Analysis of Factors That Influence Salaries in the Global AI Industry

Author: Franciszek Tokarek

1. Project Goal and Dataset Description

This project aims to conduct a comprehensive regression analysis to identify the most significant factors influencing salaries in the global artificial intelligence (AI) job market. With the exponential growth of AI-driven industries, understanding what drives compensation in this field is not only relevant for professionals and hiring managers, but also for policymakers and global tech investors.

The analysis is based on a dataset containing job postings and salary data for AI-related roles across various countries, employment types, and experience levels. The dataset includes both **numerical variables** (e.g., salary, years_of_experience, benefits_score) and **categorical variables** (e.g., job_title, country, employment_type, company_size), making it well-suited for multivariate regression analysis.

The **dependent variable** (also known as the response or target variable) is salary.

The analysis aims to assess how well this outcome can be predicted based on a set of **independent variables**, such as:

- Job title (e.g., Data Scientist, ML Engineer)
- Experience level (e.g., Junior, Senior, Executive)
- Geographic location (country)
- Company size and employment type
- Quantitative company-level indicators such as benefits score or remote ratio

Through data cleaning, exploratory analysis, and the development of multiple regression models, the project ultimately seeks to **build a predictive model** and **derive actionable insights** into the AI salary structure worldwide.

2. Data Preparation

To ensure the reliability and accuracy of the regression analysis, several preprocessing steps were applied to the raw dataset. First, duplicate entries were identified and removed to prevent data leakage and ensure model integrity. Missing values were systematically addressed — either imputed using appropriate strategies or removed if deemed non-informative or sparse.

Categorical features such as job_title, country, employment_type, and company_size were encoded using **one-hot encoding**, enabling their use in regression models while preserving category relationships. Numerical variables, where necessary, were **standardized or normalized** to ensure consistent scale and improve model convergence.

Furthermore, data types were explicitly converted for proper downstream processing (e.g., categorical → category, date fields → datetime). Outliers and extreme values in the salary column were retained intentionally, given their relevance in identifying compensation extremes within the global AI sector.

The final, cleaned dataset was saved as **ai_job_dataset_clean.csv** and used throughout the exploratory and modeling stages of the project.

3. Exploratory Data Analysis

To understand the structure and distribution of salaries in the AI industry, exploratory data analysis (EDA) was conducted using both numerical and categorical variables. This step provided critical insights before modeling, including feature relationships, skewness, outliers, and group-level effects.

3.1. Salary distribution

Descriptive statistics for the target variable salary revealed a right-skewed distribution. The majority of salaries were concentrated between $40,000 and $120,000 USD, with a long tail indicating a smaller proportion of high-paying roles.

