Insights on Data Science Jobs

1st Darshit Jagetiya   
*Dept.of Applied Science*  
*San Jose State University)*San Jose,California  
*darshitkumaromprakash*.jagetiya@sjsu.edu  
  
 2nd Rui Xu  
*Dept.of Applied Science*  
*San Jose State University*San Jose,California  
rui.xu01@sjsu.edu  
  
 3rd Pin-Hsing Lai  
*Dept.of Applied Science*  
*San Jose State University)*San Jose,California  
*pin-hsing.lai@sjsu.edu*  
  
l 4th Laxmi Sindhu Surnam  
*Dept.of Applied Science*  
*San Jose State University*San Jose,California  
*laxmisindhu.samudrala@sjsu.edu*

*Abstract*—As the internet becomes highly available and as technologies are improving rapidly, we are creating a huge number of data each day by each person. To meet the needs of professionals who have the abilities to process and analyze these data, many data-related jobs have emerged and increased significantly. The data-related jobs usually include four main types of jobs, including data analyst, business analyst, data scientist, and data engineer. In this paper, Tableau is used to create visualizations with insights about data-related jobs. We have investigated skills, salaries, locations, job levels, and industries for each of the jobs. A final dashboard has been created to summarize important insights we found from the data. The aim of our research is not only finding insights about data jobs, but also helping people who are struggling in choosing among different kinds of data jobs and to find the right job for them.

Keywords—Data scientist, Data Engineer, Business Analyst, Data Analyst

# Introduction

In this modern era, every device you use generates data. With more devices being launched every day and more of the world’s population learning these technologies, the user generated data is constantly increasing. To get these data in an usable format, professions such as Data Analyst, Business Analyst, Data Scientist, and Data Engineers have emerged. These professions are not emerging only in the Information Technology (IT) industries, but in various industry verticals such as Healthcare sector, Financial & Insurance Institutions, Business Services, Government services, and Manufacturing Industries. There are different tasks involved in utilizing the data generated in each of these domains that is further used for their business improvement and customer satisfaction. Each of these tasks are done by different professions and each one has its own skills requirements. As there are different skill requirements, there is a noticeable difference in the pay grade. To analyze all the discussed aspects above, we have created a Tableau Dashboard that shows the number of job openings in various professions of data science and openings at different locations in the United States, Industry Sectors using Data Science, skills requirements for these jobs, and the pay grade for various positions.

To create a Dashboard which gives current insights, a large dataset with all the required features is required. If a dataset is old, it would not show the recent changes in the job scenario. Thus, the various challenges to create a meaningful dashboard becomes: (1) Scrape the current job openings data for different locations and for all the four positions from the web, (2) Extract required features from the job descriptions, (3) Perform data wrangling and make it ready to use for Tableau, and (4) Create various Tableau visual charts to give insights. When all these challenges are solved, the result is a Tableau Dashboard that gives information about the various positions that is helpful for a person in selecting a job type that aligns with the skills, location, and interests.

# Related Work

As the world is becoming digitalized, a lot of data is being generated and these data are being used for analytical purposes. So, these data are very crucial in all aspects. Nowadays, in order to transfer money, it just requires once click to transfer from one account to the other. If we want to do any shopping or order anything, just one click is sufficient. Data is generated from every click and these data can be used by data analysts and data scientists to analyze the data and provide recommendations based on the search criteria. In this digital era, there are many different data-related jobs available, such as data scientist, data engineer, business analyst, and data scientists. There are many different papers which discuss the skills required for these kinds of jobs. Cegieelski et al. [1] presented comprehensive research about the BDA knowledge, skills, and abilities by analyzing many undergraduate business analytics programs. This research identifies, using expert ranking of those programs, the important components of knowledge, skills, and ability in three main knowledge domains required for the BDA profession, i.e.:

* Business domain – includes communication skills, project management, client orientation, time management, etc.
* Analytical domain – incorporates problem definition and problems solving skills, predictive analysis, integrative analysis, and some others.
* Technical domain – is divided into three subgroups:

Applications – Excel, SAS, Tableau, etc.

