Is Data Science still the sexist job of the 21st Century?

Michael Zelaya

Mzelaya@bellarmine.edu

2/05/2024

Introduction:

My data set is about salary trends in data-related careers. I found this data set on Kaggle (https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data/data). I chose this data set because I am interested in the rapidly evolving data field and want to explore further the future salaries of the various data-related careers. In this report, you can find further information regarding and understanding more about the dataset by including a variety of visualizations and tables in more detail.

Data Set Description:

Narrative summary: This data set contains 9,356 samples with nine columns, with most data being nominal and only two columns being interval data types.

Table 1: Data Types, Range of Values, and Missing Data

Variable Name	*Data Type*	*Range of Values*	*Missing Data (%)*
Work Year	Interval	[2020-2023]	0%
Job Category	Nominal	[Data Engineering, Data Architecture and Modeling, Data Science and Research, ML & AI, Data Analysis, Leadership and Management, BI and Visualization, Data Quality and Operations, Data Management and Strategy, and Cloud and Database	0%
Salary (USD)	Interval	[15k – 450k]	0%
Employee Residence	Nominal	Worldwide	0%
Experience Level	Nominal	[Entry-level, Mid-level, Senior, or Executive]	0%
Employment Type	Nominal	[Full-time, Part-time, Contract, or Freelance]	0%
Work Setting	Nominal	[Remote, In-person, or Hybrid]	0%
Company Location	Nominal	Worldwide	0%
Company Size	Nominal	[S, M, or L]	0%

Data Set Summary Statistics:

- **work_year:** This is the year it was recorded and is essential for understanding salary trends over time.
- **job_category:** A classification of the job role into broader categories for more straightforward analysis.
- **salary_in_usd:** The gross yearly salary converted to US dollars (USD) helps in global salary comparisons and analyses.
- **employee_residence:** The employee's country of residence can be helpful to explore geographical salary differences and cost of living variations.
- **experience_level:** Classifies the professional experience level of the employee; it can provide insight into how experience influences salary in data-related roles.
- **employment_type:** Specifies the type of employment; it can help analyze how different employment can affect salary.

- **work_setting:** The work setting or environment, such as "Remote," "In-person," or "Hybrid." It can reflect the impact of work settings on salary levels in the data industry.
- **company_location:** The country the company is in, which can help analyze how the company's location affects salary.
- **company_size:** The size of the employer's company can help analyze how a company's size can influence salary.

Table 2: Summary Statistics for Jobs in the Data Field

Variable Name	*Count*	*Mean*	*Min*	*25th*	*50th*	*75th*	*Max*
Work Year	*9,355*	*2022.76*	*2020*	*2023*	*2023*	*2023*	*2023*
Salary (USD)	*9,355*	*150,299.50*	*15,000*	*105,700*	*143,000*	*186,723*	*450,000*

Table 3: Proportions for Jobs in the Data Field by job_category

Job Category

Category	*Frequency*	*Proportion (%)*
Data Engineering	*2261*	*2261/9356= 24.17%*
Data Architecture and Modeling	*260*	*260/9356= 2.7%*
Data Science and Research	*3015*	*3015/9356= 32.23%*
ML & AI	*1429*	*1429/9356= 15.27%*
Data Analysis	*1458*	*1458/9356= 15.58%*
Leadership and Management	*504*	*504/9356= 5.39%*
BI and Visualization	*314*	*314/9356= 3.36%*
Data Quality and Operations	*56*	*56/9356= 56/9356= 0.6%*
Data Management and Strategy	*62*	*62/9356= 0.7%*
Cloud and Database	*6*	*6/9356= 0%*

Data Set Graphical Exploration:

