Final Report

Betzalel Moskowiz, Pete Schultz, Matt Griffith

Introduction

For our project we decided to web scrape data off of dice.com to find information about job postings for information science majors over the past few years. The information we used came from two sources, a csv file from 2017 on dice.com as well as regular job postings from 2019 off of dice.com. The first source we will be looking at is DataWorlds dice_com-job_us_sample.csv file. This file consists of 22,000 rows and 12 columns. The file was created as a subset based off of 4.6 million job postings across Dice.com. The data included within it is all data related to job postings for information science positions in 2017. Our second source of data is Dice.com. We webscraped the rest of dice.com to find relevant data to compare to that which we have on the csv file mentioned above. The reason we wanted to work on this project is because we are all information science majors preparing to enter the workforce soon. Having access to this kind of information could be very beneficial in our job search as well as being very beneficial for anyone else searching for a job in this department.

The main goal of our project was to determine what the most valuable skills that information science students should acquire before entering the job market for data and analyst related positions, as well as analyse any trends in the workplace over the past several years. After conducting our analysis, we discovered many interesting trends. First, we conducted an exploratory data analysis with descriptive statistics for all of our data. This revealed statistics for several different areas such as the percentage of job type that falls into a designated category. Secondly, we conducted topic modeling analysis using LDA for all of our data to reveal the topics whose themes most closely relate to skills in the full_job_description category. Finally, we ran an analysis on the data/analyst job market. This included four different categories and revealed common trends in the job market.

This report will contain 4 main sections. The first section we will discuss the data we used as well as how we obtained it. Next, we will explain the methods we used both for gathering the data as well as how we ran our analyses. The third section will feature descriptions about what types of analyses we conducted and our methodology. Finally, we will include a results section pointing out all of our findings.

Data

For our project we worked with two main sources of data. The first source of data came in the form of a csv file from dice.com (dice_com-job_us_sample.csv). This file consists of 22,000 rows and 12 columns. The file was created as a subset based off of 4.6 million job postings across Dice.com. The data included within it is related to job postings for information science positions in 2017. The second source of data was dice.com directly. Our team scraped any relevant data related to "data" and "analyst". After scraping and cleaning we had a new csv file that served as our 2021 data. Our final product is a dataset that contains 13 columns: company, employmenttype_jobstatus, state, job_title, shift, skills, remote, full_time, contract, other, and year, title, and full_job_description. There are 7,895 entries in the dataset.

Unnamed: 0	company	employmenttype_jobstatus	state	job_title	job_description	shift	skills	remote	full_time	contract	other	year	title	full_job_description
11	CSI (Consultant Specialists Inc.)	Full Time, Contract Corp-To-Corp, Contract Ind...	CO	9001 Data Security Administrator - Unix & IAM	Must have skills: UNIX, IAM (Identity Access M...	Telecommuting not available|Travel not required	Unix, IAM, Scripting knowledge, OIM, Windows, ...	0.0	1.0	1.0	0.0	2017	NaN	Must have skills: UNIX, IAM (Identity Access M...
17	VanderHouwen & Associates, Inc.	Full Time, Full time	OR	Principal Application Analyst-Supply Chain Lawson	VanderHouwen has more jobs you may like! Find...	Telecommuting not available|Travel not required	Lawson Supply Chain	0.0	1.0	0.0	0.0	2017	NaN	VanderHouwen has more jobs you may like! Find...
20	Genesis10	Full Time	GA	IT Business Analyst	Genesis10 is looking for a Business Analyst fo...	Telecommuting not available|Travel not required	Analysis, Analyst, Application, Business Analy...	0.0	1.0	0.0	0.0	2017	NaN	Genesis10 is looking for a Business Analyst fo...
34	VanderHouwen & Associates, Inc.	Contract W2, Contract	OR	Business Systems Analyst	VanderHouwen has more jobs you may like! Find ...	Telecommuting not available|Travel not required	AI, Maya, Browzwear, 3D	0.0	0.0	1.0	0.0	2017	NaN	VanderHouwen has more jobs you may like! Find ...
42	Amazon	Full Time, Fulltime	TX	Software Development Engineer, Big Data	Do you want to help build a highly personalize...	Telecommuting not available|Travel not required	DESCRIPTION Do you want to help build a highly...	0.0	1.0	0.0	0.0	2017	NaN	Do you want to help build a highly personalize...