Languages – R, Python, SQL, and Java.

* Infrastructure – Hadoop, Casandra, Oracle, Linux, MapReduce, Hive, and Pig.

Then there was a continuation for this paper which shows the skills required to data scientists. Mills et al. [2] adopted four pillars of skills in educating data scientists including:

* Data preprocessing, storage, and retrieval – e.g., NoSQL, data modeling, and data warehousing skills.
* Data exploration – statistical analysis and visualizations.
* Analytical models and algorithms – machine learning, data mining, and natural language processing.
* Data product – data and information organization, knowledge representation, and application development.

Another research paper on identifying DS skills [3] presents the DS framework as a combination of four components:

* Big Data Infrastructure – includes big data technologies such as Hadoop ecosystems, NoSQL databases, in-memory computing and cloud computing.
* Big Data Analytics Lifecycle – covers all stages of data analysis, data understanding, data preparation, and integration, model-building, evaluation, deployment, and monitoring.
* Data Management Skills – incorporate traditional data modeling and relational database knowledge.
* Behavioral disciplines – contain soft skills related to people and business such as abilities to think critically, to communicate with domain experts, and to make project outcomes relevant to business.

Overall, the papers were explaining that there are no distinguishing factors between the group of skills required for BDA and DS. Those papers were predominately based on the analyses of the academic programs and surveys of big data professionals. In this paper,we collected data from indeed (https://www.indeed.com/) by scrapping the pages and preprocessing descriptions to fetch the skills required for a particular job, salary, and location. In addition, some insightful visualizations are made with these data.

# Methods

In this paper, we will discuss the following aspects:

* The location of jobs.
* Skills requirements.
* Salary estimates.

For these purposes, we scraped data from Indeed (https://www.indeed.com/) in September using crawler. The keywords we used are ‘data scientist’, ‘data analyst’, ‘data engineer’, and ‘business analyst’.

We merged all the datasets and did data cleaning. We drop some columns and drop the rows that do not have Job\_Description. The job description is the most important feature and contains much valuable information. After data cleaning, we have 62,734 records. The main columns are Job\_Title,Company\_Name,Job\_Description, Salary\_Estimate, Industry, Sector, Location. Then we extract information from the Job\_Title,Job\_Description, Salary\_Estimate, and Location. Before extracting, we use python to do a word count and create a word cloud based on the Job\_description (see Fig. 1). From the word cloud, we have a better understanding of the job requirements. Based on the skills and knowledge required for those jobs, we choose our targets of extracting.

1. Word Cloud of Job\_Description
2. Extracting information

We use Regular Expression to extract information on Python. From the Job\_Title, we can find some letters, like jr, sr, II, and III, representing the job level. Salay\_Estimate is the salary range of the job, for example, 56K - 89K for a year. Most are annual salary, but some are monthly salary or hourly salary. So we extract and calculate the mean annual salary.

1. Extracting information

After Extracting information, we have:

Job\_Title - the title of a job position, some contain position level.

Job\_Level - position level is extracted from the Job\_Title, e.g., junior, senior, and lead.

category - the category of the job extracted from Job\_Title. The values of category are DS, DE, DA, BA, and other, representing Data Scientist, Data Engineer, Data Analyst, Business Analyst, and Other, respectively.

Company\_Name - the name of the company.

Location - the location of this job contains the state and city. Some of the jobs are remote.

State - the state of the job position, extracted from the Location.

City - the city of the job position, extracted from the Location.

Industry - an industry of the job or the company, e.g., IT Services, Healthcare Services & Hospitals, Staffing & Outsourcing, Computer Hardware & Software, etc.

Sector - sector of the job or the company, e.g., Information Technology, Business Service, Finance, Retail, etc.

sel\_mean - The mean annual salary is extracted from the Salary\_Estimate. The value is integer and the unit is thousand dollars.

Work\_experence - the number of years of work experience required for the job, integer.

Flag\_bachelor, Flag\_master, Flag\_phd - the education level required for the job. 1 means this education level is required for the job, 0 means not required.