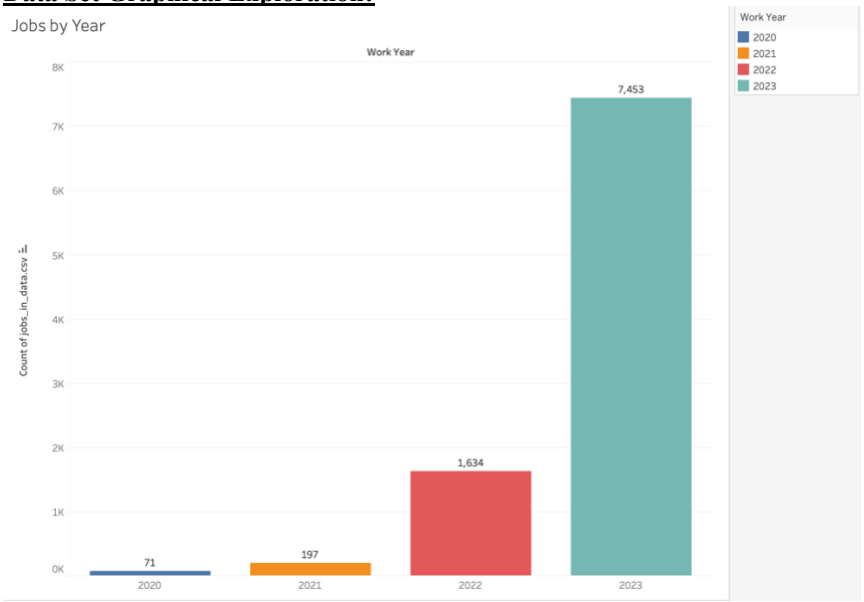

Figure 1: Looking at jobs by year.

Observation:

Bar Chart

- Approximately left-skewed, there's more data for 2023 compared to the rest of the other years.
- Meanwhile, for the year 2020, it had minimal to no data.

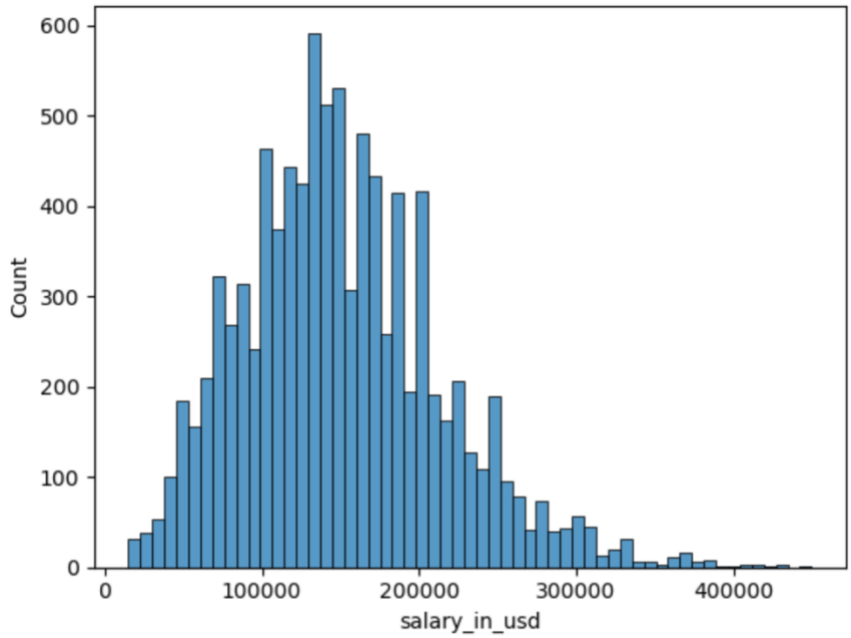

Figure 2: Looking at Salary Distribution

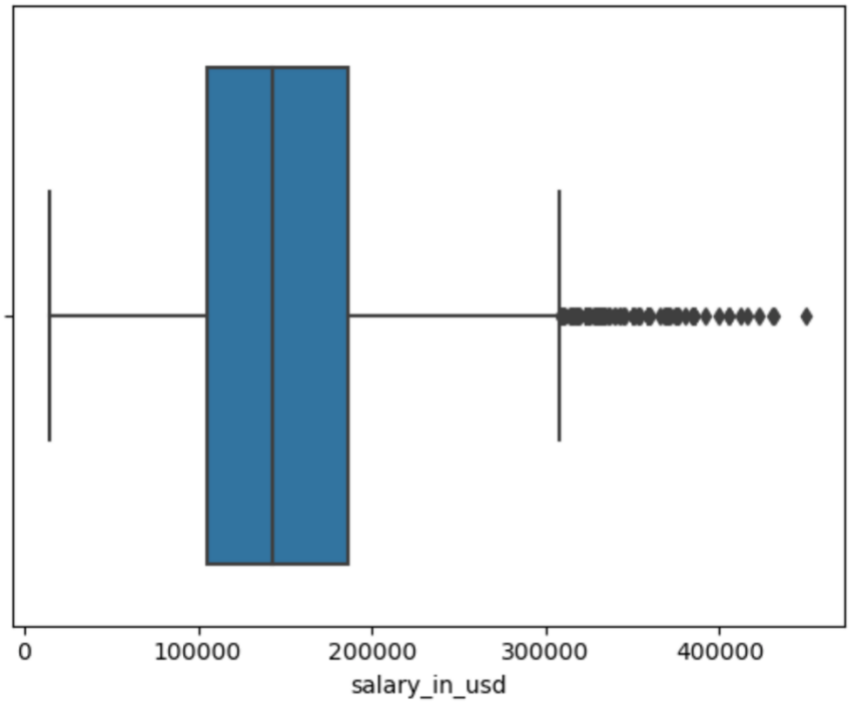

Figure 3: Looking at Salary Distribution

Observation:

Histogram Plot and Box Plot

- The two graphs demonstrate an approximately right-skewed distribution, with possible outliers after the $300,000 salary range.

