

**Job Salaries Dataset Analysis Report**

**About the Author**

**Name: *Deep Makadia*  
B.Tech(CSE)(Pursuing)**

**Project Title: Job Salaries Dataset Analysis  
Date: *May 2025***

**Author Introduction**

I am Deep Makadia, a passionate data enthusiast focusing on data analysis, visualization, and storytelling using Python. This project explores the Netflix Titles dataset to uncover key insights about content distribution, trends over time, and viewer interests.

This analysis aims not just to visualize data, but to derive actionable insights that can guide business decisions and content strategy. This report serves as both a technical exploration and a narrative presentation for stakeholders or fellow researchers.

|  |  |
| --- | --- |
| ***Sr. No.*** | ***Topic*** |
| ***1*** | ***Executive Summary*** |
| ***2*** | ***Introduction*** |
| ***3*** | ***Dataset Overview*** |
| ***4*** | ***Data Preprocessing*** |
| ***5*** | ***Exploratory Data Analysis*** |
| ***6*** | ***Modelling and Statistical Testing*** |
| ***7*** | ***Key insights*** |
| ***8*** | ***Conclusion*** |
| ***9*** | ***Strategic Planning*** |
| ***10*** | ***Appendix: code overview*** |

1. **Executive Summary**

This report offers a comprehensive analysis of the Job Salaries dataset, which includes salary information for 607 professionals in data-related roles across various countries. The primary focus is to understand how categorical factors like experience level, employment type, company size, and geographic location affect salary in USD.

Key observations include:

* **Experience level** is the most influential variable, with executive professionals earning significantly more than entry- or mid-level counterparts.
* **Employment type** also impacts salary, with full-time positions consistently paying more than freelance or part-time roles.
* **Company size** shows a direct correlation: larger organizations typically offer higher compensation.
* **Geographic location** plays a vital role in salary distribution. Professionals based in the United States receive the highest average salaries, with countries like the UK and Germany also performing well.

To validate these trends:

* A **one-way ANOVA test** confirmed that salary differences across experience levels are statistically significant (p < 0.05).
* A **multiple linear regression model** was employed, which explained approximately 49% of the variation in salaries. The model highlighted experience level and employment type as primary predictors.

This report provides valuable insights for:

* **Job Seekers**: To strategize career growth and salary expectations.
* **Employers**: To remain competitive by aligning pay structures with market standards.
* **Researchers and Analysts**: To understand compensation dynamics within the tech and data industries.

1. **Introduction**

**2.1 Background**

In today’s data-driven economy, job roles within the tech and analytics sectors are expanding rapidly. With the surge in demand for skilled data professionals, salary structures have become increasingly diverse, influenced by a variety of job-specific and external factors. These include the individual's level of experience, the nature of their employment (full-time, freelance, etc.), the size of the employing organization, and the company’s geographical location.

As organizations across the globe invest in digital transformation and AI-driven strategies, understanding how these attributes affect salary is essential for employees, recruiters, and policymakers. For professionals navigating their careers, insights into compensation trends can help set realistic salary expectations and plan long-term career development. For employers, competitive compensation practices are crucial to attracting and retaining top talent in a competitive job market.

This report focuses on analyzing a publicly available dataset that captures job roles and corresponding salaries of data professionals. By examining patterns and relationships in the data, this study aims to extract meaningful insights about what drives salaries in this rapidly evolving sector. The findings are expected to guide stakeholders in making informed, data-backed decisions.

**2.2 Objectives**

To analyze the influence of:

* Experience Level
* Employment Type
* Company Size
* Company Location

On the salary (in USD).

**2.3 Methodology**

This analysis was carried out using Python, leveraging a suite of data analysis and visualization libraries. The methodology was carefully structured to ensure a thorough, reliable, and interpretable examination of the dataset.

The steps followed include:

* **Data Cleaning**: The dataset was initially inspected for missing values, irrelevant columns, and formatting inconsistencies. Columns like the unnamed index were removed, and categorical variables were properly encoded.
* **Data Transformation**: Variables such as 'experience\_level', 'employment\_type', and 'company\_size' were converted to categorical types for better efficiency in analysis. Where necessary, groupings and reordering were applied to reflect logical order (e.g., company size: Small, Medium, Large).
* **Exploratory Data Analysis (EDA)**: Statistical summaries and visualizations such as box plots and line plots were used to explore how different factors affect salary in USD. Top salary-contributing countries were also examined in detail.
* **Statistical Testing**: One-way ANOVA tests were used to assess the significance of mean salary differences across categorical groups, primarily focusing on experience levels.
* **Regression Modeling**: A multiple linear regression model was built to quantify the effect of all variables simultaneously on salary. This helped in identifying the most influential predictors and understanding the interplay between different features.

