How to Get Hired as a Data Scientist: A Comprehensive Analysis of Glassdoor Job Postings

Jason Katz, Xi Luo, PhD

Brown University

Abstract

The field of data science has exploded in recent years and many aspiring data scientists are struggling to gain the proper skills to enter the job market. With many buzzwords and a wide range of topics, it is difficult to decipher which skills are most important when creating one's learning path. Prior research has been conducted to assess the relevance of different skills for data science. This paper extends the work of previous studies, using a larger and more recent dataset from Glassdoor, one of the most popular job sites in the country. Unlike previous work, in this study, analysis into software and skills goes beyond frequency, considering correlation between different skills and also assigning financial value to the various software, skills, and qualifications using job salaries. Results suggest that Python is ahead in its competition with R for the most relevant data science language, big data and machine learning skills are amongst the most valuable, and graduate level work is becoming necessary for many jobs. This study will help guide the next generation of data scientists in their development of the skills sought after most by employers.

Introduction

Data science dates back many decades [1], but with the advancement of computers in terms of storage and processing power, the field of data science has been grown growing rapidly [2, 3], with some even labeling it as the “Sexiest Job of the 21st Century” [4]. Despite the huge demand [5], many people are struggling to get into the field. Much of this has to do with the skills gap between employers and qualified data scientists [6]. For those looking to get into the field, there are many ways to get started, including formal education, bootcamps, online tutorials, and data competitions. It can be confusing on where to start, as there are so many potential paths to follow [7, 8, 9]. With buzzwords such as “Deep Learning”, “Natural Language Processing”, and “Genetic Algorithms”, it seems like there is always something new to learn. With the goal of educating the next generation of data scientists, previous studies have looked at what skills are in the highest demand by analyzing job postings. There have been a few case studies by individuals looking at the frequency of skills in job postings [10, 11, 12], but the data size was relatively small and all the work was from two to four years ago, and in a field evolving as fast as data science, results from that long ago may not represent its current state. CrowdFlower did a review of data science skills, analyzing job postings from LinkedIn [13], but their sample size was again small and the analysis was limited. Another interesting approach to providing insight into the job market came from a study analyzing the public LinkedIn profiles of data scientists [14]. This gave insights from the perspective of current data scientists but couldn't speak to the point of view of companies. The most comprehensive study to date was conducted by Glassdoor’s research team, analyzing 10,000 job postings from their site [15]. That study provided more insight into skills and salaries, but could still be improved upon by using the entire job market for data science (they used approximately half of all postings), using more recent data (they used jobs from one to one and half years ago), and also providing a more extensive analysis. This study aims to improve upon previous work, using more recent data, and a more complete set of job postings to provide valuable insights for aspiring data scientists. Unlike in past studies, software and skills will not only be analyzed based on their frequency in job postings, but also based on their correlations with each other in addition to financial considerations, using salaries from the job postings. The goal is to draw clear conclusions from quantitative data, stating which software is the most relevant in terms of job qualifications and also the financial value associated with expertise in particular software. Insight into the debate of “Python vs R” [16, 17], as well the comparison of big data skills with machine learning skills will be provided. Analysis of job prospects and value in terms of location and formal education will also be discussed. In the end, this study should provide an overview of what goes into being a data scientist and what areas those interested in the field should explore next, taking into account the financial relevance of such decisions.

Methods

Web Scraping

All job data was scraped from Glassdoor using the selenium package in python. ChromeDriver was used as the driver of choice for selenium. The scraping took place from April 22 - April 24, 2018. The job search was restricted to the 50 U.S. states and Washington D.C. For each search, a location was entered, and the keyword used was “Data Scientist”. Only jobs posted within the last month were used (the parameter: “fromAge=30” was added to the search URL). The parameter “radius=0” was also used in the search URL to restrict results to the search’s exact location. For states where there were 900 jobs or less, the maximum number of pages was extracted (30 results per page) and the web scraper went through each page one by one (manipulating the search URL). If there were more than 900 jobs, the search was subset by industry, and then the same process for iterating through pages was followed for each industry.

For each page, the URLs of the job postings were extracted and then visited. The HTML fields were used to find the job title, job description, company rating, and estimated median Glassdoor salary (if listed). Jobs with hourly salary estimates were converted to yearly estimates by multiplying the hourly rate by 2,000 (40 hours a week for 50 weeks). During the scraping, a random amount of sleep time was given in between HTTP requests. A total of 20,537 jobs were scraped in 40.6 hours.

Frequency Analysis

Using regular expressions to determine if a particular word appeared in the job description, counts were calculated for a list of popular software and skills, as well as various qualifications such as having a Bachelor's degree, Master's degree, and Ph.D. (collapsed for both "PhD and Ph.D.). Correlations for all attributes were calculated. Additionally, all the correlations for the 15 most frequently occurring software were used to construct a network graph.

Salary Analysis

The mean salaries (only including postings where a Glassdoor salary was present) were calculated, subset by qualification, software, and skill (only the top 25 appearing software and skills were used). The data was restructured to a single data frame, where each row was a job posting and the columns contained the salary and 103 binary features for software, skills, qualifications, and states. A non-negative least squares regression model was fit to this data. The non-zero coefficients were extracted and used to estimate the added value of a particular attribute.

Results

Frequency Analysis

In total, there was 49 software analyzed. The one that appeared in the most job descriptions was SQL with 7,004 occurrences, followed by Python and R with 6,463 and 6,125 occurrences respectively. Figure 1 below shows the number of occurrences for the 25 most common software.

