Navigating the Data Science Job Market: Insights for Statistics Majors

Jieying Ma, Feini Pek

I. Introduction:

While relatively new, data science has rapidly emerged as one of the fastest-growing professions within the statistical field, with employment opportunities for data scientists projected to increase by 35.2% over the next decade (Bureau of Labor Statistics). Consequently, data science has become one of the most sought-after professions in the industry. Acknowledging the significance of this field, our project embarks on a comprehensive study of job demand in data science, aiming to dissect the dynamics of this rapidly evolving job market. Through this endeavor, we seek to provide clarity and actionable insights beneficial to students, educators, and job seekers by addressing the following questions:

1). How has the demand for data science evolved over time?
2). Which organizations actively seek data science talent?
3). In which fields and sectors are data science skills highly sought after?
4). How do data science job opportunities differ across the United States?
5). What salary range can data scientists typically expect?

These insights will empower students and individuals interested in the field to align their skills effectively with market needs, make informed decisions about our career paths, and negotiate compensation with confidence. Through this project, we endeavor to bridge the gap between academia and industry, facilitating the transition of statistics majors into successful professionals in the dynamic realm of data science.

II. Data Acquisition and Processing

1. Finding data sources

With the interest in studying job demand within the field of data science, and recognizing the significance of obtaining accurate and up-to-date information, we turned to prominent job search platforms such as Handshake, Indeed, Glassdoor, and Simply Hired. These platforms are widely recognized for their extensive databases of job postings, encompassing a diverse range of industries, job roles, and geographic locations. Leveraging the vast repository of listings available on these platforms, we managed to gather a robust dataset that would enable us to analyze trends, identify patterns, and draw meaningful insights regarding job demand in the dynamic field of data science. Namely, we were able to extract job titles, company name, location, salary from Indeed, company information from Glassdoor, and job descriptions from Simply Hired. However, the process of getting the data was not without challenges. We learned that obtaining data from job posting websites poses challenges due to access restrictions, limitations on data scraping techniques, dynamic website structures, bot detection measures, rate limiting mechanisms, and legal and ethical considerations. These platforms often restrict access to protect user privacy and may block automated scraping attempts, making it difficult to gather large volumes of data. We will explore these challenges in the following section.

2. Accessing the data

When scraping information, we mainly employed two methods simultaneously: web requests and selenium.webdriver. In our case, most job search websites utilize a JavaScript-rendered network structure. Typically, the left side of the webpage displays basic information cards for each job, while the right side provides detailed job information. Due to this structure, the URL remains unchanged when navigating between pages, necessitating the use of tools like selenium.webdriver to simulate real browser behavior, such as clicking buttons to paginate. This design also aids in preventing web scraping by implementing anti-bot measures.

i). Handshake

While scraping Handshake, several challenges were encountered. Firstly, the process required manual login, which added complexity and increased the time required for scraping. Additionally, extracting detailed information from the website's right sidebar necessitated clicking the "More" button multiple times to load all the text, thereby extending the scraping process. Despite efforts to simulate human-like behavior by adjusting operation time intervals and incorporating random pauses, the website's anti-scraping mechanisms remained formidable, often detecting and thwarting scraping attempts. Furthermore, the dynamic nature of the website's content, influenced by anti-scraping measures, posed difficulties as it could alter the URLs' content, resulting in discrepancies between the HTML retrieved and the content displayed in the WebDriver window. These challenges collectively interrupted the scraping loop, requiring further adjustments and strategies to overcome. After a week of trial, we decided to move to another job search engine.

