Spring 2021

An In-Depth Analysis of the Data Analytics Job Market

Dagny Elaine Wilkins
University of New Hampshire, Durham

University of New Hampshire
Peter T. Paul College of Business and Economics

An In-Depth Analysis of the Data Analytics Job Market

Dagny Wilkins

Advisor: Kholekile L. Gwebu

University of New Hampshire

Peter T. Paul College of Business & Economics

May 2021

Table of Contents

Introduction

In recent years, the importance of data analytics in organizations has grown substantially. Coupled with this growth is the need for professionals with skills that allow organizations to effectively collect, process, visualize, and analyze data. Having employees with data analytics skills enables organizations to make optimal decisions effectively and efficiently, and in the case of businesses, it could ultimately lead to a competitive advantage. This has resulted in a large demand for employees who have data analytics skills. Given this high demand, it's critical that colleges and universities help students develop skillsets that best fit the needs of the job market.

This thesis seeks to understand which analytic skillsets are most sought after by organizations. The findings from this study will provide valuable insights to students, organizations, and universities alike. For students, the findings will clearly point out which skill subcategories they should focus on mastering while in college thus allowing them to become more competitive on the job market. For organizations that are hiring for analytics roles, the findings will provide them with an understanding of the mix of job skills that are most in demand as they compete for hires. Finally, universities with data analyst-related degrees or classes will be better able to align their curriculum and course offerings to the necessary skills most sought after in the job market.

The remainder of this thesis is organized as follows. The next section reviews the research pertaining to the knowledge and skills that data analysts need to possess. Thereafter, the methodology adopted to collect and analyze the data used in this study is described in detail. Next, the findings for the study are presented. Finally, the findings are discussed and potential directions for future research are presented.

Literature Review

In recent years, many researchers have sought to determine the knowledge and skills that data analysts need to possess. Consequently, this has garnered a lot of attention and been explored by researchers in the US and around the world. For instance, in a study titled "Emerging trends in data analytics and knowledge management job market: extending KSA framework", Chang, Wang, and Hawamdeh (2019) analyzed 390 job advertisements from January 1 to May 28, 2017, in the analytics and knowledge management domains from LinkedIn using both quantitative and qualitative methods. While the study does provide useful insights, it does not primarily focus on analytics jobs, rather it considers both analytics and knowledge management positions. Moreover, the sample used is rather small, and the authors conclude by stating that "Although the results of this study are not conclusive based on the five-month data set collected from LinkedIn and may require longitudinal data to verify the job trends, the efforts of finding the association between analytics and KM are certainly essential to transform data into useful knowledge and insights" (Chang, Wang, & Hawamdeh, 2019).

Other researchers have also investigated this issue; however, they have followed a slightly different approach. For instance, Pejic-Bach et al. (2020) used a text mining approach that combines topic modeling, clustering, and expert assessment to explore 25,104 job postings that appeared on Indeed, Monster, and Glassdoor. The goal of the study was to identify the types of job roles, knowledge, and skills that make up the field of data science. The study provides some interesting findings. For example, the team used text mining to identify the most frequent phrases in the job advertisements which allowed them to easily visualize the results. Additionally, they used clustering and classification which enabled them to see which traits and skills are required for specific positions and which positions are the most closely related. The

tool that they created allows them to "track the changes in relevant knowledge and skills required" in the different positions which is very useful (Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž., 2020).

Similarly, Radovilsky, et al. (2018) collected data on 1,050 business data analytics and data science job postings from LinkedIn, Monster, Dice, Indeed, Glassdoor, and CareerBuilder and used text mining methods to develop a model that identifies the knowledge domains and skill sets for jobs in the analytics field. Nevertheless, as Pejic-Bach et al. (2020) acknowledged with text mining approaches "any bias in the algorithms' parameter settings could distort the results of the topic modeling and clustering approach." Consequently, these findings would need to be validated.

Perhaps the closest of the studies to the current one is titled "An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements" by Verma et al. (2019). This study used content analysis to analyze 1,235 job postings across four states (Arkansas, Florida, Kansas, Missouri) collected from Indeed between December 2016 and February 2017. They seek to determine if there are differences between the top five skills between Business Analyst (BA), Business Intelligence Analyst (BIA), Data Analyst (DA) and Data Scientist (DS) jobs. They find that "decision-making, organization, communication and structured data management are key to all job categories". Additionally, they point out that "technical skills like statistics and programming skills are in most demand for Data Scientist." While these are interesting contributions, the study only focuses on four states, and the authors call for future research to include more states. Moreover, the job market in the analytics domain is rapidly evolving and the job market/requirements from the 2016-2017 time period may have changed.

Research Questions

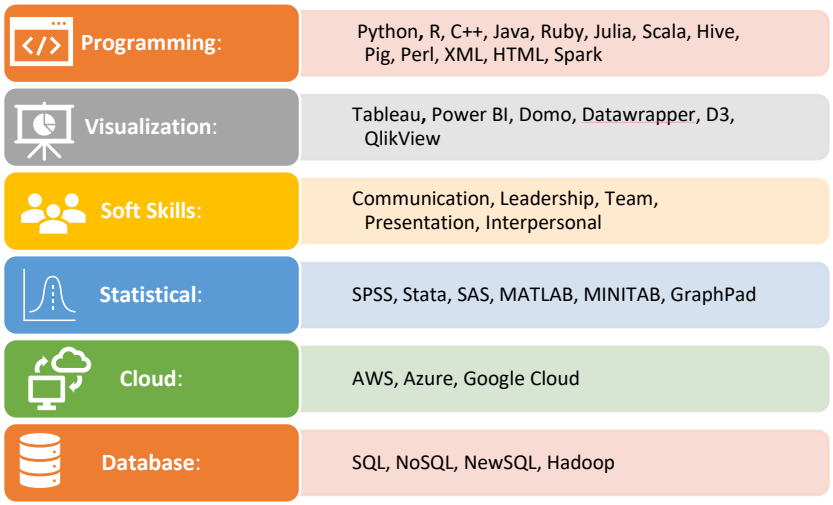
While extant literature has provided important insights into the type of knowledge and skills that the job market demands from data analytics professionals, as discussed in the preceding section, many studies have methodical limitations such as small sample sizes or the use of algorithms that have the potential to be biased if mis specified. Consequently, this study seeks to address some of these limitations and answer the following important questions:

