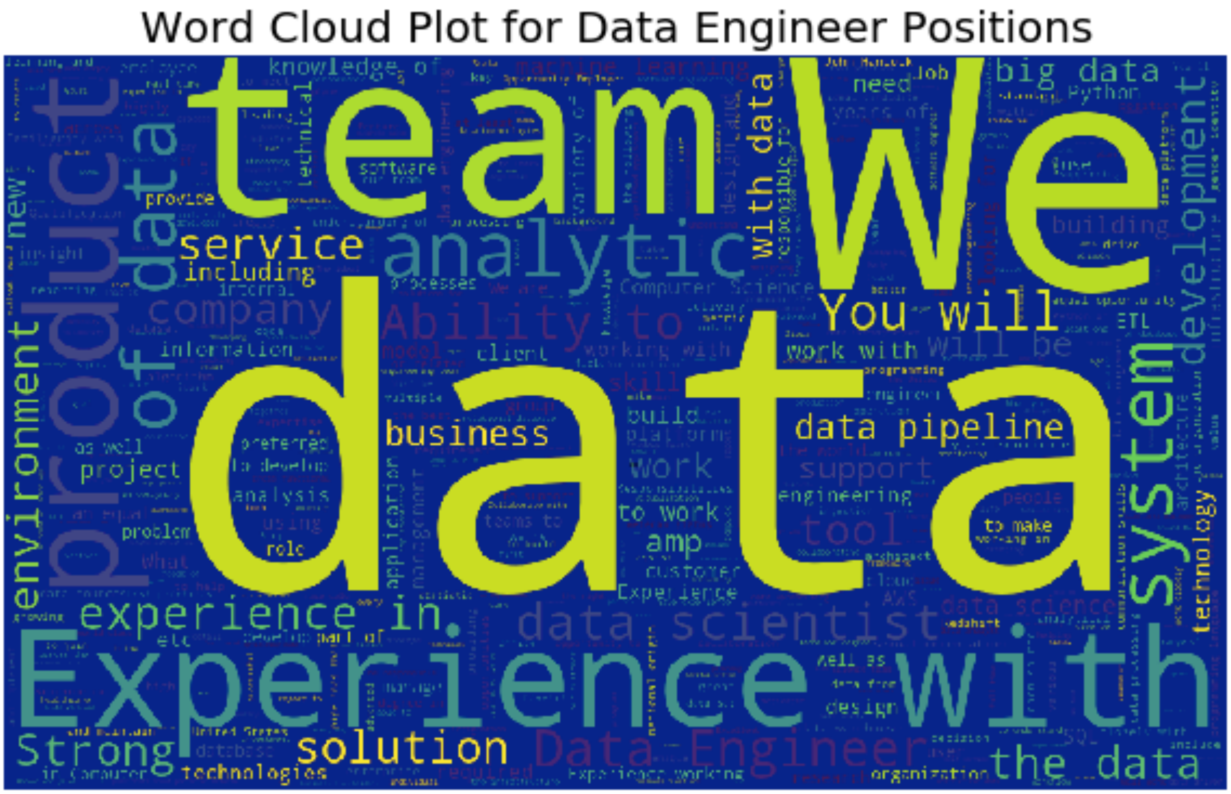


Figure 12 Word Cloud of Skills Required for Data Engineers

A bar plot is plotted to show what percentage of Data Engineer job postings require the most common skills for the industry. We see that the must-have skill for a data engineer is Python. SQL, Java and ML are secondary but equally important skills. And none of these skills are rarely required for a data engineer. Compared to other job postings, data engineer jobs seem to require a combination of these skills, similar to the case of a full stack developer, so to say.