ii). Indeed

Our next choice of source is Indeed.com, which does not require logging in. The webpage structure of Indeed resembles that of Handshake, with 15 job cards per page. We observed a systematic change in the URL when paginating, where the parameter named "start" increases by 10 with each page turn. Therefore, we chose to use requests for data retrieval. However, due to the inability to click on job cards to switch the detailed content on the right side, we first extracted the URLs of each job's detail page and basic information such as job title, company name, location, and salary from the cards. During data retrieval, we encountered the need to add additional content to the header and observed that the cookie expires approximately every ten minutes, necessitating periodic updates. Consequently, we opted for segmented data retrieval, refreshing the cookie after every 200 pages. However, after obtaining basic information and URLs for over 6000 jobs, a new button appeared for robot verification, possibly due to our extensive visits being flagged as abnormal. Thus, we attempted to utilize WebDriver window, but the robot verification continued to impede progress. Despite trying various strategies such as switching IP addresses using different Wi-Fi networks, we were unable to bypass the verification. Consequently, we proceed our analysis with our existing data.

iii) Glassdoor

Due to being blocked by Indeed, we turned to another login-free job portal, Glassdoor. This website provides additional data about the companies offering the positions, such as industry and company size. Since the website does not alter the URL when fetching more job cards, we utilized WebDriver window. While crawling this webpage, we encountered a challenge where clicking buttons would randomly trigger pop-ups for registration and promotions. To address this, we implemented an if statement to detect and close any pop-ups that appeared after each click. Additionally, unlike other platforms, Glassdoor doesn't paginate to display more job cards; instead, it utilizes a "show more" button to load additional job cards. After clicking this button 24 times, the website ceased to load more job cards. Nevertheless, we successfully gathered ample information on companies recruiting data scientists.

iv). SimplyHired

To facilitate text analysis of job details, we opted to use SimplyHired over Glassdoor due to the need for fewer clicks to access job details and better content quality. Initially, we employed a WebDriver window to gather detailed information pages for 2600 job positions. Subsequently, we utilized requests to extract the job descriptions. By employing appropriate headers and time intervals, the requests were highly successful and efficient. Moreover, SimplyHired offers qualifications tags, which provide additional insights for analysis purposes. Overall, this approach enabled streamlined data acquisition and enhanced efficiency in obtaining job details for analysis.

3. Processing the data

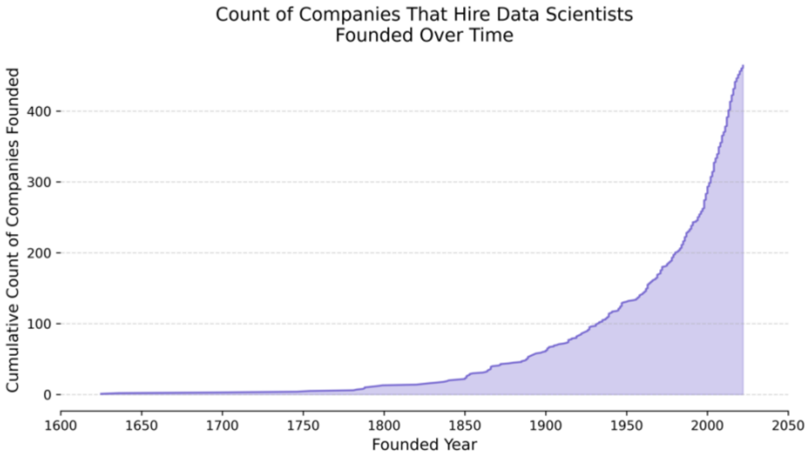
The data retrieved from Indeed, stored in a JSON format file, initially presented a challenge due to its unstructured nature. Each page of job listings, represented by 440 items in the dictionary, appeared as a wall of text, requiring further processing to extract pertinent information. Leveraging techniques learned in STA 220 course, we systematically parsed through the raw data to identify and extract relevant elements. We utilized HTML parsing techniques using beautiful soup to navigate through the nested structure of the data and to extract elements by identifying class tag that contains our variable of interest, then we utilized regular expressions to parse the text and extract key information such as job titles, company names, locations, salaries, and job types. Through meticulous examination and data cleaning, we successfully constructed a structured dataset comprising a total of 6,591 job listings. This dataset includes five primary variables: job title, company, location, salary, and job type. Each variable was carefully extracted and formatted to ensure consistency and accuracy across all entries.