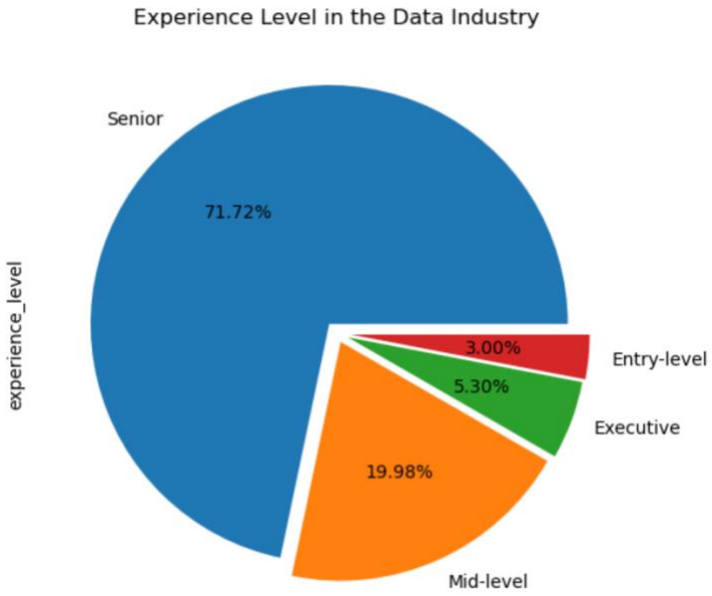

Figure 4: Experience Level in the Data Industry

Observation:

Pie Chart

- In the pie chart, we can observe how companies hire and focus more on senior-level roles, approximately 72%.
- On the other hand, only 3% of companies recruit Entry-level positions.

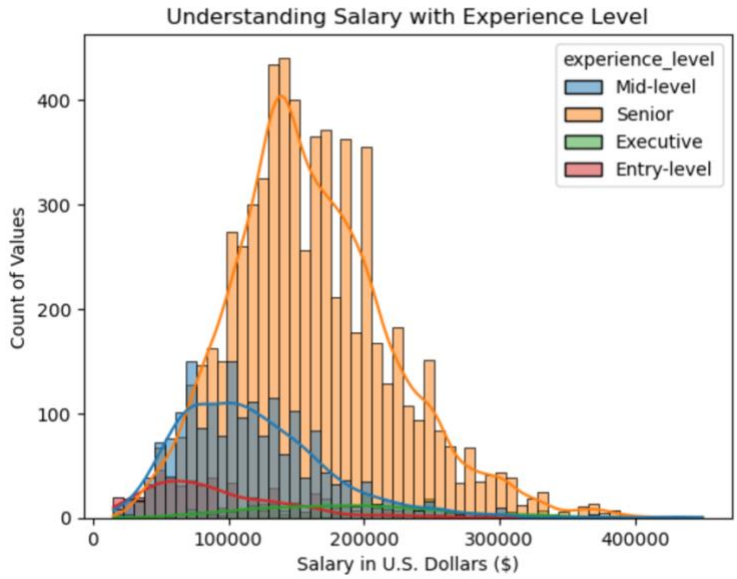

Figure 5: Understanding Salary with Experience Level

Observation:

Histogram

- A senior position receives an average of approximately 160k and a max of 400k in pay salary compared to other experience levels.
- A mid-level position receives an average of $100k; however, a few receive more than a 400k salary.

- As for Entry-level positions, the average is about 80k per year, whereas the max can be up to 250k.
- An executive experience level can average 189k and max around 400k.
- Lastly, most experience levels can demonstrate a right-skewed distribution.