- What are the most sought-after analytics skills?
- Which skill sets have the most influence on salary?
- Is Python more popular than R in the job dataset that was analyzed?
- What specific languages should individuals learn from each broad skill category?
- Does the pay vary by job type?
- What is the most popular education level by job type?
- Do the required skills vary by job type?
- Student perspective: which is better to learn?
- Which sectors currently have the most jobs available?

Methodology

The initial plan for data collection was to use ParseHub to scrape job data from Indeed.com. Nevertheless, this approached proved to be cumbersome and prone to error. Ultimately, four different datasets dealing with data analytics jobs were identified in the data science repository Kaggle. These datasets were initially collected by a user named Kenarapfaik from posting Github (https://github.com/picklesueat/data_jobs_data) then subsequently organized and archived on Kaggle. Each dataset included positions for the different job categories including data analyst, business analyst, data scientist, and data engineer. The datasets

were compiled into one master spreadsheet in Excel. After removing 34 duplicates, 12,748 unique job postings remained. The data was then cleaned, and an index was added so as to uniquely identify each posting. The job type for each position was added. Within the job postings, the variables of interest were the job title, job description, job requirements, job skills, experience, and salary, among others. Columns were created for all the important fields that needed to be collected. A formula was used to search through the job description and return a "1" if the value was found. In some instances, it was necessary to search for various versions of a specific skill such as ETL: extract, transform, load; Extract, Transform, Load; etc. Some alternatives may have been missed since job description wording varies so much. Additionally, broad skill categories and the corresponding skill subcategories were identified as previously mentioned. The chart below shows the different broad categories and subcategories that were identified.



After cleaning the data, removing errors and duplicates, recoding variables, verifying data accuracy, the data was summarized and then visualized. A handful of charts were created to help address the research questions. There are a variety of bar charts that easily compare the different

fields that will be included in the Results section. A crosstabulation chart and regression analysis were also created. The crosstabulation investigated whether the broad skill categories were required for each of the jobs. If at least one of the subcategories was required, it was coded "yes" otherwise "no". For the regression analysis, a model that includes all the broad skill categories as independent variables and size and education levels as control variables was created. This model was used to determine the impact of these variables on salary.

Findings

The results showed that the majority of the hypotheses that were formulated were supported. Below, each research question is listed in italics followed by the hypothesis if applicable. Each question includes a visualization of the key finding. A brief summary explaining the finding is provided.

What are the most sought-after analytics skills?

Hypothesis: Database and programing skills are more sought after by companies compared to other skills.

Crosstabulation Chart

		Skill							
Skill Required?		*Cloud*	*Database*	*Programming*	*Project Management*	*Soft Skills*	*Stats Packages*	*Viz Packages*	*Total*
No	Count	9726a	6386b	5711c	10212d	1405e	10963f	11101f	55504
	% within Yes/No	17.50%	11.50%	10.30%	18.40%	2.50%	19.80%	20.00%	100.00%
	% within Skill	76.30%	50.10%	44.80%	80.10%	11.00%	86.00%	87.10%	62.20%
Yes	Count	3022a	6362b	7037c	2536d	11343e	1785f	1647f	33732
	% within Yes/No	9.00%	18.90%	20.90%	7.50%	33.60%	5.30%	4.90%	100.00%
	% within Skill	23.70%	49.90%	55.20%	19.90%	89.00%	14.00%	12.90%	37.80%
Total		12748	12748	12748	12748	12748	12748	12748	89236

Each subscript letter denotes a subset of Skill categories whose column proportions do not differ significantly from each other at the .05 level.

The results show that Programming and Database Skills are the most sought-after analytics skills. This result confirmed the hypothesis. One unexpected finding was the popularity of Soft Skills in comparison to analytic and database skills. Across jobs requiring various kinds

of technical skillsets, one constant was an additional requirement for Soft Skills of some kind. This shows that although technical skills are very important, companies are seeking well rounded candidates with the requisite that people skills are required to effectively communicate with key stakeholders. The chart shows that 49.9% of companies require Database Skills; this equates to 6,362 companies out of the total 12,748 companies. There are 55.2% of companies that require Programming Skills; this equates to 7,037 out of the total 12,748 companies. Finally, there are 89% of companies that require Soft Skills; this equates to 11,343 companies out of the total 12,748 companies. All the skill category column proportions vary significantly except for Stats Packages and Viz Packages. These columns have very similar counts in the "No" and "Yes" rows.

Which skill sets have the most influence on salary?

Hypothesis: Programing, database, cloud, statistical package, visualization, project management and soft skills each have a positive and significant impact on the salary offered regardless of the education level and company size.

The following regression model was developed to assess this hypothesis.