Due to Glassdoor's data structure, where each parameter corresponds to a quantity, we opted to utilize DataFrames for storing the data. This approach facilitates easy browsing and statistical analysis. When analyzing company data, it was necessary to filter out duplicate company entries. Additionally, DataFrames offer convenient methods for handling missing values and removing duplicate rows within a column. For instance, in the "company size" column, the "employees" descriptor does not need to be repeated for each value, so we removed it. Moreover, when examining the count of samples in each category, missing values often affect calculations, which can be efficiently handled using the Pandas library.

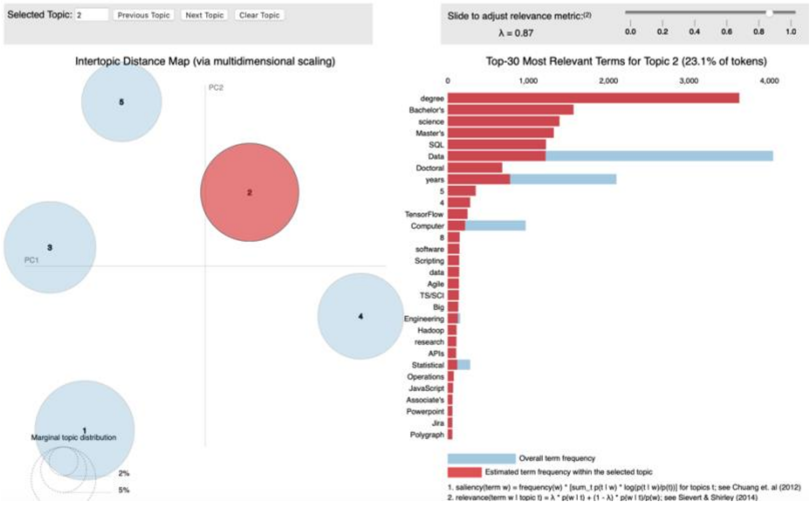
When extracting text-based information, Beautiful Soup was employed to segment paragraphs based on indicators such as newline characters, allowing us to retrieve the content corresponding to each parameter. In the case of data obtained from SimplyHired, where one parameter (e.g., benefits and qualifications) may have multiple tags, we used dictionaries to store lists of values. This approach facilitated efficient storage and retrieval of information for subsequent analysis.

III. Exploring and visualizing the data

1. Demand in data science



In the plot above we want to observe the job demand of data scientist by understanding the increasing number of hiring companies. The plot shows founded year of the companies on the x-axis vs the number of companies founded during that year. Note that the companies we are considering here are those who hires data science roles. Notice that the frequency count per year demonstrates exponential growth, which implies a significant and accelerating demand for data science talent for the past decades. This trend indicates a growing recognition of the value of data science in various industries and sectors. Moreover, it implies that data science has evolved from being a niche field to a mainstream requirement for businesses seeking to leverage data-driven insights for strategic decision-making and innovation. This growth also suggests that a wide range of employment opportunities are available in the field, spanning across diverse industries such as technology, finance, healthcare, e-commerce, and more as we will observe in the next section. This allows opportunities for statisticians to collaborate across various disciplines.