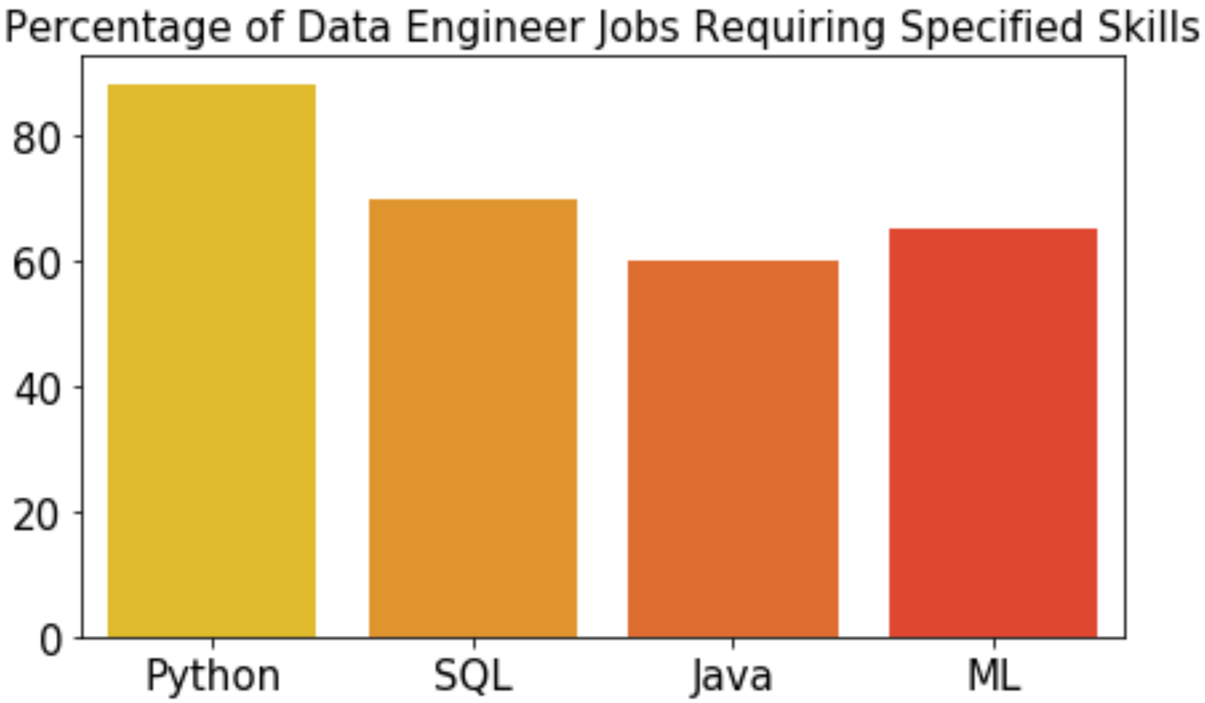


Figure 13 Percentage of Data Engineer Jobs Requiring Python, SQL, Java & ML

d. Data Analyst Jobs

Data Analyst jobs are examined in the same manner. Only the position named "Data Analyst" was considered for this job position. The volumes of Data Analyst jobs were plotted by city. The volume of job posting for Data Analyst jobs was pretty low, so I believe it distorted the analysis slightly. For example, we don't have any data from Seattle, which was a popular city for other DS related positions. Still, large cities like New York and San Francisco are on the top of the rankings.