Salary = $\beta_0 + \beta_1(CompanySize) + \beta_2(Education) + \beta_3(Programming) + \beta_4(Database) + \beta_5(Cloud) + \beta_6(StatsPackages) + \beta_7(VizPackages) + \beta_8(ProjMgt) + \beta_9(SoftSkills) + e$

Regression Results

	Variables	Parameter Estimate
Control Variables	Size	-.015*
	Ph.D.	.184***
	Masters	-.006
	Bachelors	-.006
Independent Variables	Programing	.149***
	Database	.039***
	Cloud	.058***
	Statistical Packages	-.046***
	Visualization Packages	-.013
	Project Management	-.033***
	Soft Skills	.022**

*p<0.05 , **p<0.01 , ***p<0.001; R^2=0.084

As explained above, this model includes all the broad skill categories as independent variables and size and education levels as control variables. If at least one of the skill subcategories were required for the job, it was marked that the broad skill category was required. The results reveal that the Visualization Packages variable was not statistically significant, and it showed that it would have a negative impact on salary. Another finding that was surprising was related to the Statistical Packages variable. This one was highly significant, but it also showed that it would have a negative impact on salary. This could just be the result of the types of jobs that were looked at. It is worthy of an additional investigation. The rest of the variables were significant and had a positive impact on salary.

Is Python more popular than R in the job dataset that was analyzed?

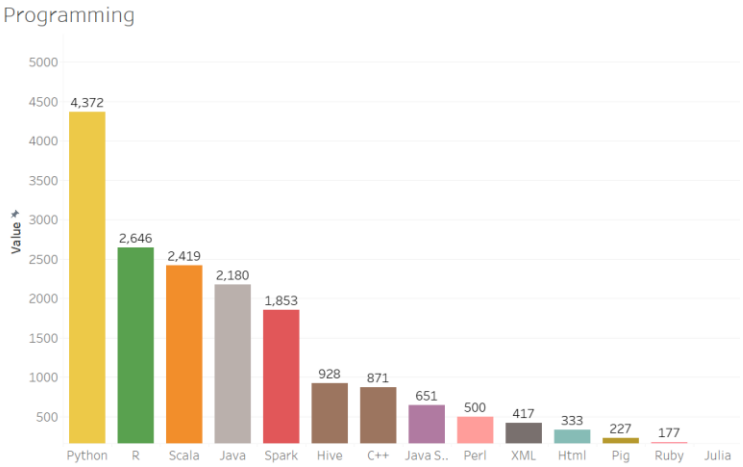
R and Python are both open-source programming languages and free to use. They share many similarities with slight variations. R is a widely used tool for statistical analysis and graphics. It is easy to learn and used in many college classrooms. R was released in 1995. It is used by statisticians and data miners. The statistical and analytical power of the program is unmatched (Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, B. (2017). Recently, it has been gaining market share in data analytics. Although it has been gaining market share, Python still has more than 7 times, 30.8%, the market share compared to R (PYPL PopularitY of Programming Language Index). Python was released a few years before R was. It is mostly used by software engineers and developers. People who use Python within their professions "are more likely to focus on the coding aspects of the language". Python is easier to read and follow. The syntax of the program also allows for easier coding and debugging (Gallagher, M., & Trendafilov, R. (2018).). Python is also highly utilized in college classrooms as professors realize the importance of teaching this program (Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, B. (2017).

According to R VS. Python: Ease of Use and Numerical Accuracy, Python attracts younger, undergraduate students who are trying to effectively market themselves to potential employers (Gallagher, M., & Trendafilov, R. (2018).). They see coding skills as an important skill in order to secure a career in data analytics (Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, B. (2017). Many also see Python as the easier option to learn due to its phonetic syntax. Python is based on a software development language, C, whereas R is based on a mathematical, statistical language, S (Gallagher, M., & Trendafilov, R. (2018).). The differences between the two languages are part of what makes Python easier to read. The Wilkinson's

Statistics Quiz was conducted on both programs. R has a slight advantage when it comes to precision since a Python user must make adjustments to receive the same results as R does. A Python user must know what they are trying to achieve and make the adjustments as necessary (Gallagher, M., & Trendafilov, R. (2018).).

Overall, R and Python attract different types of users. R has a statistical focus whereas Python is attractive to developers and coders. Many undergraduate students studying data analytics opt to learn Python first. This is partly due to its phonetic syntax and less intimidating structure; consequently, it would seem logical that more careers would require Python rather than R.

Hypothesis: Python will be more sought after than R.

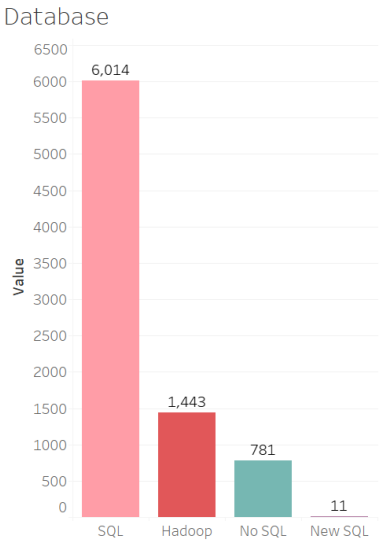


In the chart above, it shows that Python is the most popular programming language overall. Out of the 12,748 total job postings analyzed, 4,372 of them prefer a candidate to have knowledge or experience with Python, almost double the next closest language, R, at 2,646.

What specific languages should individuals learn from each broad skill category?

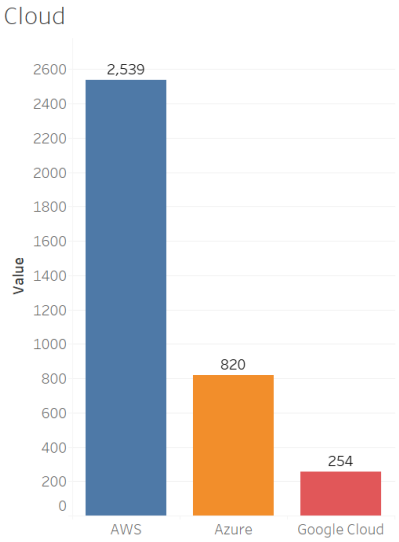
The charts show the different subcategories listed in order of most popular to least popular.

