Professor TingTing Nian

BANA 212 Data and Programming

30 November 2021

**Data & Programming Final Project Report**

**“Scraping for desired skills on JobBoards”**

**Team 23:**

Kate Zagrebneva

Prashasti Sharma

My Han Mai

Noor Zia

Samira Shafiei

**Table of Contents:**

[**Executive Summary**](#_dn0j5vc01gcu) **1**

[Objectives:](#_ki2a96qvzele) 1

[Programs Utilized:](#_xxkm7hn0i7l9) 1

[**Business Idea**](#_fdhitvmhnppy) **2**

[**Exploratory Data Analysis**](#_ib0qles7hgv9) **3**

[**Project Analysis**](#_meryyppzmtcr) **9**

[**Results**](#_3ugj0ev3j7fc) **10**

[Categorization of skills, based on Data Analyst Functions Areas:](#_mx70bz1gj2jj) 12

[Conclusion](#_1oeaivwarptv) 15

[**Bibliography**](#_h95ml0vhzjnm) **17**

[**Appendix**](#_rkeql7kqnbkq) **18**

# 

# **Executive Summary**

A common struggle of all job seekers whether they be fresh college graduates or someone already in the workforce seeking new opportunities, is whether they match all the necessary skill requirements of the job they want. Being Business Analytics graduate students, we can attest that, on a daily basis, we worry whether we have the desired skills and experience to attain the data-based careers we all joined this program to pursue. After several discussions, the general consensus was that one of the intimidating aspects of job seeking is figuring out what skills we need to acquire and develop given the vast array of skill sets pertaining to data analytics, data science, and business analytics. We came across the article, “40 Data Analyst Skills to Include on Your Resume,” from Indeed’s Career Guide, and found that it included many soft skills but did not dive into the most useful technical skills. That led our team to the question: what among our current and expected-to-gain skill sets are actually relevant to the job market?

We decided to develop a project that tackles precisely this question and aim to assist other business analytics and data analytics students in their career development journey by identifying and ranking actual desired skills listed in role-targeted, real-world job listings. The project is based on web-scraping job listings from ten cities that are known for having many tech-based careers. We used the web scraping software platform, Apify, that automatically scrapes and extracts from websites to pull key skill sets from the aforementioned job-listings and run them through N Grams. The plan is to locate the technical skills, find out how often these skills are mentioned among all the listings, and rank the top skills needed. In doing so, we hope to cut out the debris and offer up the main programs and languages that companies are seeking and that students and job-seekers in general will need to attain in order to be better prepared for the types of jobs available in the analytics industry.

## **Objectives**:

1. Identify the best cities with data-focused job positions
2. Identify the desired skills listed by employers
3. Rank skills based on how often it occurs across all job listings
4. Find the best skills combination for different types of data roles

## **Programs Utilized**:

Throughout the project we will be using Apify, a free, online web-scraping program, and Python and python packages, specifically nltk for N-grams and NLP.

# **Business Idea**

While simple, our project can offer a variety of benefits. The main goal is to offer students and data-focused job-seekers a list of skills to hone in on and develop in order to position themselves in the best possible way for employers. Basically, this meant that identifying the skills would help increase people’s potential when they are preparing for their career.

In a lot of cases, resumes are screened by algorithms before recruiters even get a chance to look at them. If the resume does not have specific keywords that the company is looking for, the algorithm will automatically reject the applicant. The results of this project will provide a set of skills that are mandatory for data analyst applicants to mention in their resume.

The results of this project would be beneficial to the aspiring data analysts who are also applying to be a marketing data analyst, financial analyst or operations data analyst. Our analysis shows the difference between the skill sets that each of these jobs usually require. This information will help people tailor their resume to have higher chances of landing an interview with a company they are applying to.

The project also provided a count of job-listings by city, which, in turn, can pinpoint which cities has the most opportunities and would be good to target when applying. Another way, we found this could assist job-seekers, was to separate desired skills into different categories: hard skills, soft skills, and technical tools/programs. While the project initially wanted to just focus on technical skills, we found that being holistic and including soft skills would be more well-rounded and helpful as compared to its original objective.

Though elementary, this project can be adapted to a broader scope in future applications by including more cities, more titles, or even specializing into specific cities or job titles. The great thing about our project is that it can be adapted to the individual. Say, if a student wanted to target only one state. That student could run the project code and adapt the job listing search to just the state they want and still pull all the job postings and skills related to such a search. That being said, the current application will be a bit more general to capture a variety of locations and open positions in hopes that it remains relevant to a wider audience.

## 

# **Exploratory Data Analysis**

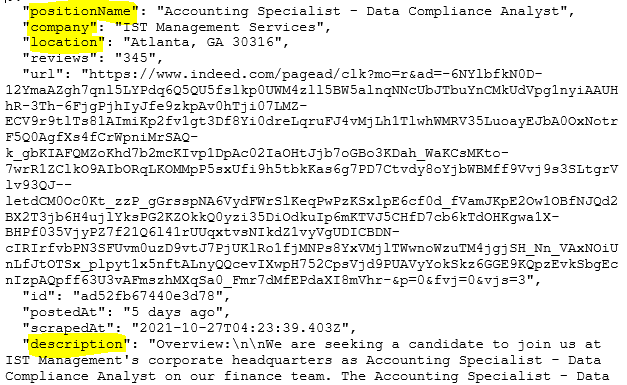
**Data source:**

We sourced the data from a popular job posting portal: [www.indeed.com](http://www.indeed.com). We decided to work with job openings in some of the major areas within the United States, such as Chicago, New York, San Francisco, et cetera, in total ten of them. Since we decided to keep the scope of our project to jobs within the Data Analytics field, we used the keyword ‘Data Analyst’ to filter the job postings.

