Job Market Analysis

Data collection: June 12th, 2018

Team

Hanying Li, Timothy Murphy, Michael Nguyen, Qixian Zhao

Code Repository

URL to Github: https://github.com/qxzhao1/STA160-Project

Website

URL to Website: https://lhyhyfwjwf.github.io/index.html

Part I: Abstract

Entering the job market is intimidating to many job seekers and newly graduated college students. It does not have to be this way. Our goal is to provide job seekers knowledge they can use to make job seekers more competitive. When someone is searching for an employment opportunity, they take into account the industry, company, duties, salary, location, and many other features. The purpose of this project is to gather, explore, and analyze job listing data to provide job seekers with insight on the skills required for a full-time data scientist, data engineer, data analyst, and business intelligence positions across the United States. Our team decided to collect data by scraping job postings from the Cybercoder and the Simplyhired. Through the use of Natural Language Processing (NLP) and Text Mining (TM) methods, we obtained insights and recognized patterns that are unknown to the typical job seeker. We took a particular interest in the skill-set information as well as salary estimates provided by the job websites and then related our findings to geographic regions by comparing and contrasting skill requirements in different states, cities, and broader areas such as the east and west coast.

Part II: Data Collection

Our initial goal for collecting job data was to scrape the Indeed website (https://www.indeed.com) since it is the most popular job website and has the largest population of listings. However, after our team scraped approximately 500 listings and attempted to clean the data, we encountered a large number of issues that would be too difficult to tackle in our 10-week timeline. It is important to highlight some of these issues:

1.) Salary information is mostly absent on this website, which is likely due to employers not wanting to release this data to the public. Also, salaries are often variable between candidates and usually negotiated before the acceptance of a job offer.

2.) The job descriptions lacked structure, meaning that most of the descriptions for the listings did not have a consistent pattern on the backend HTML scripts. Therefore we would have needed to scrape chunks of commonly structured descriptions using a different algorithm each time, and this was not conducive to our timeline. Even if we did proceed down this path, most of the

descriptions included information that is not useful to our projects focus. Thus, making it not worth our limited time and resources.

After exploring Indeed (https://www.indeed.com) and other job sites we discovered two that our team felt aligned with the scope of our project, held a clean structure for scraping, and contained the information that allowed us to perform our analysis. Therefore, we landed on Cybercoder (https://www.cybercoders.com), and Simplyhired (https://www.simplyhired.com). These are suitable for our project because they contained the following data, which met our criteria: Titles, Job Descriptions, Preferred Skills, Locations, Salaries.

Figure 1.1: Cybercoder Data Set

CyberCoder Data set				
Subset	Title	Number of tuples	Attributes	Missing data
Subset 1	Data Scientists	72	Description, Skills, Location, Cities, States, Latitude, Longitude, Min_Salary, Max_Salary, Mean_Salary	No
Subset 2	Data Analysts	49	Description, Skills, Location, Cities, States, Latitude, Longitude, Min_Salary, Max_Salary, Mean_Salary	No
Subset 3	Daata Enginner	349	Description, Skills, Location, Cities, States, Latitude, Longitude, Min_Salary, Max_Salary, Mean_Salary	No
Subset 4	Business Intelligence	49	Description, Skills, Location, Cities, States, Latitude, Longitude, Min_Salary, Max_Salary, Mean_Salary	No

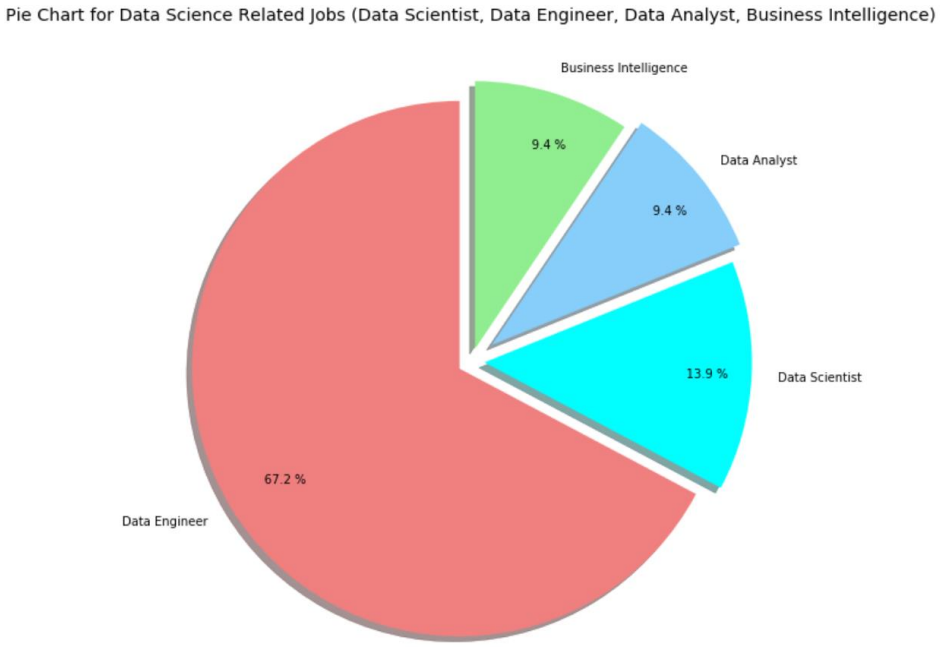


Figure 1.2: Simplyhired Data

Simplyhired Data set				
Subset	Title	Number of tuples	Attributes	Missing data
Subset 1	Data Scientists	17,229	Job title, Company, Location, Salary, Snippets and Description	Missing Salary data ~ 100
Subset 2	Data Analysts	76725	Job title, Company, Location, Salary, Snippets and Description	Missing Salary data ~ 100
Subset 3	Data Enginners	120327	Job title, Company, Location, Salary, Snippets and Description	Missing Salary data ~ 100

