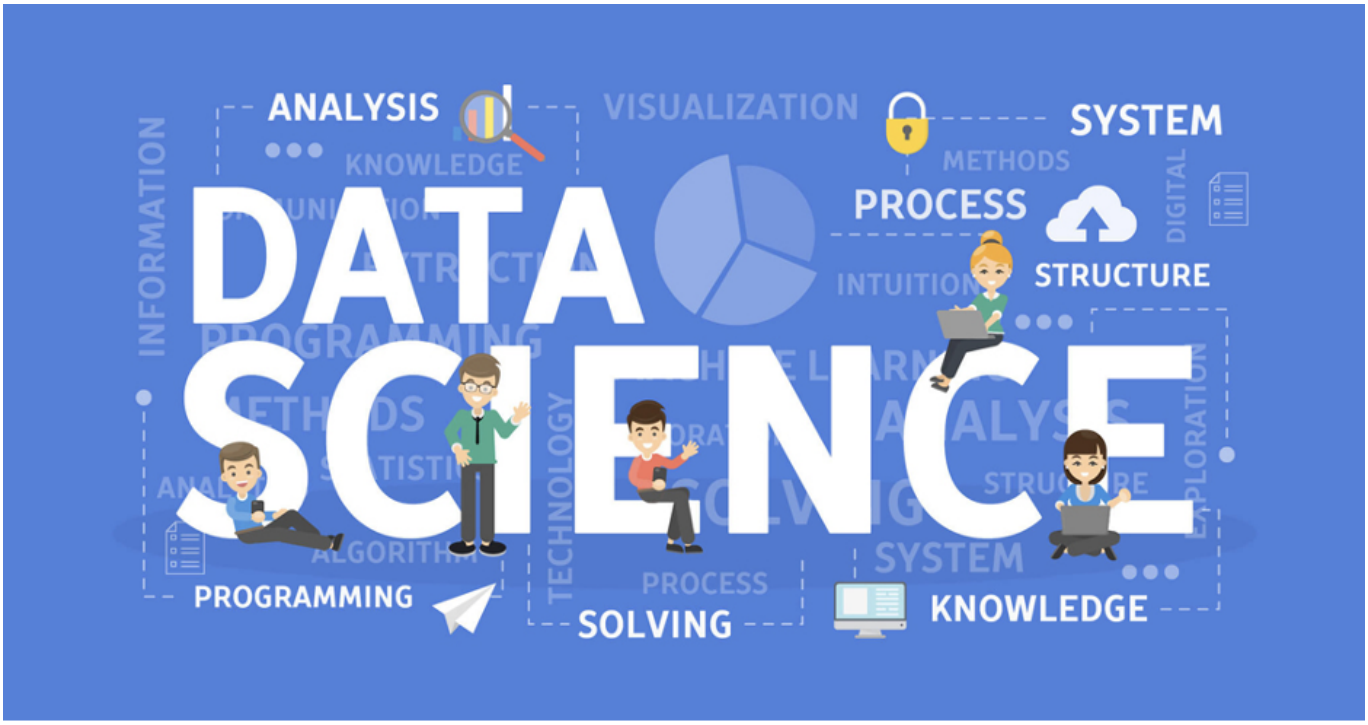
Site Bai

NLP on Data Science Related Job Postings — What You Can Do to Make Your Resume Pop and Find Your Dream Job at the Highest Salary

Site Bai Jun 11 · 16 min read

Written by: Site Bai, Dido Chang, Sammi Chang, Kevin Cheung, Song Han

> *a meaningful and tactical NLP exploration into job postings that will help our cohort find their dream jobs*



At the University of California, Irvine, my teammates and I were asked to choose an NLP dataset to derive insights from. As our cohort of almost 100 find ourselves busy in the midst of capstone projects, other final projects, graduation, and — the most stressful of them all — finding a job, we chose, as our dataset, Glassdoor job postings.

We soon realized that this was not just a normal course project but a meaningful and tactical exploration into job postings that will help our cohort find their dream jobs. This will help us and our classmates customize our resumes with the pop words that companies are looking for. It will also shed light into popular states that are hungry for data science jobs, and the certain features of job postings that may hint at higher salaries.

> *In fact, we all have an idea of what these are, just by reading job postings. But how do you know if you are correct? With NLP, we are able to read thousands of job postings at a time and aggregate our analysis over two years worth of job postings. We will take you through some facts that we discovered that are backed up by data.*

Objective

The goal of this project is to extract which hard and soft skills are needed for four different kinds of jobs: data analyst, data scientist, data engineer and business analyst and form a recommendation for the reader. Then, with the extraction of these skills and several other company-dependent variables, we will use an OLS regression model on salary to measure which skills or features have the greatest (or lowest) effect.