The data was scraped using a free, online web-scraping platform called Apify.com. The extracted data was downloaded in JSON format. There was a limitation of extracting approximately 1000 job postings from each city. All the data was scraped on 27th October 2021.

**Data Description:**

The complete dataset is a combination of ten JSON files extracted from [www.indeed.com](http://www.indeed.com). Each of the JSON file consisted of multiple job postings presented in the below format -

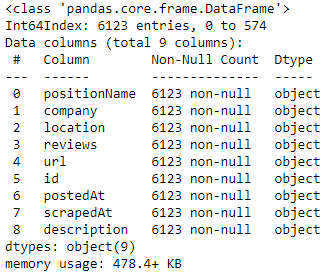


These JSON files were converted to python data frames for further data manipulation. All ten dataframes (one for each city) were concatenated to form the final dataset in python for further analysis.

The raw dataset contains a total of 9 attributes. The description of each attribute is provided below:

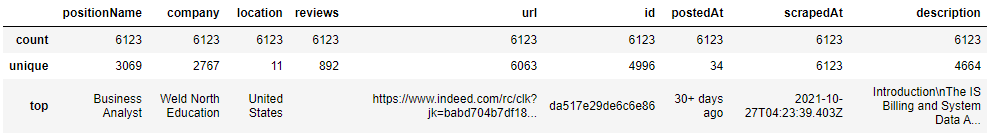
| **Attribute Name** | **Description** |
| --- | --- |
| positionName | Job Title |
| Company | Name of the company posting the job opening |
| Location | Location of the opening |
| Reviews | Number of reviews given by people working the company |
| URL | URL of the job opening |
| ID | Unique ID generated at the time of web scraping |
| Posted At | TimeStamp of job posting |
| Scraped At | TimeStamp of web scraping |
| Description | Description of the job opening provided by the employer |

The data types of the above attributes are as follows:



All the attributes are categorical in nature.

The descriptive statistics are as follows:



Out of the nine attributes, only four were picked for further analysis because the other five attributes did not contribute additional information towards our focus. The attributes picked were - positionName, company, location and description.

Below table provides insight on the size of each dataframe:

| **City Name** | **Data Frame** | **Shape** |
| --- | --- | --- |
| Atlanta | df\_atlanta | (894, 9) |
| Austin | df\_austin | (440, 9) |
| Chicago | df\_chicago | (853, 9) |
| Houston | df\_houston | (474, 9) |
| Los Angeles | df\_la | (724, 9) |
| New York | df\_ny | (861, 9) |
| Orange County | df\_oc | (441, 9) |
| San Diego | df\_sandiego | (354, 9) |
| Seattle | df\_seattle | (507, 9) |
| San Francisco | df\_sf | (575, 9) |
| Combined Dataframe | df | (6123, 9) |

**Data Cleaning**

Before combining all the data frames together, we updated the location column to evenly reflect the location information in the format of - City, State, e.g. Atlanta, GA. This was done to get rid of entries which included area code. This update was done for all the entries except where the location was set to ‘United States’.

Post concatenating the data frames, below data cleaning steps were performed on the data -

1. Dropped insignificant columns - reviews , id , postedAt , scrapedAt , url
2. Removed case sensitivity from the data
3. Removed duplicates
4. Checked for missing values

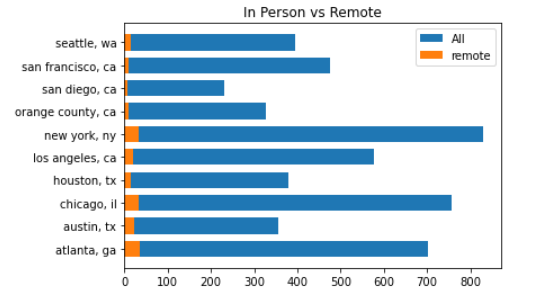
There were no null values in the dataset. But we had 1056 duplicate entries which we dropped. The final shape of our dataset was (5067, 9).

We divided the dataset into Senior vs Non-Senior positions using the keywords ‘senior’ and ‘sr.’. We also divided it into Remote vs On-site work location using keywords such as ‘remote’, ‘virtual’ and ‘united states’. We made an assumption that a location set as ‘United States’ would offer flexibility to work remotely.

