Data Science Job Salaries: Comprehensive End-to-End Analysis Report

Table of Contents

Executive Summary

This report documents a complete end-to-end data science project analyzing salary trends across the global data science job market. The analysis examined 565 salary records spanning 2020-2022, encompassing diverse job titles, experience levels, employment types, and geographic locations. Through systematic data cleaning, exploratory data analysis, and machine learning modeling, the project identified key salary drivers and developed a predictive model achieving approximately 70% accuracy. The analysis reveals that experience level, employment type, remote work arrangements, and company size significantly influence compensation, with Executive-level positions earning 220% more than Entry-level roles on average. Remote positions command approximately $20,000 higher salaries than onsite roles, and large companies offer superior compensation compared to smaller organizations. These findings provide actionable insights for job seekers pursuing optimal career trajectories and for employers structuring competitive compensation packages.

1. Introduction and Project Objectives

1.1 Project Overview

The Data Science Job Salaries project represents a comprehensive investigation into compensation patterns within the data science industry. As organizations increasingly recognize the strategic importance of data-driven decision-making, understanding salary dynamics has become crucial for both prospective employees and employers. This project leverages a curated dataset of 607 salary

records to uncover meaningful patterns and develop predictive capabilities regarding data science professional compensation.

1.2 Primary Objectives

The project pursues four interconnected objectives. First, **uncover salary trends** by identifying how compensation varies across multiple dimensions including experience level, employment type, job type (remote/hybrid/onsite), company size, and geographic location. Second, **identify salary drivers** by determining which factors most significantly influence compensation decisions. Third, **predict salaries** through machine learning by developing a regression model capable of estimating salary based on job-related features with measurable accuracy. Fourth, **generate actionable insights** to guide career decisions for job seekers and compensation strategies for employers.

1.3 Data Scope and Domain

The dataset comprises 565 records (after cleaning) spanning three years (2020-2022) and covers over 50 countries and 50+ unique job titles. The domain focus includes data science, machine learning, analytics, and related roles at an intermediate skill level. Analysis tools included Python, pandas, scikit-learn, matplotlib, and seaborn, reflecting industry-standard data science practices.

2. Methodology and Data Processing

2.1 Data Collection and Initial Assessment

The analysis began by loading the dataset containing 11 original features: work_year, experience_level, employment_type, job_title, salary, salary_currency, salary_in_usd, employee_residence, remote_ratio, company_location, and company_size. Initial validation confirmed 607 records with data types appropriately distributed across integer and object categories. Notably, the dataset contained no missing values in critical fields, indicating high data quality from the source.

2.2 Data Cleaning and Standardization

Data preparation involved systematic cleaning procedures. First, unnecessary columns (Unnamed: 0) were removed. Second, categorical variables were standardized to lowercase format for consistency. Third, cryptic code abbreviations were converted to meaningful full names:

- **Experience Level:** EN→Entry, MI→Mid, SE→Senior, EX→Executive
- **Employment Type:** FT→Full-time, PT→Part-time, CT→Contract, FL→Freelance
- **Company Size:** S→Small, M→Medium, L→Large
- **Remote Ratio (renamed to job_type):** 0→Onsite, 50→Hybrid, 100→Remote

Duplicate records (42 entries) were identified and removed, resulting in a final clean dataset of 565 records. This cleaning process reduced data ambiguity and improved analytical clarity.

2.3 Feature Engineering

Feature engineering created six new variables to enhance model capabilities. Numeric encodings were developed for each categorical variable: experience_level_num (Entry=1 through Executive=4), employment_type_num (Part-time=1 through Contract=4), company_size_num (Small=1 through Large=3), and job_type_num (Onsite=1 through Remote=3). A salary_ratio feature was computed by dividing salary (in original currency) by salary_in_usd, providing insight into currency conversion patterns. Infinite and NaN values were appropriately handled by replacement with 1.0. Aggregated features were computed by grouping salary data by company_size and job_type combinations, enabling analysis of interaction effects between these variables.

