Module 1: Exploratory Data Analysis of Data Science Salaries Dataset

Mohammed Saif Wasay (002815958)

Masters of Professional Studies in Informatics, Northeastern University

ALY 6040: Data Mining Applications

Harpreet Sharma

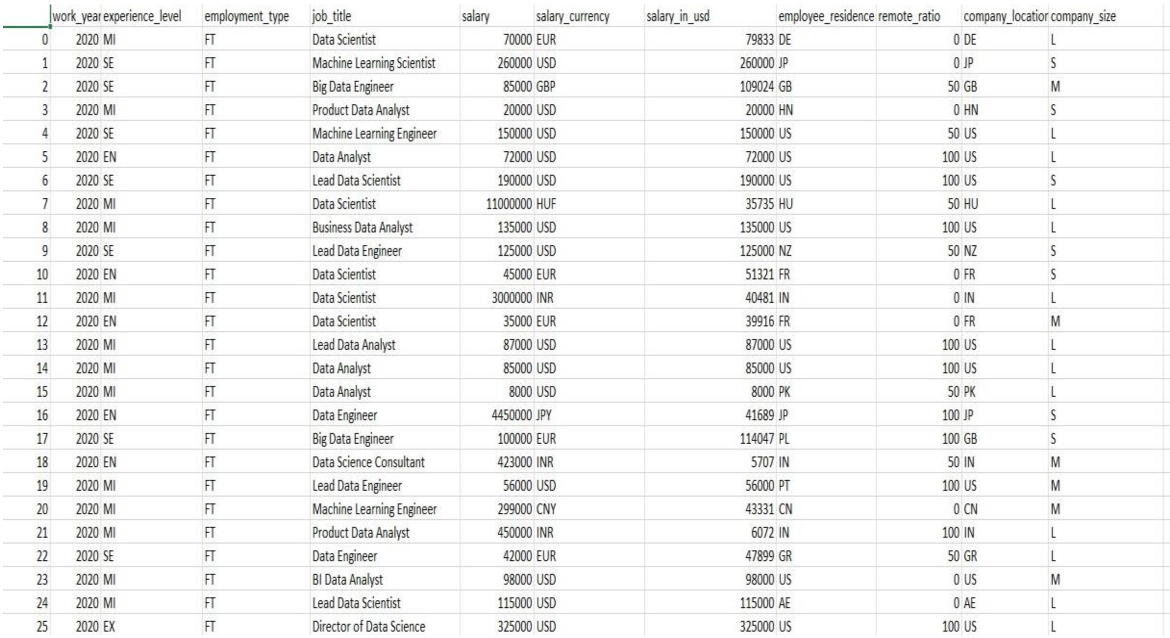
September 23rd, 2024

Introduction and Key Findings

This report provides an exploratory analysis of a dataset focused on data science roles and their corresponding salaries. The dataset includes various features such as job title, salary in USD, experience level, employment type, employee residence, company location, remote ratio, and company size. Given the growing demand for data science professionals and the increased prevalence of remote work opportunities, this dataset offers a valuable snapshot of the current state of the job market in this field. The analysis aims to explore the structure, cleanliness, and potential outliers in the data while summarizing key findings and proposing next steps for data cleaning and further analysis.