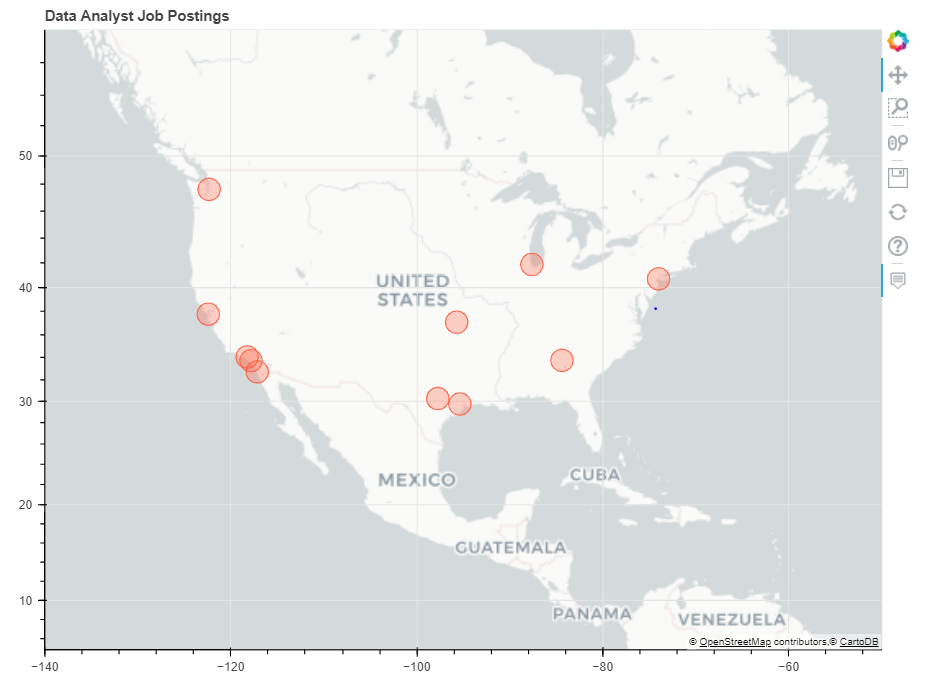
For the relevance of the MSBA program audience, we decided to work on job positions offered for non-senior roles. Hence, we dropped the rows where ‘senior’/ ’sr.’ keywords were used in the positionName column. The shape of the non-senior data set was (3896, 9).

**Data Exploration**

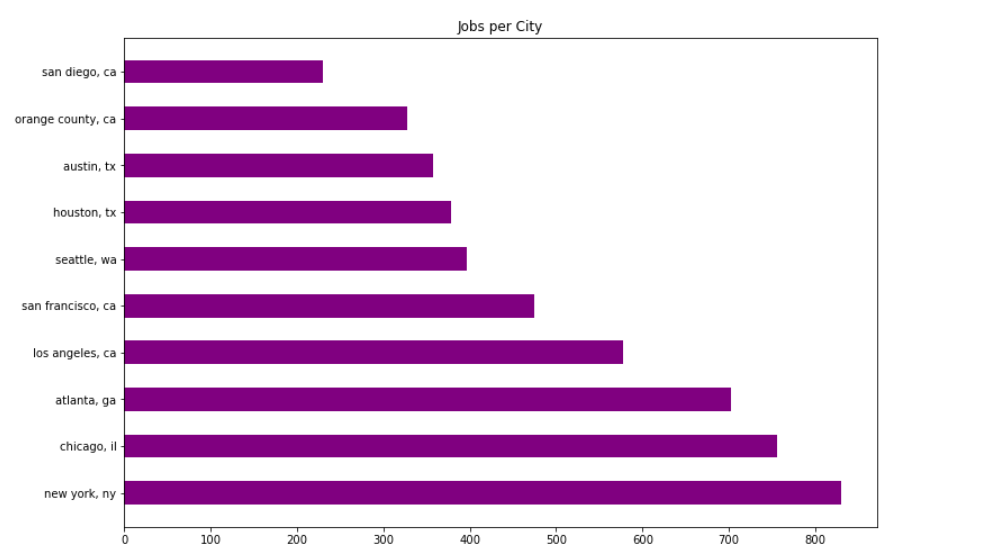
To see what the data is like and get a better understanding of the dataset in general, we performed some exploratory analysis. We saw how many jobs in each city are remote vs in-person and visualized it: all cities were relatively similar in the remote vs in person jobs percentage.



We also created a map to portray the cities that we covered visually and see how the count and distribution of jobs per city decreased after cleaning the data.



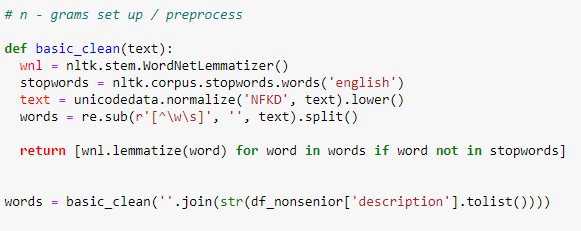
We plotted a bar graph to find out the highest number of job postings.



Our data showed that New York, NY has the highest number of job postings, followed by Chicago, IL and Atlanta, GA.

**Data Preprocessing**

Next, we performed additional steps of preprocessing to make it ready for N-grams to extract the skills from the description column. We removed tabulation, punctuation, stop words, and did lemmatization. We chose lemmatization instead of stemming, as the accuracy of the analysis result was extremely important and the data was not large enough for lemmatization to take too long. We also tokenized the text to be able to search for n grams.

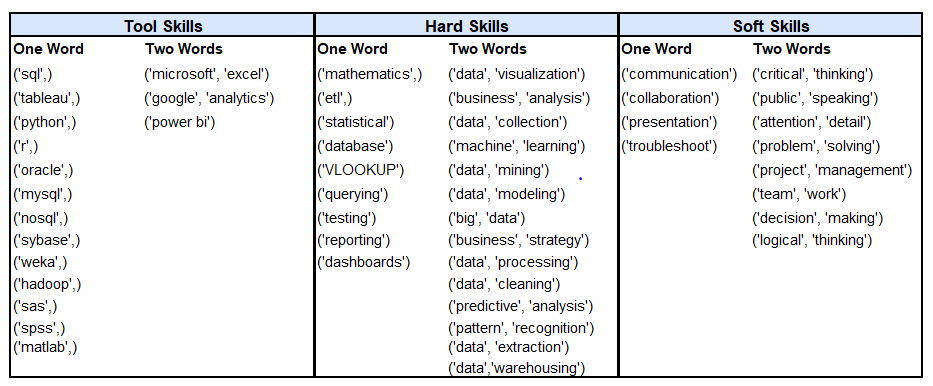


# **Project Analysis**

**N - Grams**:

We firstly performed unigram, bigram, and trigram analysis on the data to see which n grams are most often used in the descriptions. The results showed us a lot of unnecessary n grams and very few useful ones. If we went with those results, we would have had to do a lot of cleaning and pick out the ones that are considered a skill set.

