Comprehensive Salary Data - Exploratory Data Analysis Report

Executive Summary

This report presents a comprehensive exploratory data analysis (EDA) of salary data containing 15 variables across multiple dimensions including job roles, compensation, demographics, and company characteristics. Key findings include significant salary distribution skewness, 14,486 duplicate records, missing data in education fields, and diverse compensation structures across different job titles and experience levels.

1. Dataset Overview

1.1 Dataset Structure

Total Variables: 15 columns analyzed

Key Variable Types:

Categorical features :job_title, experience_level, employment_type, company_size, company_location, salary_currency, currency, education

Numerical features:remote_ratio, years_experience, base_salary, bonus, stock_options, total_salary, salary_in_usd

Data Quality Issues: 14,486 duplicate records identified

Missing Data: Education and skills columns contain null values

1.2 Variable Summary

Categorical Variables: 8 (job_title, experience_level, employment_type, company_size, company_location, salary_currency, currency, education)

Numerical Variables: 7 (remote_ratio, years_experience, base_salary, bonus, stock_options, total_salary, salary_in_usd)

2. Univariate Analysis - Categorical Variables

2.1 Job Title Distribution

Unique Values: 12 distinct job titles

Key Roles: Data Analyst, DevOps Engineer, Research Scientist, Software Engineer, Data Scientist, Machine Learning Engineer, Software Engineer

Distribution: Relatively balanced across most roles with some concentration in core data and engineering positions

Business Impact: Strong representation of technical roles, indicating focus on data-driven and engineering functions

2.2 Experience Level Analysis

Categories: 4 levels (Mid, Lead,Senior, Junior)

Distribution: Fairly even distribution across experience levels

Notable: Presence of null values in experience level data

Insight: Balanced mix of experience levels suggests diverse team composition

2.3 Employment Type Breakdown

Categories: 4 types (Contract, Full-time, Intern, Part-time)

Distribution: Relatively balanced across employment types

Observation: Multiple employment arrangements indicating flexible work structures

2.4 Company Size Distribution

Categories: 3 sizes (Medium, Small, Large)

Distribution: Approximately equal representation across all company sizes

Implication: Dataset represents diverse organizational contexts

2.5 Geographic Distribution

Company Locations: 6 locations (Germany, India, UK, Canada, USA, Remote)

Distribution: Balanced international representation

Remote Work: Significant remote work component

Global Scope: Multi-national dataset with strong geographic diversity

2.6 Currency Analysis

Salary Currency: 5 currencies (INR, GBP, EUR, CAD, USD)

Currency (Secondary): 5 currencies (USD, EUR, INR, CAD, GBP)

International Scope: Multiple currency representations reflecting global dataset

3. Univariate Analysis - Numerical Variables

3.1 Remote Work Ratio

Data Type: Integer (0-100 scale)

Mean: 49.9 (approximately 50% remote work)

Standard Deviation: 45.72

Range: 0-100

Distribution: Tri-modal distribution with peaks at 0%, 50%, and 100%

Insight: Clear segmentation between fully remote, hybrid, and fully on-site work arrangements

3.2 Years of Experience

Data Type: Integer

Mean: 10.01 years

Standard Deviation: 6.06 years

Range: 0-20 years

Distribution: Relatively uniform distribution across experience range

Insight: Workforce spans from entry-level to highly experienced professionals

3.3 Base Salary Analysis

Data Type: Float64

Mean: $273,915

Standard Deviation: $602,824

Range: $344,337 - $3,274,126

Distribution: Highly right-skewed with extreme outliers

Critical Finding: Significant positive skewness requiring transformation

3.4 Bonus Compensation

Data Type: Integer

Mean: $5,000.53

Standard Deviation: $2,891.50

Range: $0 - $9,999

Distribution: Relatively uniform distribution

Insight: Consistent bonus structures across roles

3.5 Stock Options Analysis

Data Type: Integer

Mean: $15,014.53

Standard Deviation: $8,664.14

Range: $0 - $29,998

Distribution: Uniform distribution pattern

Insight: Stock compensation is a significant component of total compensation

3.6 Total Salary Distribution

Data Type: Float64

Mean: $105,185

Standard Deviation: $36,335.19

Range: $13,732 - $196,335

Distribution: Right-skewed with normal-like characteristics

Key Finding: Less skewed than base salary but still requires attention

3.7 Salary in USD (Standardized)

Data Type: Float64

Mean: $103,339.36

Standard Deviation: $146,128.71

Range: $221.07 - $2,354,628

Distribution: Extremely right-skewed with significant outliers

Critical Finding: Requires log transformation for analysis

4. Data Quality Assessment

4.1 Missing Data Analysis

Education Column: Completely null (no data available)