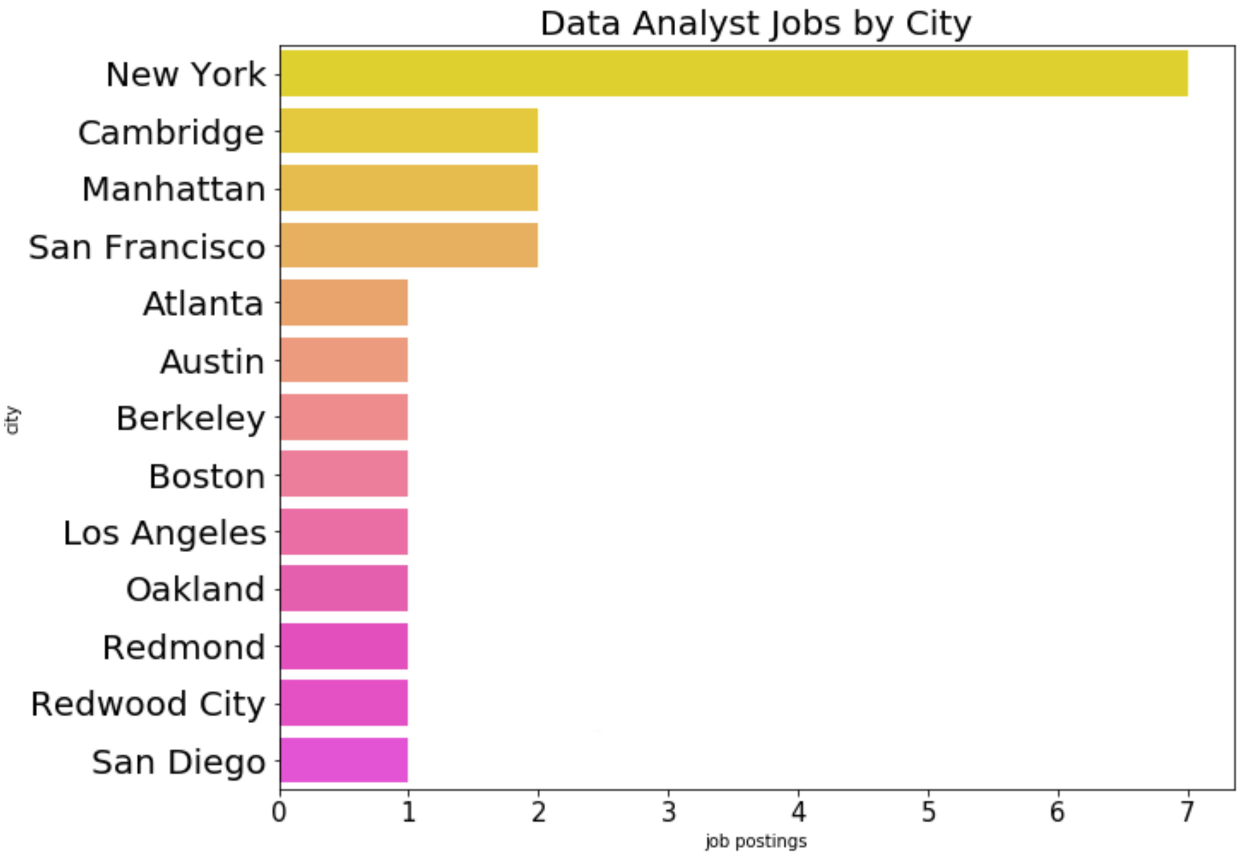


Figure 14 Data Analyst Job Postings by City

The word cloud generated for the skills asked by employers to data analyst candidates includes a low number of large keywords. This might be due to not having a large amount of data for this position. Again, the size of keyword “data” is very large in this word cloud. I believe that it shows us that not a large diversity of skills is sought from data analyst candidates. The skills asked are more straightforward and general. “Business”, “team”, “we”, “marketing”, “SQL”, “analytic”, “research” are the other most important keywords. Another insight that could be gathered from here is based on the sizes of keywords “team” and “we”. These positions usually are looking for great team players who can join their data science teams.

