Data Science Career Trends and Salary Prediction: Roles, Skills, and Market Growth

Abdul Moiz Affan Malik Fahad Hussain

I.INTRODUCTION

Data science has emerged as one of the most transformative fields in the 21st century, driving innovation and decision-making across industries. With applications ranging from healthcare to finance, the demand for skilled professionals in this domain continues to grow exponentially. This study explores the evolving land- scape of data science careers, focusing on roles, skills, salary trends, and market growth.

A. Motivation

As organizations increasingly adopt data-driven strategies, understanding the dynamics of the data science job market becomes essential. Analyzing job roles, re- quired skills, and compensation patterns provides in- sights into trends shaping career opportunities. The ”Data Science Jobs 2024” dataset serves as a rich source of information for examining these aspects and identifying the current and future state of data science professions.

B. Data Collection

The data used in this study was collected from Glassdoor, a popular platform for job listings and salary insights. The data collection process was carried out using Selenium, a powerful web scraping tool that automates the process of browsing and extracting information from websites. By utilizing Selenium, I was able to navigate through Glassdoor's job listings and extract detailed information on job titles, company details, salary estimates, job descriptions, and other relevant features for data science roles. This approach allowed for efficient gathering of a large dataset that includes various job-related attributes, providing a robust foundation for the subsequent analysis and modeling.

C. Dataset Description

The ”Data Science Jobs 2024” dataset contains detailed information on 5745 rows of job listings. Key features include job titles, salary estimates, descriptions, company details, and industry affiliations. Salary data, presented in annual and hourly formats, offers flexibility for comparative analysis. Additional fields, such as company size, founding year, type of ownership, and revenue, provide valuable organizational context. The dataset includes some missing values, ensuring robust analysis. It captures roles across sectors like information technology, retail, and manufacturing, reflecting the widespread adoption of data science.

II. SALARY ESTIMATION

A.Data Preprocessing

Salary data underwent several preprocessing steps for consistency and meaningful analysis. Salaries were classified as either Glassdoor Estimates or Employer Estimates based on their source. Non-numeric elements such as “$,” “K,” and “Per Hour” were removed to standardize the format. Minimum and maximum salary values were extracted to quantify variability. Hourly wages were converted into annual salaries using the assumption of a 40-hour workweek and 50 working weeks per year (2,000 hours annually). We addressed missing values by using targeted imputation strategies. For categorical columns like Size, Founded, and Revenue, missing values were filled with the mode, calculated within groups defined by Company Name. Industry and Sector missing values were imputed using the mode grouped by Refined Job Title, ensuring relevance to specific roles. For other columns, we used the median, grouped by Refined Job Title, to ensure a more accurate estimate. Missing values in the Average Salary column were imputed using the mean salary within groups defined by Refined Job Title, ensuring consistency with similar job roles. These strategies ensured that the dataset remained robust and contextually accurate for analysis and modeling.

B. Derived Metrics

To facilitate detailed analysis, new columns were introduced: Minimum Salary, Maximum Salary, and Aver- age Salary, calculated as the midpoint between minimum and maximum salaries. Hourly rates were adjusted to annual equivalents.

Table 1: Sample Salary Data

Job Title	Min Salary	Max Salary	Avg Salary	Per Hour
Data Analytics	$38	$45	$83K	Yes
Data Scientist	$120K	$130K	$125K	No
Data Engineer	$105K	$160K	$132.5K	No
Data Science Co-Op	$84K	$109K	$96.5K	No
Data Scientist	$66.8	$80.5	$147.3K	Yes

C. Key Observations

Hourly roles, such as data analytics positions, exhibit high annualized salaries when converted to yearly equivalents, reflecting the competitive nature of temporary data science work. Full-time roles displayed consistent salary ranges, with senior positions reaching

$160,000 annually. The flexibility in salary structures highlights diverse employment opportunities.

III. LOCATION ANALYSIS

A.Data Preprocessing

Job location details were refined and standardized. City and state pairs (e.g., “Tucker, GA”) were split, and state names were mapped to abbreviations (e.g., “Georgia”→ “GA”). Remote jobs were tagged as a distinct cate- gory.