Flag\_python, Flag\_sql, Flag\_ml, Flag\_r, Flag\_visualization, Flag\_hadoop, Flag\_big\_data, Flag\_java, Flag\_tableau, Flag\_nlp, Flag\_database, Flag\_datamining, Flag\_communication\_skill, Flag\_statistics, Flag\_problemsolving, Flag\_aws, Flag\_Databricks, Flag\_snowflake - the required skills and softwares for the job. The values are 1 and 0. 1 means it is required, and 0 means the skill is not required for this job.

From the above Dataset information, we can see the records miss some features. Some records don’t have Industry, Sector, sal\_mean, and those jobs don’t post this information on job posting websites.

We use Tableau to do the visualization and tell the story about how to find a data-related job.

* n something that alternates).

# Results

We have created ten visualizations in total. We will be explaining what we did for each visualization.

For the first visualization (Fig. 4), we created a pie chart showing the percentage of job levels for data related jobs. As we can see, about 70 percent of data-related jobs are at a senior level while the percentage of jobs at junior and lead level are about the same.

1. Pie Chart on Job Levels

Fig. 5 is showing the annual salary in thousands for different types of jobs. We can see that data scientists have the highest annual salary compared to all other kinds of jobs. The mean annual salary is about $116,000 annually for data scientists. Data engineering has the second highest annual salary, which is about $101,000 a year. Data analysts’ jobs and business analysts’ jobs have comparatively lower annual salaries compared to salaries of data engineers and data scientists, which is about $80,000 a year.

1. Annual Salary Table

Next, Fig. 6 shows the distribution of years of work experiences required by employers. From the bar chart, we can clearly see that the majority of data jobs require about 2 to 5 years of work experiences while most jobs require about 2 to 3 years of work experiences. All other years of work experiences are about the same. There are also a lot of jobs that require more than 10 years of experience, which should be for jobs at higher levels.

1. Donut chart on Work Experience

The next, Fig. 7, is also a bar chart. This chart is showing the percentage of minimum education required by employers. About 57 percent of the data jobs require only a bachelor's degree while about 22 percent of jobs require a master’s degree. Only a small portion of data jobs require a Phd degree, so it should be pretty sufficient to have a master’s degree to meet most of the jobs’ minimum requirement for data related jobs.

1. Bar chart on Minimum Education Requirements

This (Fig. 8) is a bar chart that shows the tools and skills that are required by data-related jobs. When looking at all of the data jobs, we can see that almost half of the jobs require knowledge of SQL. And the second most popular skill is communication skills. And the third most popular skill required by employers is database skill, which is a broader knowledge of SQL. So, we can see that database knowledge and SQL skill is pretty important to data jobs. We could also filter the chart by different job categories to see whether the result is significantly different from results of all the jobs.

1. Tools and Skills required

Fig. 9 shows the tools or technologies that are required by employers for data engineering’s jobs while Figure 10 shows the tools or technologies required by employers for data analyst’s jobs. From these two charts, we can see that SQL, Python, communication skills, and database knowledge are still the most popular tools that are required by employers. However, Java and Big data knowledge are more popular for data engineering jobs while Tableau, machine learning, and statistics knowledge are more important for business analyst’s jobs.

1. Tools and Technologies for DE
2. Tools and Technologies for BA

The map (Fig.11) is a combination of choropleth and proportional maps below shows the distribution of data related jobs in different states in the United States. It clearly shows that California has the most job vacancies. Texas has the second most jobs available with New York and Illinois are in the third and fourth place, respectively. The blue dots on the map show the number of jobs in the city. We can see that there is a big dot in New York, Illinois, and Washington state. It means that a specific city in each of these states has a comparatively high number of data related jobs compared to other cities in the United States. We also had a filter to see whether the results differ a lot for different job categories, but the proportion of the number of jobs in each state are about the same for each category.