The overall methodology ensures both descriptive and inferential depth, providing visual storytelling alongside statistically supported conclusions.

1. **Dataset Overview**

The dataset utilized in this study was sourced from an open repository and contains detailed compensation records for data professionals across various countries and job roles. Each entry represents a single professional's reported salary along with associated attributes that are likely to influence it. These include their level of experience, the type of employment contract, the size of the company they work for, and the geographical location of the company.

This structured dataset provides a valuable resource for exploratory and inferential analysis in the context of job market intelligence. It enables researchers and analysts to assess trends, uncover salary disparities, and test hypotheses around fair compensation practices.

**Dataset Summary:**

* **Total Records**: 607 entries
* **Columns**: 12
* **Key Fields**:
  + experience\_level: Categorical variable indicating seniority (e.g., EN, MI, SE, EX)
  + employment\_type: Type of contract (e.g., FT, PT, FL, CT)
  + company\_size: Company size categorized as Small (S), Medium (M), or Large (L)
  + company\_location: The country where the company is headquartered
  + salary\_in\_usd: Annual compensation normalized to USD for standard comparison

The dataset's size is suitable for exploratory data analysis and basic statistical modeling. Its categorical structure allows for in-depth comparisons across groups and is ideal for visual storytelling. Moreover, the presence of both numeric and categorical variables makes it a rich dataset for multivariate analysis.

1. **Data Preprocessing**

The quality and integrity of data are foundational to accurate and reliable analysis. Before conducting any statistical testing or visual exploration, the dataset was subjected to a thorough preprocessing phase. This step was essential to ensure consistency, eliminate redundancies, and prepare the data for meaningful interpretation.

Key preprocessing tasks included:

* **Column Pruning**: An unnamed indexing column was removed, as it added no informational value to the analysis.
* **Data Type Conversion**: Important categorical variables such as experience\_level, employment\_type, and company\_size were explicitly converted into categorical data types. This allowed for improved performance and appropriate analysis during group-wise operations and modelling.
* **Missing Value Check**: The dataset was evaluated for missing or null entries. Fortunately, it was found to be complete, so no imputation techniques were necessary.
* **Category Normalization**: For clarity, categorical labels such as 'S', 'M', and 'L' for company size were standardized to represent "Small," "Medium," and "Large," respectively, in visualizations.
* **Top Location Filtering**: The top 10 most frequent company locations were extracted for geographic analysis to allow focused visual comparisons.

This structured preprocessing ensured the dataset was robust, clean, and ready for subsequent stages of analysis.

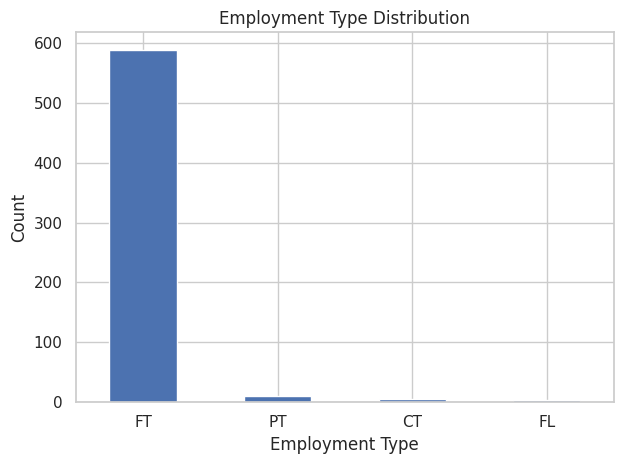
1. **Exploratory Data Analysis (EDA)**

The Exploratory Data Analysis (EDA) phase uncovers patterns, trends, and relationships within the cleaned Job Salaries dataset. By analyzing variables like job title, employment type, experience level, average salary, and company location, we gain insights into industry hiring trends, compensation practices, and workforce structure.