That is why we created lists of skills: tool skills, hard skills, and soft skills. We also divided them into lists of one word and two words to be able to search for them in our unigram and bigram dataframes:



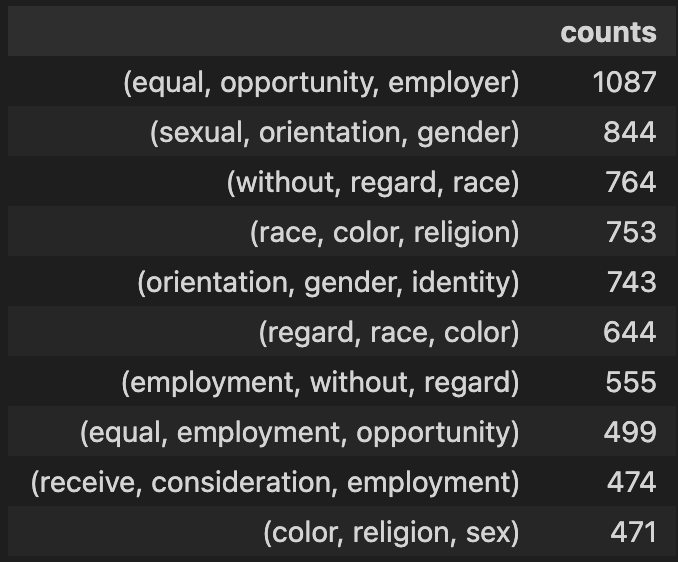
Using these lists, we found the counts of each skill and the difference between the counts and, therefore, popularity of those skills in job descriptions.

We wanted to get a bit more information and insight from the analysis and, therefore, did the same analysis on finance data analyst, marketing data analyst and operations data analyst to see the differences in the results. We documented the top five tools, hard and soft skills for each and pointed out specific skills that are more important to finance, marketing, and operation roles.

# **Results**

To find out which skills are the most desirable by employers, we started with running n grams. We ran unigrams, bigrams and trigrams. This was done to see the most popular words without making assumptions of choosing and searching for specific skills that we came up with.

The results of that initial analysis showed that bigrams performed the best, giving us the greatest amount of information with two word grams. Unigrams showed too little, since one word like ‘data’ does not tell us much, and it is most likely part of the two words skill like ‘data analysis’ or ‘data cleaning’. Looking at bigrams results, it shows us some of the skills that would be beneficial for us like ‘data analysis’. Trigrams mostly showed the legal obligations of job posting that each employer needs to mention and not the actual skills. Overall, running simply n-grams does not give us enough of skill information and has a lot of keywords that are not relevant to our analysis:



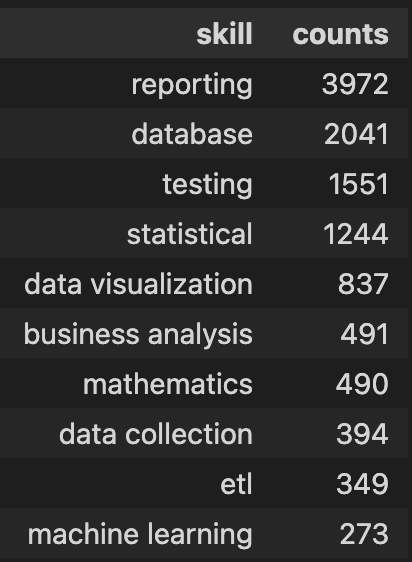
**Searching for specific keywords in n-grams:**

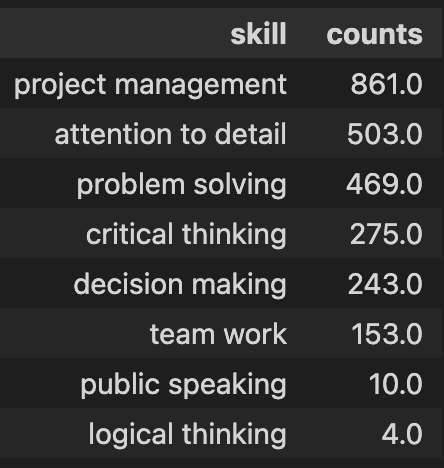
We came up with three sets of skills based on the type of skills: tool skills, hard skills and soft skills. Since our skills only contained one or two words, we only needed the results from unigrams and bigrams. We then searched in the unigram dataframe for the tool skills, hard skills and soft skills that contained only one word. After that, we searched for the skills of two words in the bigram dataframe. We then combined the results to make three data frames with each skill type: tool skills, hard skills and soft skills.

The top 10 tool skills ended up being the skills mentioned below. We can see a huge difference between ‘sql’ count and all others, so it is safe to say that going into the data analyst job market would require you to know SQL.



For hard skills, we also had one skill that stood out: ‘reporting’. While it might seem like a very obvious skill, based on this analysis, it is important to mention this skill in the resume.



