**DATA SCIENCE JOB POSTING - WHAT CAN WE LEARN ABOUT THE DATA SCIENCE JOB MARKET FROM A SET OF ADVERTISED POSITIONS-**

**Oliver Posewitz**

Applied Data Science

oeposewi@syr.edu

**Bruno de Paula Luiz**

Applied Data Science

bdluiz@syr.edu

1. **Abstract**

Considering a brief analysis of advertised data science positions, we answer questions related to the job market including availability, job market description, required skills, and distinctive groupings. Using descriptive statistics, visualizations, feature extraction techniques, and clustering techniques, our Exploratory Data Analysis (EDA) resulted in the following conclusions; There is a higher demand for data scientists with more experience compared to junior/entry level positions, soft skills are more important than technical skills alone, and technical skills, while less important than soft skills, are still a requirement for the majority of data science positions. These findings reinforce the conventional wisdom that data scientists are well rounded individuals with skills that are in high demand.

**Disclaimer:** This analysis is purely for educational purposes and is not meant to be taken as a rigorous study. Please don’t entirely base any career decisions on this article

1. **Introduction**

With the rise in technology and use of data as an important driver in business decisions, data science roles are not as mature as compared to other positions. The wide range of career paths for data scientists maintains a relatively high level of ambiguity for early career professionals. This study aims to reduce the ambiguity and shed light on what employers are looking for when they advertise new positions with data science in the title. This information will hopefully bridge the gap between employer communication of their needs and candidates seeking employment opportunities in a growing industry.

1. **Ethical Considerations**

There are several things to consider while conducting this analysis. First, this analysis is based on data produced by a web crawler presumptively authored by jobspikr. There are underlying costs associated with using web crawlers for web scraping as they can adversely affect the websites being scraped. For example, an automated crawler may overload a server hosting the website by sending more requests than the server can handle, which may cause reduced availability of said website resulting in a decline of user experience for the website's human clients. Web crawlers should include pauses between requests to reduce the impact of performance on the servers hosting the data. Since the data has already been scraped by another source, this study assumes that this data was collected in a responsible way that didn’t harm any other party.

Another consideration is the fact that the information being processed may contain sensitive information associated with the advertising business that they may not want to be posted for analysis. For example, a small business may have included their telephone number with the intention of potential candidates being the only ones seeing it. While this information was posted on a public domain, the advertising agency may not have included certain information had they known it would be used in this type of analysis.

Lastly, the data itself may contain bias that would affect the conclusions of this analysis. The biases would translate into potentially misrepresentative conclusions about the job market and mislead its readers. For example, we extracted a set of key features from the job descriptions to conclude that most data science positions require a set of specific technical skills, but the web crawler may have used those same features for detection in its data collection process, resulting in false conclusions that a set of skills are required when in reality there may be plenty of positions available without these requirements. While the risk of this misrepresentation may be small, the assumptions this analysis makes in order to come to the conclusions should be made clear to the reader so that they can form a better informed opinion.

1. **The Data Science Job Market**

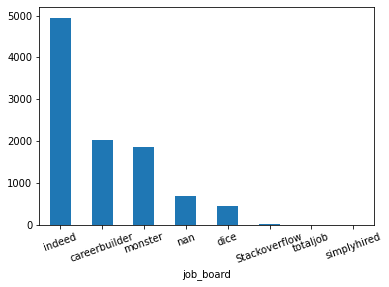
It is well documented that the data since job market has grown and diversified at a rapid rate, with growth comes a new variety of data science career opportunities, both for new college graduates and experienced data professionals who want to adjust their job responsibilities or find a progressing subfield of data science to explore

With *Glassdoor* ranking data science as the #2 jobs in America for 2021, exploring demand trends within the field can only lead to beneficial insights. After all, finding patterns and observing correlations are exactly the skills that make these data professionals so valuable in business. Understanding why the field of data science is growing will help to see where data scientists can make the most of their opportunities

1. **Scope and Data**

The scope of this analysis will be highly dependent on openly available data for analysis. The data set, downloaded from the website *data.world* (jobspikr, 2019), includes 10,000 data science job postings including job title, job description, and several other features. This study will focus on extracting information from the text of the aforementioned features to draw conclusions.