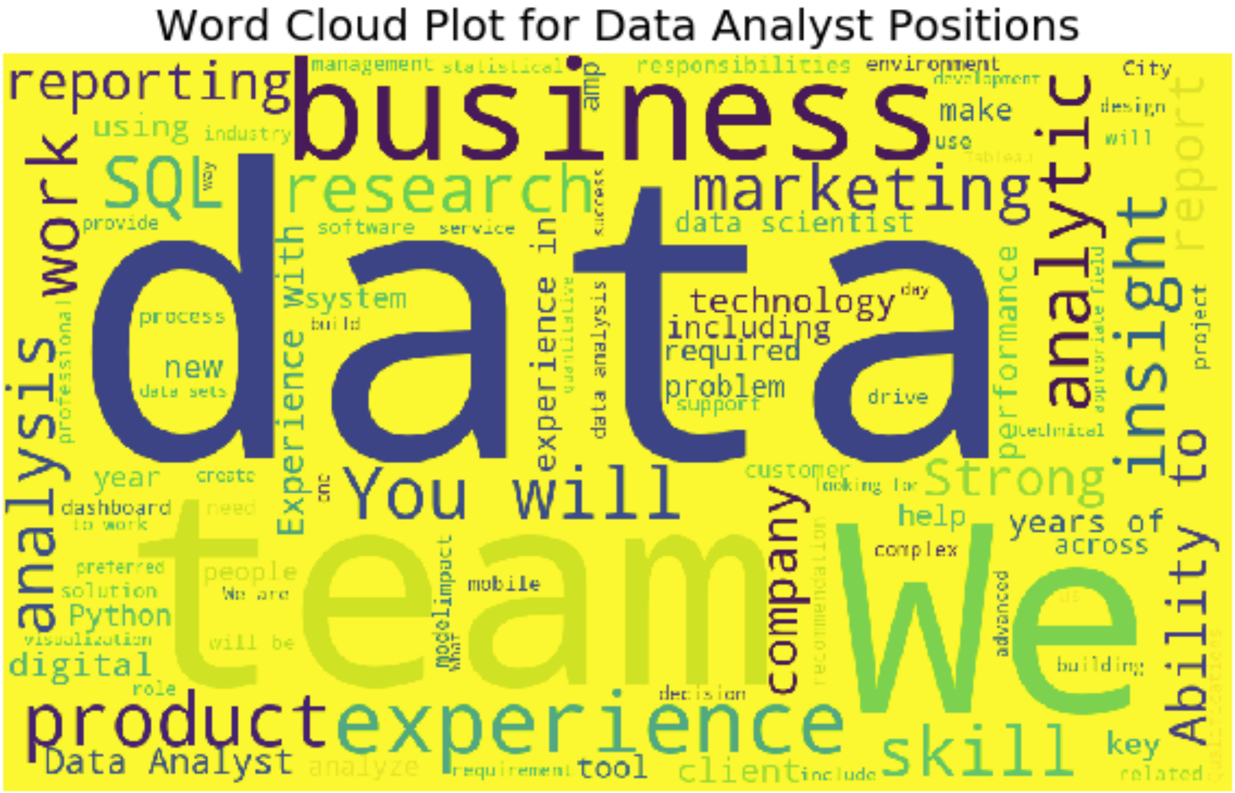

Word Cloud Plot for Data Analyst Positions

Figure 15 Word Cloud of Skills Required for Data Analysts

Looking at how essential each of these most common DS skills or data analysts, we can easily see that SQL is the leading skill for data analysts. Almost 90% of the job postings explicitly required candidates to have SQL knowledge. For this job position, Python and ML are important skills too. However, it is obvious that a very small fraction of these postings (~15%) required Java. Therefore, data analyst position looks like the easiest one to break into the field for someone without a computer science background.

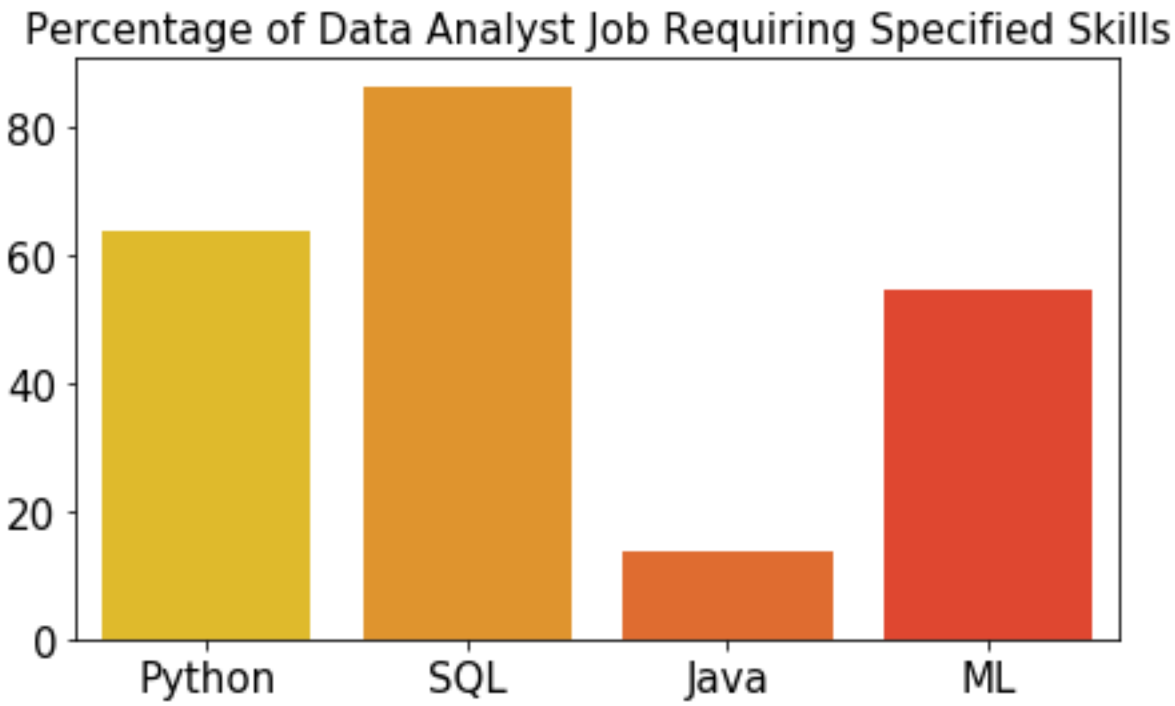


Figure 16 Percentage of Data Analyst Jobs Requiring Python, SQL, Java & ML