Dataset

The data was sourced from two different datasets on Kaggle: the jobspikr dataset, which consists of 10,000 rows of data scientist job listings across different cities in the U.S; and the andrewmvd dataset which consists of about 12,000 rows of data analyst, data scientist, data engineer and business analyst job postings scraped from Glassdoor. The two datasets were then inner-joined, arriving at a little over 20,000 rows. The variables that we had to work with are: the job title, job description (which contains the skills and qualifications), salary (an estimate), city, state, and company name. For general EDA,

stratified balance between our four job types and to prevent a bias from being introduced.

Data Pre-Processing and Manipulation

To start off, much of the data was not organized in the way that we wanted. We used regex to extract the data embedded in the strings.

Firstly, we used regex to remove all text and symbols from the salary column to only keep the numbers. Then, we separated it into salary low, salary high, salary range, and salary mean, of which the latter was used to feed into our models.

A preview into our data.

To extract the state, we cleaned up the column with regex and made sure all state names were in the form of abbreviations:

```
# Cleaning state column
data['state'] = data.state.str.replace(r'\d', '')
data['state'] = data.state.str.replace(r'\(.*\)', '')
data['state'] = data.state.str.replace(r'Virginia', 'VA')
data['state'] = data.state.str.replace(r'New Mexico',
'NM').replace(r'New Jersey', 'NJ').replace(r'Idaho',
'ID').replace(r'Massachusetts', 'MA').replace(r'Louisiana',
'LA').replace(r'Hawaii', 'HI').replace(r'Washington State', 'WA')
data['state'] = data.state.str.replace(r'\s*', '')
data['state'] = data.state.str.replace(r'\-', '')
data['state'] =
data.state.str.replace(r'(Computerorinternet|Remote|WorkatHome)',
'WFH')
data['state'] = data.state.str.replace(r'(NN|om)', 'UK')
data['state'] = data.state.str.replace(r'Engineeringorarchitecture',
'TX')
```

```
# tokenizing job title and removing stopwords/digits/punctuation
stop_words = stopwords.words('english')
to_remove = stop_words + list(string.punctuation) +
list(string.digits)
data['job_title_token'] =
data['job_title'].apply(word_tokenize).apply(lambda x: [item for
item in x if item not in
 to_remove])
```

Upon having the cleaned job titles, we extracted specific job levels from the title and categorized them into different levels of seniority as shown below:

```
# Creating levels for different levels
data['level'] = data['level'].replace("(vice president|vp)", 'exec',
regex = True)
data['level'] = data['level'].replace("(chief|director)", 'head',
regex = True)
data['level'] = data['level'].replace("(lead|principal|manager)",
'lead', regex = True)
data['level'] = data['level'].replace("(senior|iii|sr)", 'senior',
regex = True)
data['level'] = data['level'].replace("(associate|ii|sr|staff)",
'mid', regex = True)
data['level'] = data['level'].replace("(junior|jr|entry)", 'junior',
regex = True)
data.level.fillna('none', inplace = True)
```

We were then able to look at the distribution of job titles in our data:

Counts of roles in our data. Data scientist comes out on top because of our ds10k dataset join.

We removed any jobs that did not have a job title and imputed salary NA values with the mean based on level and role. From the company name, we also remove the '\n' and the company rating. The rating could not be used as the other dataset that we used to join did not have this information.

A preview into how our company column looks like.

```
# Removing the '\n' from company name
data['company'] = data['company'].str.split('\n',expand=True)[0]
```

Lastly, we move on to the job description, which is our area of interest. Below is an example of a typical job posting. As you can see, there were many '\n' that had to be removed. Other job descriptions included links and emails, which we included in our targeted removal. Also, each job description contained the legal blob of text that read, somewhere in the lines of: "We are committed to equal employment opportunity regardless…". However, since this blob was slightly different for every job posting, we decided to forego its targeted removal, and later realized that its removal was not mandatory for the analysis that had to be done.

Our messy job descriptions before any pre-processing.