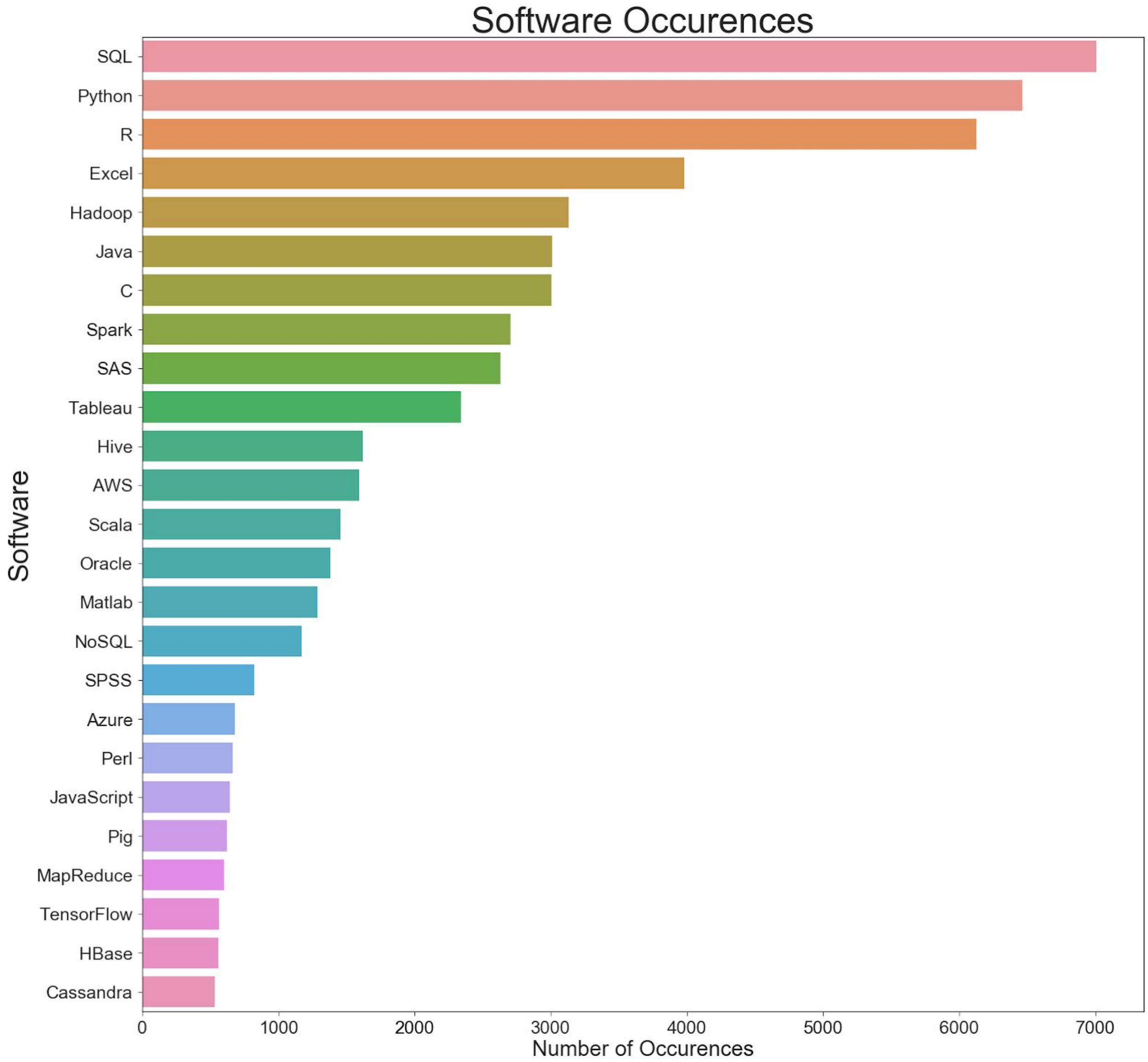
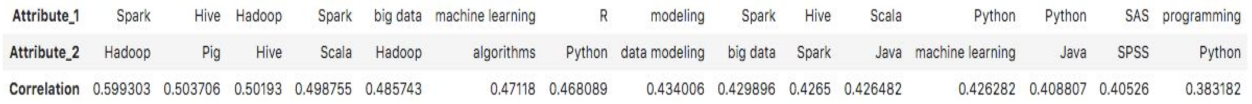


Figure 1: The 25 most common Data Science software

The qualifications that a job description mentioned were broken down into either a Bachelor's degree, Master's degree, or Ph.D. The most asked for qualification was a Ph.D. with 4,917 occurrences, followed by a Bachelor's and Master's with 4,909 and 3,202 occurrences respectively. Of all the job postings, 7,233 mentioned either a Master's or Ph.D. (or both), and 3,270 mentioned only a Bachelor's. Searching for correlations between different attributes occurring together, Spark and Hadoop had the highest correlation at .600, followed by the combination of Hive and Pig at .504 and Hadoop and Hive at .502. The top 15 correlations are shown in Figure 2 below.



Attribute_1	Spark	Hive	Hadoop	Spark	big data	machine learning	R	modeling	Spark	Hive	Scala	Python	Python	SAS	programming
Attribute_2	Hadoop	Pig	Hive	Scala	Hadoop	algorithms	Python	data modeling	big data	Spark	Java	machine learning	Java	SPSS	Python
Correlation	0.599303	0.503706	0.50193	0.498755	0.485743	0.47118	0.468089	0.434006	0.429896	0.4265	0.426482	0.426282	0.408807	0.40526	0.383182

Figure 2: The correlations of related attributes

Figure 3 below displays the network graph of the 15 most frequently occurring software. The size of each node corresponds to the software's number of occurrences and the width of each edge in the graph corresponds to the correlation (with respect to occurrences) between each software. The top 10 correlations are highlighted in blue.

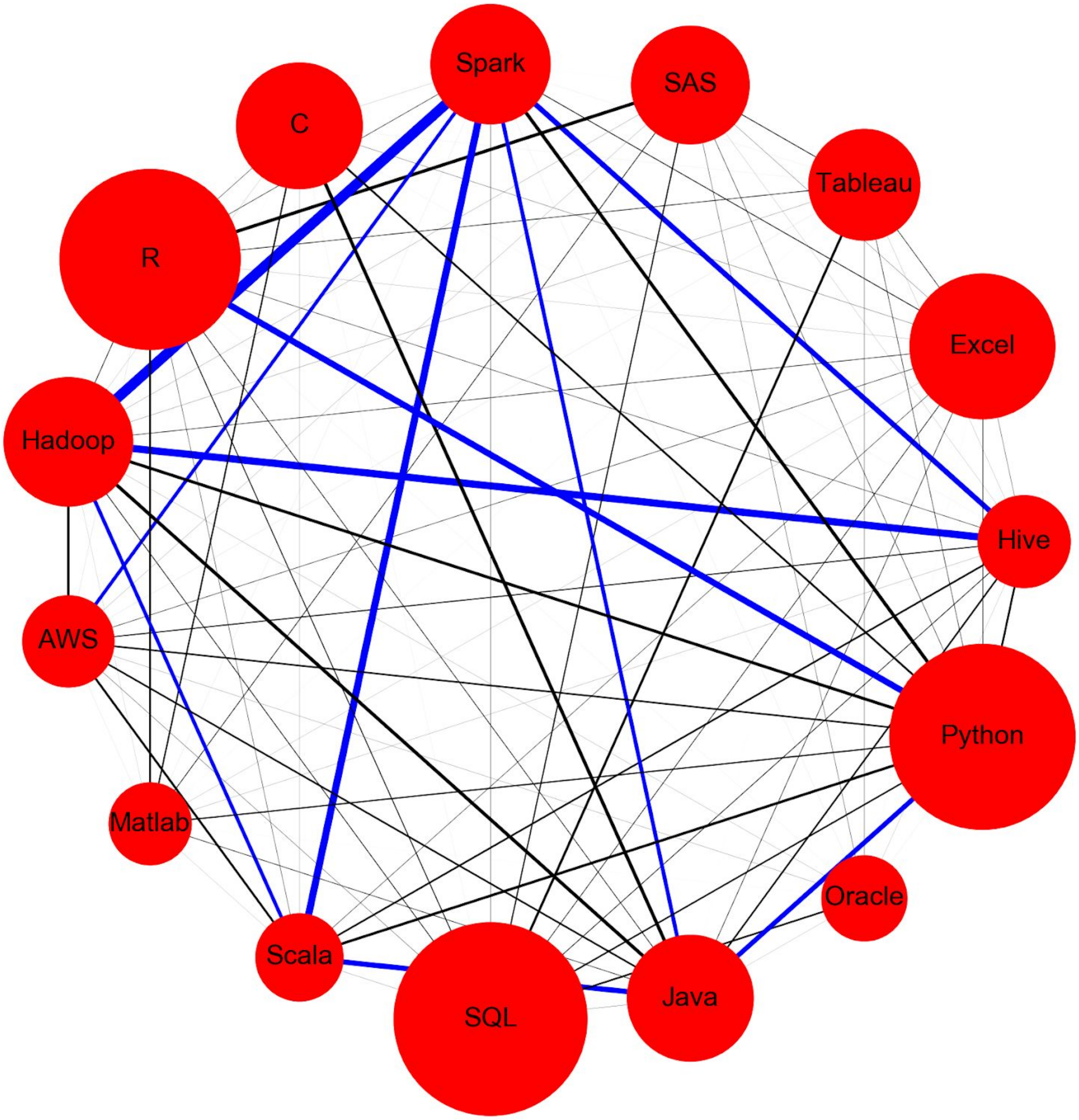


Figure 3: The network graph for the 15 most frequently occurring software

Salary Analysis

Of the 20,537 job postings, 12,468 had an estimated salary associated with it

For software, the highest salary occurred in jobs asking for TensorFlow, with a median salary of $119,000, next were Scala and Spark with salaries of $114,000 and $112,000 respectively.