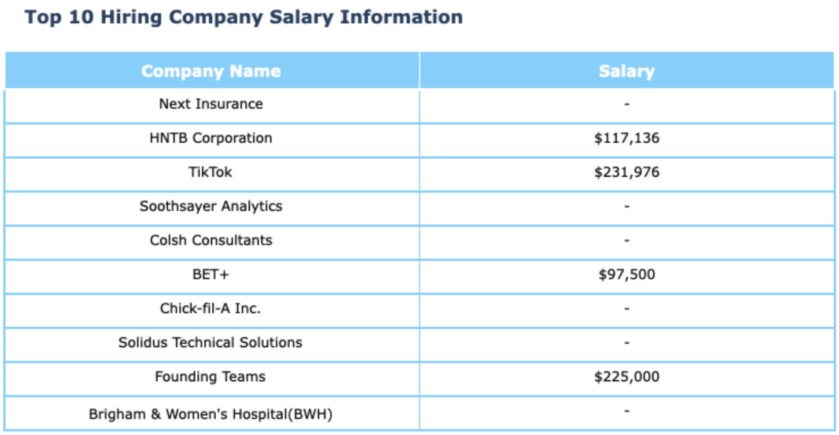


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Interactive plot : here

Through LDA analysis, we further inspect the job descriptions to categorize and identify the requisite skills. Across educational qualifications, a majority of positions necessitate Bachelor’s and master’s degrees. Proficiency in statistical software, particularly R and Tableau, is highly sought after. Additionally, familiarity with databases and object-oriented programming languages like Python and C++ are deemed essential for data scientist roles, addressing inquiries about the most desired skills in this field.

2. Company Hiring Trends



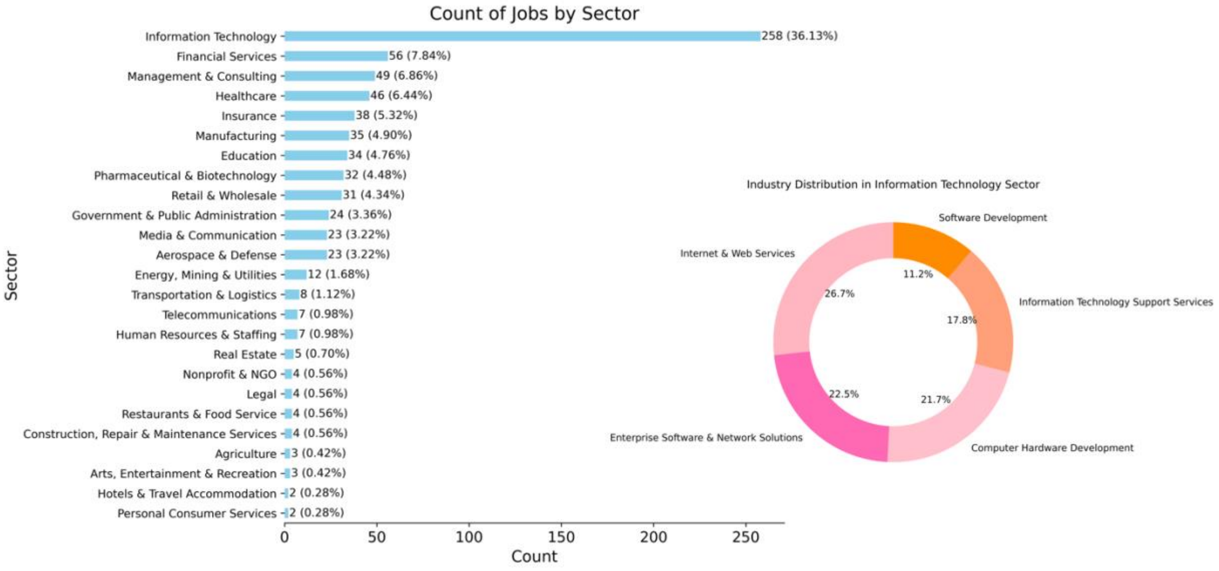
Top 10 Hiring Company Salary Information

Company Name	Salary
Next Insurance	-
HNTB Corporation	$117,136
TikTok	$231,976
Soothsayer Analytics	-
Colsh Consultants	-
BET+	$97,500
Chick-fil-A Inc.	-
Solidus Technical Solutions	-
Founding Teams	$225,000
Brigham & Women's Hospital(BWH)	-