Methods

The first step in our project was to acquire the data that we would be working with. As stated before the data came from web scraping dice.com. To accomplish this our group wrote a Python script that scraped links off search results on dice.com for terms ‘data’ and ‘analyst’. Because the content we were trying to scrape was loaded onto the site using AngularJS, BeautifulSoup did not work. Our solution was to use the requests_html package to instantiate an HTML session that rendered the page and accessed the raw html, which could then be parsed using BeautifulSoup. After that we wrote a function to scrape the number of results to determine the number of pages necessary to loop through the entire search results. Then, for each search term we looped through all of the pages and scraped each link related to the job pages. These links were added to a master link of links that were dumped into a .pkl file for serialization. Next, we loaded in the list of links to scrape from links.pkl. All of the urls were then scraped for the data in the categories of company, job-title, state, remote, skills, description, full-time, contract, and other. These are the categories that we deemed most important for our study. We then created a dataframe from the scraped data to serve as our 2021 data that we will be analyzing.

In order to analyze all of the data as a whole, we had to combine the two Datasets together from 2017 and 2021. All of the columns are the same in these two DataFrames except for the jobdescription section. We had to rename these columns to have the same name before we could concatenate these dataframes together. We also had to combine the job description and skills column into one column called 'full_job_description' in order to ensure that we would have skills for as many observations as possible. This is necessary since companies are inconsistent and sometimes put the skills in the job description and other times put the skills in the skills field.

Analysis

Our analysis was broken down into three main parts. The first type of analysis was an Exploratory Data Analysis with Descriptive Statistics. The descriptive statistics were run to develop a better understanding of our data. The descriptive statistics were run on the entire dataset in addition to the subsets of the dataset by year. The image below is of our descriptive statistics generated for the entire dataset. There are 7,895 observations in our full dataset. Because the only values for remote, full_time, and contract, and other are 0 and 1, the mean of each of these variables can tell us what percent of the jobs fall into these categories. This analysis was also run on just the year 2017 data as well as another for just 2021 data.

	remote	full_time	contract	other	year
count	7895.000000	7895.000000	7895.000000	7895.000000	7895.000000
mean	0.223559	0.571501	0.436859	0.064978	2019.333629
std	0.416656	0.494893	0.496029	0.246502	1.972102
min	0.000000	0.000000	0.000000	0.000000	2017.000000
25%	0.000000	0.000000	0.000000	0.000000	2017.000000
50%	0.000000	1.000000	0.000000	0.000000	2021.000000
75%	0.000000	1.000000	1.000000	0.000000	2021.000000
max	1.000000	1.000000	1.000000	1.000000	2021.000000

(Figure 1: Descriptive Statistics for All Data)

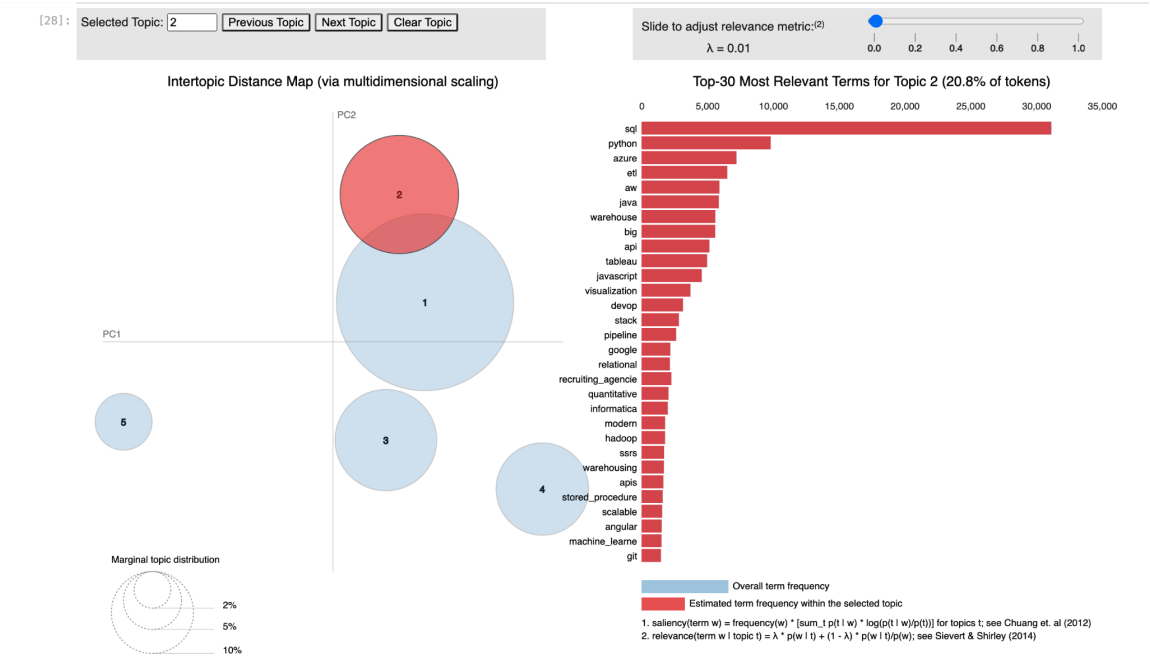
	remote	full_time	contract	other	year
count	3289.000000	3289.000000	3289.000000	3289.000000	3289.0
mean	0.008817	0.537549	0.504409	0.065369	2017.0
std	0.093500	0.498664	0.500057	0.247214	0.0
min	0.000000	0.000000	0.000000	0.000000	2017.0
25%	0.000000	0.000000	0.000000	0.000000	2017.0
50%	0.000000	1.000000	1.000000	0.000000	2017.0
75%	0.000000	1.000000	1.000000	0.000000	2017.0
max	1.000000	1.000000	1.000000	1.000000	2017.0