Figure 4 below displays the median salaries for the top 25 software.

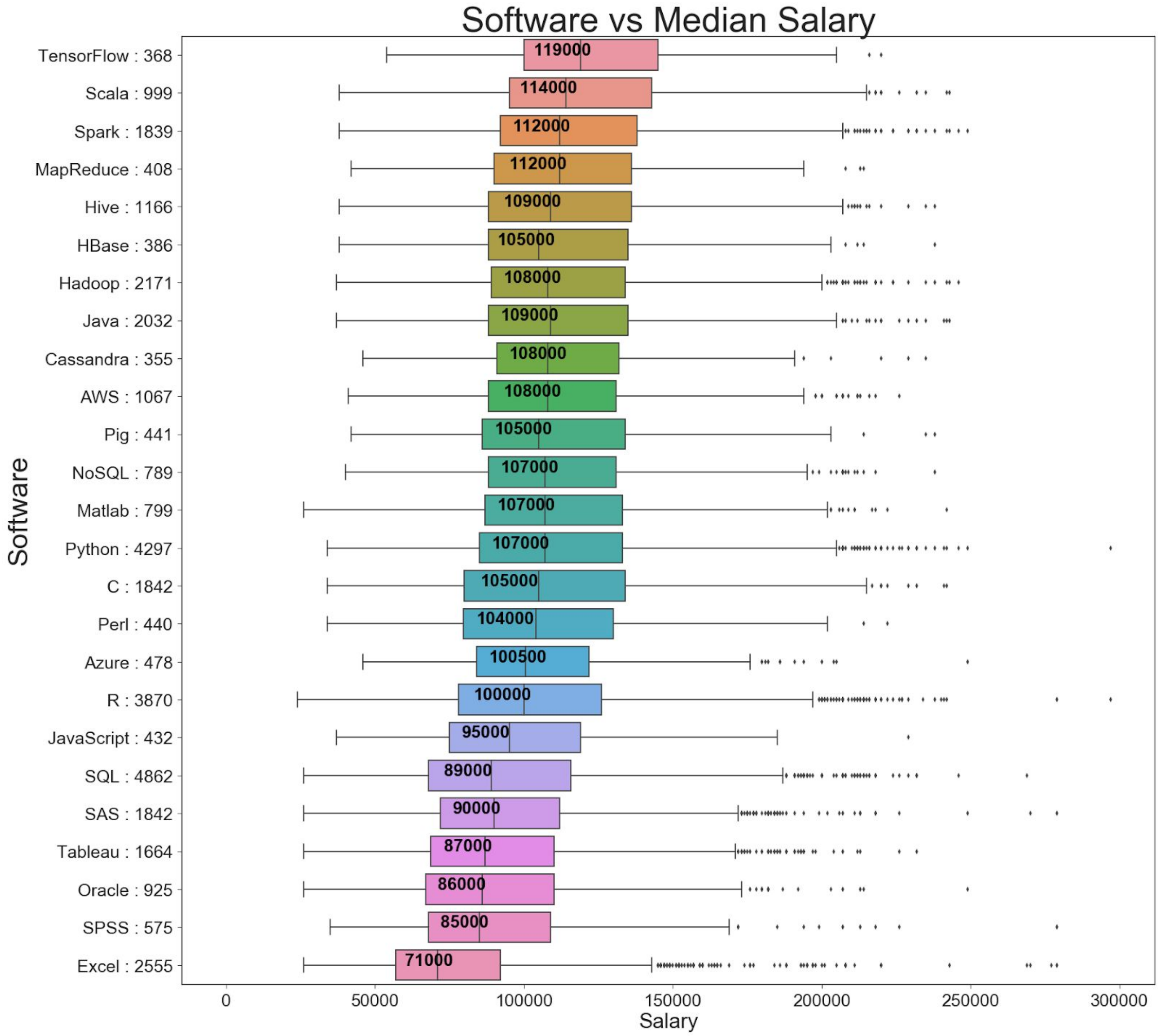


Figure 4: Boxplots of the salaries for the top 25 software. The number inside the box is the median of all salaries listed on Glassdoor, the number next to the software is the number of job postings for that software with a listed salary.

For skills, the highest salary occurred in jobs asking for Deep Learning, with a median salary of $119,000, next were Natural Language Processing and Artificial Intelligence with salaries of $115,500 and $115,000 respectively. Figure 5 below displays the median salaries for the top 25 skills.