This study will focus exclusively on data science positions in the United States that were posted in 2019. There are also a limited number of websites that were crawled to create the data set, which may be a source of implicit bias in the dataset. The main job boards are shown below:



The job boards that were used for the collection of data could be an underlying source of bias. For example, some job boards could be more popular than others in certain geographical locations, or some job boards could be more focused towards experienced positions while other websites that weren’t included (like *handshake.com*) may be more focused on advertising entry level positions for college graduates.

For the sake of this study, we will be operating under the assumption that the collective observations contained in the data set are a random sample of all the job positions posted in the job market as a whole. This assumption is necessary to conclusively characterize the data science job market in general, but puts limitations on any conclusive arguments this analysis may result in.

1. **Descriptive Statistics**

First, we import the dependencies we need, and we load our data:

*figure\_\_session5\_1*

*figure\_\_session5\_2*

We use data.info() to inspect how many null values are in our data, and we get the following:

*figure\_\_session5\_3*

We can see several null values in columns that we will not use. The code below removes the unwanted columns and columns with null values.:

*figure\_\_session5\_4*

*figure\_\_session5\_5*

In sequence, we confirmed that there are no duplicate rows in the data.

*figure\_\_session5\_6*

**Job boards -** Which boards have the most openings?

To answer this, we use a Counter data structure to count the names of each row’s Job Title attribute. We then display the findings with the following bar chart:

*figure\_\_session5\_7*

**Salaries** - What is the range of salaries?

*figure\_\_session5\_8*

**Popular Jobs**- What are the most popular jobs?

*figure\_\_session5\_9*

**Popular Companies**- What are the most popular companies?

*figure\_\_session5\_10*

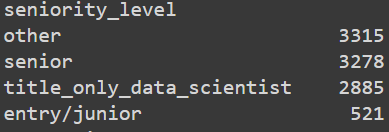
**Hiring States** - What are the states with the largest job openings?

*figure\_\_session5\_11*

1. **Grouping by Seniority**

A large effort in this study was dedicated towards separating the documents by title to identify distinguished groups of observed job postings for further analysis. This task proved to be difficult as employers often include things like pay, location, and proprietary information about the position within the job title which resulted in thousands of results when grouped by title, even after regular expressions were used to consolidate equivalent phrases (i.e. Senior Data Scientist and Sr. Data Scientist). Furthermore, a large portion of the observations were labeled ‘Data Scientist’ in the job title regardless of the job description information. This could be an indication that the original web scraper that was used to collect the data defaulted to filling the job title with this label if no title was detected by the script. Thus, further methods were included to more accurately group the observations.

The groupings that were attempted were distinguished by seniority. The reasoning behind this methodology was that a large portion of the job titles included a level associated with the position (1, 2, 3, or I, II, III) and key words in the job description could be used to infer the seniority of the position being advertised. The resulting groups included Senior and Junior/Entry Level, observations where ‘Data Scientist’ was the only text in the job title, and Other. The objective of this task was to accurately group as many observations into the ‘Senior’ and ‘junior’ groups and minimize the other groups as these groupings have the most promise for interpretable results after further analysis. While searching for key descriptive words that would further indicate specific levels of seniority, like searching for ‘entry level’ in the job description successfully captured significantly more observations, the groupings were still imbalanced (see below).



Most notably, there are very few entry level/Junior positions detected. This could be due to the fact that employers prefer to hire data scientists that are already trained and want to avoid the cost of training new hires. This result is in agreement with conventional wisdom the data science job market that it’s hard to break into the industry, but once you have some real world experience, the market flips to favor the professional with more experience.

1. **Skill Extraction**

Which skills are employers looking for?

We used spacy Matcher to identify token patterns as seen below. However, when you have a phrase to be matched, using spacy Matcher will take a lot of time and is not efficient.