(Figure 2: Descriptive Statistics for 2017 Data)

	remote	full_time	contract	other	year
count	4606.000000	4606.000000	4606.000000	4606.000000	4606.0
mean	0.376900	0.595745	0.388624	0.064698	2021.0
std	0.484662	0.490801	0.487490	0.246019	0.0
min	0.000000	0.000000	0.000000	0.000000	2021.0
25%	0.000000	0.000000	0.000000	0.000000	2021.0
50%	0.000000	1.000000	0.000000	0.000000	2021.0
75%	1.000000	1.000000	1.000000	0.000000	2021.0
max	1.000000	1.000000	1.000000	1.000000	2021.0

(Figure 3: Descriptive Statistics for 2021 Data)

The second form of analysis we did was topic modeling analysis of the category full_job_descrption using LDA. We used Natural Language Processing to extract the skills from the full_job_description column. The idea was to create LDA models and find the topic whose theme most closely relates to skills. This allows us to visualize these models to get an idea of which skills are most frequent in the skills topic, revealing which skills are the most important and widely used for that particular dataset. As we did earlier, we did this three different times, once to look at the combined data, again to look at just the 2017 data, and finally just to look at the 2021 data. The idea was to see the most relevant skills in each time period and make comparisons between years based. The models we used contained 5 topics. These topics can be selected to show the common themes within. A slider is included that adjusts the amount of times something appears vs its relevance to the topic. The images below show the models when used.

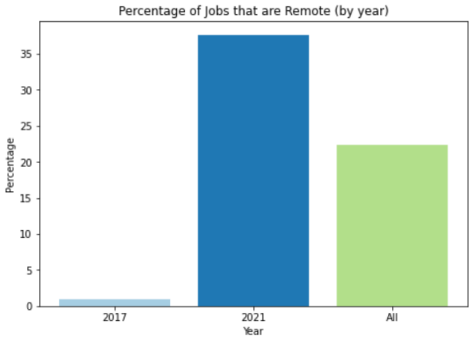


(Figure 4: Topic Modeling Analysis for Combined Data)

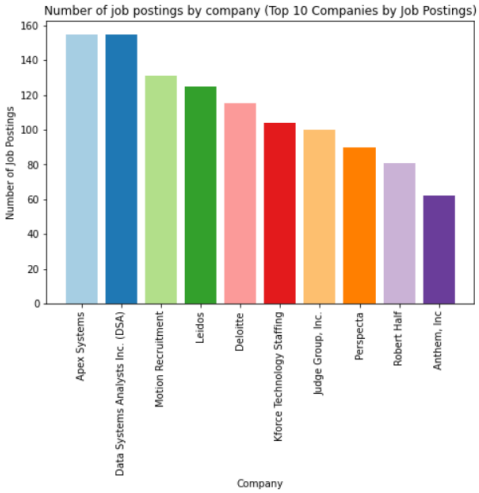
The final analysis we conducted on our data was our analysis of the Data/Analyst Job Market. We looked at four main categories, “Remote Jobs”, “Companies”, “Employment Type”, and “Job Locations". First we looked at remote jobs so we could see the percentage of jobs that were listed as remote and how this percentage changed over time. We displayed this in both a table as well as a bar graph. To analyze the companies who are posting these job descriptions, we first determined the top 10 companies by number of job postings on dice.com. After that we were able to conduct our analysis and create a bar graph visualization depicting which companies had the most listings. For employment type, there are three types of employment in our dataset - full time, contract, and other. We wanted to analyze how employment type has changed overtime among job postings. To show our results we first created a table for the three years followed by a bar graph depicting by color which category had the most opportunities. Finally, we conducted an analysis of job locations. To analyze job locations, we had to count the number of job listings by state. However, some of the states are written out, so we had to make sure that they are all in their code form. The last step was to create a data visualization to help better visualize the information. This includes a bar graph depicting the number of job postings per state as well as a table depicting the same.