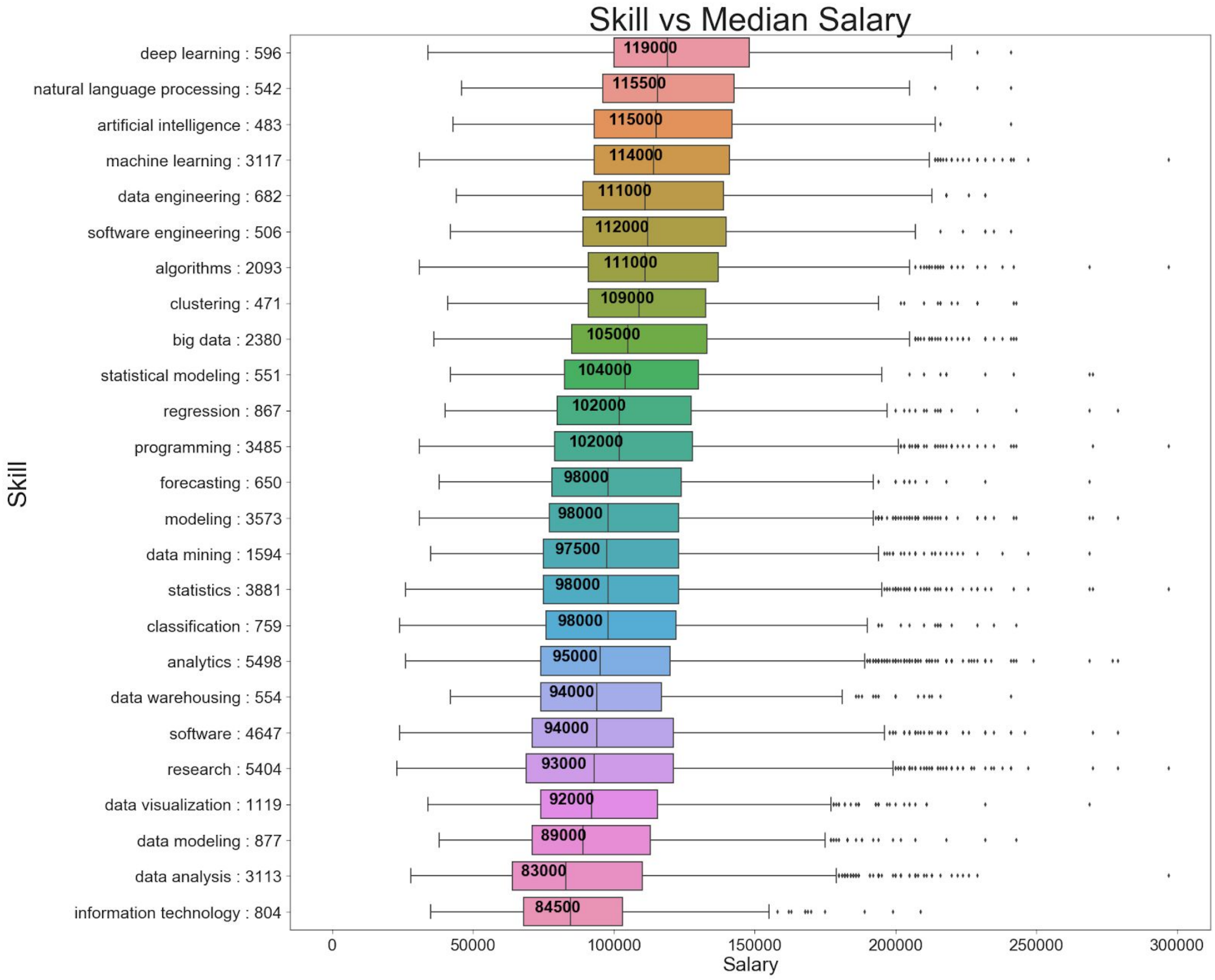


Figure 5: Boxplots of the salaries for the top 25 skills. The number inside the box is the median of all salaries listed on Glassdoor, the number next to the skill is the number of job postings for that skill with a listed salary.

Jobs asking for a Ph.D. had a median salary of $107,000, whereas jobs asking for a Master’s or Bachelor’s had salaries of $88,000 and $78,000 respectively. For the non negative least squares regression, of the 25 software used, 12 were non-zero. C had the highest coefficient at 5,012, followed by Spark and Python at 4,066 and 4,052 respectively. The coefficients of the 12 software are shown in Figure 6 below.

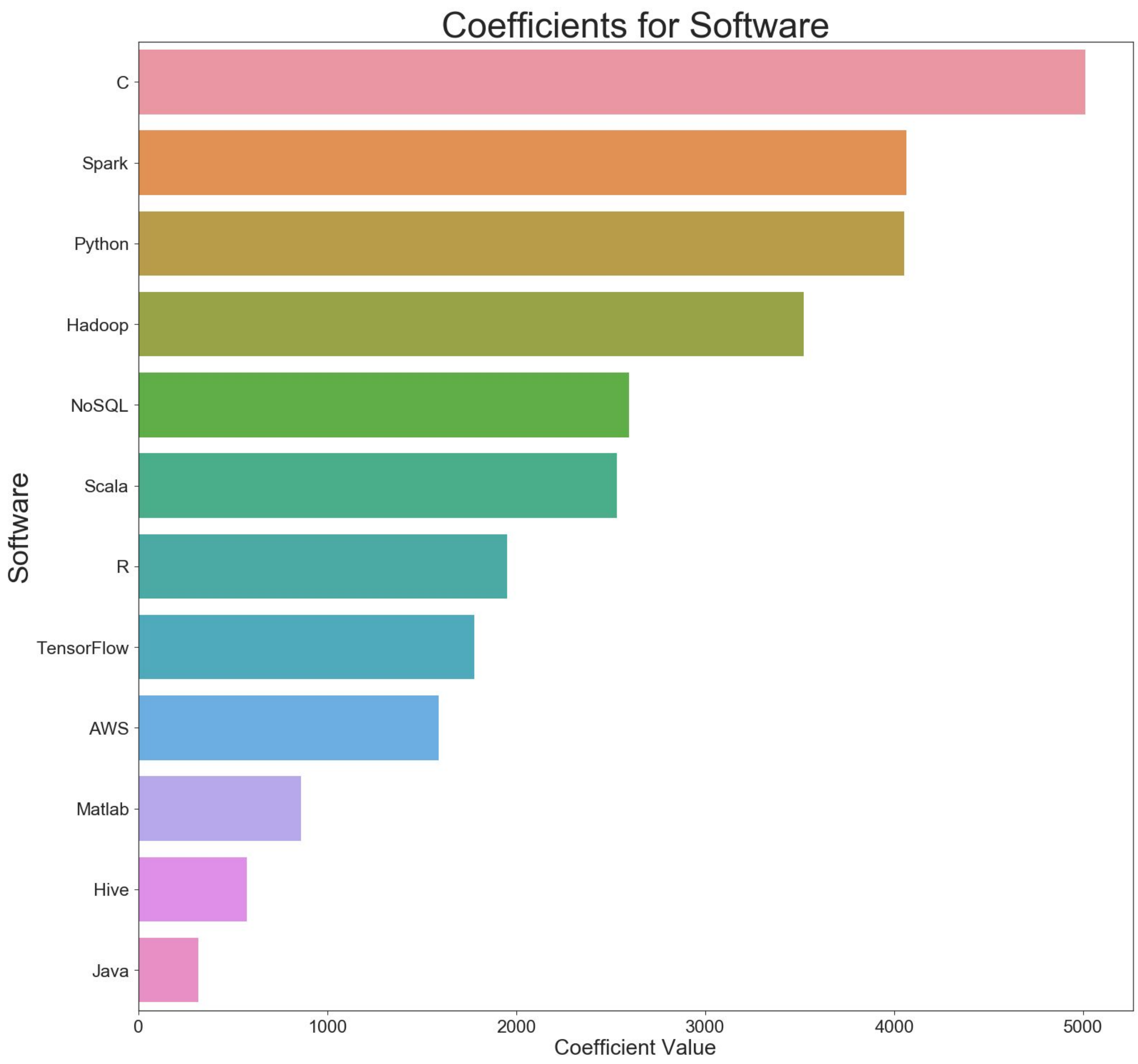


Figure 6: The coefficient values for different software

Discussion

Frequency Analysis

There have been previous studies done, analyzing the frequency of different software mentioned in job postings [11, 12, 13, 14, 15]. All of the studies had very similar findings with R, SQL, and Python being in the top five in all the studies, and Hadoop and Java each being present in the top five in all but one. The results of this study support previous work, with SQL, Python, and R being the top three software and Hadoop and Java both in the top six. The rivalry between R and Python for data science is a hot topic [16, 17] and this study suggests that there are more job opportunities for those who know Python compared to R (6,463 vs 6,125), although SQL still remains the leader in terms of the most asked for software. One result that differed from this study compared to past studies was the occurrence of Excel, which ranked number three, but was much lower and even not listed in some of the past studies. This might be due to the fact that Excel is not a true programming language like the others and might not have been searched for at all. It is also possible that other studies did not include as many "Data Analyst" roles, which would be significant since jobs that had a title containing "Analyst" asked for Excel in 41% of job postings, whereas all other roles only asked for Excel 12% of the time. For different qualifications, it appears that having some graduate level work (Master's or Ph.D.) is very important, as more than twice as many job postings mention either a Master's or Ph.D. compared to only mentioning a Bachelor's degree (7,233 vs 3,270). Analyzing correlation among different attributes, it appears that associated software is mentioned together frequently (i.e., Spark and Hadoop, Hive and Pig, Hadoop and Hive). Other related attributes frequently mentioned together include Python and R, Hadoop and "big data" Python and "programming", and R and "statistics". These results support the idea that job postings often give a list of example software they would expect experience in when asking for a specific skill set.