The last set that we got was the soft skill dataframe. While we did not have a specific skill that was ranked much higher than others, we could see that skills like ‘public speaking’ and ‘logical thinking’ were ranked extremely low, and therefore are not worth mentioning in the resume.  
  


We wanted to visually represent all the skills combined to see which ones are worth mentioning on the resume, since no one could possibly fit every single keyword on a resume and make it readable at the same time. Therefore, we created a word cloud with all skills combined. The size of the word represents the frequency of the skill in the job descriptions. While we can mostly only see tools and hard skills on here, which are most likely required if the resume was scanned by an algorithm, it is important to talk about soft skills in your resume as well to account for when a recruiter actually reads it.



## **Categorization of skills, based on Data Analyst Functions Areas**:

To find out which skills are more relevant to specific data analyst functions, we created separate data frames for each area: Finance, Marketing, and Operations. For example, to create the Financial dataframe, we searched for the positions that included keywords such as finance, financial, account, and cost. Then we implemented n-grams preprocessing on the above dataframes. Lastly, we used the n-gram technique to find the most required skills in each section. There are three skill sets; tool skills, hard skills, and soft skills. The table below shows the top five of each skill set.

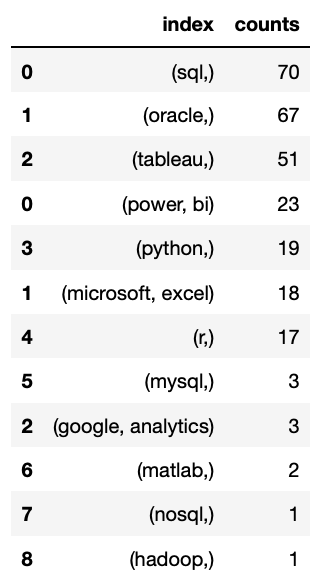
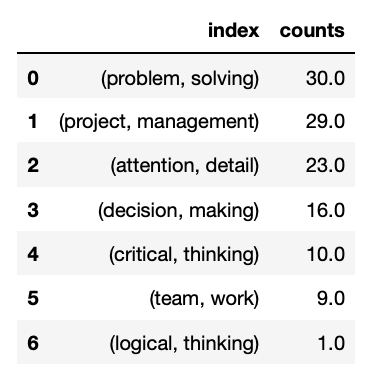
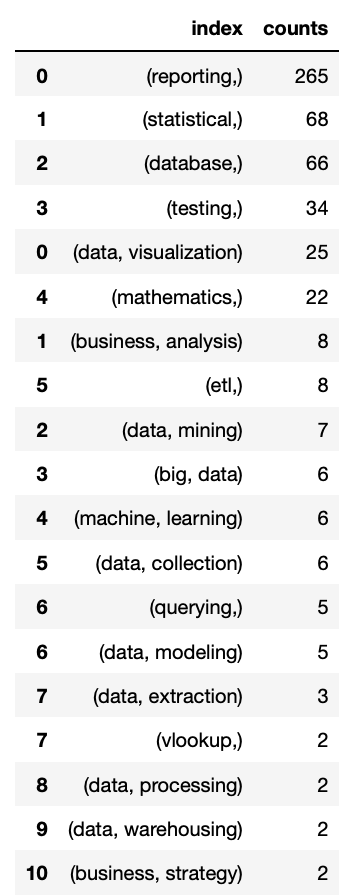
| **Function Area** | **Tool Skills** | **Hard Skills** | **Soft Skills** |
| --- | --- | --- | --- |
| **Finance** | Sql, Oracle, Tableau, Power bi, Python | Reporting, Statistical, Database, testing, Data Visualization | Problem Solving, Project Management, Detailed Oriented, Decision Making, Critical Thinking |
| **Marketing** | Sql, Tableau, Google Analytics, Python, r | Reporting,Database, Data Visualization, Statistical,testing | Attention to Details, Problem Solving, Project Management, Team Working, Critical Thinking |
| **Operations** | Sql, Tableau,Power bi, Python, Microsoft Excel | Reporting, Database, testing, Statistical, Data Visualization | Project Management, Problem Solving, Attention to Details, Critical Thinking, Team work |

***Finance:***

Similar to the “all skill” rankings, finance's highest ranked tool skill is SQL. Oracle, however, is ranked higher in finance than in general data analyst roles. Also, Power BI is more important than programming languages like Python and R.

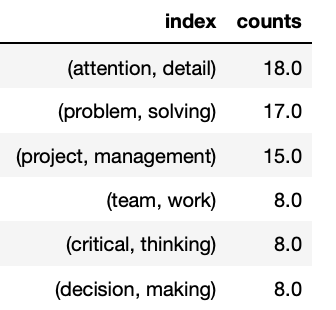
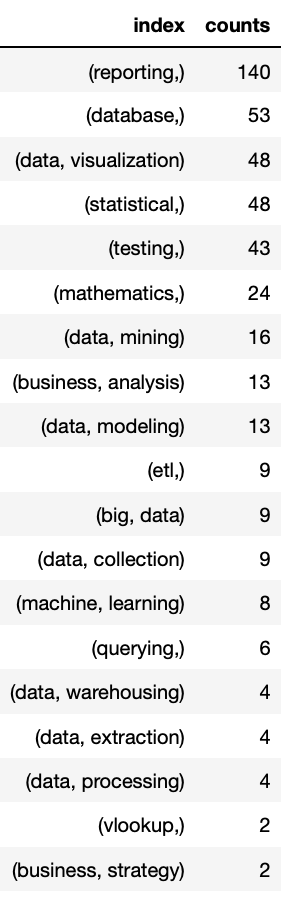
Reporting has the highest count out of all hard skills, same as general data analyst. However, ‘statistical’ is ranked higher, and, therefore, is definitely worth mentioning on the resume.