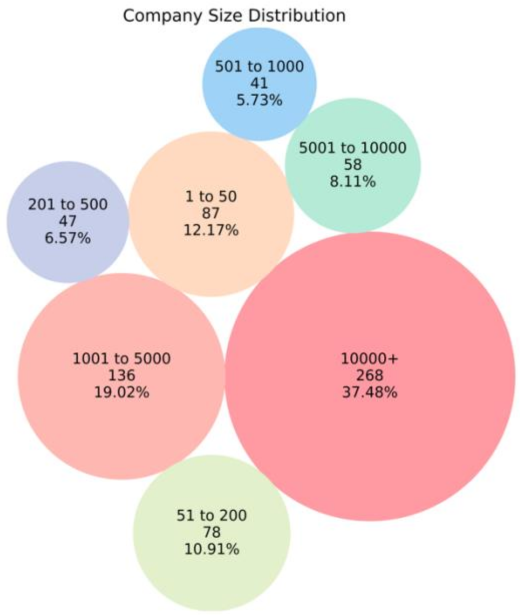
In this section, we want to identify the company hiring pattern by first looking at the top 10 companies among the 6591 job listings. Our analysis on these companies revealed that a diverse array of organizations actively seeking statistical talent across various industries. Notably, we have Next Insurance, a player in the insurance sector; HNTB Corporation, specializing in architecture, civil engineering, and construction management; TikTok and BET+, operating within the entertainment sphere; Chick-fil-A, a major player in the food business; Solidus Technical Solutions and Founding Teams, focusing on IT and artificial intelligence solutions; Brigham & Women's Hospital, an esteemed teaching hospital affiliated with Harvard Medical School; and Soothsayer Analytics and Colsh Consultants, representing consulting firms. While well-established corporations such as Next Insurance, HNTB Corporation, and Chick-fil-A demonstrate a consistent demand for statistical talent, startups like Solidus Technical Solutions and Founding Teams also play a significant role in data science. The mix of startups and established corporations suggests a dynamic job market, with opportunities for statistical professionals to contribute to a wide range of industries and sectors. However, as shown in the table above, many job listings lack salary information, making it challenging to compare these companies further. Despite this limitation, based on the available data, it appears that these positions are well compensated.

Additionally, we observed another pattern in the data by performing the keyword analysis. Using natural language processing, we were able to analyze job titles, and identified ‘data’, ’scientist’, ‘engineer’, ‘machine’, and ‘learning’ as the top 5 words across all titles. This finding highlights the importance on data-related tasks, machine learning techniques, and engineering principles within the job market for these roles, and showed a consistent demand for professionals proficient in these skills across various industries and sectors.

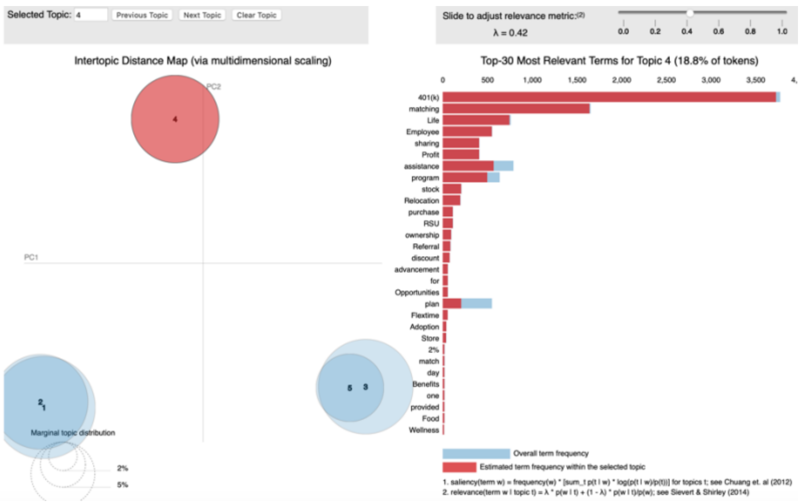
3. Company Information, Industry and Sector Preferences