*figure\_\_session7\_1*

spaCy also provides a PhraseMatcher which can be used when you have a large number of terms(single or multi-tokens) to be matched in a text document. Since writing patterns for Matcher is very time consuming for this project. PhraseMatcher solves this problem, as you can pass Doc patterns rather than Token patterns

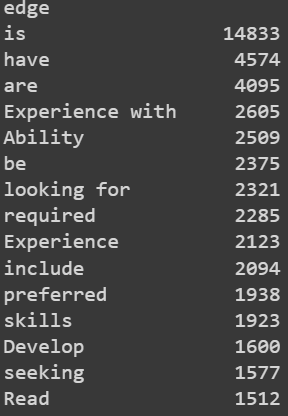
The PhraseMatcher lets you efficiently match large terminology lists. While the Matcher lets you match sequences based on lists of token descriptions, the PhraseMatcher accepts match patterns in the form of Doc objects.

We created two subsets of phrase matchers, the first is named “computer\_languages\_list”, which includes hard skills such as *Python, R ,SQL* while the second is “ personal\_skills\_list” which includes phrases such as machine learning, cloud computing.

*figure\_\_session7\_2*

1. **Syntactic Knowledge Mapping**

Another method of systematically extracting knowledge from the text of each job posting is identifying parts of speech and dependencies using spaCy’s dependency mapping model. The framework has a model that predicts part of speech and word dependencies for each document. These predictions were used to identify the root word and entities for each sentence in each document. In this case, the root word is considered the relationship between the two entities in each sentence which results in a heuristic process to capture patterns of descriptive characteristics established by employers. Below are some of the most frequent relationships:



Several of the top frequent relationships were tested for their top resulting dependencies. We decided to use ‘Experience’ (also including ‘experience with’) for one knowledge graph and the phrase ‘looking for’ as another. Trying to map the entire results of the knowledge mapping would be too noisy and the resulting graph wouldn’t be effective. For each filter, we took the top 25 most frequent entities that were related to the root word. This would capture a large amount of the instances where employers listed descriptive characteristics of candidates.

The ‘experience’ filter resulted in a large list of ‘hard skills’ like tools (i.e. Tableau and Excel), coding languages, and common tasks. (See attached kg\_Experience.png)

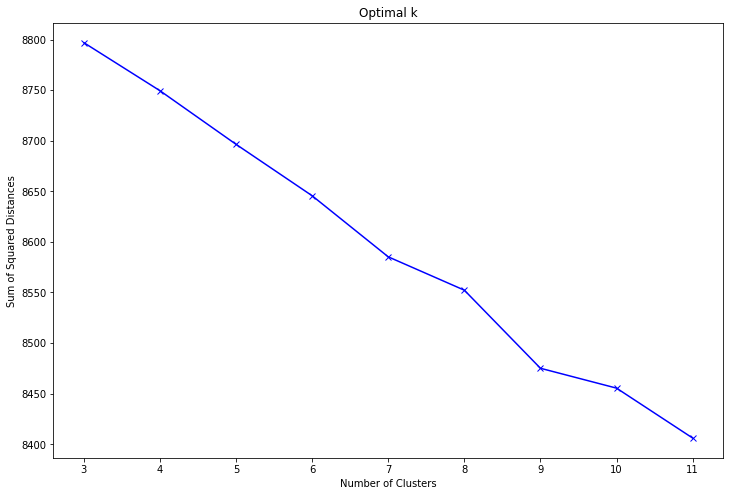
The ‘looking for’ filter, resulted in more descriptive characteristics of candidates. The idea behind this process was to capture the soft skills that employers may be mentioning are important in a candidate aside from coding languages, tool knowledge, etc. Some examples of the descriptive characteristics by this filter include; team, data discovery, actionable insights, and talented data team. (see attached kg\_looking\_for.png)

These knowledge graphs resulted in conclusive insights on the characteristics of candidates the employers are mentioning. This information could be helpful for those who are wondering about where to focus their efforts in training for a certain position and what kind of information to highlight when they are applying for these types of positions.

1. **Clustering**

Since grouping the job postings by job title requires more time and resources than what was afforded for this study, clustering by Kmeans on the job descriptions were conducted in an effort to identify different types of positions that are being advertised in this data set.