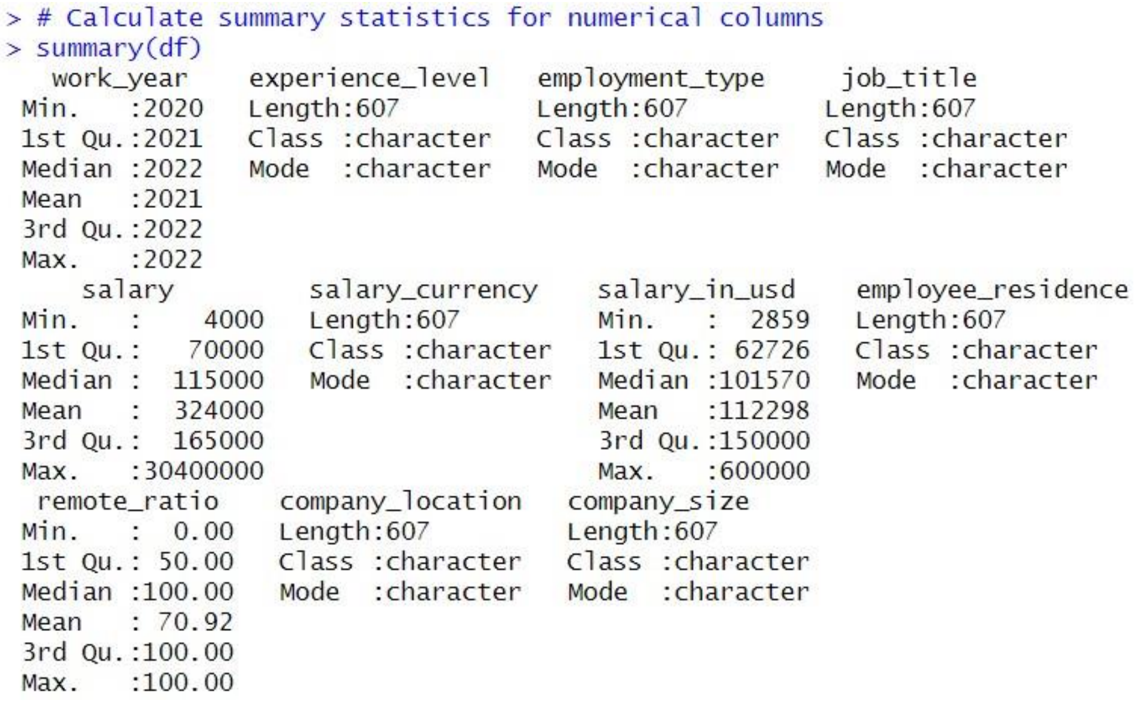
Data Exploration Process

1) Firstly, the dataset consists of 12 columns with 607 entries, where each row represents a data science role with associated details such as salary, job title, experience level, and remote work ratio. The primary goal of this phase was to explore the data structure, assess its cleanliness, and identify any potential outliers or issues that could impact the quality of analysis. Based on the initial exploration of your dataset, there were **no missing values** in any of the columns. All 607 rows had complete data for each feature, meaning that the dataset was relatively clean in terms of missing data. However, it's always good practice to double-check for any potential issues when conducting further analysis or cleaning



	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_locatior	company_size
0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S
2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S
4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	L
5	2020	EN	FT	Data Analyst	72000	USD	72000	US	100	US	L
6	2020	SE	FT	Lead Data Scientist	190000	USD	190000	US	100	US	S
7	2020	MI	FT	Data Scientist	11000000	HUF	35735	HU	50	HU	L
8	2020	MI	FT	Business Data Analyst	135000	USD	135000	US	100	US	L
9	2020	SE	FT	Lead Data Engineer	125000	USD	125000	NZ	50	NZ	S
10	2020	EN	FT	Data Scientist	45000	EUR	51321	FR	0	FR	S
11	2020	MI	FT	Data Scientist	3000000	INR	40481	IN	0	IN	L
12	2020	EN	FT	Data Scientist	35000	EUR	39916	FR	0	FR	M
13	2020	MI	FT	Lead Data Analyst	87000	USD	87000	US	100	US	L
14	2020	MI	FT	Data Analyst	85000	USD	85000	US	100	US	L
15	2020	MI	FT	Data Analyst	8000	USD	8000	PK	50	PK	L
16	2020	EN	FT	Data Engineer	4450000	JPY	41689	JP	100	JP	S
17	2020	SE	FT	Big Data Engineer	100000	EUR	114047	PL	100	GB	S
18	2020	EN	FT	Data Science Consultant	423000	INR	5707	IN	50	IN	M
19	2020	MI	FT	Lead Data Engineer	56000	USD	56000	PT	100	US	M
20	2020	MI	FT	Machine Learning Engineer	299000	CNY	43331	CN	0	CN	M
21	2020	MI	FT	Product Data Analyst	450000	INR	6072	IN	100	IN	L
22	2020	SE	FT	Data Engineer	42000	EUR	47899	GR	50	GR	L
23	2020	MI	FT	BI Data Analyst	98000	USD	98000	US	0	US	M
24	2020	MI	FT	Lead Data Scientist	115000	USD	115000	AE	0	AE	L
25	2020	EX	FT	Director of Data Science	325000	USD	325000	US	100	US	L