The two charts Figure 1.1: Cybercoder Data consisting of (Data Scientist, Data Engineer, Data Analyst, and Business Intelligence) data and Figure 1.2 Simplyhired Data consisting of (Data Scientist, Data

Engineer, and Data Analyst) data show a brief overview of the two datasets that we have scraped. As we can see here, we have collected mostly complete information for each of the jobs types. With a complete dataset (lacking null values) we can conduct a more in-depth exploration of the relationships and lack thereof these job types have with each other. Also, it is worth noting that the Cybercoder dataset after cleaning contains descriptions, skills, location, and salary estimates but lacks companies and industry data. The lack of company information is most likely because staffing agencies use Cybercoder and their client's information is kept confidential to job seekers unless contacted by the recruiter. The lack of this company information is a downfall of this dataset; however, for our purposes, the data is still suitable. Missing company information is not the only drawback of the Cybercoder data; it is not big enough as it only has 519 postings making it relatively smaller compared to the Simplyhired data. Also, the proportions of job types within the population skew in one direction due to 67.2% of the listing data distributed to the Data Engineer position. Unfortunately, we don't know if this is true to the real world. Meaning we do not know if the demand for Data Engineer positions is several times higher than other data related positions.

However, when we turned to the Simplyhired dataset, we can see that it is enormous. With a massive number of job postings, we can gain greater insight into the relationship the preferred skills have with other fields and be able to more precisely analyze how job types are stratified geographically, thus making it a more robust analysis. However, unlike the Cybercoder website where salary information is known to the recruiter, we later discovered that the salary data are estimates calculated by the Simplyhired website and are therefore not going to be representative of the actual values.

Part III: Cybercoder Analysis

How many listings are represented for each job type (Data Scientist, Data Engineer, Data Analyst, and Business Intelligence) within each state?

When first exploring our data, it is essential to understand the distribution of not only the job types but of the fields themselves. However, as was noted in "Data Collection" section, our data is heavily skewed due to the dependency on the number of listings posted for each job posted on a career website at any given time. In our case, 67.2% of our data for the Cybercoder dataset consists of Data Engineer positions, meaning this can have a relationship with the geographic distribution of these jobs. Intuition says that in most cases a particular job is going to be in higher demand in different places across the USA. On the other hand, since our scope is on data related jobs, this assumption can be violated since analytics departments within companies have teams of these job types. To summarize this point a typical data science team for these positions consists of the following:

Figure 1.3: Background on the Four Cybercoder Job Types

Data Scientist: Someone who knows how to extract patterns otherwise unseen from data and interpret it to member across an organization. Data Scientists utilize tools and methods from statistics and machine learning, as well as creativity and intuition only found in humans. They also spend much time in the process of collecting, cleaning, and wrangling data, because data is unfortunately almost never clean.

Data Engineer: Someone who is responsible for the creation and maintenance of analytics infrastructure that enables almost every other function on the data science team. Data Engineers are responsible for the development, construction, maintenance, and testing of architectures. These responsibilities apply to company databases and large-scale processing systems.

Data Analyst: Someone who is responsible for translating as well as understanding metrics and numbers that a business collects data, whether it's sales figures, market research, logistics, or transportation costs. Data Analysts take that data and use it to help companies make competent business decisions.

Business Intelligence: Someone who is responsible for using the results and data found by the Data Scientist and Data Analyst to provide the company with easily understood visualizations. Business intelligence (BI) is more generally an umbrella term that includes the applications, infrastructure, tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.

In other words, for analyzing the distribution of listings geographically by state and city, the massive proportion Data Engineer positions represented in our data can be ignored. The assumption that location and the job type distribution are independent within our data is because these job types rely on each other as a team and companies hire for each one regardless of location.