Database Hypothesis: SQL



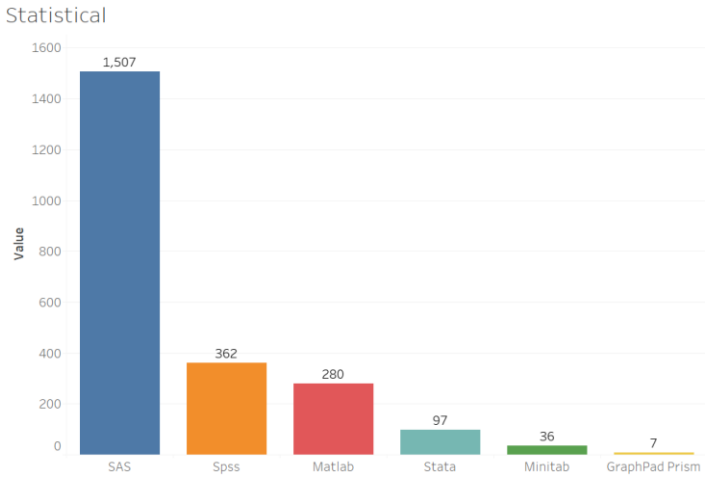
This chart shows that SQL is the most popular followed by Hadoop and No SQL. Only 11 job postings mention New SQL. Interestingly, SQL was the overall most popular skill that jobs require.

Cloud Hypothesis: AWS



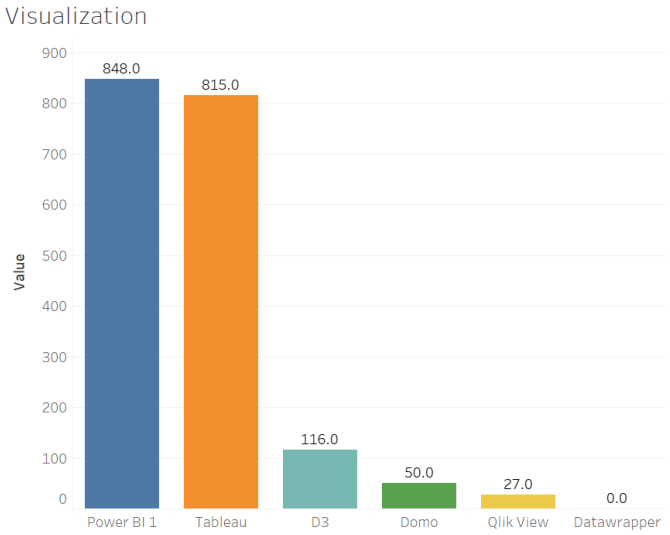
For this category, it shows that the AWS cloud storage is the most popular option. However, having general cloud storage knowledge is applicable to the different cloud storage companies.

Statistical Hypothesis: SAS



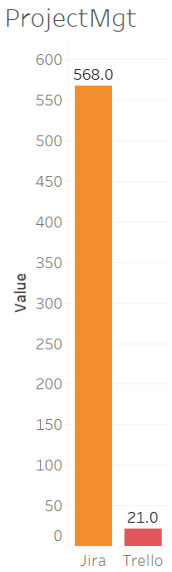
SAS was by far the most popular statistical software.

Visualization Hypothesis: Tableau



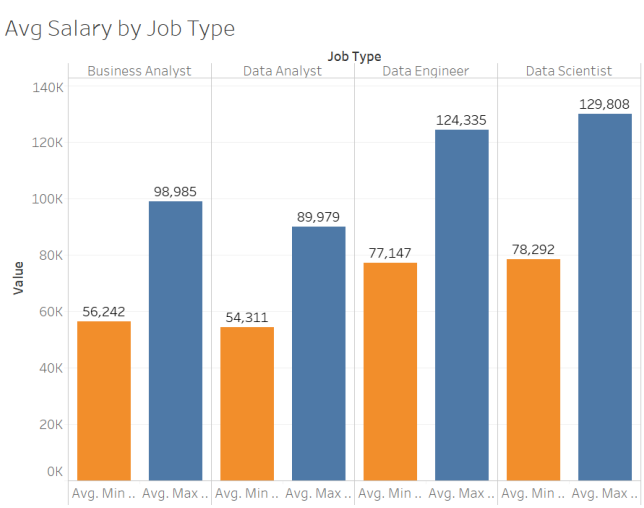
Power BI was the most popular with Tableau as a close second.

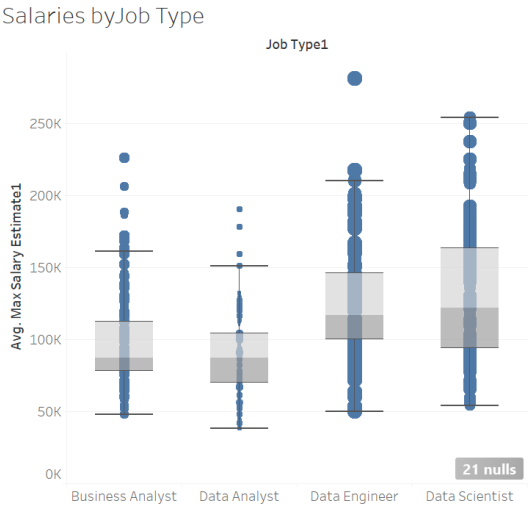
Project Management Hypothesis: Trello



Project management tools are very helpful with an agile methodology, and after conducting additional research, it was determined that Jira is the most commonly used tool when managing agile projects.

Does the pay vary by job type?