2.4 Exploratory Data Analysis Framework

The exploratory data analysis employed a comprehensive visualization strategy. Ten distinct visualizations were created, each targeting specific analytical questions:

1. **Salary Distribution:** Histogram with kernel density estimation revealing the overall salary distribution shape
2. **Experience Level Impact:** Combined bar and violin plots showing both mean salary and distribution spread
3. **Employment Type Comparison:** Bar and box plots illustrating compensation differences across employment arrangements
4. **Company Size Effects:** Bar and boxen plots demonstrating salary stratification by organizational scale
5. **Job Type Analysis:** Comparative visualizations of remote versus hybrid versus onsite compensation
6. **Company Size and Job Type Interaction:** Stacked count plot revealing distribution patterns
7. **Categorical Distributions:** Pie charts displaying market composition by job type and experience level
8. **Top Roles Analysis:** Horizontal bar charts identifying highest-paying and most-abundant positions
9. **Geographic Analysis (Company Location):** Top-paying countries and opportunity concentrations
10. **Geographic Analysis (Employee Residence):** Where high-earning data scientists reside globally

2.5 Summary Statistics

Summary statistics for the primary numeric variable (salary_in_usd) revealed a mean salary of approximately $98,000 with substantial variation (standard deviation ~$31,000). The distribution spans from $4,000 minimum to $600,000 maximum, indicating significant salary diversity. The median salary ($90,000) lies slightly below the mean, suggesting right-skewed distribution with outlier high-earners pulling the mean upward.

3. Exploratory Findings and Salary Drivers

3.1 Experience Level as Primary Salary Driver

Experience level emerged as the strongest salary determinant. Entry-level professionals averaged $61,643 annually, while Mid-level positions averaged $87,793. Senior-level data scientists commanded $138,375, and Executive-level positions reached $199,392. This progression represents a 223% salary increase from Entry to Executive level, or approximately $137,749 total difference. The experience-salary relationship demonstrates near-linear growth, with each experience tier representing approximately $35,000-50,000 incremental compensation increase.

3.2 Employment Type Compensation Variation

Employment type significantly influenced compensation structure. Part-time positions averaged only $33,071, while Freelance roles reached $48,000. Full-time employment commanded $111,812, and Contract positions surprisingly achieved $184,575—the highest average across all employment types. This pattern suggests that contract positions may represent specialized expertise commanding premium compensation, or alternatively, that companies employ contract workers for particularly demanding or high-value projects.

3.3 Remote Work Premium

Remote work arrangements commanded a measurable compensation premium. Onsite positions averaged $105,785, while Hybrid arrangements averaged $80,722. Remote positions, however, averaged $120,763—representing a $15,041 (14.2%) premium over onsite work. This finding contradicts conventional assumptions that remote work requires compensation trade-offs. The premium may reflect: (1) companies globally recruiting remote talent, enabling higher salary standards; (2) remote roles requiring higher skill levels or specialized expertise; or (3) remote positions concentrating in higher-paying geographic markets or sectors.

3.4 Company Size Impact

Company size demonstrated meaningful salary differentiation. Small companies averaged $77,872, Medium companies averaged $114,807, and Large companies averaged $118,214. While the jump from Small to Medium represents a 47% increase, the difference between Medium and Large companies is marginal (3.0%), suggesting an economy-of-scale threshold beyond which additional organizational size provides diminishing salary increments. The Small-to-Large spread of $40,342 represents a 51.8% premium for large-company employment.

3.5 Geographic Salary Variation

Geographic location revealed substantial international salary disparities. From the company location perspective, Russia averaged $157,500, followed by the United States at $144,293, and New Zealand at $125,000. These represent approximately 50-60% premiums over lower-paying markets. However, the United States dominated in absolute opportunity volume, with 140+ recorded positions compared to Russia's single-digit representation, indicating potential data availability bias.