Salary Analysis

Previous studies did not conduct detailed salary analysis; therefore, the results of this study were analyzed independently. The salary data used in this study comes from Glassdoor estimates. Sometimes, the estimates are directly supplied by the employer, but when they are not, Glassdoor estimates the salary using known company salaries, competitor salaries, and other market indicators [18]. Analysis of salaries segmented by software had interesting results. Despite being at the very top for software mentioned in postings, SQL, Python, and R all fell in the bottom half for median salaries. This might be due to the fact that they are all mentioned so frequently and required for so many jobs that having experience with that software is more of a general requirement and expectation as opposed to something that sets someone apart. The software with the highest salary is TensorFlow, being one of the most widely used libraries when it comes to deep learning [19] (which happened to be the skill with the highest salary). Interestingly, the majority of the software with the highest salaries are all related to big data (Scala, Spark, MapReduce, Hive, Hadoop, etc.). With data becoming much bigger in recent years [20], the need for big data skills is becoming more valuable. It also makes sense that Excel has the lowest salary, as it is very common software and also mostly associated with "Analyst" jobs, which are lower paying. Looking at which skills having the highest salaries, it appears that common data science buzzwords are associated with the highest salaries. Topics such as "deep learning", "natural language processing", and "artificial intelligence" are more than just buzzwords, as they also yield higher salaries. Analyzing salaries for different qualifications, there is a clear order to which degrees are most valuable, with median salary increasing by $10,000 going from a Bachelor's to a Master's and then increasing another $19,000 when earning a Ph.D. In addition to many jobs asking for graduate degrees, salaries increase significantly the more time someone spends in school. These numbers have strong relevance, as the financial impact of attaining higher qualifications is often forefront in a person's decision to invest in higher education. When analyzing different attributes, the value comes from both the frequency the attribute is asked for in a job description, as well as the median salary associated with those jobs. In an attempt to combine the meaning of frequency and salary into an added value of a particular attribute, a non-negative least squares regression was fit. An ordinary regression would yield negative coefficients, which wouldn’t make sense as an added value of an attribute. The coefficients which end up as zero means that attribute does not add any significant value to one’s job prospects. The rest of the coefficients can loosely be interpreted as the value a particular attribute adds to one’s net worth as a data scientist. It is interesting to note that C is at the top of added value for software. One reason could be that experience with a low level language like C often indicates the ability to work well with higher level software. Looking at the other top software, R and Python are towards the top of the list, as they are the two languages often used in a data science setting [16, 17]. The value added for Python is about twice that of R, which suggests it is winning the battle for the main data science language. Similar to median salary analysis, big data tools such as Spark and Hadoop add a lot of value. It should be noted that SQL does not add any value, which might be due to the fact that it is mentioned so frequently it is thought of as a standard requirement. It could also revolve around the simplicity of SQL and the ability of for someone to learn it much faster than other software [21]. When deciding what software to learn first (or next), this list could be a very useful resource for aspiring data scientists.

Limitations and Future Research

One limitation of this study was the quality of the salary data used. The vast majority of the salaries were estimated by Glassdoor. To get a better understanding of how different attributes factor into a salary the true salary for the positions would have to be used. This is very difficult, as many companies keep their salary information private. Another limitation of this study was the raw extraction of keywords from job postings. Not all mentions of a particular software, skill, or qualification should be weighted equally. For example, some attributes are listed as a requirement for a job, whereas others are only preferred. Additionally, employers often post a list of relevant skills, not expecting all of them to be met. Future studies should perform more advanced text analysis to determine the weight of different attributes in the job postings. Additionally, the use of non-negative least squares regression to determine the value added of particular attributes is not a true mapping of the value added of learning a particular skill. More advanced methods involving an individual's current skill set should be performed to determine how much value is truly added to an aspiring data scientist's net worth. It would also be interesting to see how job postings from other sites such as Indeed and LinkedIn compare to the ones from Glassdoor.

Conclusion

There is a great need for quantitative assessment of the various aspects of data science. The demand for data scientists far outpace the current supply and many are trying to get into the field. With a plethora of online resources and a wide range of relevant topics, it can be overwhelming for aspiring data scientists to find a place to start. There have been previous attempts to quantify this job market in terms of relevant skills, but none have used the entire corpus of job postings, nor have any connected the value of different skills using job salaries. A review of this study should help give an idea of what employers are looking for and provide a guide to what software and topics are most relevant for those trying to gain experience in the field. A more informed population will allow individuals to master the skills most desired by companies and help advance the field of data science, generating more breakthroughs and allowing companies to make better use of their vast amount of data.