B. Geographic Distribution

The dataset identified jobs in 33 states and included remote positions. High concentrations of roles were found in California, Texas, and New York. Remote jobs comprised a significant portion of listings, highlighting flexibility in data science roles.

Table 2: Sample Location Data

Job Title	Location	State
Data Analytics	Tucker, GA	GA
Data Scientist	Remote	Remote
Data Engineer	Dallas, TX	TX
Data Science Co-Op	Bridgeton, MO	MO
Data Scientist	Malvern, PA	PA

IV. JOB TITLE AND DESCRIPTION ANALYSIS

A. Refinement of Job Titles

Job titles were standardized into categories such as Data Scientist, Data Engineer, and Data Analyst based on key phrases. Each title was mapped to seniority levels (e.g., Junior, Mid-Level, Senior, or Leadership) using keywords.

B. Job Description Preprocessing

Job descriptions were cleaned to extract skills. Non-alphanumeric characters and stop words were removed, and lemmatization was applied. Relevant skills, such as “Python proficiency” and “machine learning expertise,” were highlighted for further analysis.

V. SKILLS ANALYSIS AND INSUDTRY INSIGHTS

A. Skills Analysis:

TF-IDF (Term Frequency-Inverse Document Frequency) was applied to rank skills. Common phrases such as machine learning, data analysis, and SQL ranked highest.

Table 3: Top Skills by TF-IDF Score

Skill	Score
Analysis	0.0712
Data	0.0531
Machine Learning	0.0365
Management	0.0338
Data Analysis	0.0316

B. Company Insights

Most organizations were private and fell into the IT sec- tor, with revenues typically between $5M and $25M. Startups and established firms both contributed to the dataset.

VI. Conclusion and Future Directions

This analysis highlights key trends in data science careers, emphasizing skills such as machine learning, data analysis, and technical proficiency. Future work could explore salary variations across regions and emerging skill trends over time.

VI. EXPLORATORY DATA ANALYSIS

A.Salary Comparisons by Job Title

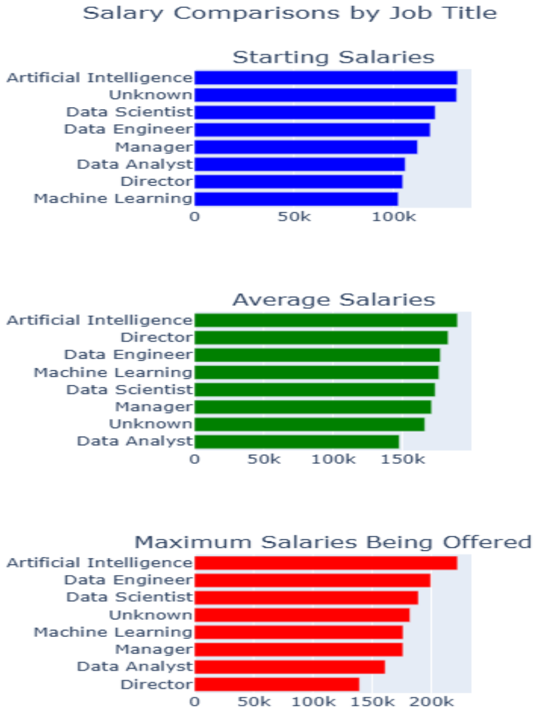
The analysis revealed significant variations in salaries based on job roles. Starting salaries for roles in Artificial Intelligence (AI) are the highest, exceeding $140,000, while technical roles like Data Engineer, Data Scientist, and Manager offer competitive starting pay between $120,000 and $140,000.

Roles in Machine Learning and Data Analysis were observed to have comparatively lower starting salaries.

For average salaries, AI-focused roles lead the market with averages near $150,000, closely followed by positions like Director and Data Engineer. Data Scientists, Managers, and Machine Learning professionals also maintain competitive average salaries, slightly below $150,000. Data Analysts, however, fall on the lower end of the spectrum.