2) The exploration process involved an initial review of the dataset structure, where I examined both numerical and categorical variables. Key variables included "work_year," which refers to the year the data was recorded (ranging from 2020 to 2022), "experience_level," which captures the level of experience for each role (e.g., Entry-Level, Mid-Level, Senior-Level, and Executive-Level), and "employment_type," which categorizes the job contract (e.g., Full-Time, Part-Time, Contract, and Freelance). The dataset also contains "salary_in_usd," which is the primary numerical variable of interest, representing the salary converted to USD. Another critical feature is the "remote_ratio," indicating how much of the job can be done remotely, with values ranging from 0% (on-site) to 100% (fully remote). Lastly, "company_size" classifies organizations as Large (L), Medium (M), or Small (S).



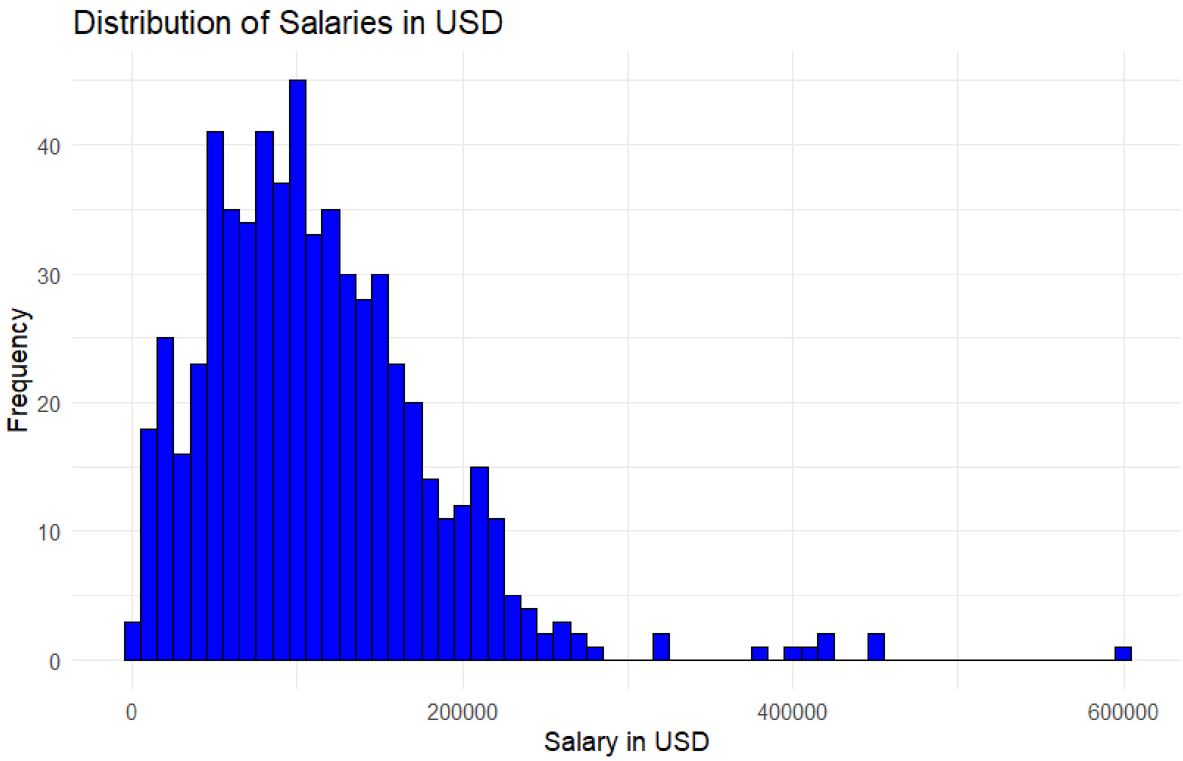
```
> # Calculate summary statistics for numerical columns
> summary(df)
   work_year    experience_level   employment_type     job_title        
 Min.   :2020   Length:607         Length:607         Length:607        
 1st Qu.:2021   Class :character   Class :character   Class :character  
 Median :2022   Mode  :character   Mode  :character   Mode  :character  
 Mean   :2021                                                           
 3rd Qu.:2022                                                           
 Max.   :2022                                                           
     salary         salary_currency    salary_in_usd    employee_residence
 Min.   :    4000   Length:607         Min.   :  2859   Length:607        
 1st Qu.:   70000   Class :character   1st Qu.: 62726   Class :character  
 Median :  115000   Mode  :character   Median :101570   Mode  :character  
 Mean   :  324000                      Mean   :112298                     
 3rd Qu.:  165000                      3rd Qu.:150000                     
 Max.   :30400000                      Max.   :600000                     
  remote_ratio    company_location   company_size      
 Min.   :  0.00   Length:607         Length:607        
 1st Qu.: 50.00   Class :character   Class :character  
 Median :100.00   Mode  :character   Mode  :character  
 Mean   : 70.92                                        
 3rd Qu.:100.00                                        
 Max.   :100.00                                        
```