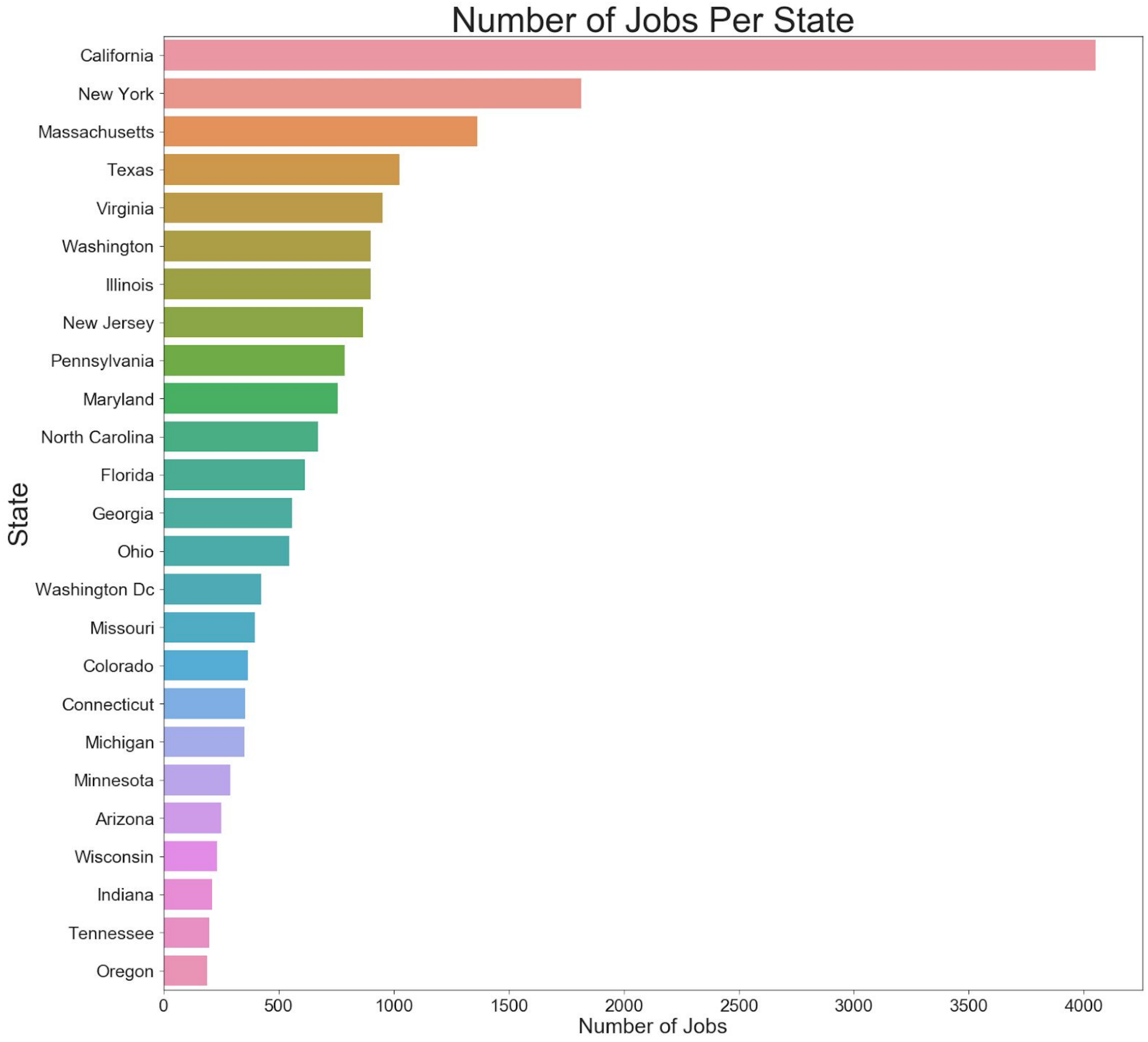
References

1. Press, Gil. “A Very Short History Of Data Science.” Forbes, Forbes Magazine, 15 Oct. 2014, www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#1897b3f955cf.

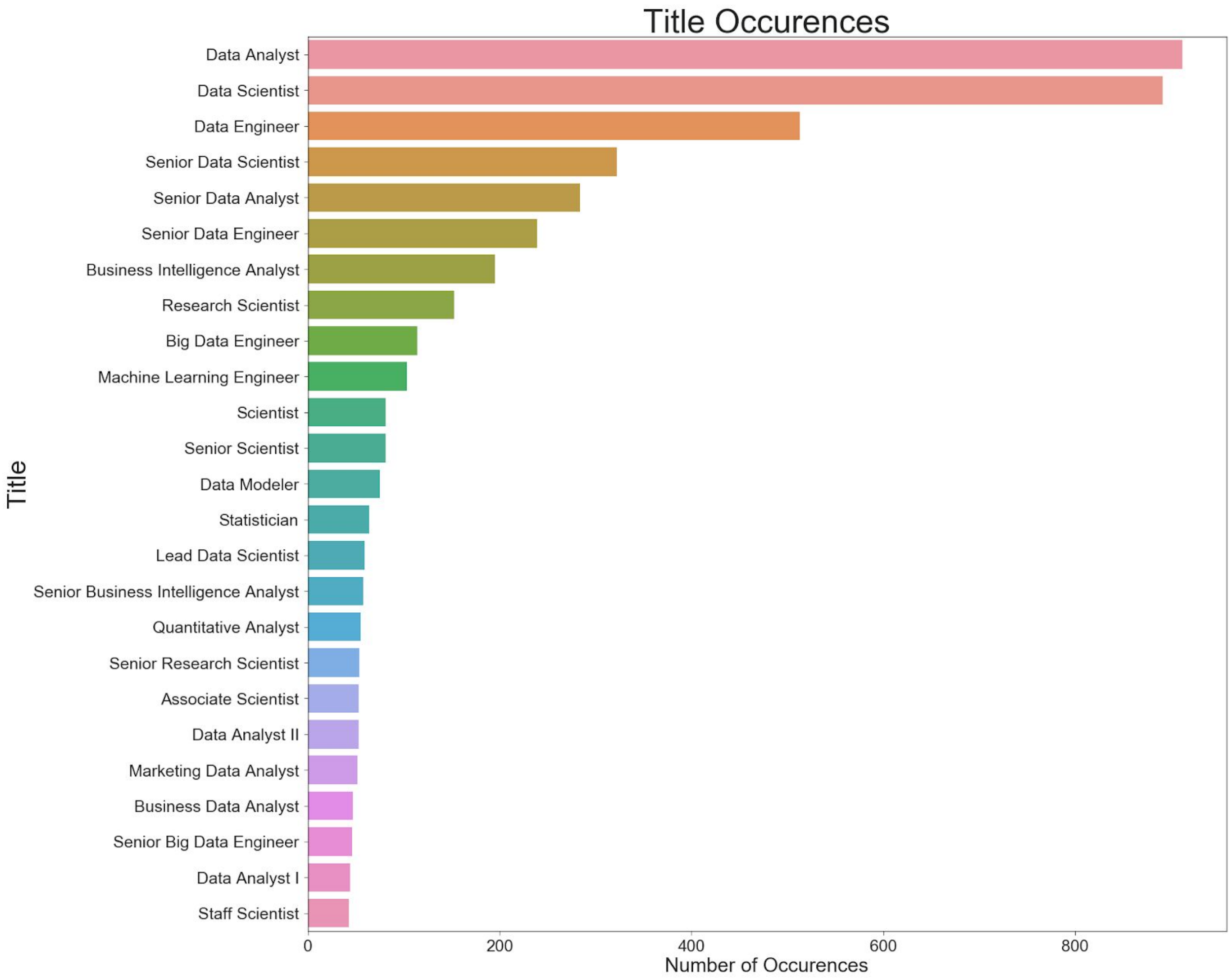
2. Pavoni, Silvia. “The Rise of the Data Scientist.” The Banker, The Banker, www.thebanker.com/Banking-Regulation-Risk/Private-Banking/The-rise-of-the-data-scientist.
3. Kelly, Meghan. “Data Scientists Needed: Why This Career Is Exploding Right Now.” VentureBeat, VentureBeat, 11 Nov. 2013, venturebeat.com/2013/11/11/data-scientists-needed/.
4. “Data Scientist: The Sexiest Job of the 21st Century.” States News Service, 1 Oct. 2012, www.highbeam.com/doc/1G1-303903551.html?refid=easy_hf.
5. Kishore. “ Quantifying the Current Demand for Data Scientists.” Deeplearningtrack - Online Data Science School, 27 Aug. 2017, www.deeplearningtrack.com/single-post/2017/08/27/Quantifying-the-current-demand-for-data-scientists.
6. Short, Eva. “What Can Be Done about the Data Science Skills Gap?” Silicon Republic, 13 Mar. 2018, www.siliconrepublic.com/careers/data-science-skills-gap.
7. Quora. “What's The Best Path To Becoming A Data Scientist?” Forbes, Forbes Magazine, 20 Jan. 2017, www.forbes.com/sites/quora/2017/01/20/whats-the-best-path-to-becoming-a-data-scientist/#208516a237d2.
8. Choudhary, Ankit, et al. “The Most Comprehensive Data Science Learning Plan for 2017.” Analytics Vidhya, 4 Apr. 2018, www.analyticsvidhya.com/blog/2017/01/the-most-comprehensive-data-science-learning-plan-for-2017/.
9. Stern, Dan. “Teach Yourself Data Science: the Learning Path I Used to Get an Analytics Job at Jet.com.” FreeCodeCamp, FreeCodeCamp, 12 Nov. 2017, medium.freecodecamp.org/a-path-for-you-to-learn-analytics-and-data-skills-bd48ccde7325.
10. Ponnambalam, Kumaran. “Popular Software Skills in Data Science Job Postings.” Big Data Science Practice, 21 Nov. 2014, kumaranpm.blogspot.com/2014/11/popular-software-skills-in-data-science.html.
11. Steinweg-Woods, Jesse. Web Scraping Indeed for Key Data Science Job Skills. 17 Mar. 2015, jessesw.com/Data-Science-Skills/.
12. De Lazzari, Diego. “Landing My Dream Job by Scraping Glassdoor.com.” NYC Data Science Academy Blog, 21 Aug. 2016, nycdatascience.com/blog/student-works/web-scraping/glassdoor-web-scraping/.
13. 2016 DATA SCIENCE REPORT. CrowdFlower, 2016, www.bing.com/cr?IG=0F5CD7A29E384FFE86698B62EE748519&CID=00C737B780F86F2C217F3C5A81576E66&rd=1&h=OcGf1kuo2hMxNjqFXqFmnghzwHnmc5AupS2rcQZrvWQ&v=1&r=http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf&p=DevEx.LB.1,5530.1.
14. “Can I Become a Data Scientist: Research into 1,001 Data Scientist Profiles.” 365 Data Science, 17 Apr. 2018, 365datascience.com/research-into-1001-data-scientist-profiles/#10.