	Percent_Remote
2017	0.881727
2021	37.689970
All	22.355921

(Figure 5: Analysis of Remote Jobs Table)



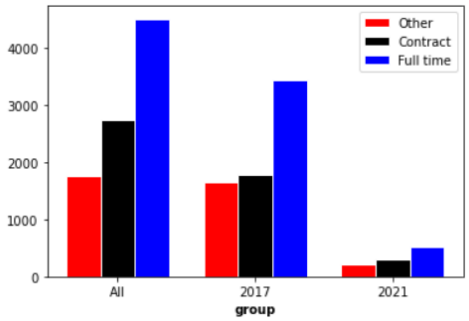
(Figure 6: Analysis of Remote Jobs Graph)



(Figure 7: Analysis of Companies Graph)

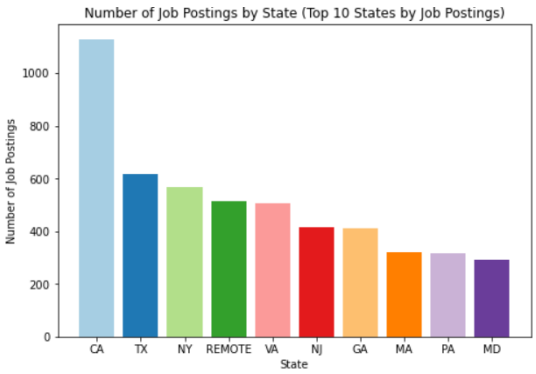
	Full_Time	Contract	Other
2017	1768.0	1659.0	215.0
2021	2744.0	1790.0	298.0
All	4512.0	3449.0	513.0

(Figure 8: Analysis of Employment Type Table)



(Figure 9: Analysis of Employment Type Graph)

	Count
CA	1130
TX	617
NY	566
REMOTE	514
VA	507
NJ	414
GA	412
MA	319
PA	318
MD	290



(Figure 8: Analysis of Job Location Table) (Figure 9: Analysis of Job Location Graph)

Results

We will break the results down into three parts. The first set of results we will be looking at is what we learned from our Exploratory Data Analysis with descriptive statistics. As stated earlier, we ran the descriptive statistics to better understand our data overall. The first dataset used was the combined 2017 and 2021 data. Here were our results:

For all jobs (2017 and 2021) it appears as that 22% are remote, 57% are full time, 44% are contracting jobs, and about 25% are other (part time). Note that it is possible for observations to have multiple 1s in these categories (they are not mutually exclusive). We ran the same tests on the 2017 data and it showed that for jobs in 2017, 0.8% are remote, 54% are full time, 50% are contracting jobs, and about 6% are other (part time). For jobs in 2021 it seems as if about 38% are remote, 60% are full time, 39% are contracting jobs, and about 6% are other (part time). Our analysis shows that in the past few years there has been a major increase in remote jobs especially, most likely due to the COVID-19 pandemic.

The goal of our topic modeling analysis was to look at all of the data and develop an understanding of what the most relevant skills for this time period were. When looking at the

different topics from our LDA model visualization, we found that topic 2 was the most related to skills for the full dataset (2017 and 2021). By setting the relevance λ = 0.01 (to sort almost exclusively by term frequency within the selected topic), we can see that the following skills appear to have the greatest estimated term frequency within the the selected topic, meaning that the following skills appear to be the most in demand for jobs in both 2017 and 2021:

1. SQL
2. Python
3. Azure
4. ETL (likely extract, transform, and load tools)
5. AW (likely AWS)
6. Java
7. Warehouse (likely Data Warehousing)
8. Big (likely Big Data)
9. API
10. Tableau
11. Javascript
12. Visualization (likely data visualization)
13. Devop (likely DevOps)
14. Stack (likely full-stack development)
15. Pipeline (likely data pipeline)
16. Google (likely Google Cloud)
17. Relational (likely Relational Database)
18. Quantitative (likely quantitative skills)
19. Informatica
20. Hadoop
21. SSRS (likely SQL Server Reporting Services)
22. Stored_Procedure (likely in the context of databases)
23. Angular (likely Angular JS)
24. Machine Learning
25. Git