3) During this phase of exploration, I calculated summary statistics to understand the data distribution and variability. The mean salary in USD was $112,297, with the lowest salary at $2,859 and the highest at $600,000. The remote ratio showed that many jobs are either fully remote (100%) or partially remote (50%), indicating a significant shift toward flexible work arrangements in the tech industry. The dataset also revealed that the majority of employees reside in the United States, with

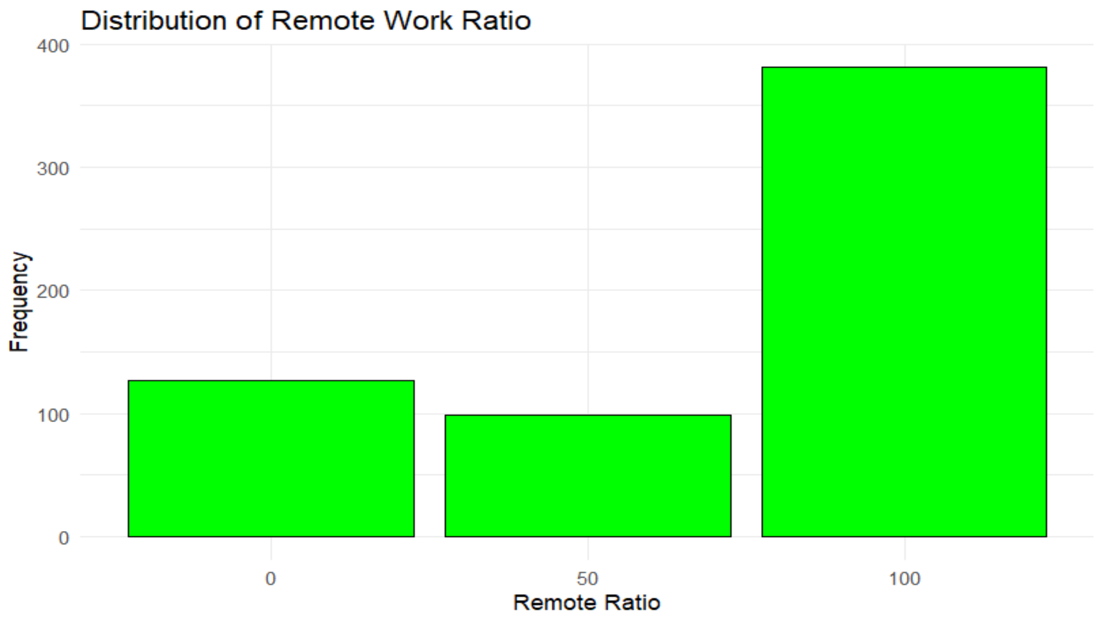
the most common job title being "Data Scientist." Furthermore, experience level was predominantly Senior-Level (SE), which aligns with the higher average salaries observed in the dataset.

4) To gain further insights, I generated visualizations, including a histogram of salaries in USD, which indicated that most salaries are clustered between $60,000 and $150,000, with a few extreme high outliers. A bar plot of job titles showed that "Data Scientist" and "Machine Learning Engineer" were among the most frequent roles. Additionally, a bar plot of the remote work ratio illustrated the growing trend toward remote work. This initial exploration provided a foundational understanding of the dataset's structure, distribution, and potential areas of concern.