15. Junco, Pablo Ruiz, and Andrew Chamberlain. “Data Scientist Personas: What Skills Do They Have and How Much Do They Make?” Glassdoor, 11 Oct. 2017, www.glassdoor.com/research/data-scientist-personas/.
16. Theuwissen, Martijn. “R Vs Python for Data Science: The Winner Is ….” KDnuggets, May 2015, www.kdnuggets.com/2015/05/r-vs-python-data-science.html.
17. Lee, Cheng Han. “How to Choose Between Learning Python or R First.” Udacity, 12 Jan. 2015, blog.udacity.com/2015/01/python-vs-r-learn-first.html.
18. “What Are Salary Estimates in Job Listings?” Glassdoor, 8 May 2018, help.glassdoor.com/article/What-are-Salary-Estimates-in-Job-Listings.
19. Metz, Cade. “GOOGLE OPEN-SOURCING TENSORFLOW SHOWS AI'S FUTURE IS DATA.” Wired, Conde Nast, 16 Nov. 2015, www.wired.com/2015/11/google-open-sourcing-tensorflow-shows-ais-future-is-data-not-code/.
20. “How Is Big Data Becoming Bigger in Today's World? [Infographic].” Social Marketing Fella, 15 Mar. 2017, socialmarketingfella.com/big-data-becoming-bigger-todays-world-infographic/.
21. Keys, Jessica. “Is SQL Hard to Learn?” Study.com, Study.com, study.com/academy/popular/is-sql-hard-to-learn.html.