Figure 1.4: State Distribution of Jobs Listings

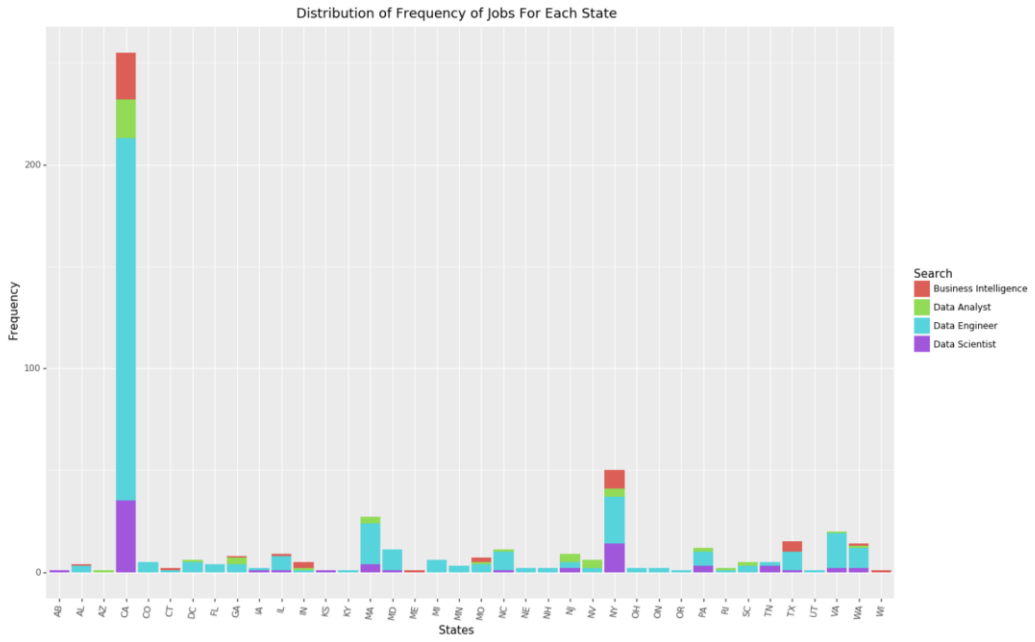
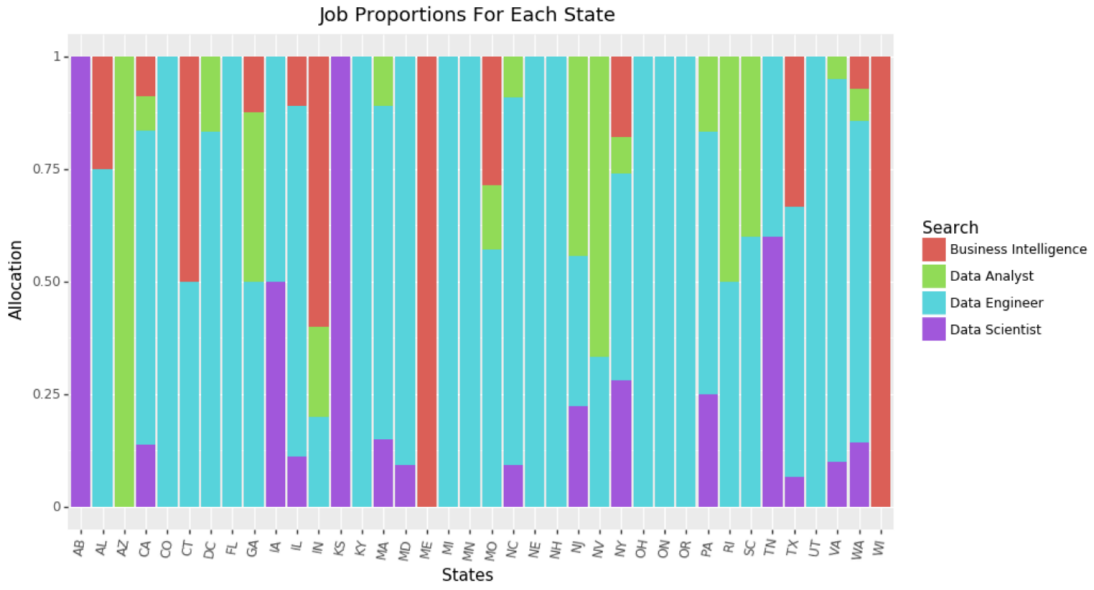


Figure 1.5: State Distribution of Job Listings as a Percentage of the Total



As you can see from Figure 1.4 that California, Massachusetts, New York, Texas, Virginia, and Washington stand out in that they represent each of the four job types. This finding makes sense because these states have a significant number of technology companies and therefore data related jobs. For example, unrelated to our data, CA with the Bay Area is the hub for everything data with companies like Facebook and Google that have warehouses holding and processing roughly 300 Petabytes of data each (322,122,547,200 megabytes). That is almost a comprehensible amount of data, and with that in mind, it makes sense that companies in California are in need of filling Data Scientist,

Data Engineer, Data Analyst, and Business Intelligence positions. It is worth noting, some states are very particular in the jobs posted in their state, and this can be related to the industry type representing these listings in that state.

Figure 1.5 shows that the states with only a solid color are only seeking a single job type. Unfortunately, we do not have information on the companies listing these jobs, so our team is not able to reach a robust conclusion relating to these abnormalities.

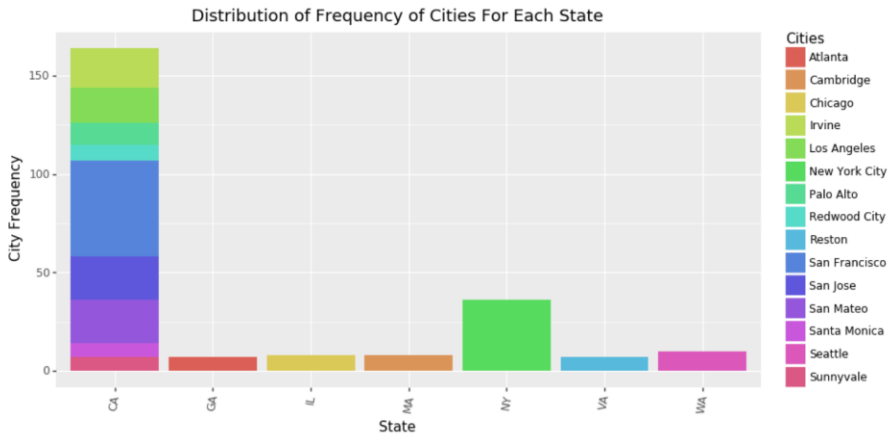
What are the twenty cities with the highest frequencies for jobs in our population and do any stand out which may indicate a hotspot for data science jobs?

After exploring the distribution of state frequencies within our Cybercoder population, it is essential to understand the distribution of the not only the states where these jobs are located but also of the cities with the highest number of these jobs. One of the most common and significant determinants of job seeker interest is where that company has their headquarters and more precisely, the particular position. That is why analyzing what cities are representing the most substantial number of listings is vital in helping a job seeker better understand where they should be looking and whether their location preferences align with the supply of these positions. Also, by taking a closer look at the distribution of cities within the states, we can determine whether the most represented states have a more diverse option for a job seekers location preferences.