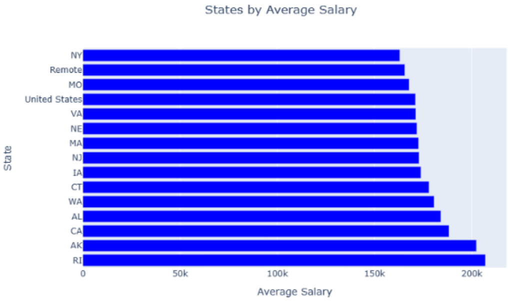
The maximum salaries demonstrate the earning potential at senior levels, particularly for AI roles, where salaries exceed $200,000. Data Engineer and Data Scientist positions also offer high ceilings, ranging from $180,000 to $200,000. Conversely, maximum salaries for Data Analyst and Director roles

are notably lower within this dataset.



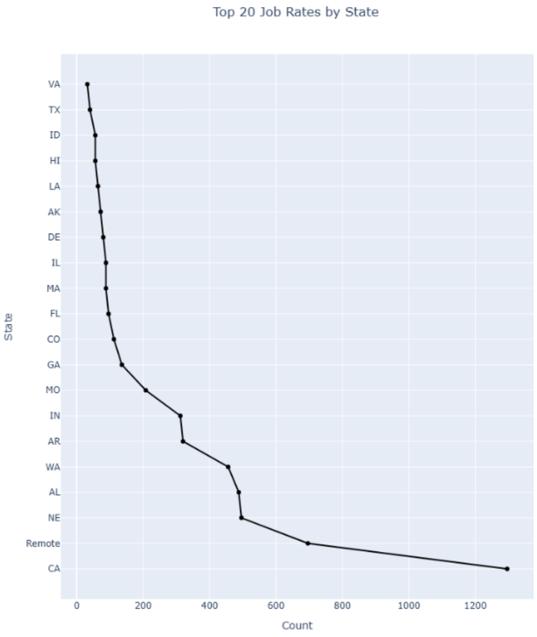
B.Salary Comparisons by State

Rhode Island (RI) emerged as the highest-paying state, with average salaries exceeding $200,000. Alaska (AK), California (CA), and Washington (WA) also rank among the top-paying regions, offering averages above $150,000. Remote roles were found to maintain competitive salaries, often comparable to those in high-paying states like New York (NY) and Massachusetts (MA). At the lower end, states like Pennsylvania (PA) and Minnesota (MN) showed average salaries below $100,000. This trend indicates significant regional disparities in salary distributions and highlights the competitiveness of remote opportunities.



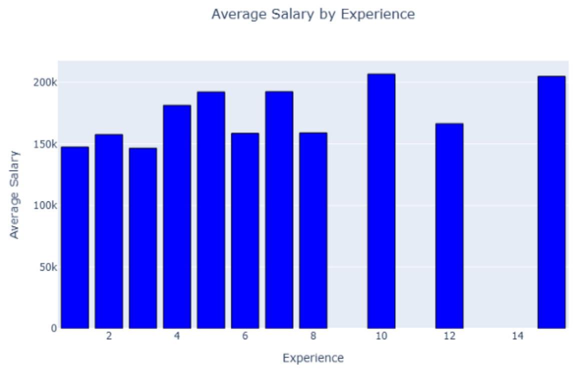
C.Job Rate Distribution by State

California (CA) stands out with the highest job rate, followed by Remote roles. Other prominent states like New York (NY), Texas (TX), and Washington (WA) also exhibit strong demand for data science professionals. Beyond these hubs, job rates gradually decline, with smaller contributions from other states, reflecting a concentrated demand in major tech ecosystems.



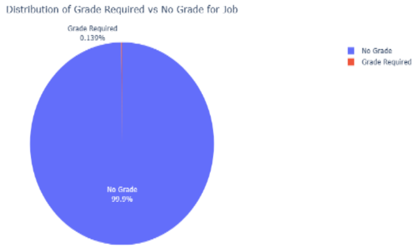
D.Average Salary by Experience

There is a clear positive correlation between average salary and years of experience. Salaries increase significantly with seniority, with a notable jump observed at the 10-year experience mark, where averages cross $200,000. Professionals with 15 or more years of experience exhibit similar high salary trends. Salaries stabilize between 4 and 8 years of experience, suggesting a plateau in mid-career compensation before a rise at more advanced levels. Entry-level professionals with less than 2 years of experience earn below the overall average, representing the typical starting range in the industry.