Skills Column: Completely null (no data available)

Experience Level: Contains some null values

Employment Type: Contains some null values

Impact: Missing education and skills data limits comprehensive analysis

4.2 Duplicate Records

Total Duplicates: 14,486 rows identified

Status: Duplicates found during data quality assessment

Action Required: Duplicate removal needed for data cleaning

4.3 Outlier Analysis

Base Salary: Extreme outliers (max: $3.27M vs mean: $273K)

Salary in USD: Extreme outliers (max: $2.35M vs mean: $103K)

Standard Deviation: Very high relative to means indicating significant outliers

Impact: Outliers severely affect distribution normality

5. Distribution Skewness Analysis

5.1 Highly Skewed Variables

Base Salary: Extreme positive skewness

Salary in USD: Extreme positive skewness

Total Salary: Moderate positive skewness

5.2 Transformation Requirements

Log Transformation Applied: Successfully applied to adjusted_total_salary

Recommended: Apply log transformation to base_salary and salary_in_usd

Benefits: Improved normality, reduced outlier impact, better statistical properties

6. Business Insights

6.1 Compensation Structure

Base Salary Range: Extremely wide ($344K - $3.27M)

Total Compensation: Includes base salary + bonus + stock options

Stock Options: Significant component (mean: $15K)

Bonus Structure: Consistent across roles (mean: $5K)

6.2 Workforce Composition

Job Roles: Heavy concentration in data science and engineering roles

Experience Mix: Balanced across junior to senior levels

Geographic Diversity: Strong international representation

Work Arrangements: Balanced remote/hybrid/on-site distribution

6.3 Company Demographics

Size Distribution: Equal representation across small, medium, large companies

International Presence: Multi-national dataset with diverse currency exposure

Remote Work: 50% average remote work ratio indicating flexible work policies

7. Critical Data Quality Issues

7.1 Immediate Action Required

Remove 14,486 duplicate records - Data cleaning priority

Handle missing education and skills data - Determine if collection is possible

Apply log transformation to severely skewed salary variables (partially completed for adjusted_total_salary)

Investigate and treat extreme outliers in salary data

7.2 Data Validation Needs

Source Verification: Investigate cause of massive duplication

Outlier Verification: Validate extreme salary values

Missing Data Strategy: Determine approach for null education/skills fields

Currency Standardization: Ensure consistent currency conversion

8. Statistical Analysis Recommendations

8.1 Preprocessing Steps

Duplicate Removal: Clean 14,486 duplicate records

Log Transformation: Apply to base_salary, salary_in_usd, total_salary

Outlier Treatment: Consider capping or transformation for extreme values

Missing Data Handling: Implement appropriate strategy for null values

8.2 Analysis Approach

Use log-transformed variables for parametric statistical tests

Apply robust statistical methods for non-normal distributions

Consider non-parametric alternatives for severely skewed data

Implement proper confidence intervals accounting for data characteristics

9. Recommendations for Business Analysis

9.1 Compensation Analysis

Salary Benchmarking: Use cleaned, transformed data for reliable benchmarks

Pay Equity Analysis: Examine compensation across demographics after data cleaning

Total Compensation Focus: Include base salary + bonus + stock options in analysis

Geographic Adjustments: Consider cost of living adjustments for international comparison

9.2 Workforce Planning

Role Distribution: Analyze current vs. desired job title distribution

Experience Level Planning: Assess experience level balance for succession planning

Remote Work Strategy: Leverage 50% remote work ratio insights for policy development

Geographic Strategy: Use international distribution for expansion planning

10. Conclusion

This EDA reveals a rich dataset with significant analytical potential, but critical data quality issues must be addressed first. The identification of 14,486 duplicate records and extreme salary skewness requires immediate attention. The dataset provides excellent coverage of job roles, experience levels, and geographic distribution, making it valuable for comprehensive compensation analysis once cleaned.

Key Priorities:

Data Quality: Address duplicates and missing data

Distribution Normalization: Apply appropriate transformations (log transformation partially completed)

Outlier Management: Implement robust outlier treatment

Analysis Framework: Establish reliable analytical foundation

Business Value:

Comprehensive compensation benchmarking capability

Global workforce analysis potential

Remote work policy insights

Strategic compensation planning support

This report provides a foundation for reliable salary analysis. Data cleaning and transformation are essential prerequisites for meaningful business insights and decision-making.