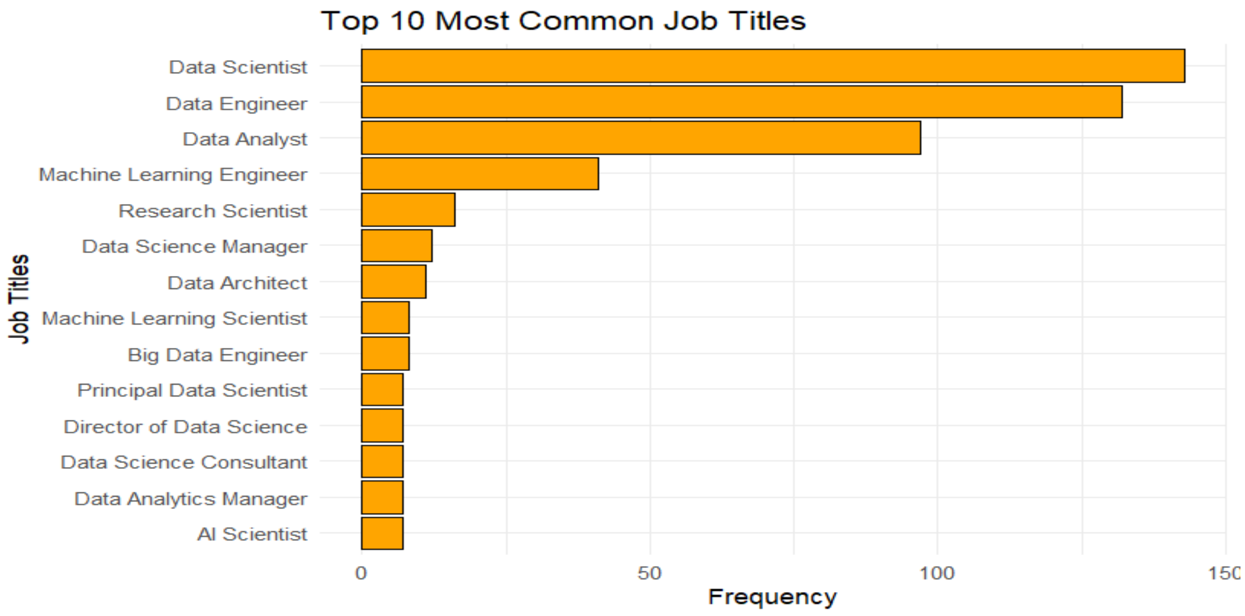
5) The salary range in data science roles is quite broad, with the majority of salaries falling between $60,000 and $150,000. Here, senior-level roles (SE) dominate the dataset, which may explain the higher average salaries. Moreover, outliers are particularly salaries exceeding $500,000, require further verification to confirm their accuracy.



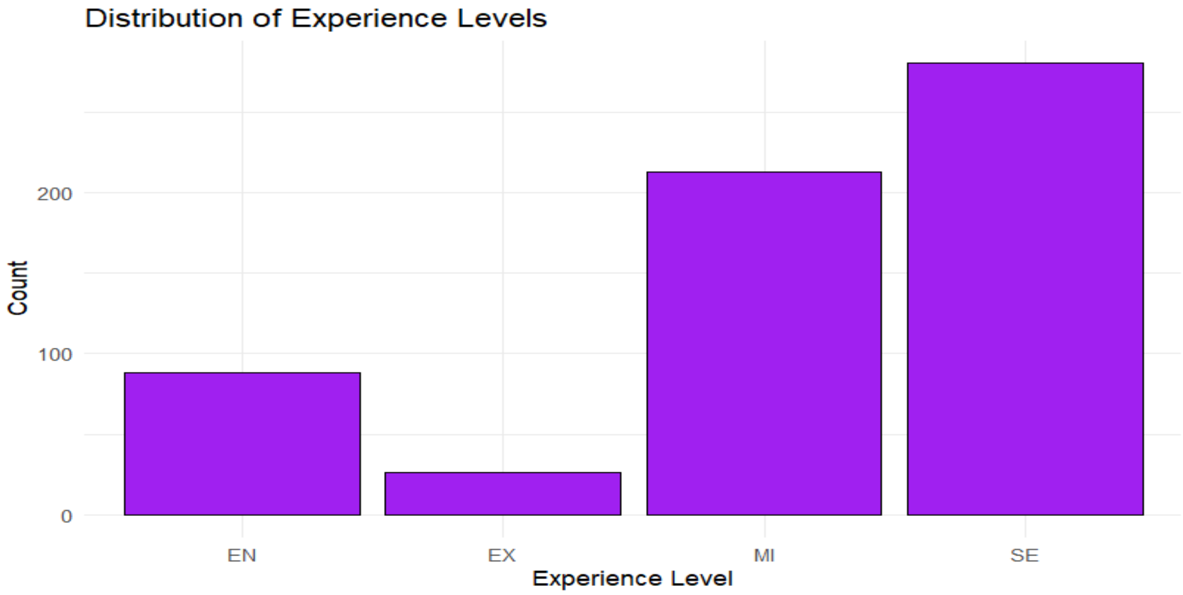
6) A significant portion of the dataset reflects a shift towards remote work, with 100% remote jobs being the most frequent. And, this trend aligns with the global shift towards more flexible work environments, particularly in the tech industry.