The job description was then tokenized and then detokenized to achieve two elegant versions of the job description: job_desc_token, and job_desc_clean. The tokenizer put each word into a list of strings, which was just what we needed.

```
# Cleaning job description
data['job_desc'] = data['job_desc'].str.lower()
data['job_desc'] = data['job_desc'].str.replace('\\n',' ')
data['job_desc'] =
data['job_desc'].str.replace('http\S+|www\S+|https\S+','')
data['job_desc'] = data['job_desc'].str.replace('\S+@\S+\.\S+','')
# Use this to remove unwanted in job description, kept digits
stop_words = stopwords.words('english')
to_remove2 = stop_words + list(string.punctuation)
#create col: tokenizing
data['job_desc_token'] =
data['job_desc'].apply(word_tokenize).apply(lambda x: [item for item
in x if item not in

to_remove2])
#create col: detokenizing
data['job_desc_clean'] =
data['job_desc_token'].apply(TreebankWordDetokenizer().detokenize)
```

We're now ready to make some visualizations and EDA.

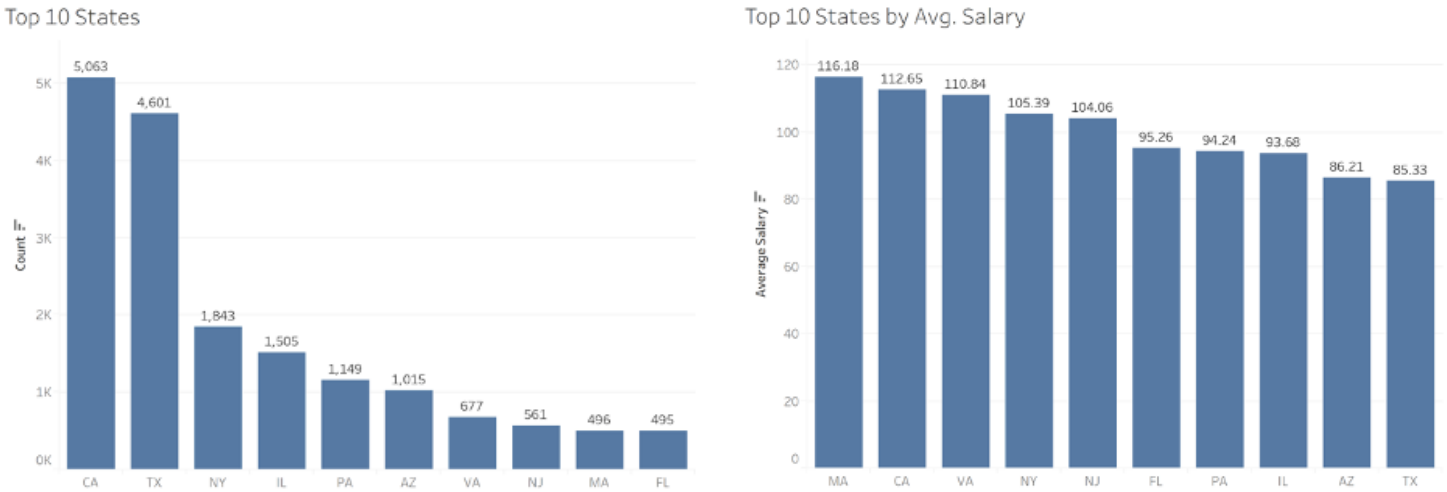
EDA/Visualizations

The WordCloud shown below was used on the entire dataset with our own list of custom stopwords. Spacy was used for lemmatization and the result was tokenized for the word

that appear the most throughout the whole dataset.

A word cloud that encompasses all that has to do with data jobs.