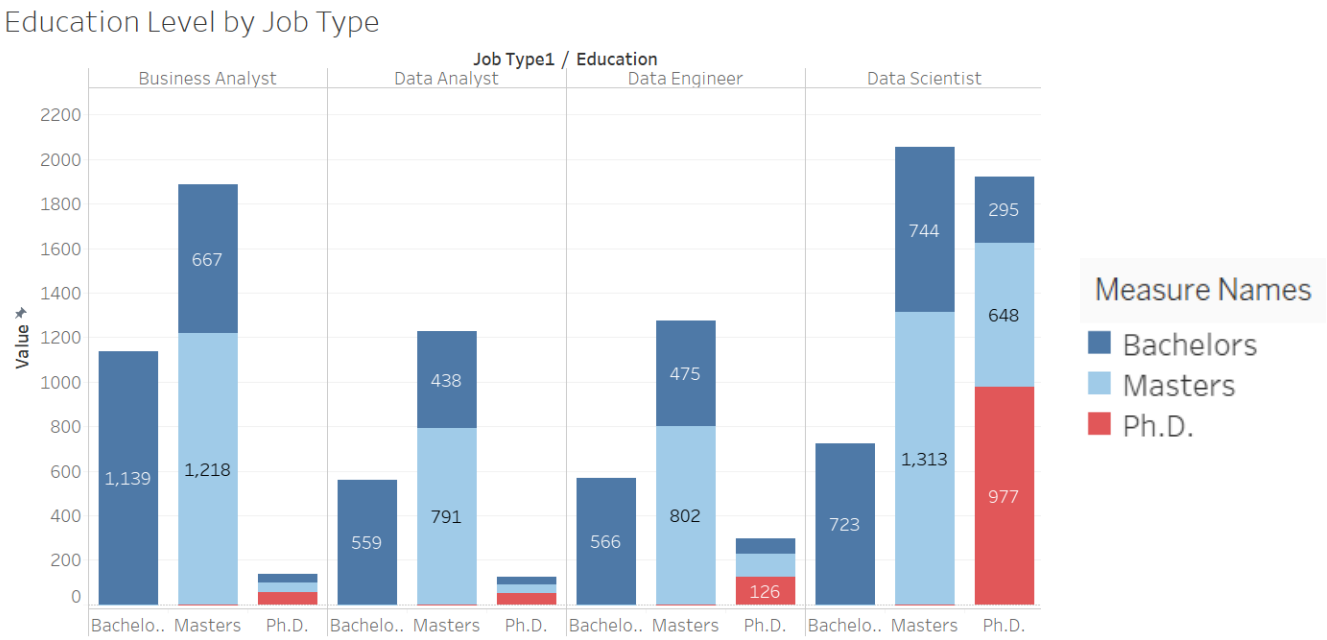




The orange bars display the average minimum salary, and the blue column displays the average maximum salary per job type. The Data Scientist position has the highest minimum

($78,292) and maximum ($129,808) salary. The Data Engineer position was close behind followed by Business Analyst and Data Analyst. The box plot displays the distribution of the salaries and the outliers which influence the average salary. This chart shows that the averages and medians still follow the same pattern with Data Scientist positions paying the most and Data Analyst positions paying the least overall.

What is the most popular education level by job type?



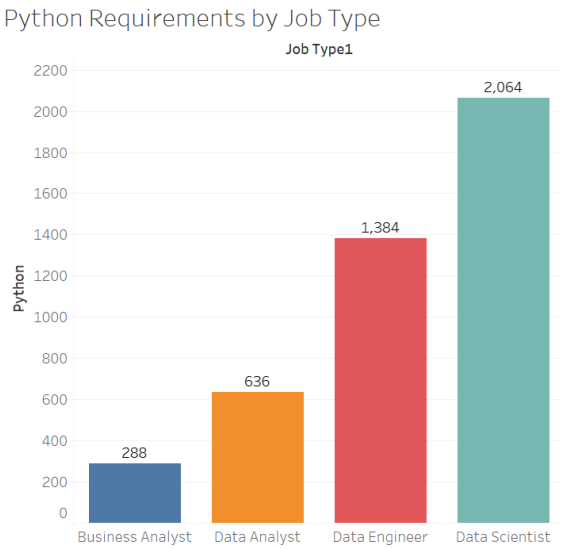
This chart shows the breakdown of the education level desired for the different job types. A large majority of the job postings allowed for a bachelor's degree but preferred a master's or Ph.D. degree. In Excel, if the job description mentioned either a bachelor's or master's degree, it was set up, so that it would show that a master's degree is preferred. The same thing was done for the Ph.D. degree as well. This is why there aren't any additional degree types in the bachelor's column. The Data Scientist position stands out because of the large number of

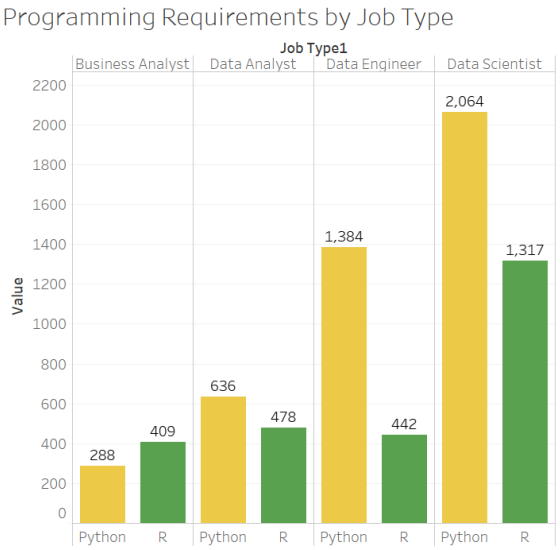
companies that prefer their candidates to have a Ph.D. or master's degree. This makes sense since it is the highest paying job type.

Do the required skills vary by job type?