Similar to our previous assumption we can ignore the fact that California is vastly represented in our population and can conclude that the following cities may be good places to search for Data Scientist, Data Engineer, Data Analyst, and Business Intelligence jobs.

Figure 1.6: Distribution of the Cities within Top 15 Frequent Locations

	States	Cities	Frequency
0	CA	San Francisco	49
1	NY	New York City	36
2	CA	San Mateo	22
3	CA	San Jose	22
4	CA	Irvine	20
5	CA	Los Angeles	18
6	CA	Palo Alto	11
7	WA	Seattle	10
8	CA	Redwood City	8
9	MA	Cambridge	8
10	IL	Chicago	8
11	GA	Atlanta	7
12	VA	Reston	7
13	CA	Santa Monica	7
14	CA	Sunnyvale	7



We can see from above that 9 out of the top 15 represented cities are located in the Bay Area, making it a great place to look for a data science related position. Figure 1.6 also reveals an exciting part of our job listing population that aligns with our previous findings. We can see that California has a diverse option for the cities where one can search for listings. Once again, this is related to the vast number of companies seeking those that have the skills necessary to understand the immense amount of data they are collecting. The other states do not have these diverse set of city options, indicating that majority of technology firms in these states are only in specific cities.

What are the most in-demand skills for each of the job types?

After understanding the geographic distribution of these listings, it is essential to understand the skills needed for these positions and which ones a job seeker should learn to make themselves the most competitive candidate. Our goal was to partition each job type (Data Scientist, Data Engineer, Data Analyst, and Business Intelligence), and then determine what the most skills are the most in-demand for each of these of these positions. As a job seeker, learning the most sought-after technical skills and emphasizing this in your resume allows the applicant to improve their ability to stand out in the automated HR systems that are popular in the job market.

Figure 1.7: Top Skills - Overall and for Each Job Type

	Overall	Frequency
0	python	172
1	java	138
2	sql	130
3	machine-learning	100
4	c	88
5	big-data	77
6	hadoop	64
7	aws	59
8	linux	47
9	javascript	45
10	spark	44
11	scala	40
12	etl	36
13	intelligence	33
14	cisco	32
15	network	32
16	cloud	32
17	r	32
18	web	30
19	kafka	30

	Data Scientist	Frequency
0	learning	66
1	machine	58
2	python	51
3	r	22
4	sql	19
5	mining	13
6	hadoop	13
7	nlp	11
8	big	10
9	java	8
10	spark	8
11	ai	7
12	predictive	7
13	statistical	7
14	product	7
15	risk	6
16	neural	6
17	bayesian	6
18	code	6
19	network	6

	Data Engineer	Frequency
0	java	120
1	python	112
2	c	76
3	big	73
4	sql	62
5	aws	51
6	hadoop	50
7	learning	47
8	linux	47
9	machine	43
10	javascript	42
11	spark	35
12	scala	34
13	cisco	32
14	ruby	29
15	web	28
16	application	27
17	cloud	27
18	kafka	26
19	android	26

	Business Intelligence	Frequency
0	intelligence	27
1	sql	19
2	bi	15
3	oracle	12
4	etl	11
5	java	9
6	product	8
7	sales	8
8	enterprise	7
9	one	7
10	net	7
11	marketing	6
12	tableau	6
13	power	6
14	saas	6
15	process	5
16	c	5
17	microsoft	5
18	publisher	5
19	ssrs	4

	Data Analyst	Frequency
0	sql	30
1	excel	18
2	tableau	10
3	financial	10
4	analyst	8
5	python	8
6	automation	8
7	marketing	5
8	e-commerce	5
9	create	5
10	project	5
11	query	4
12	process	4
13	automate	4
14	team	4
15	write	4
16	test	4
17	end	4
18	qa	4
19	ms	4

The requirements for each job type (Data Scientist, Data Engineer, Data Analyst, and Business Intelligence) differ across the population. As expected, due to the technicality required in these roles,

we found that programming languages and business applications are pivotal in being able to qualify for these job types. In particular, the most surprising being Python, which has gained some traction in the data science world and is the most demanded skill for Data Scientists and the second to Java for Data Engineers. This finding can add some evidence to the Python versus R debate, and which one is the most commonly used by businesses.

Another skill that is worth noting is the need for candidates with SQL skills. As seen in Figure 1.7, SQL is demanded across all job types and learning this is necessary for making yourself the most competitive candidate. Knowing SQL should be expected though because each of these job types requires the collection, cleaning, and use of data which almost always requires the use of databases. Thus, knowledge of SQL is a must to be a self-reliant member of your team.

How much can one expect to earn overall based on the current listing and how does this differ across the states and cities?

Understanding the expected earnings for these job types within each state is an integral part of the job search because the cost of living differs across each state and even more so between cities. Therefore once a candidate understands their skillset and has chosen a particular job type, they want to know the difference in earnings depending on the location given their skill-set is suitable.