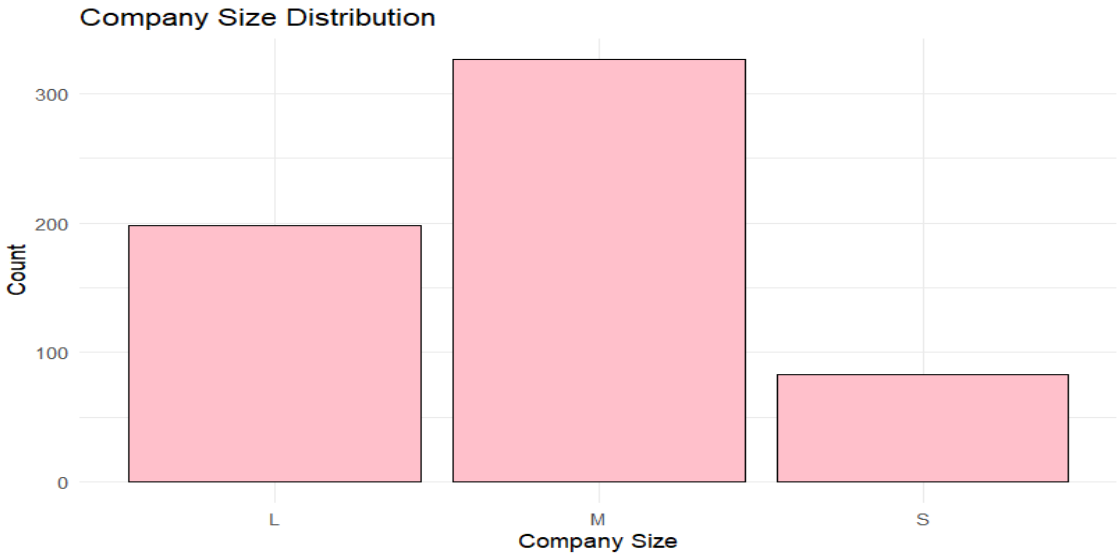
7) Data Scientist is by far the most common job title, followed by Machine Learning Engineer and Data Engineer. This concentration around specific job roles suggests that while the dataset is comprehensive, it may be more focused on certain types of data-related roles, with other roles being underrepresented.