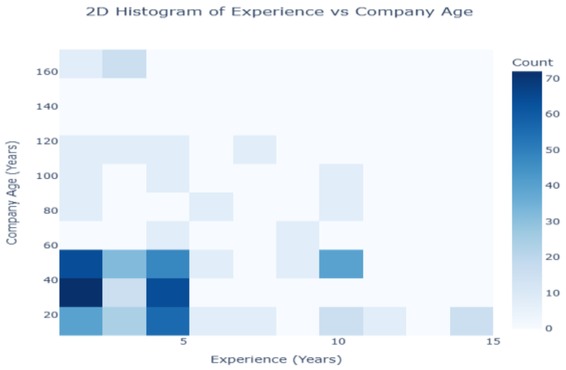
E. Distribution of Grade Requirements for Jobs

The dataset reveals that 99.9% of job postings do not require specific grades or GPAs, underscoring the industry's focus on practical skills and relevant experience over academic credentials. Only a negligible fraction (0.139%) of roles mandate a grade or GPA, likely for specialized or technical positions.



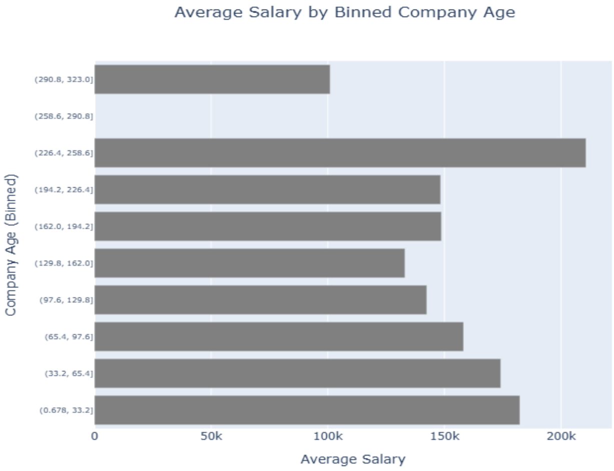
F. Experience vs. Company Age

The analysis of experience levels against company age reveals distinct hiring patterns. Younger companies (under 20 years) predominantly hire candidates with 3–5 years of experience, aligning with their preference for skilled professionals who require minimal training. Established firms, particularly those aged 20–60 years, demonstrate broader hiring patterns, targeting both entry-level and highly experienced professionals. Older firms generally balance their workforce across various levels of expertise.



G. Average Salary by Company Age

Companies aged 225 years or more offer the highest average salaries, exceeding $200,000, reflecting their financial stability and competitive compensation strategies. Moderate-aged firms (20–200 years) maintain consistent averages between $150,000 and $180,000, indicative of mature hiring practices. Startups and younger organizations exhibit more variable salaries, often below $150,000, driven by funding constraints and growth stages.