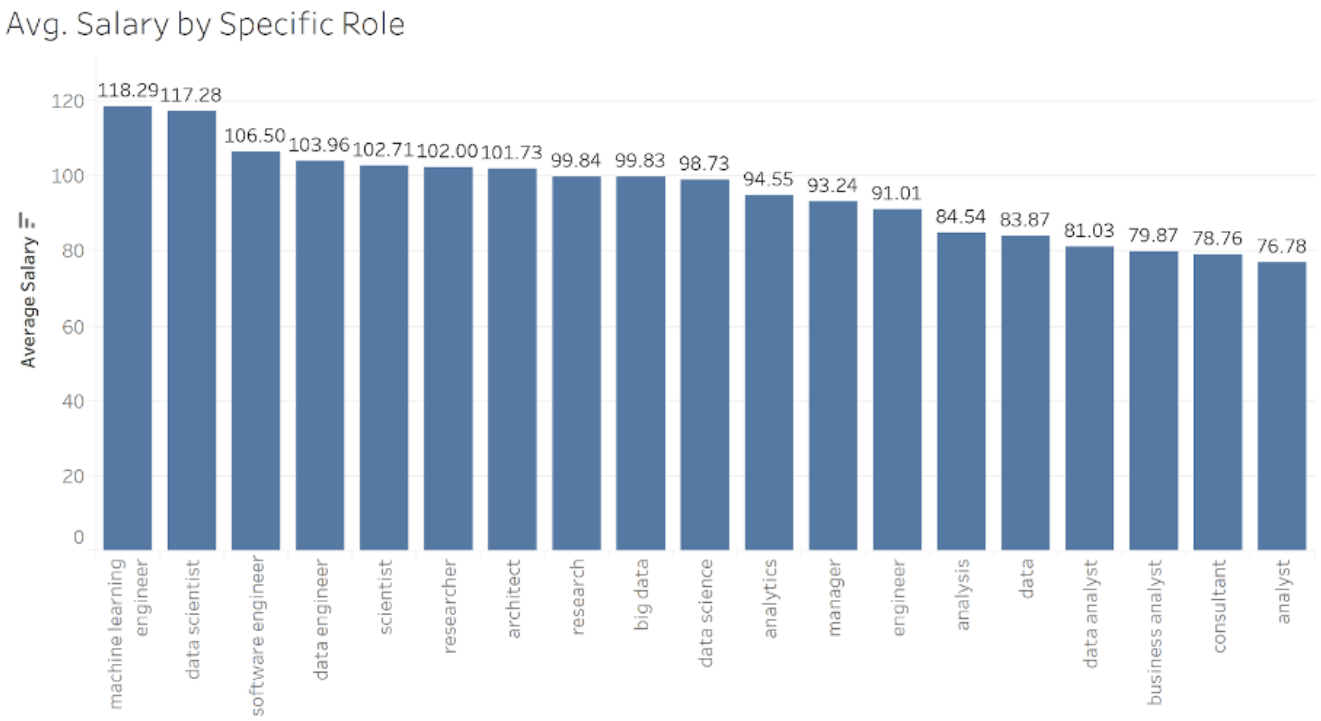
The graph on the left shows the top states based on job posting counts. California, Texas, and New York have the most job listings; but out of the top three, the average salary in Texas is the lowest, likely due to Texas' lower cost of living.



Left: Top 10 states by count (numbers in 1's). Right: Top 10 states by average salary. (numbers in thousands)

only west coast state on this graph.

As a preview into our linear regression model, here is a comparison of the average salaries for each role. It looks like machine learning might be one of the most important skills to learn or list on your resume.



Average salary by specific role. Machine learning engineers are paid the most, with “analyst” coming out on the bottom at $76k a year. (Numbers in thousands)

Not surprisingly, the salaries are greater as you move up the ladder. If the job title writes “Junior,” it would be safe for you to assume that you would be paid lower than if the job title did not contain the word. This will be useful for salary negotiations and may serve as a baseline measurement when answering the “compensation question.”

The average salary by job rank. Salary climbs as one climbs up the ladder.

Word2Vec and Custom Lists

Note: From here, we remove the jobspikr data from our dataset to keep the number of rows of all four types of jobs similar. This will prevent skews in our analysis.

We wanted to do a word count for skills of several types — for example, programming skills, soft skills and etc.. To do this, we must first form word lists for which we use to implement a count. Using Word2Vec from Gensim models, we trained a model which would input our job description tokens and output other words that are similar in domain meaning.

Here is a sample of our Word2Vec in action. For ‘python’, the closest words are ‘r’, ‘java’, ‘scala’ and so on. Thus, our word list for programming skills will include ‘r’,’ java’, ‘scala’, ‘matlab’, and ‘c++’. Other programming languages like SAS and SPSS were also added.

Programming skills that are most similar to “python.” These are then fed into our Counter.

Skill Counting

> knowing SQL is a must have skill, even for business analysts roles that require less technical skills.

Comparing Across 4 Different Roles: Business Analyst, Data Analyst, Data Engineer and Data Scientist