Our investigation into industry and sector preferences reveals that Information Technology emerges as the dominant sector for hiring data scientists. Further exploration into the distribution within the Information Technology sector highlights notable participation across various industries, including Internet & Web Services, Software Development, IT Support Services, Computer Hardware Development, Enterprise Software, and Network Solutions. This indicates a widespread recognition and demand for statistical skills within the technology sector, encompassing diverse areas such as software development, hardware engineering, and network infrastructure. The consistent hiring activity across these industries underscores the importance of statistical expertise in driving data-driven decision-making and innovation within the technology landscape – a vastly growing field as well. This findings suggest a trend towards integrating statistical analysis and data-driven methodologies across a wide range of technology-related domains, reflecting the growing reliance on data-driven insights for strategic business operations and development.



Furthermore, the analysis on company data also showed that the majority of companies hiring for data scientist exhibit considerable size, with a significant portion boasting over 10,000 employees, followed closely by those in the 1001 to 5000 employee range. Companies with over 10,000 employees are generally considered large-scale enterprises, while those with 1001 to 5000 employees fall within the category of medium-sized companies. Both sizes suggest significant operational capacity and infrastructure, with larger companies often having more extensive resources and market presence compared to their medium-sized counterparts. These findings highlight the substantial operational capacity and infrastructure of these organizations, positioning them as major players within their respective industries. Importantly, larger companies are renowned for their established market presence, diversified revenue streams, robust infrastructures, and hence, all of which contribute to greater job security and confidence their position, as well as offering a range of benefits for their employees.

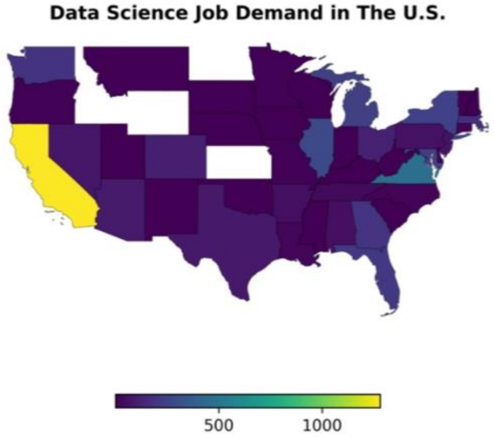


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Interactive plot: here

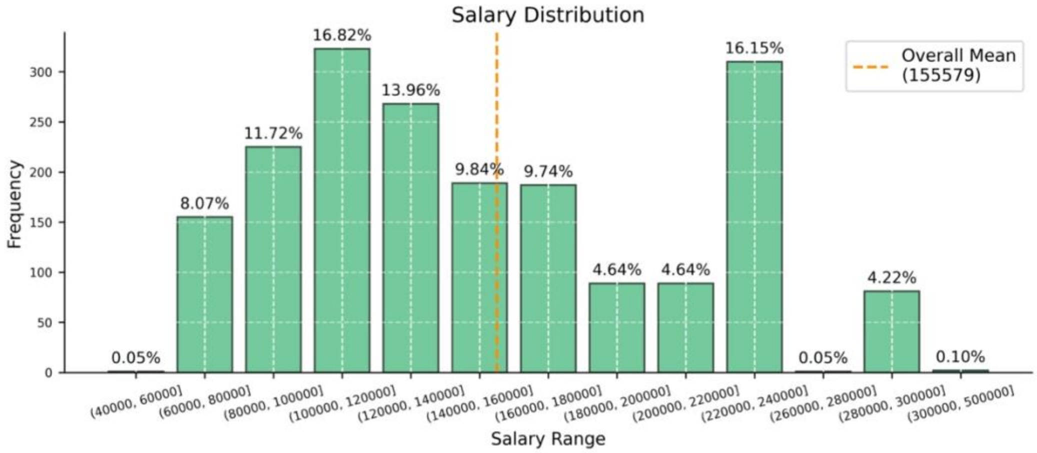
The above plot provides insights into the spectrum of benefits provided by companies, as revealed through job descriptions and another LDA. The plot illustrates distinct categories of benefits, including health, vacation, and parental, indicating the comprehensive nature of benefits packages associated with data science roles.