Answering this question will permit students to continue narrowing down the skills to focus on learning. If a student wants to become a business analyst versus a data scientist, they will need a slightly different set of skills and knowledge base. The following numbers are the counts before the 34 duplicates were removed. There are 2,253 Data Analyst positions, 4,092 Business Analyst positions, 3,909 Data Scientist positions, and 2,528 Data Engineer positions in the job dataset analyzed. These numbers are provided to show that the counts for the different job types vary. The Data Analyst and Data Engineer positions allow for the best comparison due to the count similarity. If a job type has more positions such as Business Analyst, it makes it possible for it to almost double the requirements of the Data Engineer position. This is important to remember when viewing these charts.

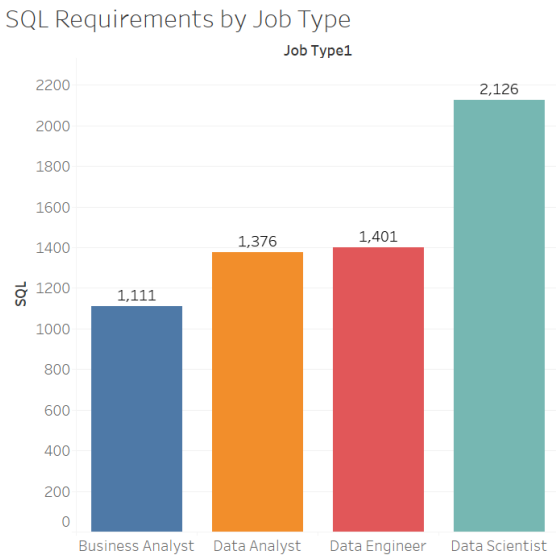
Python:





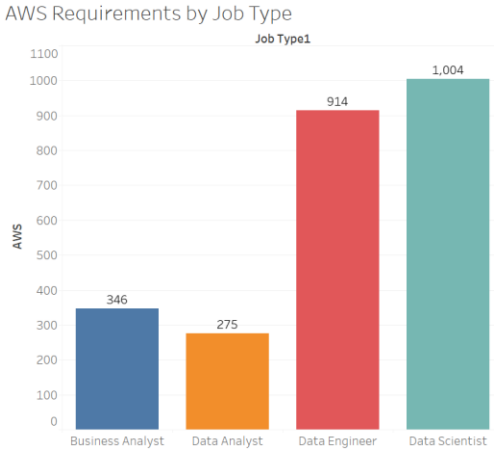
The chart on the left shows that Data Scientists are more likely to require Python compared to other job types. In an effort to determine if this was also the case for other programming languages as well, another chart that included R was created and found a similar pattern. One interesting finding is that for the Business Analyst position more companies require R versus Python.

SQL:



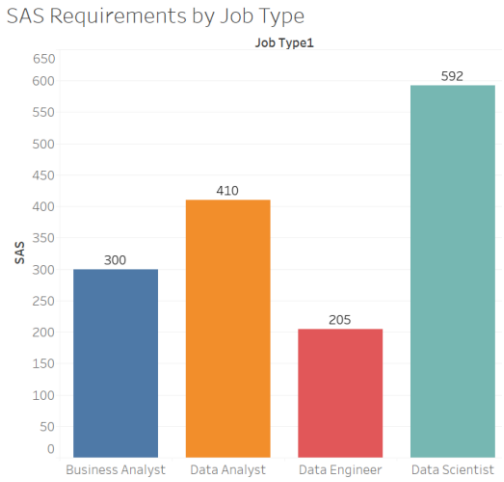
There is a similar pattern with this chart as well. This may show why the Data Engineer and Data Scientist positions pay more. They are requiring slightly more skills and knowledge compared to the other positions.

AWS:



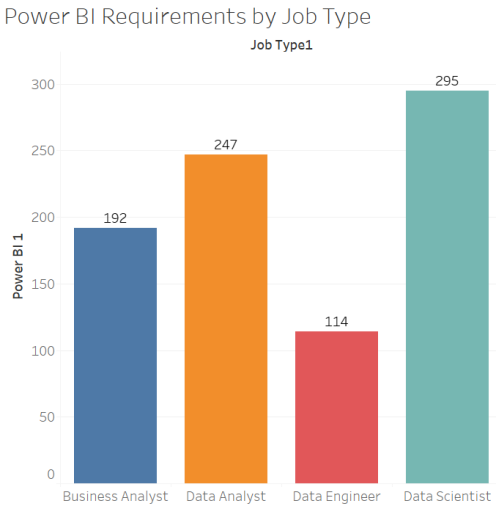
This chart showed that more Data Engineer and Data Scientist positions require AWS knowledge compared to the other positions.

SAS:



This chart shows that Data Scientist and Data Analyst positions are more likely to require SAS. This is one of the higher requirements for the Data Analyst position.

Power BI:

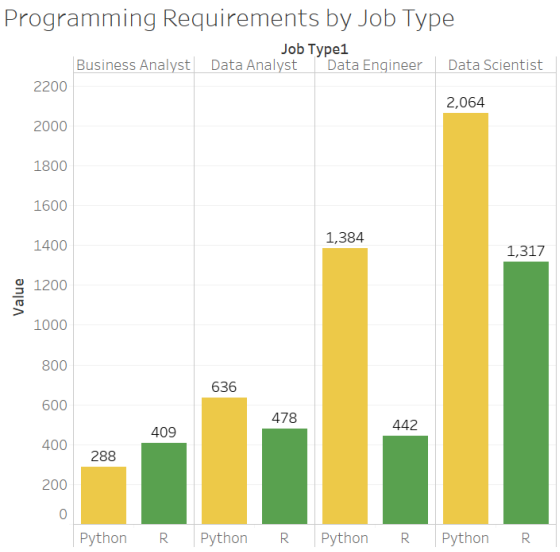


This chart shows which job types more heavily utilize Power BI. It was surprising to see how low the Data Engineer count was.