Soft skills have the most differences compared to general data analyst results. While the top 5 skills are the same, they are ranked differently. Problem solving and project management are still the two soft skills that remain as the most important.



***Marketing***:

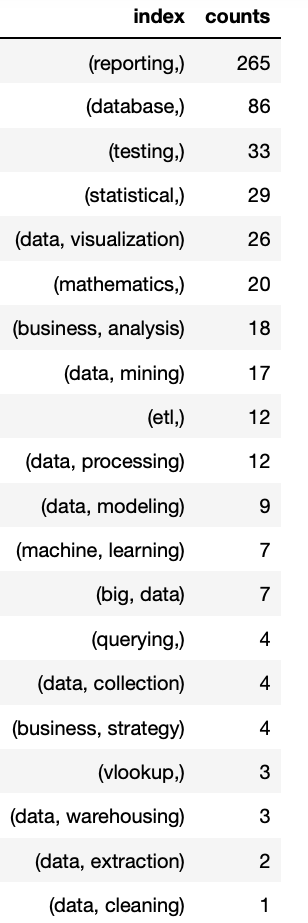
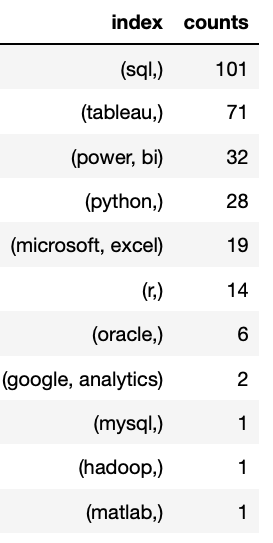
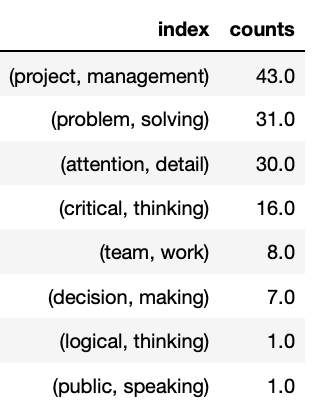
Looking at results for marketing data analysts, the biggest difference that we can see is that Data Visualization and Tableau are ranked higher compared to others. Google Analytics is also ranked pretty high compared to others. It is extremely important to mention these skills on the resume when applying for such positions. SQL is still a must, as well as reporting. When it comes to soft skills, attention to detail, problem solving and project management are the top three soft skills, with others are ranked significantly lower. It is safe to say that after mentioning these top three skills, you can use the remaining space on the resume to talk about tool skills and hard skills.





***Operation****:*

For operations roles, SQL and reporting are, again, still the strongest skills. On the other hand, Power BI has more significance compared to other roles here. Surprisingly, for soft skills, decision making is ranked way lower than in other positions. Overall, the ranking is quite similar to the general data analyst ranking that we performed in the first part of the project with the exception of Power BI.



## 

## 

## 

## 

## 

## 

## **Conclusion**

Overall, this project was designed to show the top skills for general data analysts, as well as differences between financial, marketing, and operations data analysts. Looking at the top five skills for each, it is recommended that you mention those in your resume to be considered for an interview. Even though the experience that you talk about in the resume can seem relevant and important to you, a lot of the time your resume is simply scanned for keywords that employers are looking for. That is why it is important to include those keywords on your resume.

If the function area (finance, marketing, operations) is not a deal-breaker for you and you are applying to all roles, it is important to have slightly different resumes for each. As the analysis shows, financial and marketing analysts have pretty big differences when it comes to top desired skills. Finance data analyst needs to know Oracle and Power BI, while marketing data analysts have to have experience with data visualization, Tableau, and Google Analytics. Operations data analyst also needs to mention Power BI, although the rest of the resume can be the same as those used for a general data analyst role.

**Things to improve**

While this project analysis gave some good and helpful insights, there are ways it can be improved. Firstly, we can use the API to have data from all cities in the US and not just those that were selected, making the assumption that they are the top cities for data analyst roles. We would have more data that way and the results would be more accurate and precise. Also, we could use pull listings from more sources rather than focusing on Indeed alone.

## 

## 

# 

# **Bibliography**

Grus, Joel. *Data Science from Scratch: Second Edition by Joel Grus*. O'Reilly Media, n.d..

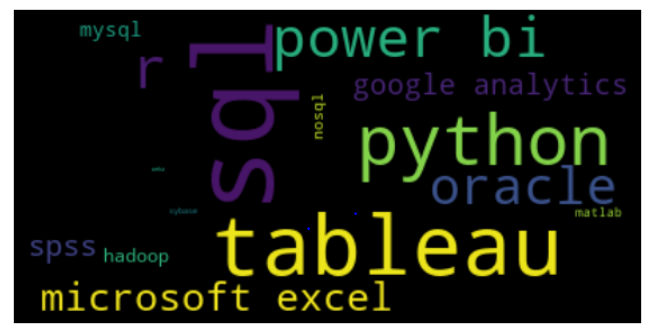
“40 Data Analyst Skills to Include on Your Resume.” *Indeed.com*, Indeed Editorial Team, 5 Apr. 2021, <https://www.indeed.com/career-advice/resumes-cover-letters/data-analyst-skills-resume>.