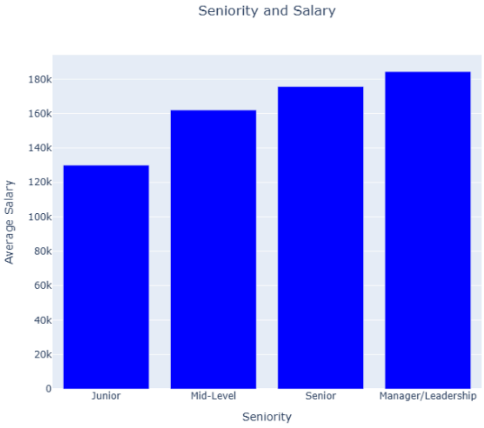
H. Seniority and Salary Trends

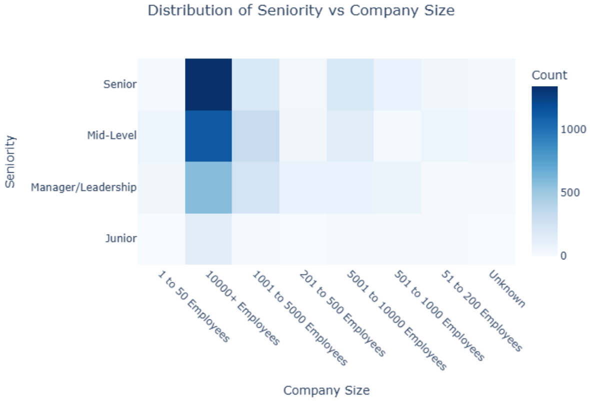
Salary trends indicate a steady increase with seniority. Entry-level roles offer average salaries of approximately$120,000. Mid-level professionals see significant salary growth, reaching averages around $150,000. Senior roles command salaries near

$170,000, while managerial and leadership positions peak at $180,000. This trend underscores the premium placed on experience and leadership in the field.

I.Job Distribution by Company Size

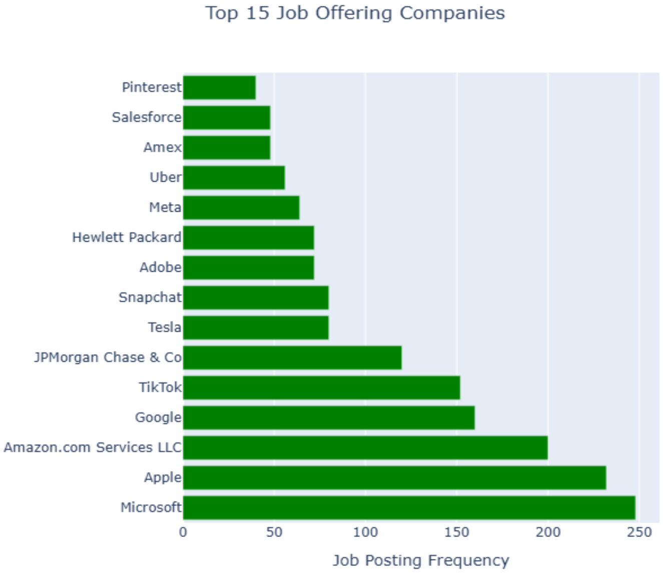
Larger companies (10,000+ employees) dominate the job market, contributing over 3,500 listings, showcasing their substantial hiring capacity. Medium-sized companies (1,001–5,000 employees) also provide ample opportunities, with approximately 700 job postings. Smaller organizations (51–1,000 employees) offer moderate listings, while very small firms (1–50 employees) account for a minimal share of the job market.





J. Top Companies Offering Roles

The analysis of job postings by company highlights Microsoft as the leading recruiter, with over 250 listings. Apple and Amazon follow closely, each offering between 200 and 230 positions. Prominent tech firms like Google, TikTok, and Tesla are also significant contributors, reflecting their aggressive hiring strategies. Financial institutions like JPMorgan Chase & Co and companies like Adobe, Meta, and Hewlett Packard further illustrate the widespread demand for data science talent across diverse industries.



VII. MODEL DEPLOYMENT AND EVALUATION

In this study, multiple regression and machine learning models were applied to predict the average salary based on features such as company age, rating, and job attributes. The models used include Ridge regression, Lasso regression, Gradient Boosting, Random Forest, Linear Regression, XGBoost, and Artificial Neural Networks (ANN). These models were evaluated and compared based on their Root Mean Squared Error (RMSE) and R-squared (R²) values.

A. Data Preprocessing

The preprocessing phase involved scaling the numerical features (Company Age and Rating) using MinMaxScaler, and encoding categorical variables using one-hot encoding. The target variable (Average Salary) was also scaled for consistency. The dataset was split into training and testing sets, with 30% of the data held out for testing.

B. Model Training and Evaluation

The following models were trained and evaluated separately:**Ridge Regression**: Ridge regression applies L2 regularization to the linear regression model. It was tuned using cross-validation to identify the best alpha value. The model achieved an RMSE of 0.0397 on the scaled data, with an R^2 of 0.9067, and an RMSE of 15,893.43 on the original scale.