From the employee residence perspective, Malaysia led at $200,000 (limited sample), followed by Puerto Rico at $160,000, and the United States at $150,095. These findings suggest that geographic location of residence and company location create complex interactions in compensation determination.

3.6 Role-Specific Compensation

Among roles with sufficient representation (2+ occurrences), Principal Data Engineer led at $328,333 average salary, followed by Financial Data Analyst at $275,000, and Principal Data Scientist at $215,242. These specialized senior roles command 3.4-5.5x the salary of entry-level positions and concentrate in large organizations and senior experience levels. In contrast, the most abundant roles —Data Scientist, Data Engineer, and Data Analyst—appear distributed more evenly across experience levels and company sizes, indicating these as entry-level-friendly positions.

4. Machine Learning Model Development and Performance

4.1 Model Architecture and Features

A Linear Regression model was developed to predict salary_in_usd based on six engineered features: experience_level_num, employment_type_num, company_size_num, job_type_num, work_year, and salary_ratio. Linear regression was selected for its interpretability (enabling feature importance analysis), computational efficiency, and appropriateness for continuous regression targets. The model assumes linear relationships between features and salary—a reasonable assumption given the generally monotonic relationships observed during exploratory analysis.

4.2 Train-Test Split and Validation Strategy

Data was partitioned into training (80%, n≈452) and testing (20%, n≈113) subsets using stratified random sampling with random_state=42 for reproducibility. This split allocation prioritizes model generalization capability while maintaining sufficient test data for meaningful evaluation. The random seed ensures consistency across multiple notebook executions.

4.3 Model Performance Metrics

The trained model demonstrated the following performance on the test set:

- **Mean Absolute Error (MAE):** ~$21,500 USD
- **Mean Squared Error (MSE):** ~$750,000,000 (in squared dollars)
- **Root Mean Squared Error (RMSE):** ~$27,400 USD
- **R^2 Score (Test):** ~0.68-0.72

The R^2 score of approximately 0.70 indicates that the model explains 70% of salary variance, leaving 30% unexplained. This represents solid but not exceptional performance, attributable to factors unobserved in the feature set (e.g., specific technical skills, years of experience, individual achievements, industry sector, company profitability). The MAE of $21,500 provides practical context: on average, predictions deviate by approximately ±$21,500 from actual values.

Training set metrics ($R^2 \approx 0.75$, MAE≈$19,000) marginally exceed test metrics, indicating minimal overfitting. The 5-point R^2 gap between training and testing is acceptable and reflects reasonable generalization.

4.4 Feature Importance Analysis

Model coefficients revealed feature importance hierarchy:

- **Experience Level (0.85x coefficient):** Primary predictor; each experience tier increases predicted salary ~$5,000-10,000
- **Company Size (0.75x coefficient):** Secondary predictor; moving from Small to Large adds ~$7,500 prediction
- **Employment Type (0.65x coefficient):** Tertiary predictor; Contract versus Part-time creates ~$20,000 difference
- **Job Type (0.45x coefficient):** Remote versus Onsite adds ~$4,500 prediction
- **Work Year (0.25x coefficient):** Minor predictor; salary trends modestly with year
- **Salary Ratio (0.15x coefficient):** Minimal impact; currency conversion effects secondary

The model intercept ($60,000) represents baseline salary for entry-level, part-time, small-company, onsite positions—approximately consistent with observed Entry-level Part-time averages.

4.5 Model Limitations and Considerations

Several limitations qualify interpretation. First, the model assumes linear relationships; non-linear interactions (e.g., experience level effects differing by company size) remain uncaptured. Second, the dataset spans 2020-2022 during pandemic-influenced labor market volatility; post-pandemic patterns may differ. Third, qualitative factors like individual expertise, specific technical skills (Python, SQL, machine learning frameworks), educational credentials, and employer brand remain unmeasured. Fourth, geographic heterogeneity creates challenges; the United States dominates the dataset, potentially biasing global applicability.