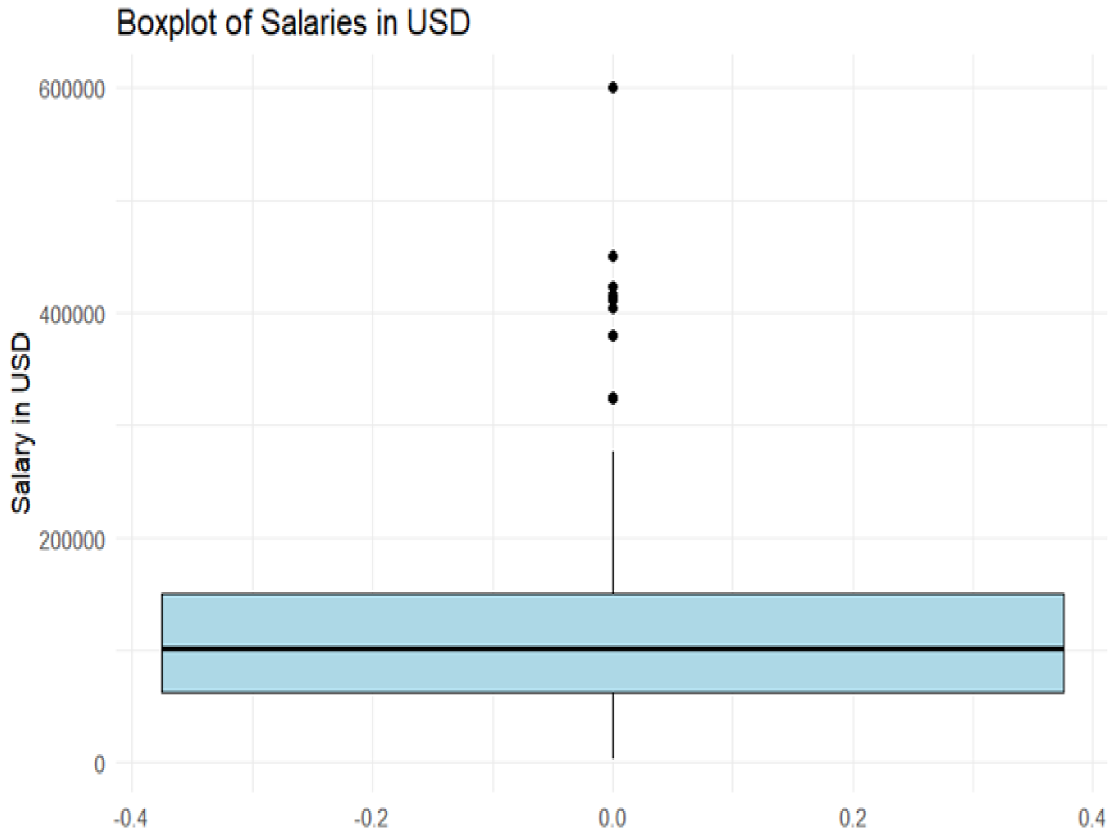
8) The experience level distribution reveals that the majority of roles are at the Senior (SE) and Mid-Level (MI), with Senior-level positions being the most common. Entry-level (EN) and Executive-level (EX) positions are less frequent, indicating that the dataset skews towards experienced professionals in data science.



9) This plot illustrates that most data science positions are at medium-sized companies (M), while fewer opportunities are available at large (L) and small (S) companies. This could suggest that medium-sized businesses are increasingly adopting data-driven strategies and investing in data science roles.