As mentioned before, W2V helped us develop custom lists for word counting. In this section, we will investigate the distribution of word counts for words within each skill group. We do this for each of the four data-related jobs. Our goal is to give recommendations based on the most popular skill. In addition to capturing unigrams, we also utilized NLTK ngrams function to extract bigrams from the job description, which was useful for detecting bigrams like “machine learning”.

```
# Custom lists extracted from W2V
programming = ['python', 'r', 'c++', 'java', 'c', 'matlab', 'sas',
'spss', 'stata']
visualization = ['chartio', 'tableau', 'looker', 'powerbi',
'visualization', 'story-telling',
                 'spotfire', 'rshiny', 'qlikview', 'domo',
'webfocus', 'qlik']
data_eg =
['agile','aws','bigquery','bigsql','cassandra','docker','dockering',
'hadoop','hbase','hdfs','hive','hivesql',

'kubernetes','mongodb','mysql','nosql','pig','pyspark','s3','scala',
'spark','sparksql','sql','teradata']
python = ['scipy', 'pandas', 'sklearn', 'statsmodels', 'scikit-
learn', 'geopandas', 'pybrain', 'scikitlearn', 'numpy',
'matplotlib']
education = ['phd', 'ph.d.', 'doctorate', 'masters', 'master', 'ms',
'm.s.', 'bachelor', 'bachelors', 'bs', 'b.s.']
experience = ['1+', '2+', '3+', '4+', '5+', '6+', '7+', '8+', '9+',
```

```
regression'), ('k', 'means'),
     ('random', 'forest'), ('naive', 'bayes'), ('pca', 'svd'),
('decision', 'tree'), ('ensemble', 'model')]
dl = [('neural','network'), ('deep', 'learning'), ('object',
'detection'), ('keras', 'tensorflow'),
     ('convolutional', 'neural'), ('tensorflow', 'keras')]
softskill = ['communication', 'interpersonal', 'verbal', 'written',
'oral', 'inter-personal', 'cross-functional',
            'cross-organizational', 'multi-functional', 'teamwork',
'collaboration']
```

The code chunk below shows how these bigrams and unigrams are counted for each tokenized job description.

```
#for programming skills
matches_programming = []

for i in data1['job_desc_token']:
    for j in i:
        if j in programming:
            matches_programming.append(j)
```

The custom words and counter pave us a path to compare which skills should be highlighted for each job type. Here, we look at the distribution of education requirements across all four job types. Data Analyst and Data Scientist roles mention advanced degrees in their job postings more often than Business Analyst and Data Engineer roles. When we look at Data Scientist specifically, doctorate degrees are mentioned more than 1 in 5 times, which suggests that having a doctorate degree will make you more competitive in the job market. To become a data engineer or business analyst, having a bachelor's degree should suffice more than half of the time.

It is an interesting comparison that, although master's degrees are mentioned about half the time for data analyst jobs, the average salary for data analysts is 30% less than that of data engineers where bachelor's degrees are more prevalent. It may be that degree levels do not have a direct correlation to salary. For discussion's sake, we take notice that the subset of data analysts that require master's degrees may in fact have higher salaries than data engineer jobs that only require a bachelor's degree. Following this line of analysis for future studies may give way to more in-depth insights.

The graph below shows the years of experience that are needed for the four jobs. The line graph spike from 2 to 5 years show that having 2, 3, 4 or 5 plus years of experience may make job searching easier.

Years of experience needed are similar across all four jobs.

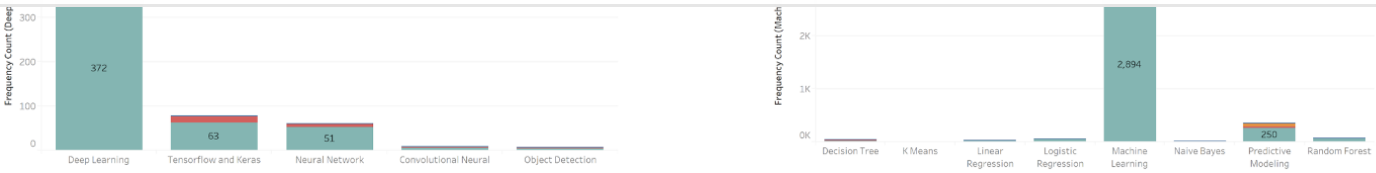
If you are wondering what kind of database management skills you may need, knowing SQL is a must have skill, even for business analysts that require less technical skills. If you are looking specifically at data engineering roles, knowing AWS, Hadoop, Hive, and Spark will make your resume stand out. Some data science roles will also ask for these skills — AWS and Spark look particularly prominent.

Data engineering skills: SQL wins.

Data scientists and data engineers once again display the most need for technical skills. Python is the most popular programming language for Data Scientists and **beats R** by a landslide. Surprisingly, Business Analysts and Data Analysts may not need an extensive coding background. Java, the old-school programming language proves still relevant, even amongst data scientists. And lastly, it seems like SAS is **just as important** for data analysts to master as is R.