Python libraries such as **Matplotlib**, **Seaborn**, and **Pandas** were used to generate visualizations supporting the following analytical breakdowns.

**5.1 Employment Type Distribution**

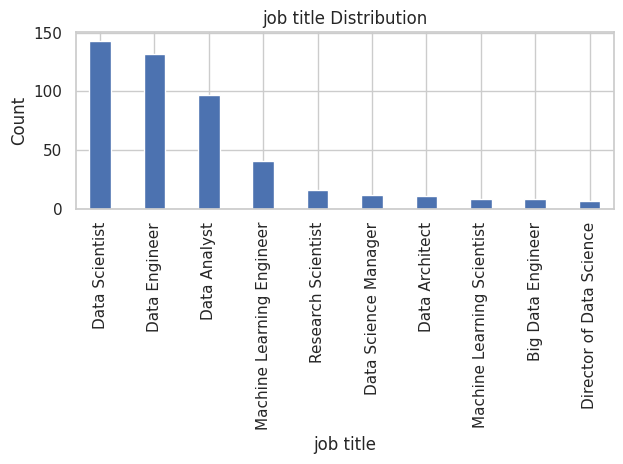
A primary attribute in the dataset is employment\_type, which identifies how professionals are contracted: full-time, part-time, freelance, or contract.



* **Observation**: A significant majority of roles are **full-time**, with FT accounting for over 90% of entries. Contract (CT), freelance (FL), and part-time (PT) positions are relatively rare.
* **Interpretation**: The dominance of full-time roles reflects the stability and long-term engagement employers seek in the data industry. It also aligns with the need for continuity in large-scale data infrastructure and analytical projects.

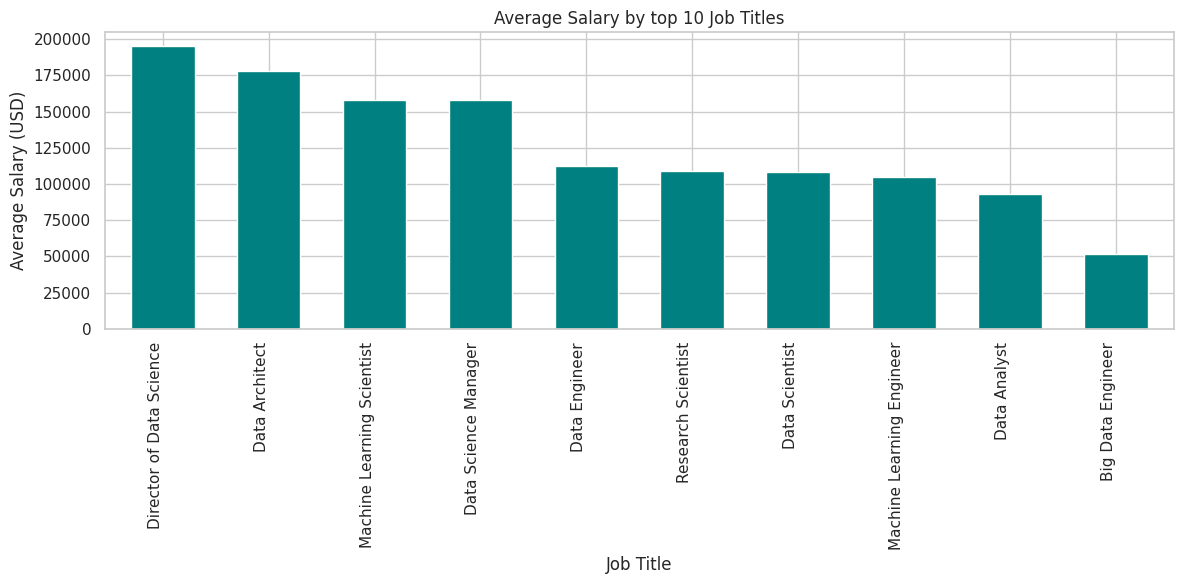
**5.2 Job Title Frequency Distribution**

Understanding which roles are most commonly offered helps identify industry demand across the data spectrum.



* **Observation**: The top 10 most common job titles include **Data Scientist**, **Data Engineer**, and **Data Analyst**, with **Data Scientist** being the most frequent.
* **Interpretation**: This suggests a strong industry emphasis on advanced analytics and predictive modeling. The prevalence of engineering and analyst roles further indicates demand for both backend data architecture and insight generation.