Figure 3.1: Histogram of salary distribution

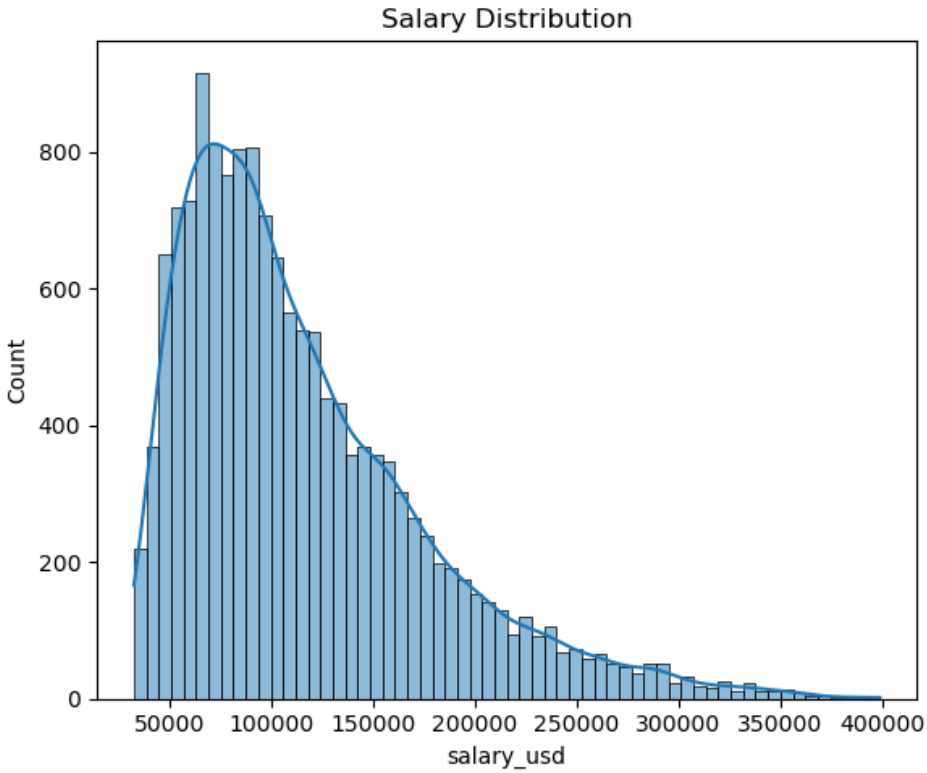
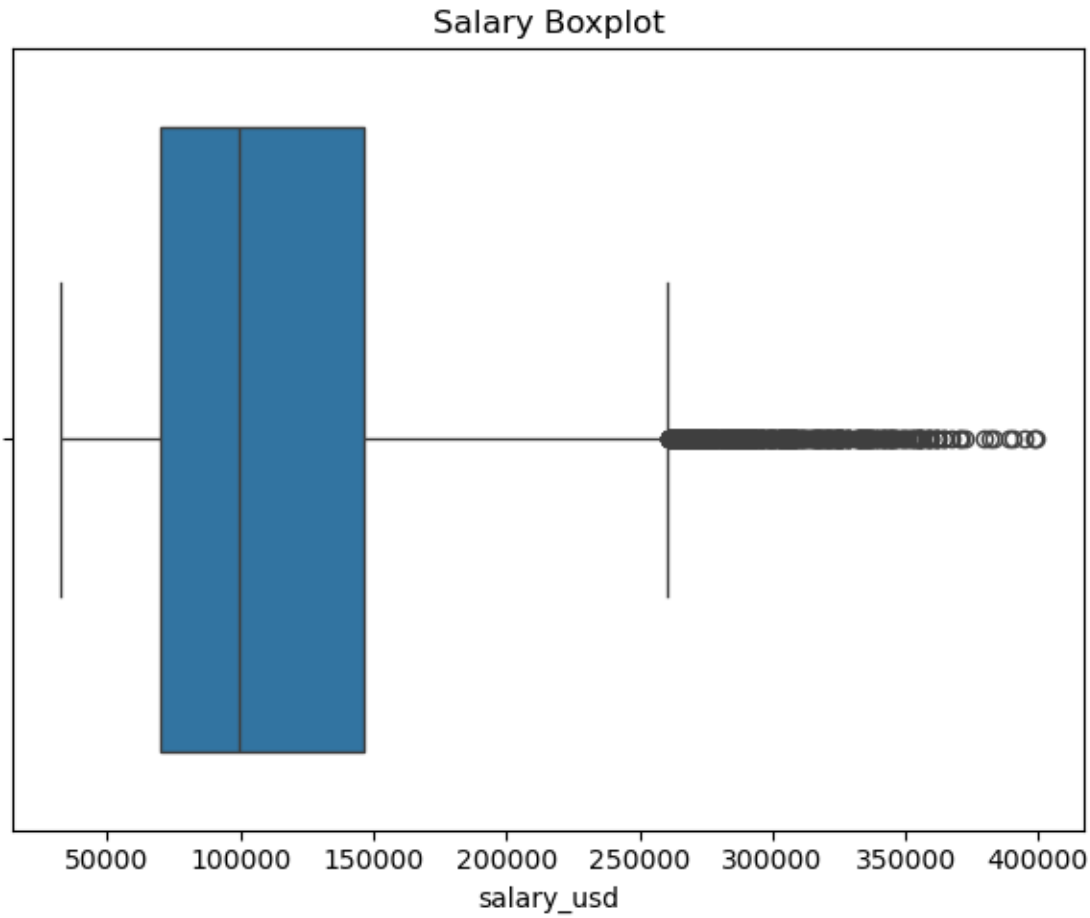