Overall, the Data Scientist position had the highest requirement counts for all the subcategories. The top three most frequently required subcategories for this position were SQL, Python, and AWS. The top three most frequently required subcategories for the Data Engineer position were SQL, Python, and AWS as well. For the Data Analyst position, it was SQL, Python, and Power BI. Finally, for the Business Analyst position, it was SQL, AWS, and SAS. This was the most unique set of top three subcategories. This shows how important SQL knowledge is to anyone who is interested in entering the Data Analytics job market.

Student perspective: which is better to learn?

Hypothesis: Python is the better language to learn due to its popularity.



This chart from above helps to answer this question. Python is significantly more popular overall, but this chart shows the breakdown by job type. R is actually more popular for the Business Analyst position; however, the rest of the job types prefer Python.

Which sectors currently have the most jobs available?

Number of Jobs by Education

Sector	Education			
	Bachelors	Masters	Ph.D.	Unknown
Information Technology	735	1,059	324	1,562
Business Services	585	805	131	985
Finance	312	320	67	277
Health Care	137	262	49	118
Biotech & Pharmaceuticals	51	128	199	49
Insurance	147	140	52	81
Manufacturing	107	128	37	80
Education	73	94	32	54
Government	41	75	47	64
Aerospace & Defense	52	104	24	45

This chart shows that the Information Technology and Business Services sectors have the most open positions. It also shows that the companies in these sectors prefer their applicants to have a master's degree compared to the other degree types. It also shows that there are still plenty of opportunities for people who only want to earn their bachelor's degree.