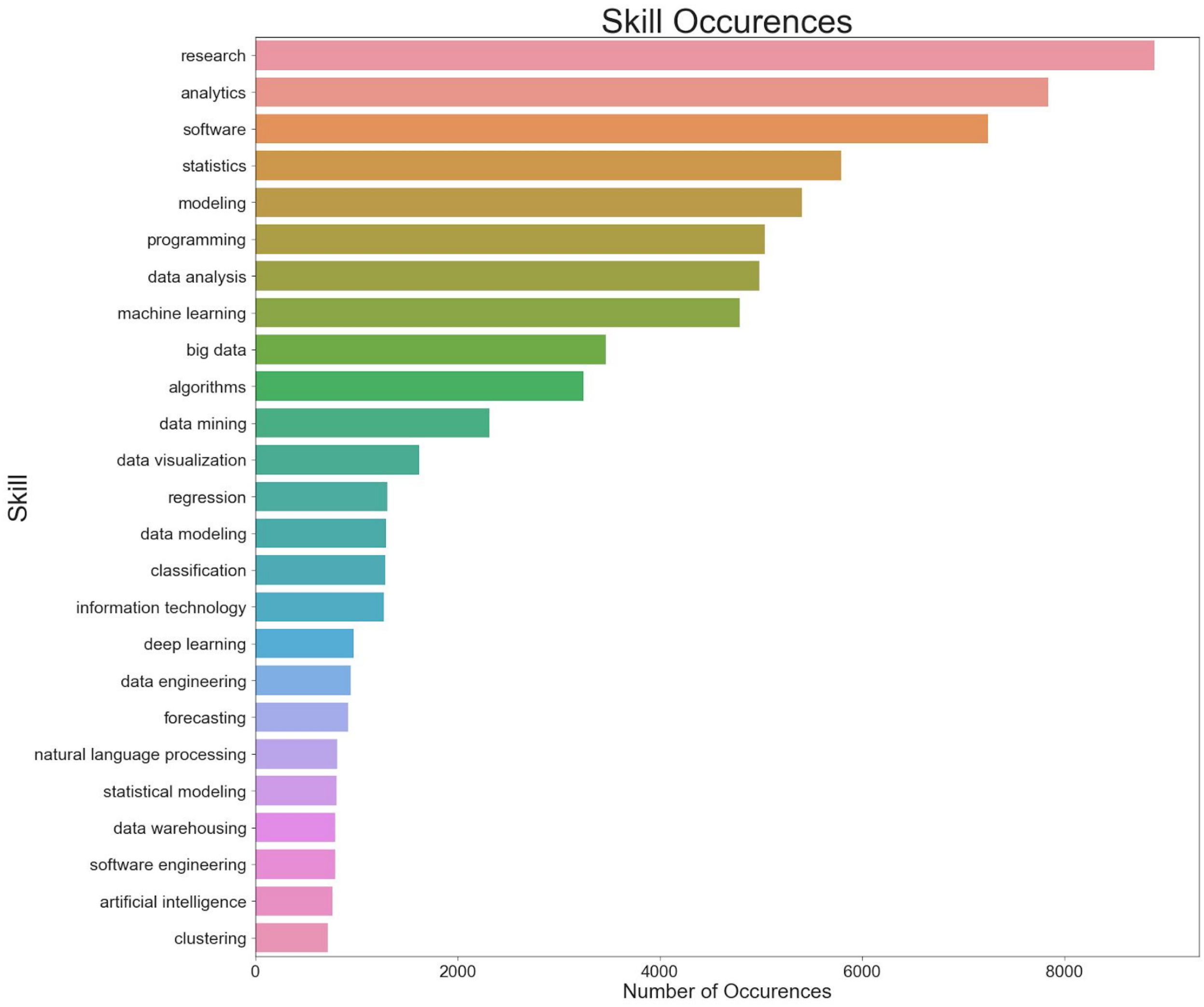
Supplementary Material



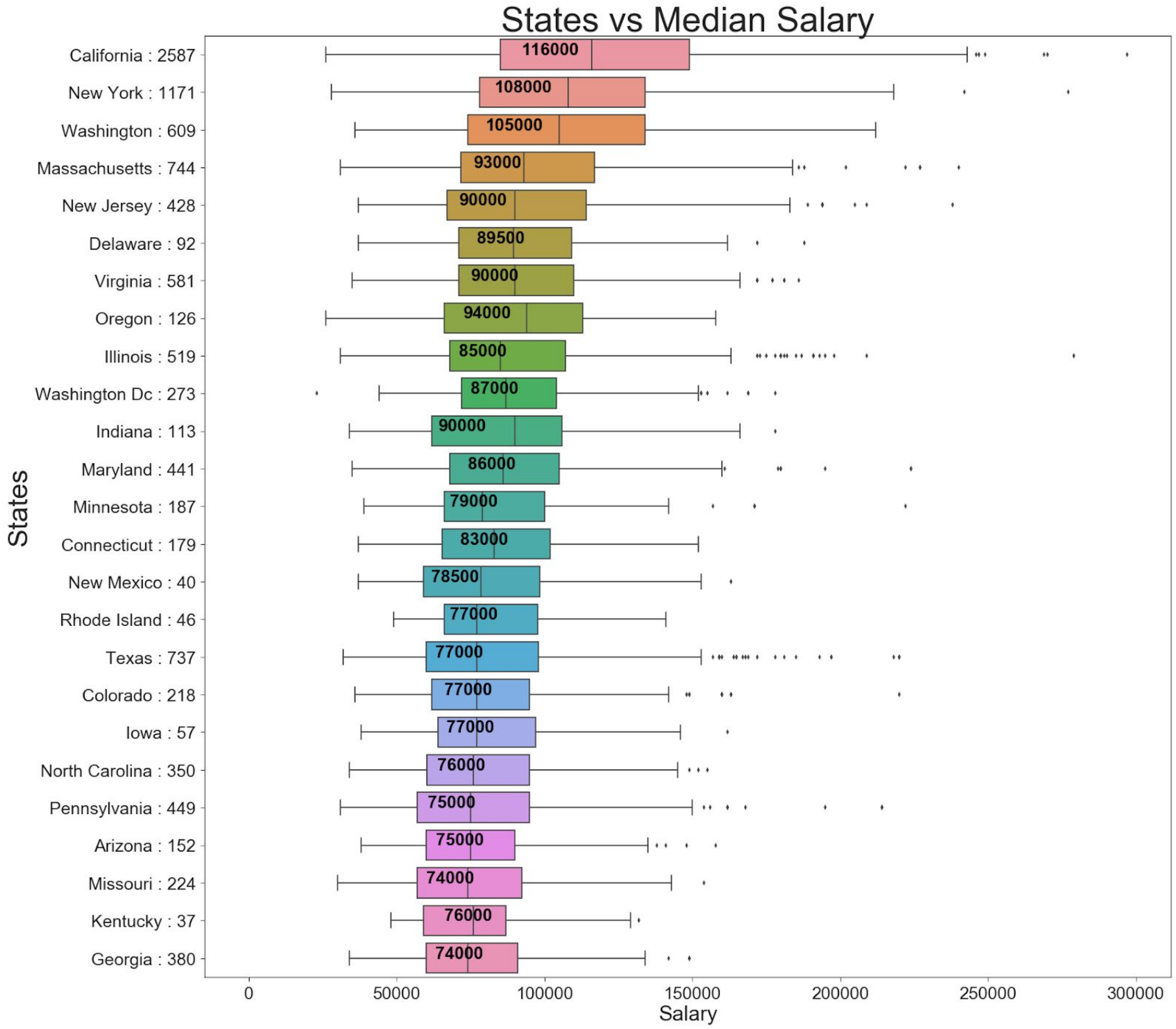
Supplement 1: The top 25 states for the number of Data Science jobs



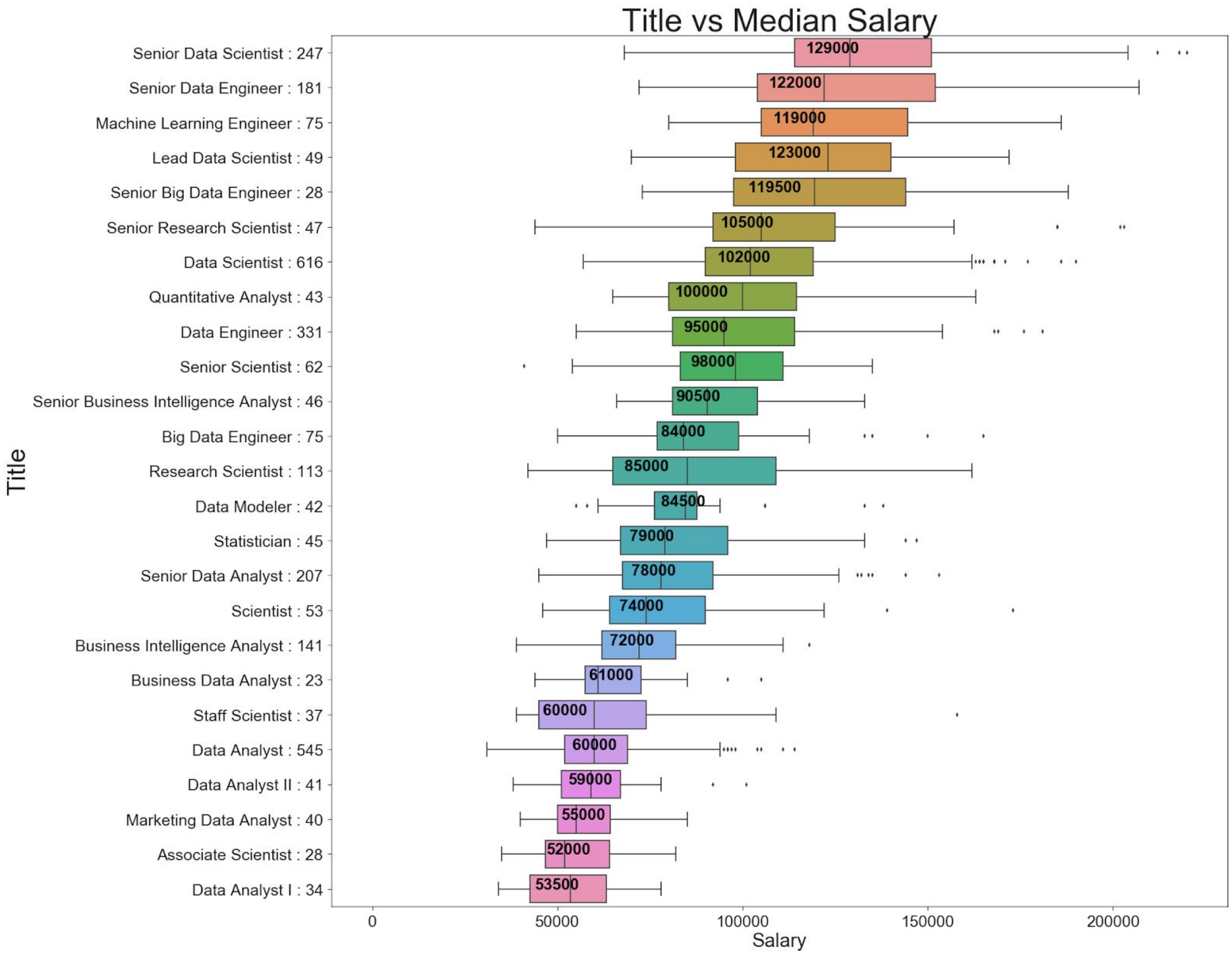
Supplement 2: The 25 most common Data Science job titles



Supplement 3: The 25 most common Data Science skills



Supplement 4: Boxplots of the salaries for the top 25 states. The number inside the box is the median of all salaries listed on Glassdoor, the number next to the state name is the number of job postings for that state with a listed salary.



Supplement 5: Boxplots of the salaries for the top 25 job titles. The number inside the box is the median of all salaries listed on Glassdoor, the number next to the job title is the number of job postings for that title with a listed salary.