Figure 1.8: Earnings - States and Cities

	States	Min Earnings
0	DC	140000
1	CA	123017
2	IL	120833
3	OR	120000
4	NJ	120000
5	CT	115000
6	TX	114285
7	NY	112750
8	WA	111428
9	VA	110454

	States	Max Earnings
0	IL	197500
1	WA	175714
2	CT	175000
3	CA	171686
4	DC	166666
5	NY	160875
6	VA	157272
7	TX	152142
8	OR	150000
9	MA	145769

	Location	Min Earnings
0	Cupertino, CA	175000
1	Palo Alto, CA	160000
2	Santa Clara, CA	160000
3	Sunnyvale, CA	158333
4	Milpitas, CA	156666
5	Brooklyn, NY	150000
6	San Mateo, CA	142500
7	Long Beach, CA	140000
8	Ventura County, CA	140000
9	Washington, DC	140000

	Location	Max Earnings
0	Cupertino, CA	325000
1	West Hollywood, CA	250000
2	Chicago, IL	220000
3	Palo Alto, CA	206428
4	Seattle, WA	205000
5	Annapolis Junction, MD	200000
6	Brooklyn, NY	200000
7	Fairfield, CT	200000
8	Santa Clara, CA	200000
9	San Jose, CA	198636

Analyzing the expected earnings from two different perspectives returns two vastly different results. We found that Washington DC and Illinois have the highest expected minimum and maximum earnings. However, when we change perspectives and look at the highest expected minimum and maximum earnings by location only, we find that California is highest earning based on city locations with 8 out of 10 in the minimum perspective and 5 out of 10 in the maximum perspective. A difference between these two subsets is likely due to some outliers with very high earnings in the state's perspective skewing the earnings in one direction. Once again, the Bay Area is the best place to look for a high salary.

What are the mean salaries for the data scientist, business intelligence, the data analyst and the data engineer jobs? Are there large differences in salaries among these jobs? If so, which two jobs comparatively have the most difference in salary?

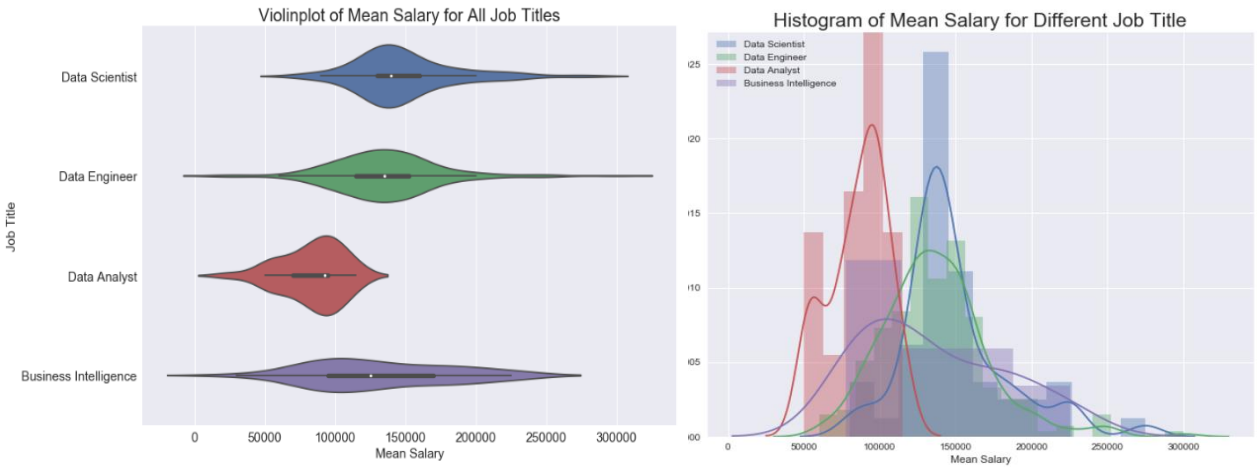
	Job Type	Min Earnings
0	Data Scientist	121800
1	Business Intelligence	115757
2	Data Engineer	114073
3	Data Analyst	73793

	Job Type	Max Earnings
0	Data Scientist	176800
1	Data Engineer	158706
2	Business Intelligence	153125
3	Data Analyst	96428

Figure 1.9: Earnings - Job Type

The Data Scientist position being the highest paid job type out of the group is not very surprising. Data Scientists have the most experience and in most cases is not an entry-level position. Whereas Data Analyst is the lowest earning, likely due to the least amount of experience required to obtain this position. An exciting finding is that the Business Intelligence and Data Engineer position switch places when going from minimum to maximum earnings. The switch between minimum and maximum earnings may indicate that Data Engineers have a higher earning a potential mid-end career.

Figure 1.10: Comparisons among Different Job Titles



The Data Engineer job type has the highest earnings variance and therefore the most fluctuation out of the group. In contrast, the Data Analyst position has the lowest earnings variance and therefore the least fluctuation out of the four positions. The Data Analyst position having the lowest variance in earnings is likely due to it being an entry-mid career position. Moreover, the Data Engineer position can vary from entry-end career resulting in broader low and high salary range. We found it surprising that the mean salary distributions for the Data Scientist, Data Engineer, and Business Intelligence are very close, but the minimum and maximum ranges are different which may support the experience level hypothesis.

Are the key words mined from the descriptions similar between jobs and do any duties overlap?

Though our focus is primarily to analyze the preferred skills, we decided to have a try at LDA (Latent Dirichlet Allocation), a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. We used the Cybercoder dataset and found that although the title of each cluster of job postings is roughly the same, the content (requirements) for candidates is different. Say if someone wants to find data science job, they may type in “data science” when searching on a website. Although every one of the job listings named mapped to “Data Scientist” will be returned, we are interested in determining how many of these meet the expectations related to the job seekers skills the most (i.e., which job postings’ requirements match his skill sets). Therefore, LDA is appropriate and used to partition job postings into 10 categories. For example, category 9 emphasizes the ability to handle projects and their ability to work out solutions. Category 10 requires candidates to know machine learning, python and to have a certain level of programming skills.

We could go more in-depth on LDA, but we chose not to because this is not our goal. However, through exploring LDA, we did get the hang of the idea that it is the content that separates each posting apart, meaning the listings are not homogeneous.