1. Choropleth and Proportional Map

The following (Figure 12) is a packed bubble chart that shows the top ten sectors for data-related jobs. The bigger bubble represents a higher number of jobs. We can see that most jobs are in the information technology or business services sector. The third, fourth, and fifth sectors are finance, biotech and pharmaceuticals, and health care, respectively.

1. Bubble chart on Sectors

|  |
| --- |
|  |

1. Tree Map on Average Salary Industry wise

The last but not least, Fig. 13, is a treemap that shows the average salary offered in different industries in thousands. The bigger area the industry is, the higher number of jobs that is available in that industry. The darker the color is, the higher the mean salary is. Salaries range from 56k to 140.5k. We can see that the Internet industry has the highest mean salary among the top industries. While IT service has the largest number of data-related jobs available, its mean salary is just at around 50th percentile for mean salaries for all data-related jobs.

# Discussion

The dashboard gives insights into various job information that a person seeking a job in data science and big data analytics might want to know. This information will help a person select a position that matches their skills, interests, and location requirements. It shows the salary estimates according to the different data science positions and describes which industrial sectors have the maximum openings. The education requirements for jobs is also visualized. Along with the states, one can even find job openings at each city in the United States. The work experience pie chart gives information on openings for professionals with different years of experience. Looking at the dashboard, data science aspirants can also decide what skills they need to focus on. Overall, this dashboard is useful for students as well as professionals.

# Futrue Work

As this data is static, we would like to build a platform that can fetch the real time data from the web and show live visualization about various openings. This also enables opportunity to show the trend in various skill requirements over time giving insights for the rising programming languages and the tools preferred by the industries. If we have past job data, we can discuss what will influence the job market. For example, before the COVID 19, there were few or even no remote jobs. Now, we can find a great number of remote jobs. Thus, future work will be creating a live visualization that gives time- series analysis and shows the trends over time along with the visualization shown in this dashboard.

# What did you learn

We have learned the way to choose the appropriate chart to help audiences to understand better about what we want to show. By doing this project, we have a deeper insight into data visualization “start with data and end with the user” [4]. As a graduate student majoring in data analytics, finding a job is an inevitable step. So we know what a user looks for from a job dashboard. We also learned and gained a deep understanding of working with Tableau and its tools. Showing the data in visual charts which is easily understood and covers all the aspects is a key requirement for a good dashboard and this project has given us all those skills.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. Cegieelski, C.G. and Jones-Farmer, L.A., “Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study”, Decision Science Journal of Innovative Education, 14(1), 2016, 91-118.
2. Mills, R., Chudoba, K.M. and Olsen, D.H., “IS Programs Responding to Industry Demands for Data Scientists: A Comparison between 2011 – 2016”, Journal of Information Systems Education, 27(2), 2016, 131-140.
3. Song, I.Y. and Zhu, Y., “Big data and data science: what should we teach”, Expert Systems, 33(4), 2016, 364-373.
4. M. O. Ward, Interactive data visualization: foundations, techniques, and applications. CRC Press Inc, 2015.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

Appendix A

Dataset for Visualization

The dataset for this Tableau visualization includes various columns that might give insights into designing an informative dashboard. These columns and its datatype are shown in the image below:

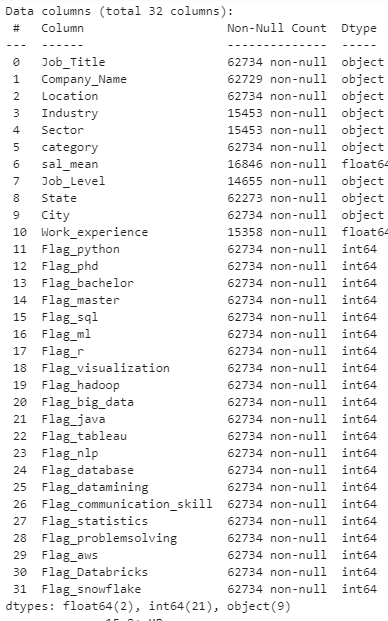


Fig. A 1 Columns and its Data Types

Appendix B

Tableau Dashboard for Insights on Data Science Jobs