For the clustering method, each document was tokenized and converted into vectors using the spaCy model. The data set was then converted to a sparse term-frequency inverse document frequency. This method was chosen over using term frequency as it resulted in better groupings when searching for the most optimal number of clusters. Below is a graph of the number of clusters that are dependent on the sum of squared differences.



Using the elbow method, there are two potential optimal K’s that can be used for this analysis. Either seven or nine clusters. For the context of this analysis, using a smaller number of clusters would be more likely to result in interpretable results, so we opted for seven clusters to analyze.

1. **Cluster Interpretation**

Since each cluster is based on the job description of each observation, and reading even the top three (smallest distance to the cluster centroid) job descriptions for each cluster closest to the medoid would result in inconcise depictions for each cluster.

The job titles of the top three documents for each cluster resulted in concise and interpretable results. For further interpretation, we created a simple word cloud for each cluster and displayed our results (see cluster\_results.docx). The combination of job titles and word cloud analysis resulted in the following labels for each cluster:

1. Job postings created by Jefferson Frank (prominent recruiter for AWS)
2. Medical Field positions
3. Employers that mention equal opportunity
4. Government entity or contractor
5. (possibly) Start-up company positions
6. Senior Positions
7. Unknown

These clusters offer some insight into the types of positions being advertised and may be helpful for those searching for positions that are in a specific industry. For example, if working for a company that is an equal opportunity employer is important for a job seeker, the positions included in cluster three may lead to better results in their job search.

1. **Conclusion**

After analyzing the descriptive statistics, job title groupings, extracting key features, and clustering the documents, we’ve learned several conclusive points. These conclusions are operating under the assumption that the data collected is a random sample of all job postings in the job market as a whole and having a domain constraint of postings in the United States in the year 2019.

Grouping by interpreted seniority, the job postings are skewed heavily towards senior level positions for candidates with more experience when compared to junior/entry level positions. This could be an indication that experienced professionals are in higher demand than junior level, less experienced professionals. It’s harder to break into the data science field, but once a professional gains more experience, the job market flips to their favor.

After extracting key features from the job descriptions, it’s apparent that employers mention soft skills more often than technical skills. Thus, we can infer that employers are prioritizing soft skills over technical skills. This makes sense as data scientists are usually working as a team tasked with communicating their results. Oftentimes, employers purchase third party solutions like Tableau that reduce the technical requirements of their employees and allows them to focus on translating modeling results into descriptive, predictive, and prescriptive results. The communication of these results in an interpretable and accessible manner seem to take higher priority than technical skills alone.

Technical skills, while concluded as less important than soft-skills, are still a requirement for the majority data scientist roles. This was shown while observing the entities in sentences that were in relation to ‘experience’, as well as observing the results extracted when using pattern matching techniques. The high frequency of positive matching by these techniques show that these technical skills are being mentioned very often by employers. Thus, we can deduce that, while not being as important as soft-skills, the majority of data science positions require a plethora of technical skills.

The results of this analysis reinforce the conventional wisdom that data scientists are expected to be well-rounded professionals, with an emphasis on soft-skills as well as technical proficiency.

**Appendix**

1. **jobspikr. (2019, December 17). *10000 data scientist job postings from the USA - dataset by jobspikr*. data.world. - Retrieved date December 14, 2021, from** [**https://data.world/jobspikr/10000-data-scientist-job-postings-from-the-usa**](https://data.world/jobspikr/10000-data-scientist-job-postings-from-the-usa)
2. **Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand** - Aankur A.Patel - ISBN: 9781492062578 - Chapter 8 - Bertology

# Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem - Duygu Altinok - ISBN 10:1800563353 - Chapters 4- Rule Base Matching

# Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning - Benjamin Bengfort - ISBN-10: 1491963042- chapter 6 Clustering for Similarity and Chapter 8 Text Visualization

1. **NLP - Natural Language Processing with Python - Udemy - Learn how to use ML, Spacy , NLTK, SciKit-Learn, Deep Learning, and more to conduct Natural Language Processing** - created by Jose Portila

# Natural Language Processing with Classification and Vector Spaces - Coursera - created by Younes Bensouda Mourri, Eddy Shyu