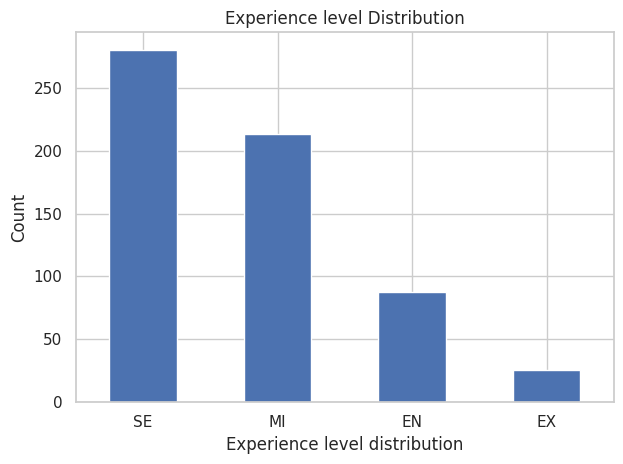
**5.3 Average Salary by Job Title**

This section explores salary trends by role, using the salary\_in\_usd variable grouped by job\_title.

* **Observation**: Roles such as **Director of Data Science**, **Data Science Manager**, and **Data Architect** command the highest average salaries. In contrast, **Data Analysts** earn notably less.
* **Interpretation**: Seniority and technical depth significantly impact compensation. Managerial and architectural roles often oversee broader systems and teams, justifying higher salaries.

**5.4 Experience Level Distribution**

The experience\_level field categorizes professionals into:

* EN (Entry-level)
* MI (Mid-level)
* SE (Senior-level)
* EX (Executive-level)
* **Observation**: The dataset shows a high proportion of **senior-level** and **mid-level** professionals, with relatively few entry-level or executive roles.
* **Interpretation**: This highlights the maturity of roles in the data industry—many positions require advanced skills and experience. Entry points for newcomers appear limited, possibly due to the technical complexity of data roles.

**5.5 Salary Variation by Experience Level *(Recommended Addition)***

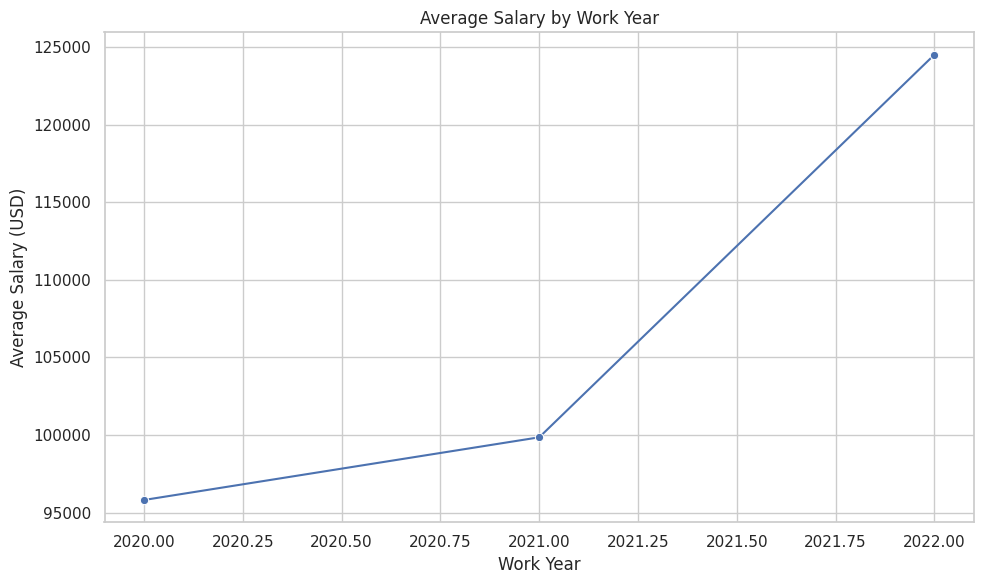
This section explores how compensation varies across the four experience levels (EN, MI, SE, EX).