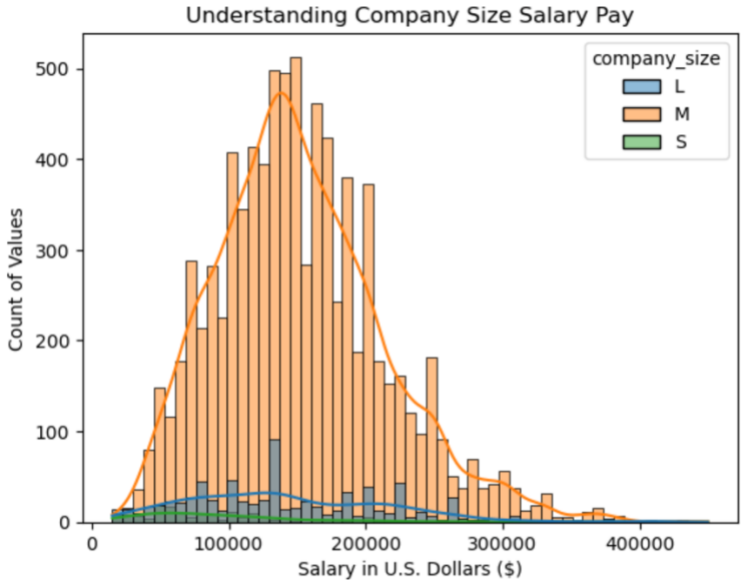

Figure 6: Understanding the Size of a Company and Salary Pay

Observation:

Histogram

- A small-sized company's average salary is around 90k, and the max is over 250k.
- A medium-sized company's average salary is around 152k, and max is 400k.
- A larger company's average salary is around 141k, with a max of more than 400k.

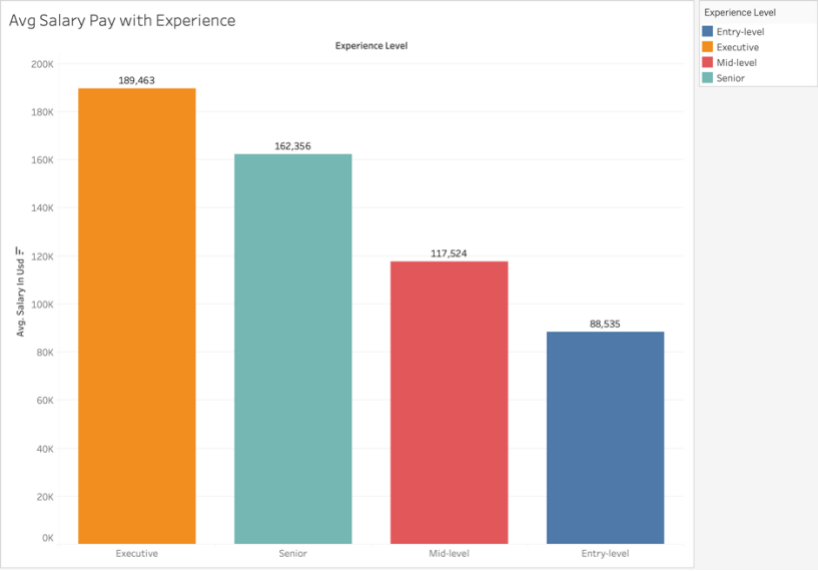

Figure 7: Understanding Average Salary Pay with Experience Level

Observation:

Bar Chart

- Executive-level positions received the highest salaries, while entry-level positions received the lowest.

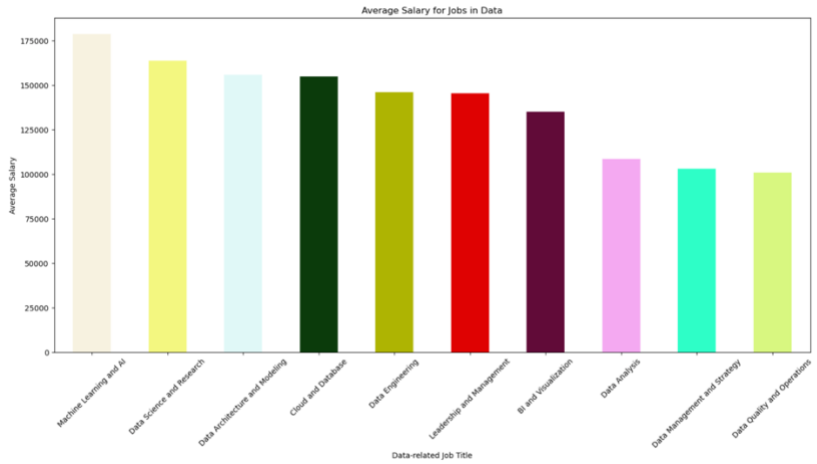

Figure 8: Understanding Average Salary in the Data Field

Observation:

Bar Chart

- Machine Learning and AI get the highest average salary compared to other roles in the data field.
- However, other job positions like Data Science and Research, Data Architecture and Modeling, Cloud and Database, Data Engineering, Leadership and Management, and BI and Visualization are right behind.