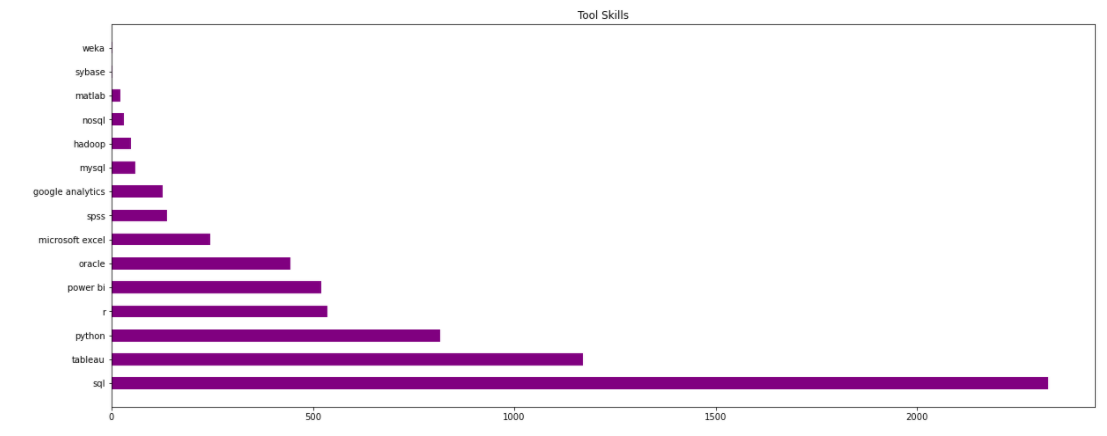
“Web Scraping, Data Extraction and Automation · APIFY.” *Apify*, [https://apify.com/?utm\_term=apify&  
utm\_campaign=BRAND-APIFY&utm\_source=adwords&utm\_medium=ppc&hsa\_acc=9303439903&hsa\_cam=12208847443&hsa\_grp=115467448485&hsa\_ad=495840157411&hsa\_src=g&hsa\_tgt=kwd-401768082175&hsa\_kw=apify&hsa\_mt=e&hsa\_net=adwords&hsa\_ver=3&gclid=Cj0KCQiA7oyNBhDiARIsADtGRZYEaQi6BTJ\_Y7lNnymcbaN9CbOFrfJXcS6i8bIRwqP0GhBiilJskiUaAu2sEALw\_wcB](https://apify.com/?utm_term=apify&utm_campaign=BRAND-APIFY&utm_source=adwords&utm_medium=ppc&hsa_acc=9303439903&hsa_cam=12208847443&hsa_grp=115467448485&hsa_ad=495840157411&hsa_src=g&hsa_tgt=kwd-401768082175&hsa_kw=apify&hsa_mt=e&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQiA7oyNBhDiARIsADtGRZYEaQi6BTJ_Y7lNnymcbaN9CbOFrfJXcS6i8bIRwqP0GhBiilJskiUaAu2sEALw_wcB).