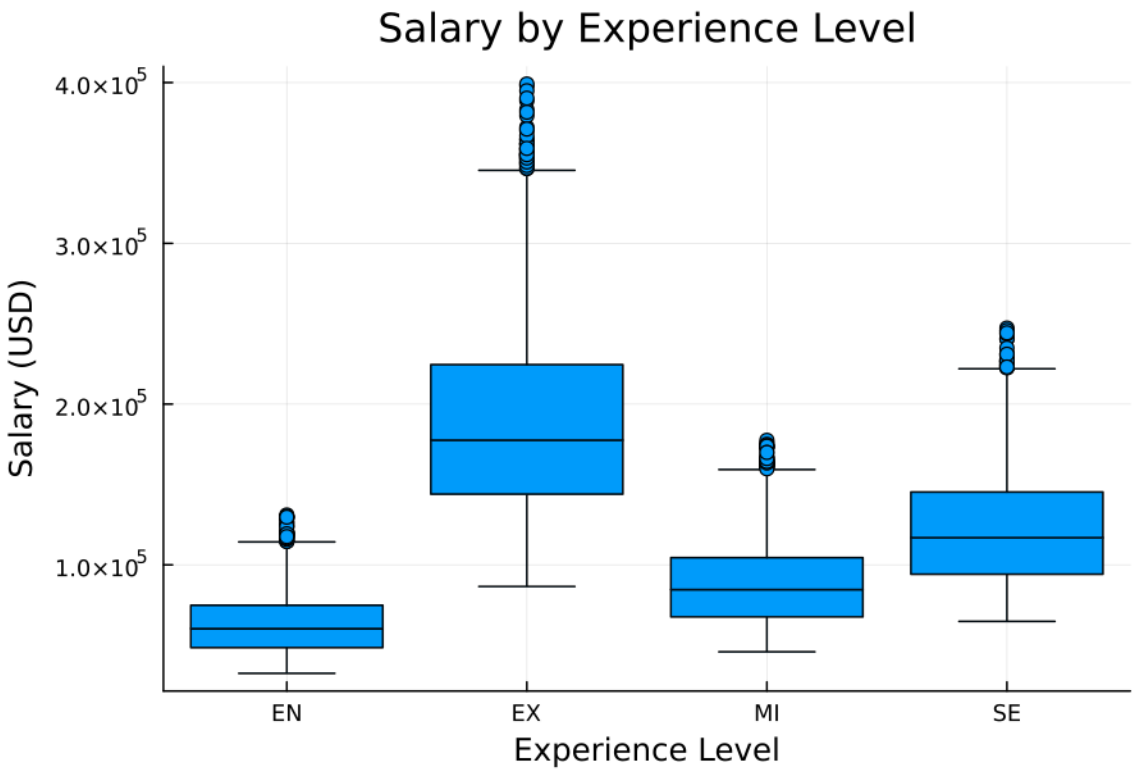
Figure 3.2: Boxplot of salary values including outliers



3.2. Salary by categorical features

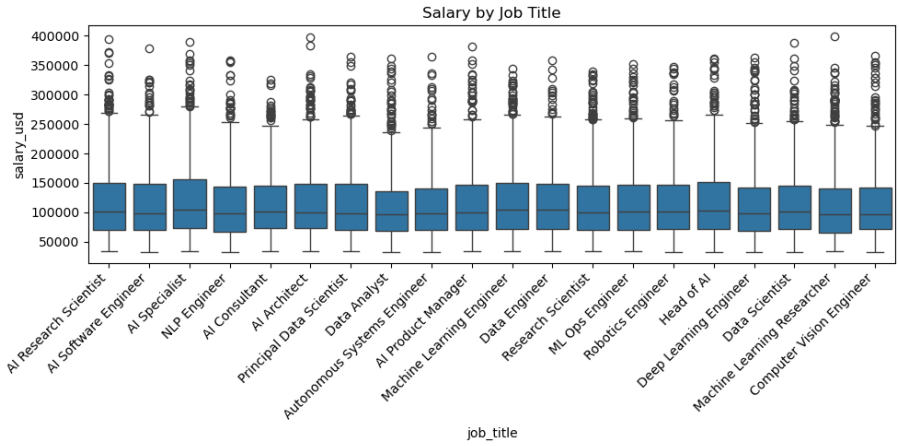
To assess how salary varies across key qualitative dimensions, group-based boxplots were generated.

Figure 3.3: Salary by experience level



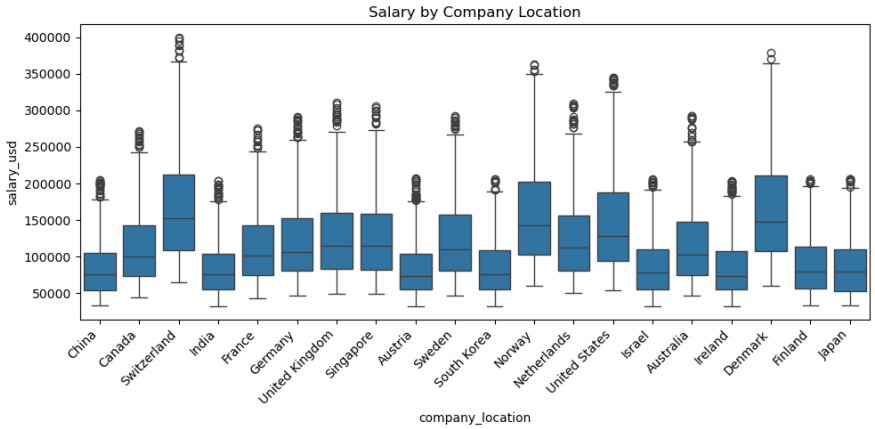
Employees with higher experience levels (Senior, Executive) showed significantly greater median salaries compared to entry-level professionals.