Part IV: Simplyhired Analysis

Step one: First exploration

We found that information about location exists in every posting and is a reliable source of information, so we decided to focus on the relationship between geographic features and job skills. We choose to use Simplyhired dataset because it is significant (as it has 17,229 postings for data scientists, 76,725 for data engineers, and 120,327 for data analysts) and the information of each posting is complete(it has attributes: job title, company, location, salary, snippets and description and seldom missing information).

We are curious about what is the distribution of job postings, i.e., which state offers most data-related jobs. We chose to plot a choropleth map because it is intuitive as well as interactive, and therefore we can include more information such as salary, job skills in the pop-up when we click on the location. Another advantage of plotting on a map is that we can quickly discover what we can do next, for example, if we can see jobs are gathering in the east coast and west coast, then we can do further analysis by grouping these two places and dive deeper. However, we only had the name of each job location and needed to figure out a way to transform to longitude and latitude to plot on a map. After doing research, we found Google API and also find a simple but effective way to plot on google map.(Google map plot:https://www.google.com/maps/d/edit?hl=en&hl=en&mid=1Pij2IqMyAutaW4R79vqlB1AP8SzMD6qZ&ll=41.67111456382234%2C-110.37450030000002&z=5) So we first grouped posts by same working locations and counted the number of postings for each location. Also, we used Google geocoding API and geocoder package to transfer the location name to longitude and latitude. We created and uploaded three files to google map, we set blue points for data scientist job postings, yellow points for data analyst job postings and red points for data engineering job postings. If you click on each point, you can see the job salary for each posting grouped by the same location, skills of those postings, counting number of each skill and the number of postings of this location.

We can see that data related jobs cluster within the west and east coast of USA compared to relatively scarce points scatter in the central part of the State. The result matches our expectation since the population in USA is distributed roughly in this way. And this can also be explained by thinking about data-job features. Data-related jobs gained popularity in recent years, it is a relatively new job type and relates to key words like high-tech, massive data, fast developing speed and so on. It is not hard to see the correlation between job distribution and economic and technology development level of that place. Due to gathering effect, newly emerged enterprises like to set their locations in places having mature infrastructure, developed economy and abundant talents. West coast, lake area, east part of USA have been well developed for a long time and meet criteria, thus can explain the distribution of the job postings.

Initially, we planned to make a 3D interactive map of the USA providing an in-depth view of all of the job listings. We first used leaflet as our base package to make a basic plot of the US that was able to examine the country as a whole and zoom deeper into states and cities to provide accurate locations if

need be. We combined this with the package shiny to make sliders in which users could adjust the maps to show results based on their desired salary ranges or location preferences. However, we decided to shift our focus away from an interactive plot. Rather than giving the user the power to do their analysis, we wanted to turn to a more direct approach in which we guided our users through our narrative of the job market, starting from a large scale approach based on the USA, down to differences among cities and states.

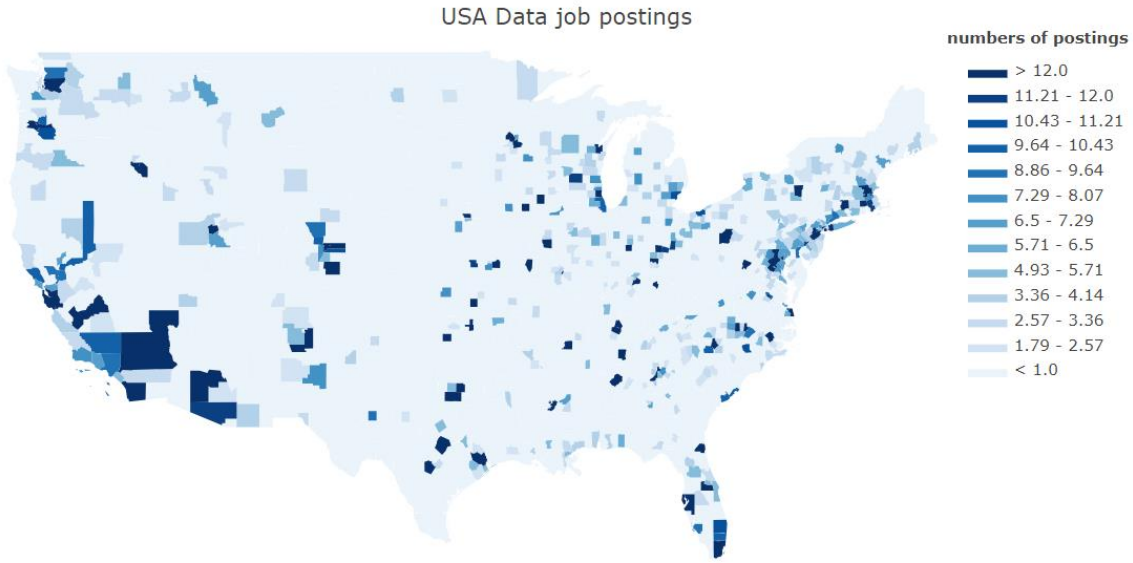
Step two: exploring intuitive way to represent results

Since the last map is not intuitive in showing the extent of the distribution of job postings (because it is grouped by location and no matter how many jobs there are, each location is represented by a single point, making it hard to see how many job postings are in this single location). So we also needed to make a choropleth map of the job postings. Since we wanted to show the number of jobs in each state, we needed some way to match cities in the job posts to their corresponding state. After searching and comparing different libraries of plotting, we found the Plotly module which can plot a choropleth graph beautifully. Unfortunately, it needs FIPS county codes (FIPS county codes provide information about each county, and the plot includes a more informative view than the state level). We were reluctant to give up this approach, as we have tried many, and this was the closest we came to success. Finally, we found a website that offers API for matching and transforming longitude and latitude to FIPS county codes. So we used the longitude and latitude acquired in step one to get the FIPS code and plot the following graph.