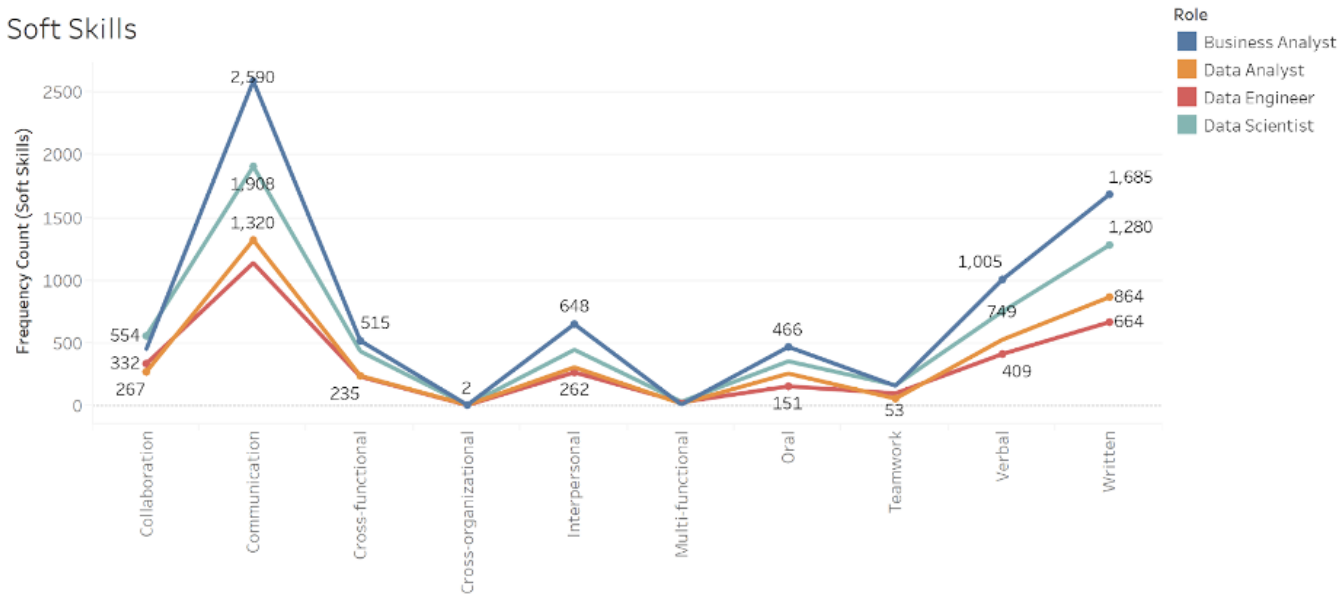
Programming skills: Python wins.

If you are wondering if you will be applying your deep learning knowledge in the workforce, there may be a place for that, especially amongst data scientists, but not so much for business analysts (and similarly for machine learning skills):



Deep learning and machine learning skills are particularly important for data scientists but less so for data analysts. It's nice to have for business analysts but not largely required.

Not surprisingly, Business Analysts lead in the soft skills. In all four job types, communication is the most important, with written skills trailing second.



Business analysts dominate soft skills, particularly communication skills.

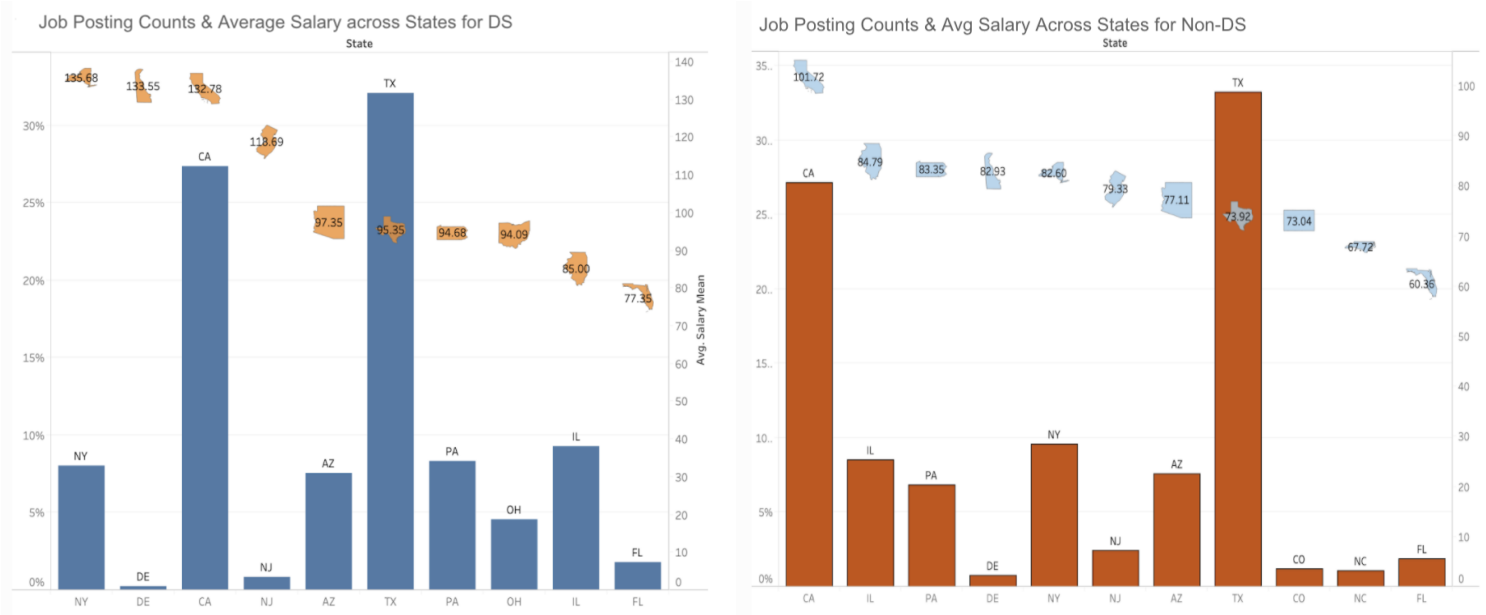
And lastly, Data Engineers do not need Visualization skills as much as the other roles. Tableau is the most commonly used visualization software, followed by PowerBI.