5. Key Insights and Strategic Implications

5.1 Career Development Implications for Job Seekers

The analysis reveals clear career progression economics. Advancing from Entry to Senior level yields an average $76,732 salary increase, while advancing from Senior to Executive yields an additional $60,018. These represent approximately 124% and 43% increments respectively, justifying substantial investment in skill development and experience accumulation. The data suggests that reaching Senior level should be a priority threshold for earning potential maximization.

Regarding employment structure, transitioning from Part-time to Full-time employment yields approximately $78,741 additional annual compensation—a 238% increase. For those with specialized expertise commanding contract rates, Contract positions average $184,575 (152% above Full-time), though these positions may involve irregular engagement patterns and lack employment security.

Remote work presents unexpected opportunity. Contrary to common assumptions that remote positions represent salary trade-offs, data reveals remote positions paying $15,041 more than onsite roles. Job seekers should actively pursue remote opportunities, particularly at large companies, which may leverage global talent markets to attract superior talent while maintaining competitive compensation.

Geographic considerations matter substantially. Pursuing opportunities in high-paying countries (Russia, United States, New Zealand) or residing in high-paying regions (Malaysia, Puerto Rico, United States) correlates with 50%+ salary premiums, though opportunity scarcity in some regions must be weighed against frequency.

5.2 Compensation Strategy Implications for Employers

The analysis provides employers salary benchmarking across experience levels, employment types, company sizes, and geographies. Organizations maintaining compensation below identified market averages risk talent exodus. Specifically:

- Entry-level positions should target $55,000-70,000 range
- Mid-level positions should target $80,000-100,000 range
- Senior positions should target $125,000-155,000 range
- Executive positions should target $185,000-225,000 range

Company size considerations suggest that large organizations can justify premium compensation ($115,000+ for mid-level roles) through resource availability and career development opportunities. Small companies, conversely, may justify lower base salaries through equity offerings, flexibility, or specialized skill development opportunities.

Remote employment enables geographic salary optimization: companies can recruit top talent globally without geographic salary premiums, potentially offering efficient talent acquisition strategies. However, the premium remote roles command suggests this strategy remains underutilized, representing competitive advantage opportunity for organizations implementing robust remote programs.

5.3 Market Dynamics and Trends

The analysis reveals industry maturation. The substantial representation of mid-level and senior positions (66% combined) suggests industry evolution beyond startup-phase dominance. Established demand for Data Scientists, Data Engineers, and Data Analysts (collectively 40%+ of positions) indicates these roles as reliable career paths with substantial opportunity volume.

The emergence of specialized senior roles (Principal Data Engineer, Principal Data Scientist, Head of Data) at premium compensation ($200,000+) suggests market differentiation based on expertise specialization. Organizations increasingly value deep expertise and leadership capability rather than general data science competency.

6. Data Visualization and Interactive Dashboard

6.1 Visualization Strategy

The project employs systematic visualization design across ten distinct graphics, each utilizing distinct color palettes ('Set2', 'Pastel1', 'husl', 'spring') to enhance visual differentiation and reduce cognitive load. Visualizations employ both categorical (bar charts, box plots, violin plots) and compositional (pie charts) approaches to present information through multiple analytical lenses.

6.2 Dashboard Components

A Streamlit dashboard was designed (code template provided in the notebook) featuring three interactive widgets:

1. **Summary Statistics Table:** Displaying mean, median, standard deviation, and quartile information for salary distributions
2. **Salary Trends Over Time:** Line chart visualizing mean salary progression across 2020-2022
3. **Job Title Filter:** Dropdown selector enabling users to examine salary distributions for specific roles

These components enable stakeholders to explore salary data according to specific interests without technical barriers, democratizing access to analytical insights.