The darker color indicates areas with a higher number of job postings and therefore high demand for data science related talent. California, Washington, Nevada, Texas, Florida, Pennsylvania, Maryland, New Jersey, Massachusetts, and Connecticut all have an ample supply of businesses hiring for these types of jobs (Data Scientist, Data Engineer, and Data Analyst).

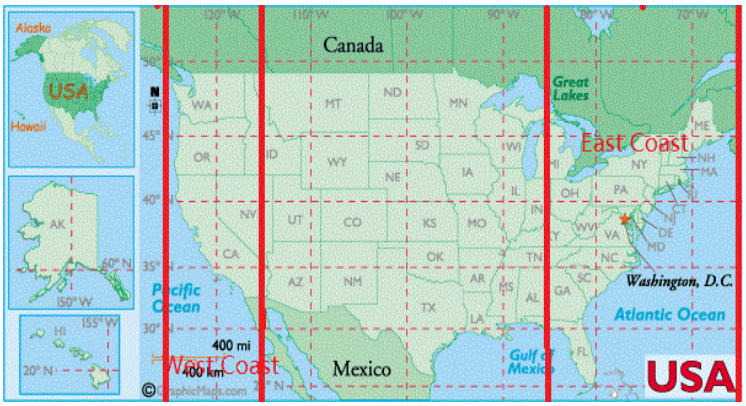
For any prospective worker, location is one of the most critical factors in deciding whether or not one might take a job offer. With a more local perspective, an in-state job seeker (i.e., someone living in Los Angeles, CA) may know that they need to relocate in advance due to many technology jobs located in the Bay Area. Location is essential in the job search, that is why conducting a detailed analysis of the geographic distribution of our population necessary.

Figure 2.0: USA - Map of Jobs Listings



Step three: Deeper dive into east and west coast comparison

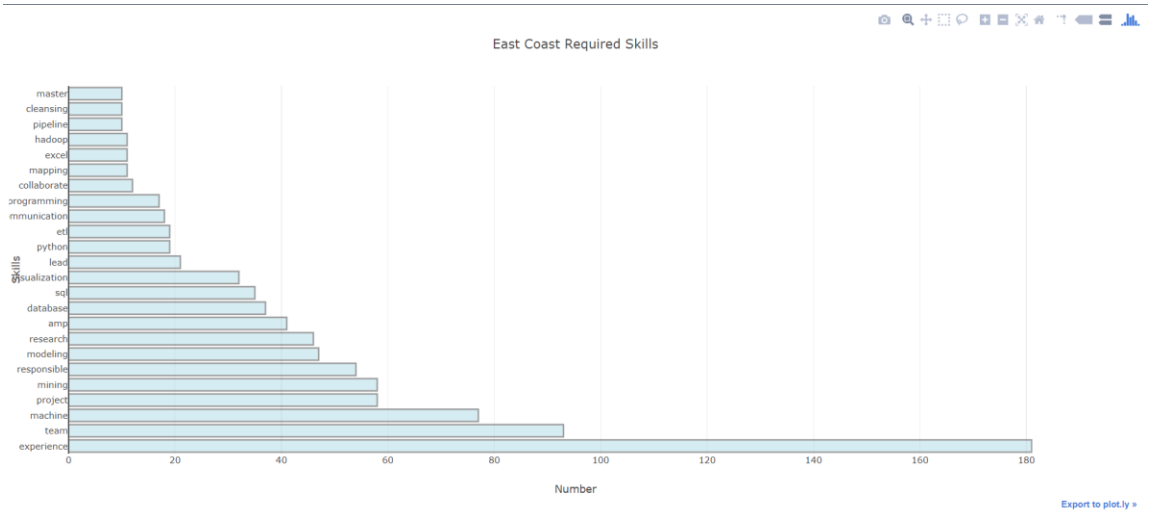
Figure 2.1: USA - Stratification of Job Listings

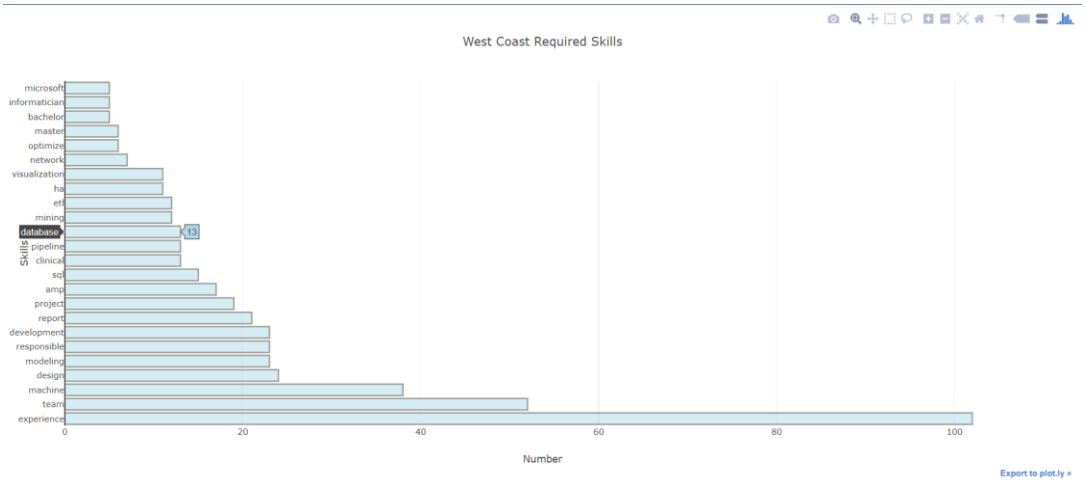


From the two previous graph, we can see that east coast and west coast are favorite places for seeking data related jobs. So we want to dig further and explore if there is any difference in job requirements for talents between east and west coast.