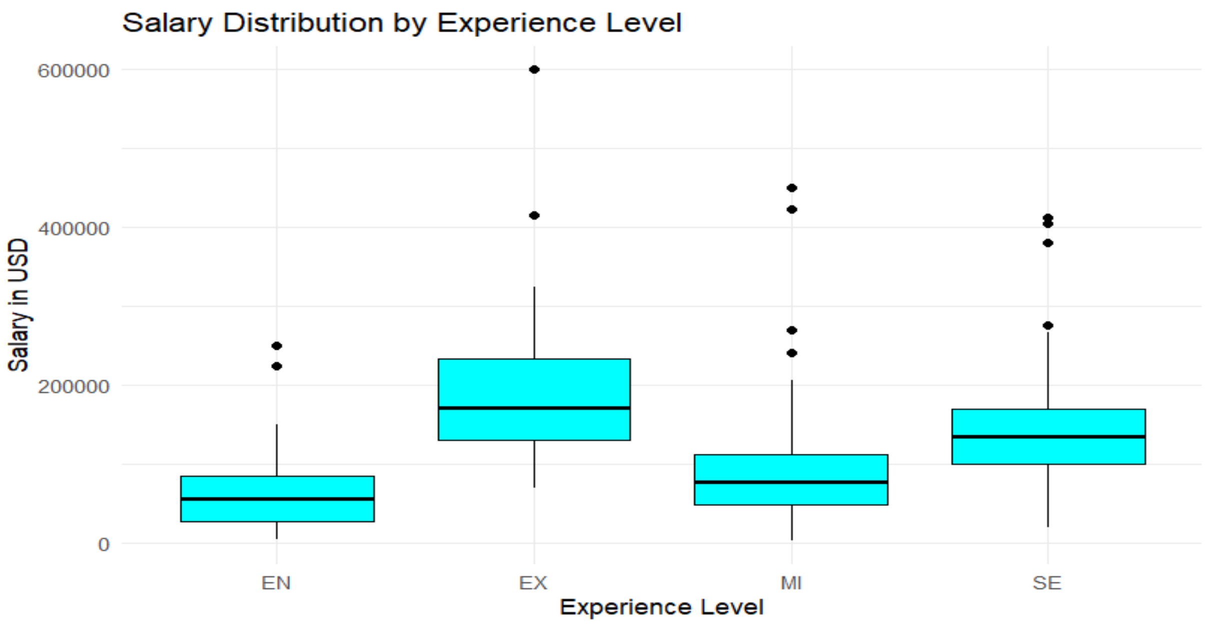


10) The boxplot reveals that the majority of salaries in the dataset range between approximately $50,000 and $200,000. However, several high outliers exist, with salaries approaching $600,000. These outliers may represent executive or specialized roles, but they require further analysis to determine their validity.



11) This boxplot illustrates how salaries are distributed across different experience levels. Executive-level (EX) roles have the highest median salary, with several high outliers approaching $600,000. Senior-level (SE) and Mid-level (MI) positions have similar salary ranges, with some outliers, although the overall salary distribution for Mid-level roles is lower. Entry-level (EN) roles show the lowest salaries, with minimal variation compared to other experience levels.

This visualization highlights the significant impact of experience level on salary, with executives earning considerably more than other levels.



Overall Key Findings:

The exploratory analysis revealed a broad range of salaries in data science roles, with most salaries falling between $60,000 and $150,000, and senior-level positions being the most prevalent, likely contributing to the higher average salaries. The presence of salary outliers exceeding $500,000 is notable and warrants further investigation to confirm their validity. Additionally, the dataset highlighted a strong trend toward remote work, with a significant portion of jobs being fully or partially remote, reflecting the tech industry's shift toward flexible work environments. Geographically, the dataset was skewed toward the United States, suggesting that it may primarily reflect North American trends despite including 57 unique countries. "Data Scientist" was the most common job title, with roles like "Machine Learning Engineer" and "Data Engineer" also well-represented, though other data science positions may be underrepresented. The large number of fully remote jobs and the extreme variation in salaries, particularly the high outliers, are intriguing, raising questions about the factors driving these salary discrepancies, such as geographic location, experience level, or specific job functions.

Proposed Next Steps:

To prepare the dataset for advanced analysis, key data cleaning steps include checking for and removing duplicate entries and addressing salary outliers, such as the extreme \$30.4 million value. Ensuring standardized currency values and consistent formatting of categorical variables like "job_title" and "experience_level" will further enhance data quality. Augmenting the dataset with additional features, such as industry sector and geographic region, will provide more granular insights. Feature engineering, such as calculating tenure or creating regional categories, will also aid in more targeted analysis. Once these steps are complete, the dataset will be suitable for predictive modeling or regression analysis to explore salary trends.

Conclusion/Recommendations

This exploratory data analysis has provided valuable insights into the structure, distribution, and quality of the dataset. While the dataset is generally clean and well-structured, there are areas that require further attention, particularly in handling outliers and verifying the accuracy of salary data. The prevalence of senior-level roles and fully remote positions highlights key trends in the data science job market, and further analysis will provide a deeper understanding of these trends. With appropriate data cleaning and feature engineering, this dataset can be used for more advanced analysis to gain insights into salary drivers and job market trends in the data science field.

Citations

R Documentation, An introduction to R. Retrieved 23rd September 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation

Albusairi, F. (2023, March 26). Mastering Simple R Visualizations: From Scatter Plots to Heat Maps. . Retrieved 23rd September 2024 from https://www.linkedin.com/pulse/mastering-simple-r-visualizations-from-scatterplots-heat-albusairi/.

Dataset Reference, Goodreads. Retrieved 23rd September 2024 from https://www.goodreads.com/.