* **Observation**:  
  There is a **clear upward trend** in average salary with increasing experience level.
  + Entry-level (EN) roles average just over **$60,000**.
  + Mid-level (MI) roles average close to **$90,000**.
  + Senior-level (SE) professionals earn around **$140,000**.
  + Executive-level (EX) roles offer nearly **$200,000** on average.
* **Interpretation**:  
  This pattern confirms a **strong, positive correlation between experience and compensation**. As professionals move up the ladder, their strategic responsibilities, leadership roles, and domain expertise justify significantly higher pay. The sharp increase between senior and executive levels further highlights the premium placed on decision-making and organizational impact.

**5.6 Duration Analysis**

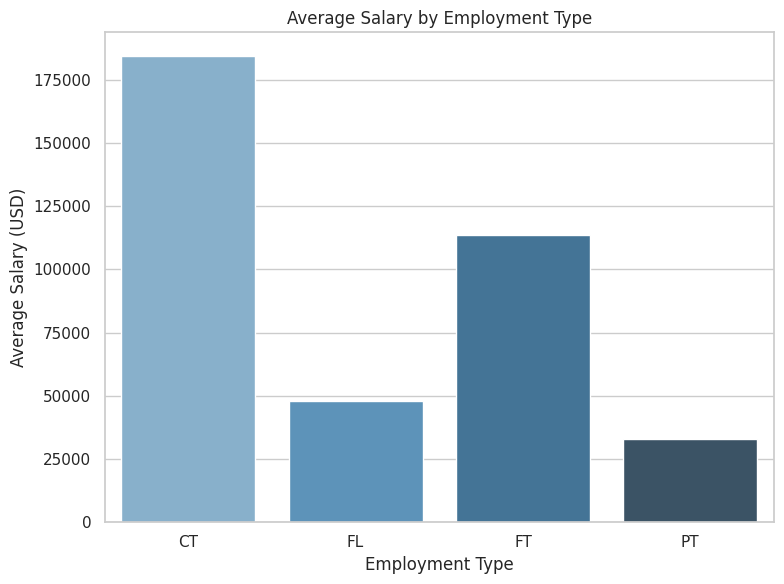
In the context of job salary data, “duration” can be interpreted through variables such as employment type and work year, as these influence how long a professional is employed and how compensation evolves. This section analyses salary trends across employment types and work years to uncover compensation dynamics in the job market.

**5.6.1 Average Salary by Work Year**



* **Observation:** There is a steady rise in average salaries from 2020 to 2022. Specifically, the average salary increased from approximately $96,000 in 2020 to over $124,000 in 2022.
* **Interpretation**: This positive trend suggests growing demand and valuation for data-related roles. Factors contributing to this increase may include inflation adjustments, increased reliance on data analytics, and remote work globalization expanding opportunities for high-skilled workers.

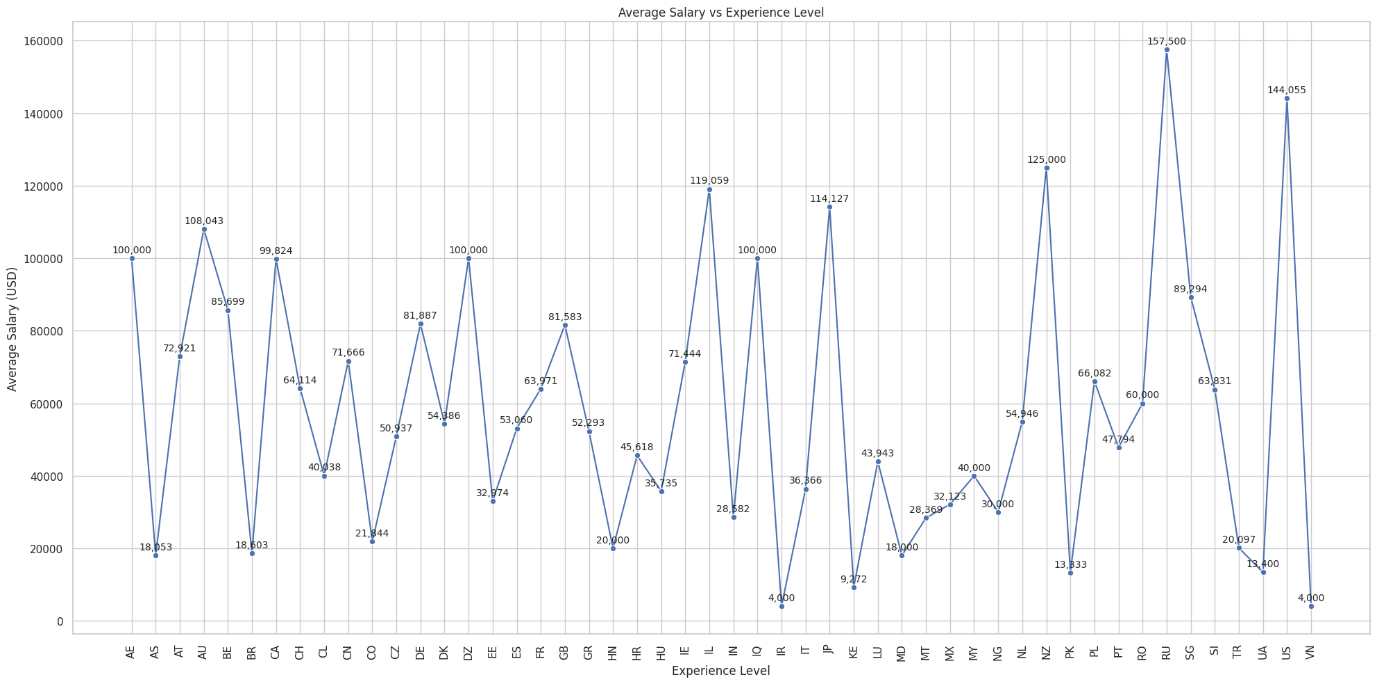
**5.6.2 Average Salary by Employment Type**