When we ran this analysis for our 2017 data. It appeared that topic 3 was the most related to job skills. Setting the relevance λ = 0.01 (to sort almost exclusively by term frequency within the selected topic), we can see that the following skills appear to have the greatest estimated term frequency within the the selected topic, meaning that the following skills appear to be the most in demand for jobs in 2017:

1. Big (likely Big Data)
2. ETL (likely extract, transform, and load tools)
3. Hadoop
4. Tableau
5. Spark (likely Apache Spark)
6. Informatica
7. Tera Data
8. NoSQL
9. Cassandra (likely Apache Cassandra)
10. SA (likely SAS)
11. Warehousing (likely data warehousing)
12. Scala
13. Algorithm
13. Algorithm
14. Hive
15. Pipeline (likely data pipeline)
16. Visualization (likely data visualization)
17. MySQL
18. HBase (likely Apache HBase)
19. Machine Learne (likely Machine Learning)
20. Datastage (likely IBM InfoSphere DataStage)
21. Python
22. SSIS (likely SQL Server Integration Services)
23. OBIEE (likely Oracle Business Intelligence Suite Enterprise Edition)
24. Predictive (likely predictive modeling)
25. AW (likely AWS)

When we ran this analysis for our 2021 data It appeared that topic 4 was the most related to job skills. Setting the relevance λ = 0.01 (to sort almost exclusively by term frequency within the selected topic), we can see that the following skills appear to have the greatest estimated term

frequency within the the selected topic, meaning that the following skills appear to be the most in demand for jobs in 2021:

1. API
2. SOA
3. Azure
4. Java
5. Python
6. AW (likely AWS)
7. Devop (likely DevOps)
8. Native
9. Angular (likely Angular JS)
10. Javascript
11. Big (likley big data)
12. Microservices
13. Apex
14. JSON
15. Stack (likely full-stack development)
16. Relational (likely Relational Database)
17. AEM (likely Adobe Experience Manager)
18. Pattern (likely pattern detection or pattern database)
19. HTML
20. Apigee (likely Apigee API)
21. Stored_Procedure (likely in the context of databases)
22. Informatica
23. App (likely application development)
24. Kubernete

Overall, we concluded that the most relevant skills over the past 4 years are:

1. **SQL**
2. **Python**
3. **Azure**
4. **ETL (likely extract, transform, and load tools)**
5. **AW (likely AWS)**
6. **Java**
7. **Warehouse (likely Data Warehousing)**
8. **Big (likely Big Data)**
9. **API**
10. **Tableau**

Our comparison between the 2017 and 2021 data highlighted that python and cloud computing knowledge and tools have become much more desirable, knowing how to work with big data has always been desirable, knowing how to work with databases (SQL) continues to be desirable, and front end tools such as HTML and JavaScript in addition to data visualization tools have also become much more popular.

From our analysis on the Data/Analyst market we noticed a number of trends. First off, when we compared the number of remote jobs in 2017 to 2021, we found that the number of remote jobs was way up in 2021, increasing by almost 37% likely due to the COVID-19 pandemic. When we did an analysis of companies we were looking to see who posts the most job listings, our top 10 results came out to be:

```
Apex Systems                          155
Data Systems Analysts Inc. (DSA)      155
Motion Recruitment                    131
Leidos                                125
Deloitte                              115
Kforce Technology Staffing            104
Judge Group, Inc.                     100
Perspecta                              90
Robert Half                            81
Anthem, Inc                            62
Name: company, dtype: int64
```

The next Results we found were for our analysis on employment type. We found that in general there are more full-time positions than contracts and other positions. This appears to be true across all job listings, especially in 2017 but a little less so in 2021. Finally, the last conclusion we made was from our analysis on which state posts the most jobs. Overall, the top 10 states with the greatest number of job postings are CA, TX, NY, REMOTE, VA, NJ, GA, MA, PA, and MD. California had easily the most with well over 1,000 postings.

Conclusion

Our group created this project with the main goal of finding out what the most valuable skills that information science students should acquire before entering the job market for data and analyst related positions are. In addition, we analyzed trends in the workplace over the past several years. Overall, our team produced the results we set out to achieve. We concluded that the most relevant skills over the past four years in the data/analyst field include SQl, Python, Azure as well as many other similar ones. Our study also revealed many other useful to know trends such that cloud computing tools are drastically becoming more desirable. The data we produced could be extremely beneficial to those looking for a job. It would allow them to see what skills they may need to learn or tell them if they are already qualified for a job.