Tableau seems to be the most popular.

Ordinary Least Squares Regression for Salary: Data Scientists vs Non-Data Scientists

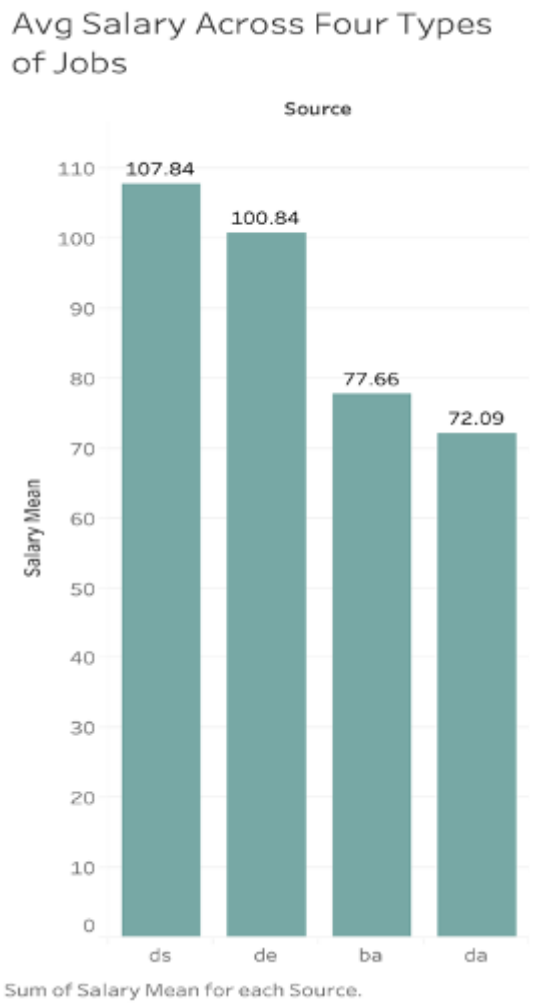
Finally, we want to introduce a salary component. We ran an OLS regression model on salary for data scientists, and also ran another for non-data scientist, data-related jobs (data analyst, data engineer, business analyst). The motivation behind this kind of separation was that we noticed a divide, in terms of education, experience and skills required, between data scientist and non-data scientist jobs.

As a preface into our linear regression model, we want to introduce some similarities and differences between the counts of job postings in each state versus the average salary. It may be easier to get a job interview in Texas, which has the most job postings due to a growing economy**. However, this does not necessarily mean that you would get a higher salary — in fact, the average salary in Texas places 30% less than that of California.**



The numbers overlaying the state icons are the average salary; the bars themselves represent the job posting count for each state. The right axis on both graphs are the percent of total job postings along table (across).

The salary also differs between the four types of jobs that we want to study, with data



From left to right: Data Scientist, Data Engineer, Business Analyst, and Data Analyst. (Numbers in thousands)

Non-DS Jobs

We ran the regression model to find the effect of various variables on salary for non-data science jobs. For non-data scientists, salary is most positively impacted by California and the least impacted by the sunshine state, Florida. Working at public companies offers a $12k higher marginal return on salary than working at private companies. Lastly, out of the skill dummy variables listed towards the bottom, programming skills have the largest effect on salary, so learning python, R or other required languages is well worth the investment. Deep learning skills like keras and tensorflow are also important for non-data science jobs.