We set longitude between -65°W to -85°W for east coast and longitude between -115°W and -125°W for the west coast (see above picture) and put job skills words into these two groups to compare and analyze. We chose the most popular skills by counting the number of appearance of job postings, and rank them, thus we got the following plot.

Figure 2.2: East Vs. West Coast - Preferred Skills





Both the east and west coast expect candidates to have experience and an aptitude for teamwork and collaboration. The ability to finish projects or experience doing projects increases the competitiveness of the candidate. Both SQL and database appeared in east and west and are ranked very high. So it is not hard to tell its importance. Both sides think candidates should know HA (High Availability), ETL (Extract, Transform, Load), pipeline, mapping, amp. Other skills include modeling, data mining, machine learning, data cleaning, and visualization.

What appeared most on the east coast is python, it seems that this region prefers candidates to have programming skills. Other skills that only appears on the east coast is Hadoop and Microsoft Excel. Skills that appear only in the west is Microsoft, but we can't tell it is a skill or the company (since its headquarters are on the west coast). The Bioinformatician only appears on the west coast. Having a high level of "responsibility" is the most emphasized personality of candidates from both east and west. Communication skills and leadership is also a plus for candidates.

However, overall, the difference between east and west coast is not apparent, they differ slightly in some specific job skills, but more often they share the same job skill requirements. It actually can be justified since globalization has led to convergence, technology shortened distances. Differences caused by geographic factor alone is not that obvious anymore. So it is normal to have same expectation for a job position. And another reason we can think of is that requirements listed in job postings are broad in themselves. For example, almost every company wants SQL talents but surely they will emphasize on different application on SQL, but they won't post in such a specific way, and we can't capture information like that, making our result of the difference between east and west coast seems similar.

Step four: possible exploration on salary of Simplyhired dataset

We still want to make use of salary, and we are curious about how the salary changes across the country from west to east and from north to south. So we create a dictionary, whose key is longitude/latitude and its value is salary, and we sort the dictionary by its key in ascending order (from east to west/from south to east), we plotted it out but didn't see any useful/visible pattern. The reason may because the salary is estimated/generated arbitrarily by the website which results invalid, so we can't find clear patterns. The other reason is that there are no patterns in itself. Moreover, a possible reason for this is that we should set a specific longitude when plotting how the relationship between latitude changes given salary. However, the reason we didn't do this is that we don't know which specific number to set and if we did we don't have enough data that has latitude information for a particular longitude.

Part V: Conclusion

After the first step of writing our project proposal, we set out to do data collection and data cleaning. We spent much time finding suitable websites. Later we discovered that it is difficult to do our initial goal, which is salary predictions since the information about it is hard to collect and its source is not reliable. At last, we found two websites provide data that meet our expectations.

Although every dataset is not perfect, we take advantage of each data source and compliment the drawbacks of one with the advantage of the other one. We used Cybercoder dataset to conduct an overall analysis. First, we looked at the composition of the four types of job postings. Then we looked at job distribution and job proportion of each state, and then we found states and cities that are popular for data-related jobs. Next, we turned to skill analysis where we wanted to know what is the most in-demand skills in each job title and the overall data jobs. In the salary section, we analyzed it in two parts, first is to compare the earnings in different states and cities, second is to compare salary between different job types. For the Simplyhired dataset, we took full advantage of its enormous quantity to do geo analysis. After discovering the distribution of job postings, we grouped job locations into two groups then compared skills between the two coast.

Overall, we found our analysis to be challenging but also rewarding. As college students, we are entering the job market soon, and after doing our analysis, we hope that our readers find our work informative.

Each job type in this field is similar, but they each hold their responsibilities, skills demanded, and salary expectations for a particular geographic location. As a result of our findings relating to the preferred skills section of our report, a job seeker can learn these skills to better prepare as a competitive candidate. Learning the top skills demanded by these positions is only one example of how our results are useful to any reader that is seeking a job.

Unfortunately, there are no easy answers to many of our questions, and we wish that we had more time because we would have done a more detailed analysis of the descriptions and salary information. All in all, our team had much fun working on this project together, and we hope that you enjoy reading and learning more about the job market for data science positions!

Part VI: Skills Learned

a. *Web scraping:* Used XPath and Beautiful soup library to scrape the data.
b. *XML and HTML:* Learned to deal with HTML format when we scrape the data, deal with a response through API from a website and also when we make our website.
c. *Plotting:* Mainly used the Plotly, Matplotlib, Plotnine - ggplot, and Google maps to show our results. Our graph types mainly include bar plots, charts, pie plot, choropleth map plot.
d. *API:* We used an API twice in this project, and the first time is to request for longitude and latitude for each job posting location name from google map API. The second time is to use the longitude and latitude to get FIPS county code for each job posting location. Both are for plotting purpose.
e. *Natural Language Processing (NLP):* Learned skills related to the cleaning of text data, such as regular expression, wordnet, and lemmatizer.
f. *LDA (Latent Dirichlet Allocation)*
g. *Website development:* When we built the website, we learned to use HTML, CSS, some javascript to add some animation, and how to embed graphs in the website.