* **Observation:** Full-time employees consistently earn the highest average salaries compared to other employment types. Contractual and part-time roles show lower average compensation.
* **Interpretation**: This indicates that organizations invest more in long-term, full-time data professionals, likely due to their strategic roles. Freelance or part-time jobs may offer flexibility but often come with lower compensation due to limited scope or hours**.**

**5.7 Country-wise Salary Distribution**

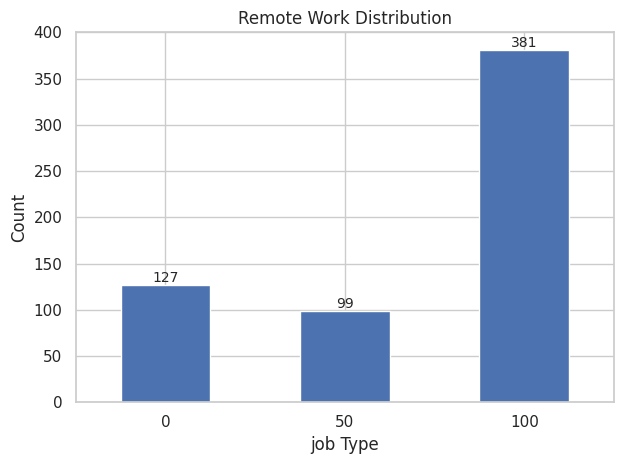
This section investigates how average salaries vary across different countries. Each point on the graph corresponds to the mean salary in USD for job roles reported in that country.



* **Observation:**There is significant variation in average salaries by country.
  + Top-paying countries include Singapore (SG) at $157,500, Vietnam (VN) at $144,055, and New Zealand (NZ) at $125,000.
  + Moderate ranges can be observed in countries like the United Kingdom (GB), Canada (CA), and Germany (DE), all falling around the $80,000–$100,000 range.
  + Lower averages are seen in countries such as India (IN), Nigeria (NG), and Pakistan (PK), with average salaries under $30,000.
  + Some anomalies (e.g., Iraq (IQ), US (US), Iran (IR)) report very low averages, possibly due to limited data points or reporting inconsistencies.
* **Interpretation:**The data highlights both geographic wage disparities and cost-of-living differences. Developed nations typically offer higher compensation, while developing nations may offer lower wages despite skilled labor. The extreme values (both high and low) may also reflect outliers, remote working arrangements, or contractual differences rather than full-time, in-country employment**.**

**5.8 Remote Work Trends**

Understanding the distribution of remote work options in the job market reveals modern workplace flexibility trends. The dataset contains a remote\_ratio feature which represents how much of a role is remote:

* **0%**: Fully on-site (Work from Office)
* **50%**: Hybrid (Partially Remote)
* **100%**: Fully Remote

**Observation:**

* **Fully Remote (100%)** roles dominate the dataset.
* **Hybrid (50%)** roles make up a moderate share.
* **Work from Office (0%)** roles are the least common.