The significant coefficients for Non-Data Science roles. It seems like machine learning is just as important.

DS-Jobs

We also ran a regression for Data Science jobs and found that the results are mostly the same; however, Delaware competes with California for having the highest effect on salaries. Machine learning is also important for data scientist jobs. For data scientist jobs, working for a public company would provide $9k marginal return of salary compared to private and government companies. Furthermore, working in a 10k+ employee company offers a positive impact on salary as well.

The significant coefficients for Data Science roles.

> *Since we were studying a subject that we were especially familiar with, it was, simply, hard to let go of our initial beliefs of what it meant to be a data scientist…*

To talk about the limitations first, perhaps one of the limitations of this dataset is that we did not have review data for the particular job and for the company. With this information, we could have used sentiment analysis to add another layer that may link the relationship between, for instance, the presence of ML skills or experience requirements on a job posting to the overall sentiment that employees feel for their companies.

An area of interest starting this project is to track the changes in job requirements imposed by the current pandemic, COVID-19. Although it would have been rewarding and contemporary, we did not have continuous time-related data across the period of 2019–2021 that would help us track these changes. However, future research into this subject may consider gathering the data appropriate for this approach.

Lastly, in our OLS regression model and EDA, an area for refinement lies in taking into consideration the income tax of each state. With this we may achieve a more normalized result across all states.

One challenge that we faced while implementing the Counter was due to messy text data. For instance, extracting the years of experience for each job post proved difficult as each company had their own standards of writing that information. Some companies would write “2–5,” while others preferred “2–5+”, while even others used “2 to 5” years of experience. To remedy this, we kept it simple and only used “2–5” and “2–5+” by assuming that the distribution of counts would not change if we included or excluded “2 to 5”. To make sure that we did not capture *non-experience* numbers that are actually part of the job description, we made sure to tag on a “+” after the numbers in our custom list.

Lastly, we want to talk about the human bias factor that surrounded us throughout the project. Since we were studying a subject that we were especially familiar with, it was, simply, hard to let go of our initial beliefs of what it meant to be a data scientist, or how the skills differ between a business analyst and a data analyst. To prevent bias, we

would not be required to learn Tableau, we still ran the Counter for it. The result was — although we were mostly right — that there were actually 10% of data engineer job postings that required Tableau.

In Conclusion

If you are flexible about location, the best states to apply for jobs might be Texas, California and New York. And if compensation is important to you, we recommend public companies in California, Illinois, Pennsylvania and New York, which may provide the highest salaries.

In terms of our word count results, we outline the important points that we hope you take away from this article:

Data Scientist: With the highest salary out of the four, the data scientist requires the most technical skills, namely Python and R. Machine learning is important to mention on your resume, but it is also important for non-data science jobs as well. Having an advanced degree such as a master's or doctorate would give you a leg up in the job market.

Data Engineer: With the second to highest salary out of the four, a data engineer role requires strong programming and database management skills. The data engineer must know Python and Java and can put less emphasis on visualization skills. In addition, the data engineer must be familiar with AWS platforms and be able to use Hadoop, Spark, Hive and most importantly, SQL.

Data Analyst: A data analyst needs to have strong SQL skills, and Tableau, Python, R and SAS will surely enhance your resume but may not be required for many job listings. *However, it is important to note that, even if a job listing does not list SQL as a requirement, it is important to include it in your customized resume as its importance ranks high on our aggregate of data analyst job postings.*

Business Analyst: Contrary to popular belief, technical skills such as Python or R are not particularly required but will enhance your resume and competitiveness. More than

A Discussion

Three years ago, I attended an interview for a non-technical analyst position straight out of college. I listed on my resume that I had basic SQL knowledge. The manager was surprised, and with eagerness and a smile asked me how much I knew about SQL. He was planning to build a data science team within his department. This shows that, although you may not have 2–5 years of experience, knowing some programming skills will take you a long way. As emphasized in this article, it is well worth the investment.

We hope that this will be useful in your job search. And if you are not looking for a job, this article will help you understand how the world looks like through the lenses of NLP, and of the job postings of data scientists, data engineers, data analysts and business analysts.

Link for Kaggle datasets:

andrewmvd: https://www.kaggle.com/andrewmvd/data-analyst-jobs

ds10k: https://www.kaggle.com/jobspikr/data-scientist-job-postings-from-the-usa

Link for Github:

https://github.com/sitebai21/projects/blob/master/NLP_JOBPOSTINGS_PROJECT.ipynb

NLP Regression Analysis