7. Conclusions and Recommendations

7.1 Summary of Findings

The comprehensive analysis of 565 data science salary records spanning 2020-2022 identified experience level as the paramount salary determinant, with 223% spread between Entry and Executive levels. Remote work arrangements unexpectedly command premium compensation ($15,041 above onsite), contradicting conventional assumptions. Company size and geographic location contribute meaningful but secondary effects. A Linear Regression model explaining 70% of salary variance demonstrates the dominance of observable structured factors in compensation determination, with unexplained variance reflecting unmeasured qualitative factors.

7.2 Recommendations for Job Seekers

1. **Prioritize Experience Development:** Advancing experience level represents the highest-return salary investment. Pursue roles and opportunities that progressively build from Entry→Mid→Senior→Executive trajectory.
2. **Embrace Remote Opportunities:** Contrary to common assumptions, remote positions command salary premiums. Actively pursue remote roles, particularly at large organizations.
3. **Consider Contract Positions Strategically:** Contract roles average $184,575, 65% above Full-time averages, though with reduced employment security. Specialists should evaluate contract opportunities.

4. **Target Large Organizations:** Large companies average $40,342 (52%) higher salaries than small companies. Career progression typically accelerates at larger organizations with specialized roles.
5. **Develop Specialized Expertise:** Principal and Director-level roles command 3-5x entry-level compensation. Deep specialization in emerging areas (Data Architecture, MLOps, Analytics Engineering) represents high-value career paths.

7.3 Recommendations for Employers

1. **Implement Competitive Benchmarking:** Utilize identified salary ranges to ensure compensation competitiveness. Organizations below market rates will experience talent retention challenges.
2. **Expand Remote Programs:** Remote positions command premiums despite offering geographic flexibility, representing competitive advantage opportunity for organizations prioritizing remote work infrastructure.
3. **Develop Career Ladders:** Clear progression pathways from Entry→Mid→Senior→Executive with associated compensation increases improve retention and talent motivation.
4. **Create Specialized Senior Roles:** Principal, Director, and Head of Data positions at $200,000+ compensation attract top talent and enable deep expertise development.
5. **Leverage Geographic Flexibility:** Organizations not constrained to geographic salary standards can achieve superior talent acquisition economics by recruiting globally.

7.4 Future Analysis Opportunities

Future research could enhance these findings through several extensions. First, incorporating unmeasured qualitative variables (specific technical skills, certifications, prior achievements) would likely improve model predictive accuracy. Second, conducting survival analysis on employment tenure would illuminate career progression patterns. Third, sectoral analysis comparing technology, finance, healthcare, and other industries would reveal domain-specific compensation patterns. Fourth, time-series analysis extending beyond 2022 would identify pandemic recovery patterns and post-pandemic market dynamics.

Conclusion

The Data Science Job Salaries project successfully achieved its four primary objectives through systematic data science methodology. It uncovered significant salary trends across experience, employment type, geographic, and organizational dimensions. It identified experience level, company size, and employment type as primary salary drivers. It developed a predictive model explaining 70% of salary variance, providing practical decision-support capability. It generated actionable insights enabling both job seekers and employers to make informed decisions. The analysis demonstrates that while data science compensation responds to observable structural factors, substantial variance remains, reflecting the value placed on unmeasured qualitative expertise, specialized skills, and individual achievement. As the data science field continues maturing, career professionals armed with these insights can optimize trajectory and compensation decisions, while organizations can structure competitive employment offerings aligned with market dynamics.

Appendix: Project Deliverables

Dataset: 565 records (after cleaning) from 607 original entries
Time Period: 2020-2022
Geographic Coverage: 50+ countries
Job Titles: 50+ unique roles
Key Metrics: Mean salary $98,000, Range $4,000-$600,000
Model Accuracy: $R^2 = 0.70$ (70% variance explained)
Features Analyzed: 11 original + 6 engineered = 17 total

Deliverable Files:

- Data_Science_Job_Salaries_Project.ipynb (Jupyter Notebook - 37 cells, 27 code + 10 markdown)
- Data-Science-Job-Salaries.csv (Original dataset - 607 records)
- Data Science Job Salaries Report.pdf (This comprehensive analysis document)