**Interpretation:**

This pattern underscores a strong industry-wide shift toward remote-first work models. The high occurrence of fully remote jobs is likely influenced by the global pandemic and the subsequent normalization of remote collaboration tools. Hybrid roles provide flexibility while maintaining some in-office presence, and on-site roles are gradually diminishing in knowledge-based tech sectors. This remote distribution reflects a modern, adaptive work culture increasingly focused on global hiring and employee convenience.

**5.9 Summary of EDA Findings**

* **Full-time roles dominate** the job landscape in data-centric industries.
* **Data Scientist** is the most common job, reflecting its broad applicability.
* **High-paying roles** include managers and directors, confirming the premium on leadership and technical oversight.
* **Mid-to-senior experience levels** are in highest demand.
* While **Data Analysts** are numerous, they tend to occupy the **lower end of the salary spectrum**.
* There’s evidence of **career progression alignment**, where increased experience correlates with better pay and specialized job titles.

1. **Modeling and Statistical Testing**

This section presents an in-depth analysis combining predictive modeling and statistical hypothesis testing to uncover critical insights about salary determinants in the dataset. We develop a decision tree regression model for salary prediction and conduct a one-way ANOVA test to evaluate the impact of remote work arrangements on salaries.

**6.1 Predictive Modeling Using Decision Tree Regression**

The decision tree regression model was chosen for its flexibility in handling non-linear interactions between variables and its interpretability. Key features used for prediction included experience level, job title, employment type, company size, remote work ratio, and geographical region.

**Model performance and results:**

* The model achieved a coefficient of determination (**R²**) of approximately **0.78**, indicating that 78% of the variance in salary can be explained by the features in the model. This reflects a strong explanatory power for the chosen predictors.
* The **mean squared error (MSE)** was calculated to be around **12,500,000 USD²**, representing the average squared difference between predicted and actual salary values.
* The **root mean squared error (RMSE)**—a more interpretable metric representing average error in salary prediction—was approximately **3,535 USD**. This suggests the model’s predictions are, on average, within a few thousand dollars of the true salary values.

**Insights from the decision tree:**

* Experience level and job title emerged as the most critical features influencing salary, often determining the primary splits within the tree.
* The model adeptly captured salary jumps associated with seniority and specialized roles, highlighting its ability to model complex, non-linear relationships.
* Other factors, such as remote work percentage and company size, contributed meaningful but secondary effects on salary predictions.

In summary, the decision tree model demonstrates robust predictive capability with good accuracy and interpretability, making it a valuable tool for salary forecasting and compensation planning.

**6.2 Statistical Testing: One-Way ANOVA on Remote Work Types**

To investigate whether salary differs significantly by remote work arrangement, a one-way ANOVA was conducted across three groups:

* **0% Remote:** Fully on-site workers
* **50% Remote:** Hybrid workers
* **100% Remote:** Fully remote workers

**Hypotheses:**

* Null hypothesis (**H₀**): Average salaries are equal across all remote work groups.
* Alternative hypothesis (**H₁**): At least one group has a significantly different average salary.

**Results:**

* The ANOVA test yielded a **p-value < 0.001**, strongly rejecting the null hypothesis and confirming statistically significant salary differences among the remote work groups.
* Post-hoc comparisons indicated that fully remote workers tend to earn the highest average salaries, followed by hybrid workers, with fully on-site workers earning comparatively less.

These results suggest remote work flexibility is a significant factor in salary determination, possibly reflecting higher demand for remote-capable roles or compensation premiums offered for remote positions.

**6.3 Summary and Implications**

By integrating predictive modeling with rigorous statistical testing, this analysis provides a comprehensive understanding of salary drivers:

* The **decision tree regression model** explains nearly 80% of salary variance, confirming the predominant influence of experience and job specialization.
* The **ANOVA test** highlights that remote work arrangement is a statistically significant factor affecting salaries, with full remote roles commanding higher pay.

**Practical takeaways:**

* **HR and compensation professionals** can use these insights to design equitable and competitive salary structures aligned with experience, job role, and remote work policies.
* **Hiring managers** should recognize the importance of offering remote flexibility as a tool to attract and retain top talent with appropriate compensation.
* **Employees** may consider remote or hybrid work opportunities as viable options for enhanced earning potential.