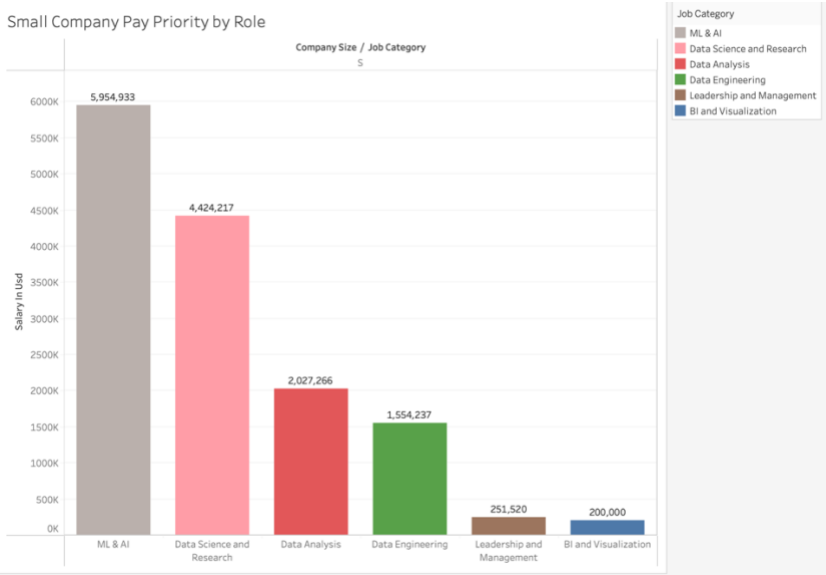

Figure 9: Understanding small company's pay.

Observation:

Bar Chart

- From the chart above, ML and AI are the highest-paying jobs in small companies.
- In addition, this can also demonstrate how ML and AI are now becoming vital roles in the data field.

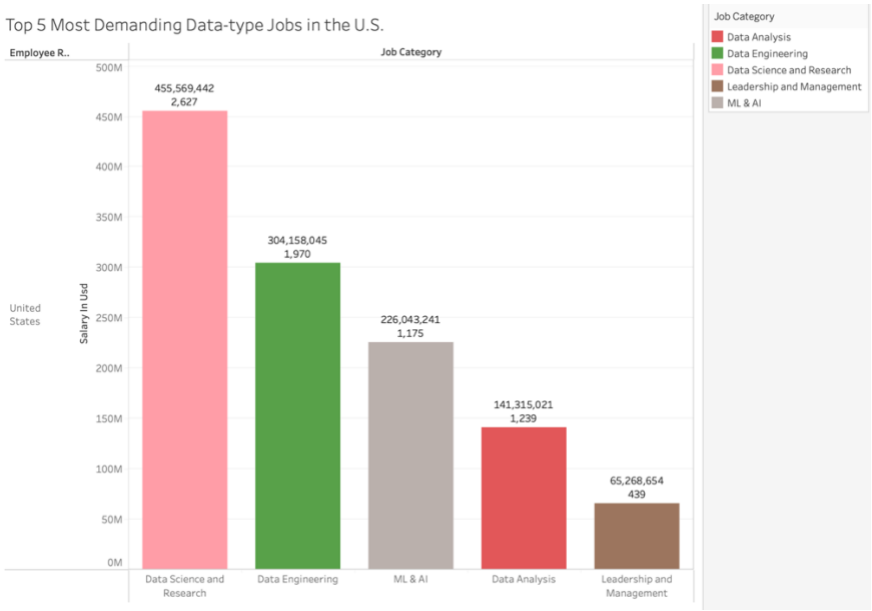

Figure 10: Top 5 Most Demanding Data Roles in the USA

Observation:

Bar Chart

- From the chart above, data science/research is considered the most important career in the data field in the United States, which also has the highest pay.

Figure 11: Location of Jobs Around the World

Observation:

Maps

- As you can see, the United States has the most data-type jobs from this data set.

Figure 12: Location of Jobs Around the World

Observation:

Word Cloud

- From this visualization, you can see how data science/research and data engineering are words that stick out the most compared to other careers in this data set.

Summary of findings:

In summary, after exploring and testing many different types of EDA, it can demonstrate how salary depends on different factors and can fluctuate in the data industry. In addition, the data set only focuses on a few interval data types like work year and salary. The rest of the columns are all nominal data, which means they must change from categorical to numerical data to make future salary predictions work and be as accurate as possible. The only way to make this change is by using the one-hot encoding, as you demonstrated in class, by transferring the categorical variables into a numerical layout before splitting and testing with the ML algorithms. Finally, I enjoyed exploring and learning more about different variables affecting salaries in the data field.