4. **Geographical distribution of data science roles in the US**



Next, we want to inspect the geographical distribution of data science roles. The plot above shows varying levels of demand for data science roles across different states in the United States. California (CA) emerges as the leading state with the highest frequency of job postings, indicating a robust demand for data science professionals in the tech hub of Silicon Valley. Virginia (VA) follows with a substantial number of job listings, suggesting a growing presence of data-driven industries in the state. Illinois (IL) and New York (NY) also demonstrate notable demand, likely driven by the presence of major metropolitan areas such as Chicago and New York City, respectively. Other states with significant job opportunities include Michigan (MI), Maryland (MD), Florida (FL), Georgia (GA), Washington (WA), and Massachusetts (MA). However, it's noteworthy that certain states such as Idaho, Wyoming, and Kansas show no job listings, indicating potential areas for growth or underrepresentation in the data science job market within those regions. Overall, the geographical distribution highlights the importance of urban centers and tech hubs in driving demand for data science roles, while also underscoring the need for further exploration and development of data science opportunities in other regions.

5. Salary Range Insights



To analyze the salary data, the above binning offers valuable insights into the expected salary range for statistical professionals. Notice that the distribution is roughly bimodal, where the majority of salaries fall within the range of $60,000 to $220,000, with one of the most frequent salary bands being between $100,000 and $140,000. This indicates that statistical professionals can anticipate competitive compensation, with a significant proportion earning salaries in the six-figure range. Notably, the mean salary across all salary bins is $155,579, further emphasizing the overall robustness of compensation within the field of statistics. Inspecting the first mode of range of $100,000 to $120,000, we found that the majority of positions are full-time roles located in California. The primary hiring employer within this range is HNTB Corporation. Similarly, inspecting the second mode, the salary range of $220,000 to $240,000, the majority of positions are full-time roles, with a significant presence in Illinois. However, there are significant amount of missing location data within this range. The primary hiring employer within this higher salary bracket is TikTok. The distribution also highlights the presence of outliers, particularly in the higher salary bands above $220,000, where salaries exceed the mean. These outliers suggest the potential for lucrative opportunities for experienced or specialized statisticians, possibly in roles requiring advanced skills or expertise.

IV. Conclusion

Data science is a growing field, with promised job security, good compensation, and extensive room to grow as a professional. Our study aims to shed light on the evolving job market of data science, offering actionable insights to guide individuals in making informed career decisions amidst the dynamic landscape of the field. The accelerating growth in demand for data scientists, as evidenced by the increasing number of founded companies, underscores the rising significance of data-driven roles in various industries. We found that sought-after skills for data scientist roles include educational qualifications such as Bachelor’s and master’s degrees, proficiency in statistical software like R and Tableau, familiarity with databases, and expertise in object-oriented programming languages such as Python and C++. Our analysis also reveals diverse hiring patterns across industries, with Information Technology emerging as a dominant sector for data scientist recruitment. Furthermore, the geographical distribution of job postings highlights the leading role of states like California in driving demand, while also indicating opportunities for growth in other regions. Importantly, our findings indicate that data scientists can expect competitive compensation, particularly as their experience increases, with salaries predominantly falling within the range of $60,000 to $220,000 and a notable frequency of salaries ranging from $100,000 to $140,000. As a future direction, this analysis could involve exploring the impact of emerging technologies on the demand for data science roles, investigating evolving job requirements and skill sets sought by employers, and conducting comparative analyses of data science job markets across regions. In conclusion, our analysis has answered some key aspects of starting a career in data science. We hope that these findings will aid individuals as they embark on their journey into the field of data science as a profession with its promising outlook.

Code can be found: here