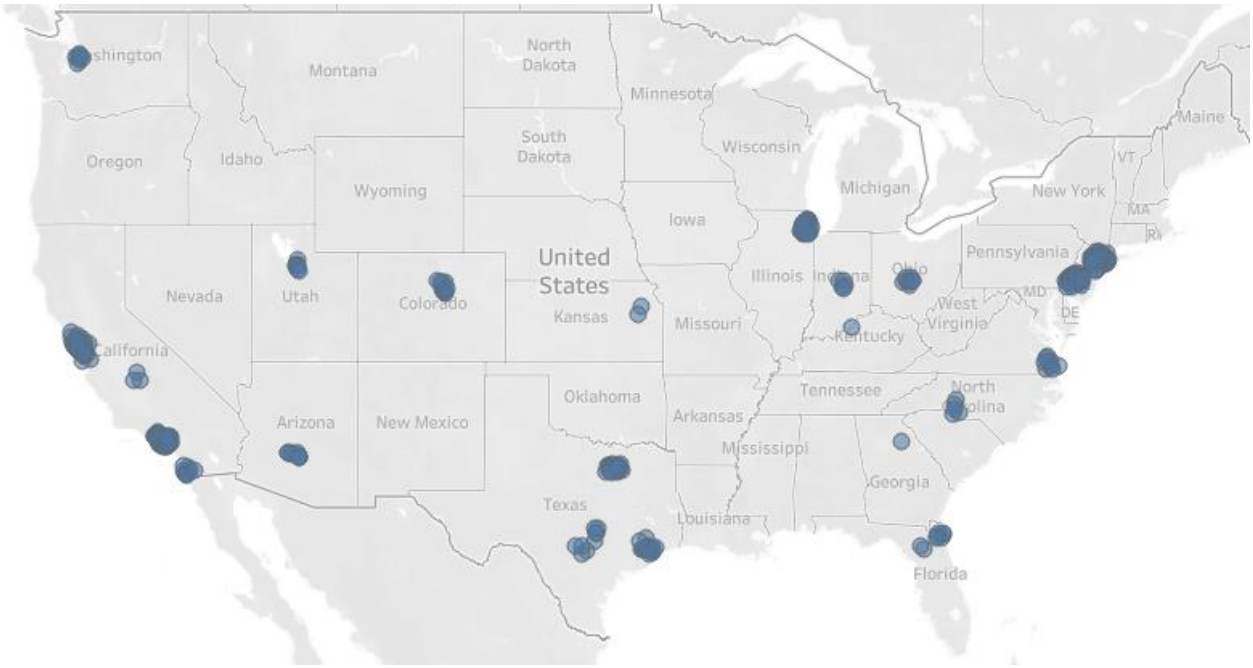
Discussion

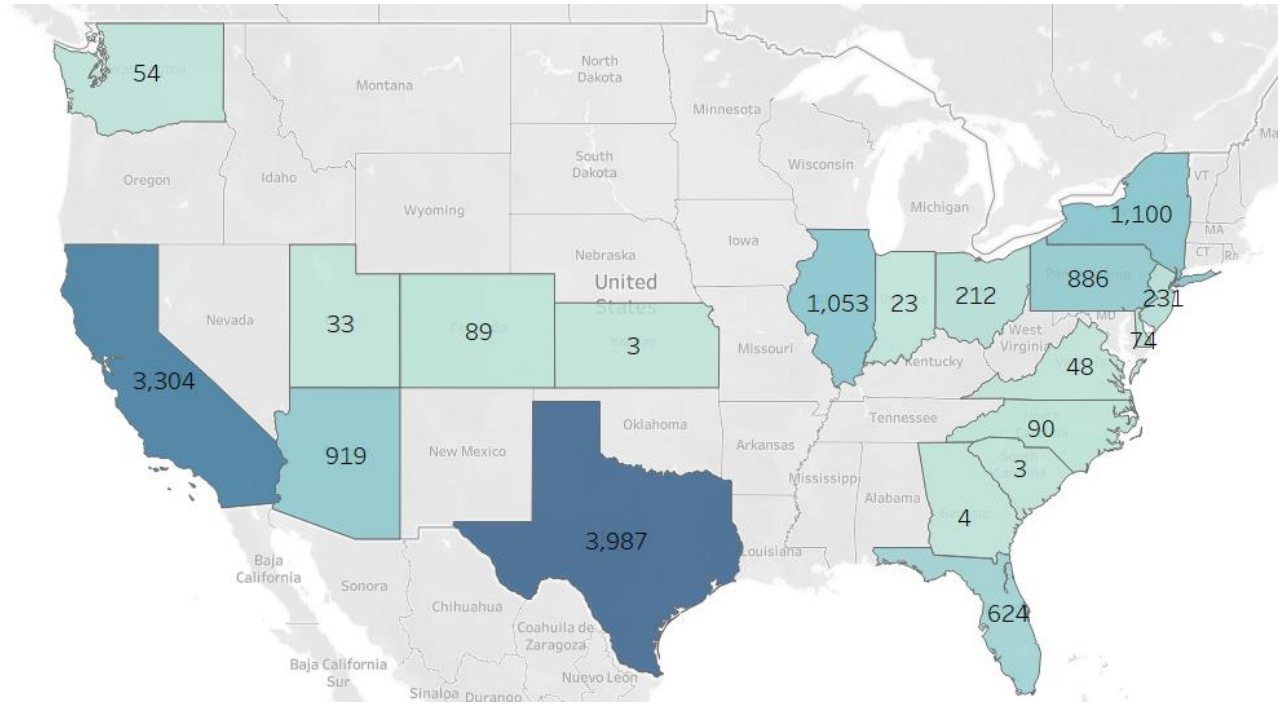
Using a sample of 12,748 job postings from Glassdoor, this study has helped to shed some light on which analytic skillsets are most sought after by organizations. The major findings of the study are:

1. Python, SQL, AWS, SAS, and Power BI are the most popular analytics tools.
2. The independent variables used in the Regression Analysis all had a significant impact on salary besides Visualization Packages. The Visualization Packages and Statistical Packages variables both had a negative effect on salary, but the rest of the variables had a positive effect.
3. The pay and preferred education level varies by job type. The Data Scientist position was the most highly paid position. A significantly higher number of companies also prefer candidates to have their Ph.D. compared to the other job types.
4. In addition to Programming and Database Skills, Soft Skills are crucial and will help individuals to be well-rounded and successful in their careers.

Like all studies, the current study had some limitations which future research may wish to address. First, although, a relatively large dataset was used, it would be worthwhile to study a larger dataset. Based on the first chart listed below, the location of these jobs seem to be where all the technology companies are headquartered, besides Boston. Datasets that have more jobs

within the different industries besides Information Technology and Business Services would be interesting and important to analyze.





It would also be useful to conduct more research into the differences between the job types in order to understand why the jobs pay differently. Additionally, it would be helpful to learn how the day-to-day tasks vary before applying for a job in the future.

Future studies could further explore the impact of Statistical Packages on salary. This was one of the most surprising results of this research, and it would be useful to definitively find out why the data shows that Statistical Packages has a negative impact on salary.

Moving forward, the goal is to finish the analysis of the top data analytics graduate programs across the country. So far, information about the curriculum of 61 different programs and all the publicly available information about the coursework and syllabi were collected. Collecting tuition information was more difficult to find than had been anticipated. If the college or university taught a specific skill, the cell was marked with a "1". This method will be similar to the one used during the original thesis work. The foundation for this research has been set up, and it will be an interesting project. This thesis and research were able to provide valuable information.

Recognition

A special thank you to my advisor Kholekile Gwebu. He was instrumental in helping me throughout this process. He consistently provided resources, guidance, and advice. It was a pleasure working with him.

References

Brittain, Jim; Cendon, Mariana; Nizzi, Jennifer; and Pleis, John (2018) "Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance," SMU Data Science Review: Vol. 1: No. 2, Article 7.

Chang, H. C., Wang, C. Y., & Hawamdeh, S. (2019). Emerging trends in data analytics and knowledge management job market: extending KSA framework. *Journal of Knowledge Management*.

Gallagher, M., & Trendafilov, R. (2018). R vs. Python: Ease of Use and Numerical Accuracy. Journal of Business and Accounting, 11(1), 117-126.

Larxel. (2020, July). Data Analyst Jobs, Version 1. Retrieved February 11, 2021 from https://www.kaggle.com/andrewmvd/data-analyst-jobs.

Larxel. (2020, July). Business Analyst Job Listings, Version 1. Retrieved February 11, 2021 from https://www.kaggle.com/andrewmvd/business-analyst-jobs.

Larxel. (2020, July). Data Scientist Jobs, Version 1. Retrieved February 11, 2021 from https://www.kaggle.com/andrewmvd/data-scientist-jobs

Larxel. (2020, July). Data Engineer Jobs, Version 1. Retrieved February 11, 2021 from https://www.kaggle.com/andrewmvd/data-engineer-jobs

Michalczyk, S., Nadj, M., Maedche, A., & Gröger, C. (2021). Demystifying Job Roles in Data Science: A Text Mining Approach.

"Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, B. (2017). MatLab vs. Python vs. R. Journal of Data Science, 15(3), 355-372."

Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International journal of information management*, *50*, 416-431.

"PYPL PopularitY of Programming Language Index." *Index*, Dec. 2020, pypl.github.io/PYPL.html.

Radovilsky, Z., Hegde, V., Acharya, A., & Uma, U. (2018). Skills requirements of business data analytics and data science jobs: A comparative analysis. Journal of Supply Chain and Operations Management, 16(1), 82-101.

Verma, A., Yurov, K. M., Lane, P. L., & Yurova, Y. V. (2019). An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *Journal of Education for Business*, *94*(4), 243-250.