# 

# **Appendix**

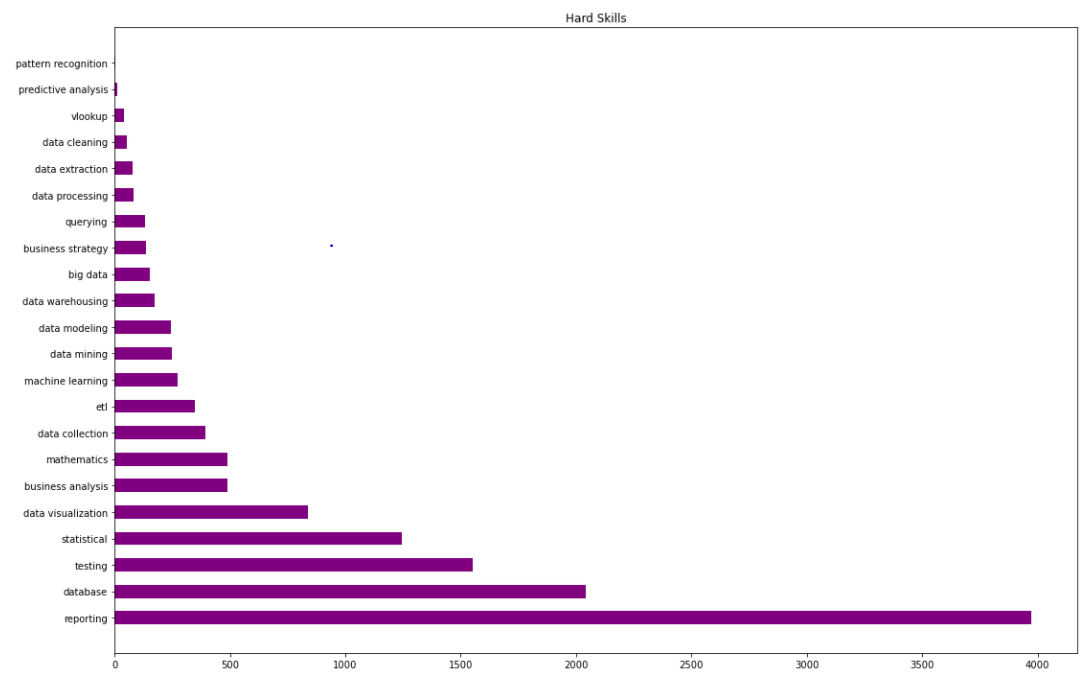
**Tool Skills: Word cloud & Bar graph**



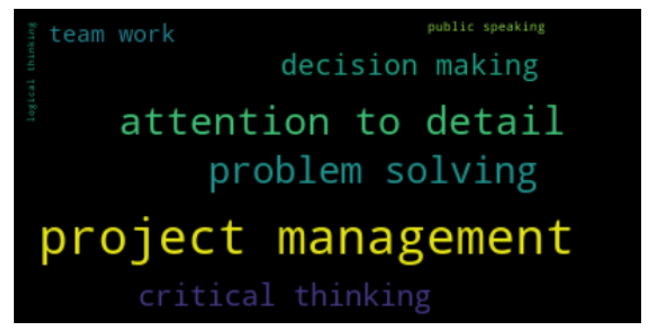


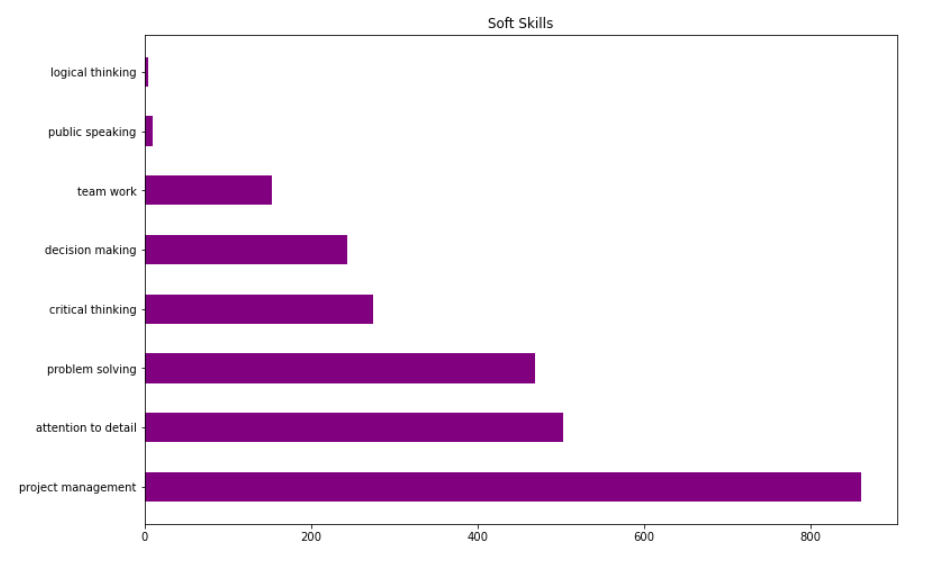
**Hard Skills: Word cloud & Bar graph**

****

****

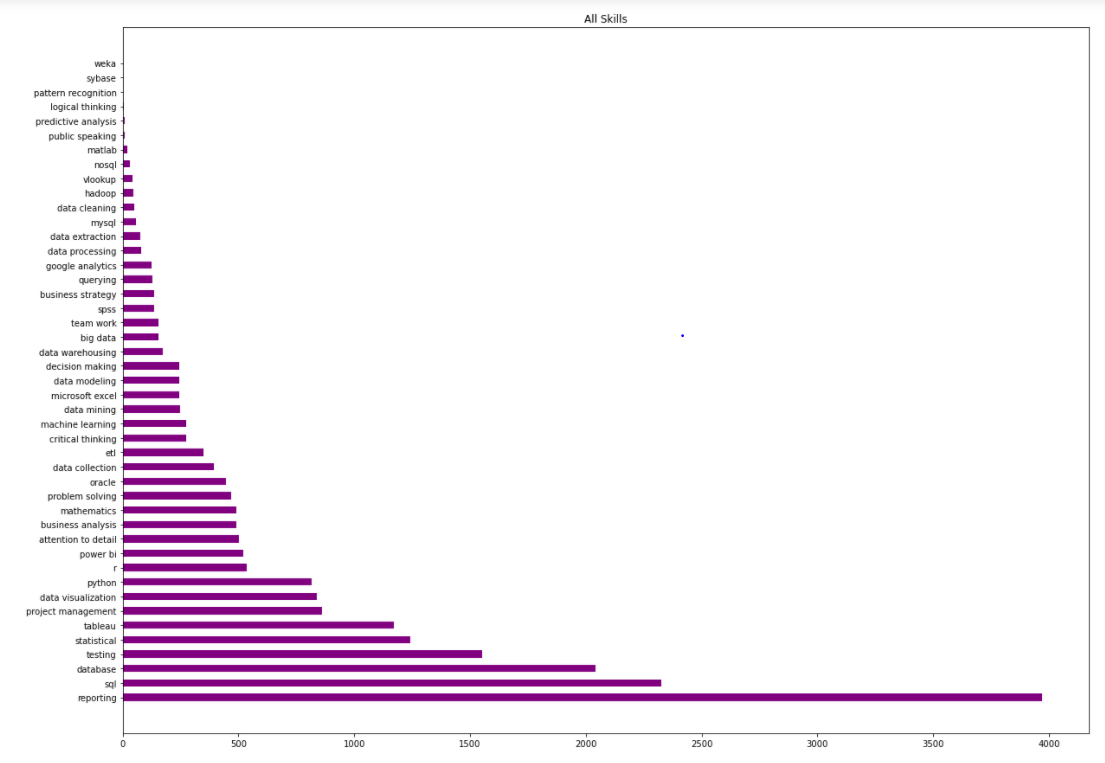
**Soft Skills: Word cloud & Bar graph**

****

****

**Top Skills in all 3 Categories:**

****

****