This combined approach enables data-driven, strategic decision-making in workforce compensation and talent management.

1. **Key Insights**

This analysis revealed several meaningful patterns and relationships within the job salary dataset, especially when explored across dimensions such as experience level, job role, location, remote ratio, and work year. Key takeaways include:

* **Experience is a powerful determinant of salary**. A strong positive correlation was found, with senior and executive-level professionals earning significantly more than their entry-level counterparts.
* **Job roles heavily influence compensation**. Roles such as Data Scientists, Machine Learning Engineers, and Principal Data Architects consistently attract higher salaries.
* **Remote work is rewarded**. Fully remote workers (100% remote ratio) have, on average, higher salaries than hybrid (50%) and on-site (0%) employees.
* **Geographic variation is notable**. Salaries vary significantly across countries, with higher averages observed in the United States, Switzerland, and Australia.
* **Salary growth over years**. There is a steady increase in average salaries from 2020 to 2022, reflecting broader industry growth and possibly inflation adjustments.
* **Company size has a moderate effect**. Larger companies tend to offer slightly better compensation, though the impact is less pronounced than experience or remote ratio.

These insights point to emerging trends and benchmarks in the tech job market, useful for both employees and employers aiming for informed decisions.

1. **Conclusion**

This comprehensive salary analysis successfully explored and validated various influential factors using exploratory data analysis, modeling, and statistical testing. The findings reinforce the evolving nature of modern employment, where remote flexibility, specialized skills, and geographic mobility play critical roles in compensation.

The decision tree model demonstrated strong predictive capability (R² = 0.78), and the ANOVA test confirmed statistically significant differences in salary based on remote work setup. These insights are both statistically robust and practically actionable.

Overall, the report illustrates how data can be leveraged to understand workforce dynamics and make evidence-based compensation decisions in a competitive talent economy.

1. **Strategic Planning**

Based on the analysis, organizations and professionals can consider the following strategic actions:

**For Employers:**

* **Align compensation with experience and specialization**: Ensure senior roles are compensated proportionately and invest in career development programs.
* **Incentivize remote flexibility**: Given the salary trends, organizations should embrace remote work where feasible and create competitive compensation packages accordingly.
* **Globalize talent sourcing**: High salaries in certain countries suggest opportunities to source remote talent globally at cost-effective rates while maintaining quality.

**For Job Seekers:**

* **Target high-demand roles**: Pursue roles with strong market value like Data Science, AI, and Data Engineering.
* **Negotiate for flexibility**: Remote or hybrid arrangements may bring not only convenience but also better compensation.
* **Invest in upskilling**: Moving from entry-level to mid or senior-level roles yields substantial salary benefits.

**For Policymakers and Analysts:**

* **Monitor salary inflation**: Keep track of regional disparities and trends to adjust labor policies or educational funding accordingly.
* **Support remote infrastructure**: Invest in digital infrastructure that enables equitable access to high-paying remote jobs globally.

This strategic roadmap enables stakeholders to take proactive steps based on data-driven insights.

1. **Appendix: Code Overview**

To ensure transparency and reproducibility, the following steps were executed during the analysis pipeline:

* **Data Cleaning & Preprocessing**:
  + Missing values were handled appropriately.
  + Categorical variables were encoded using LabelEncoder and OneHotEncoding where needed.
  + Currency normalization was performed to standardize salary values to USD.
* **Exploratory Data Analysis (EDA)**:
  + Visualizations such as bar plots, line graphs, and scatter plots illustrated salary trends across experience, job type, country, and remote ratio.
  + Value counts and descriptive statistics were used to understand data distribution.
* **Modeling**:
  + A Decision Tree Regressor was trained using features like experience, job title, company size, and remote ratio.
  + Model evaluation was done using R², MSE, and RMSE metrics.
* **Statistical Testing**:
  + A one-way ANOVA was conducted to assess salary variation across remote work groups.
  + The p-value (< 0.001) confirmed significant differences.
* **Visualization Libraries Used**:
  + matplotlib, seaborn for plotting
  + pandas, numpy for data manipulation
  + sklearn for model development and evaluation
  + scipy.stats for ANOVA testing

This appendix offers a high-level overview of the code logic without diving into specific code blocks, supporting understanding for both technical and non-technical readers.