Lasso Regression: Lasso regression applies L1 regularization, which shrinks some coefficients to zero, effectively performing feature selection. The Lasso model had an RMSE of 0.0392 and an R^2 of 0.9094, with an original RMSE of 15,665.67.

Gradient Boosting: This ensemble method showed an RMSE of 0.0414 and an R^2 of 0.8985, with an RMSE of 16,576.25 on the original scale.

Random Forest: A Random Forest regressor was trained and achieved an RMSE of 0.0429 and an R^2 of 0.8913, with an RMSE of 17,152.21 on the original scale.

Linear Regression: As a baseline model, Linear Regression performed poorly, with a very high RMSE (535,129.89) and a negative R^2 value, indicating poor fit to the data.

XGBoost: XGBoost achieved an RMSE of 0.0571 and an R^2 of 0.8801, with an RMSE of 17337.91 on the original scale.

C. Model Combination for Visualization

To better understand the performance of each model and to facilitate comparison, all models were evaluated in one cell for better visualization. This process allowed us to present the results of each individual model side by side for direct comparison. The outputs of Ridge, Lasso, Gradient Boosting, Random Forest, Linear Regression, and XGBoost were visualized together to gain insights into their relative performance.

D. Artificial Neural Network (ANN)

Finally, we implemented an Artificial Neural Network (ANN) using Keras. The ANN model was designed with the following architecture:

Input layer with 128 neurons, two hidden layers (64 and 32 neurons), and dropout layers to prevent overfitting.

The model was trained for 250 epochs using the Adam optimizer and Mean Squared Error (MSE) loss function.

Model Comparison

Model	**RMSE (Scaled)**	**R^2 (Scaled)**	**RMSE (Original)**
Lasso	0.0392	0.9094	15,665.67
Ridge	0.0397	0.9067	15,893.43
Gradient Boosting	0.0414	0.8985	16,576.25
Random Forest	0.0429	0.8913	17,152.21
Linear Regression	535,129.89	-169236914	214,051,958
XGBoost	0.0571	0.8801	17337.91
ANN	0.0422	0.8949	15,536.91

However, the Lasso regression model emerged as the best-performing method, achieving the lowest RMSE of 0.0392 (scaled) and an R^2 of 0.9094, demonstrating its ability to provide accurate predictions while maintaining simplicity. Ensemble models like Gradient Boosting, Random Forest, and XGBoost also delivered competitive results, emphasizing their utility in handling high-dimensional data. In contrast, the Linear Regression model performed poorly, reaffirming that simplistic linear approaches are unsuitable for complex datasets like this.

The evaluation of these models underscored the value of both regularization techniques, such as Lasso regression, and advanced machine learning methods, particularly ensemble and neural network approaches, for predicting salaries. The insights provided by these models highlight key factors influencing salary trends, offering actionable knowledge for hiring and talent management.

CONCLUSION

This study conducted a detailed exploration of salary predictions in data science using diverse machine learning models. Starting with preprocessing and feature scaling, we tested traditional models such as

Ridge and Lasso regression alongside advanced methods like Gradient Boosting, Random Forest, XGBoost, and Artificial Neural Networks (ANN).

Lasso regression emerged as the best-performing model, achieving the lowest RMSE and highest R^2, showcasing its balance between accuracy and interpretability. While ANN provided robust predictions, the simplicity and efficiency of Lasso make it a particularly effective choice. Ensemble methods like Gradient Boosting and Random Forest were strong alternatives, proving valuable for handling complex data patterns. In contrast, Linear Regression fell short, underscoring the importance of using more sophisticated techniques.

The findings from this study demonstrate the effectiveness of machine learning for salary prediction, offering valuable insights into compensation trends. These methods can guide organizations in making informed decisions regarding workforce planning and salary structuring. Future research could incorporate additional features, such as geographic and industry-specific factors, and explore advanced neural network architectures to further enhance predictive accuracy.