Figure 3.4: Salary by job title



Technical and specialist roles such as Machine Learning Engineer and AI Researcher reported higher compensation than business or analyst roles.

Figure 3.5: Salary by company location



Geographical differences were substantial. U.S. and Swiss companies offered the highest salary medians across roles.

Figure 3.6: Salary vs years of experience

The relationship between salary and experience is positively correlated, albeit with diminishing returns beyond 10–15 years.

Figure 3.7: Salary vs benefits score

Higher benefits score tends to correlate with higher salaries, suggesting that overall compensation packages are interconnected.

These exploratory analyses highlight key structural patterns in the AI job market: compensation is not only a function of technical skills and seniority, but also strongly influenced by geographical and organizational context. These findings informed the feature selection process for regression modeling.

4. Statistical Verification

To validate the statistical assumptions necessary for regression modeling, the distribution of the target variable (salary) was first examined for normality.

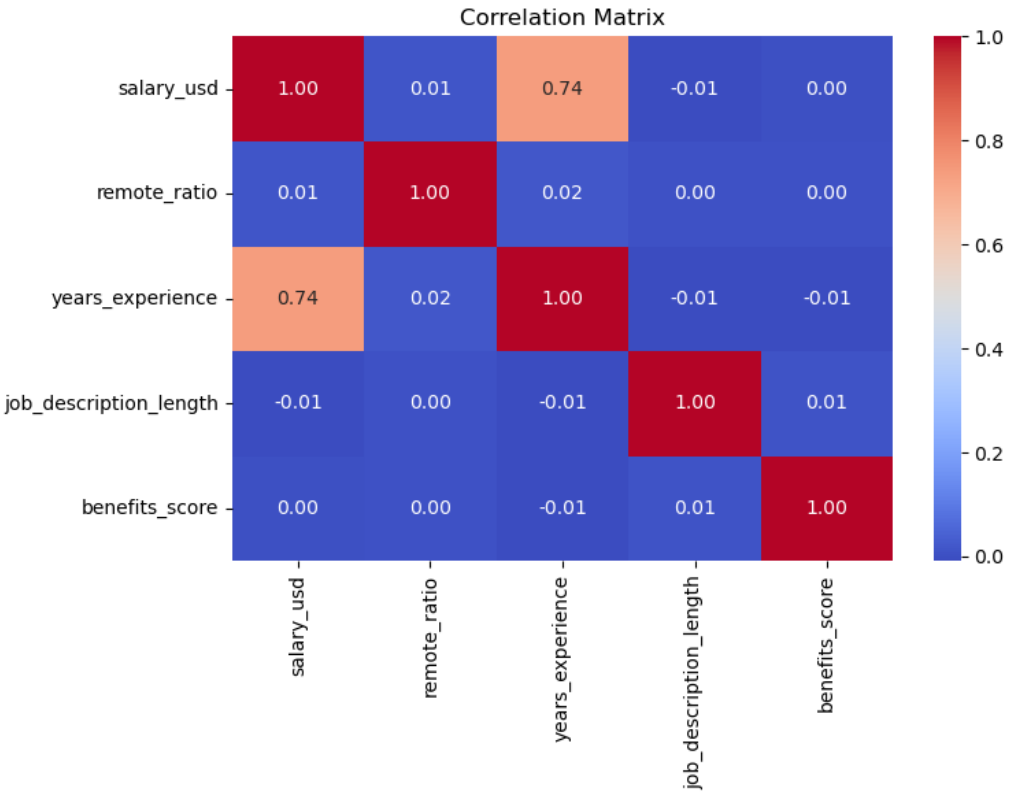
The **Shapiro-Wilk test** was applied, yielding a p-value well below 0.05, which suggests that the salary variable does **not follow a normal distribution**. This was visually confirmed using a **QQ-plot**, where the points deviated significantly from the theoretical normal line—particularly in the tails.

Figure 4.1: QQ-plot of salary distribution

Interpretation: The non-normality confirms the need for robust models less sensitive to distributional assumptions, such as decision trees.

In addition to examining the target variable, pairwise correlations were computed between numerical features. A **correlation matrix** and associated **heatmap** highlighted moderate relationships, such as between years_of_experience and salary, as well as between benefits_score and salary. While no variables were strongly collinear, these relationships informed feature selection for regression modeling.

Figure 4.2: Correlation matrix heatmap



Interpretation: The matrix shows meaningful, but not redundant, relationships — supporting the inclusion of all features in initial models.

5. Regression Modeling

To evaluate the predictive power of various features on salary, the cleaned dataset was split into training and testing subsets using an 80/20 ratio. Multiple regression models were implemented to capture both linear relationships and complex nonlinear patterns in the data.

The following models were trained and compared:

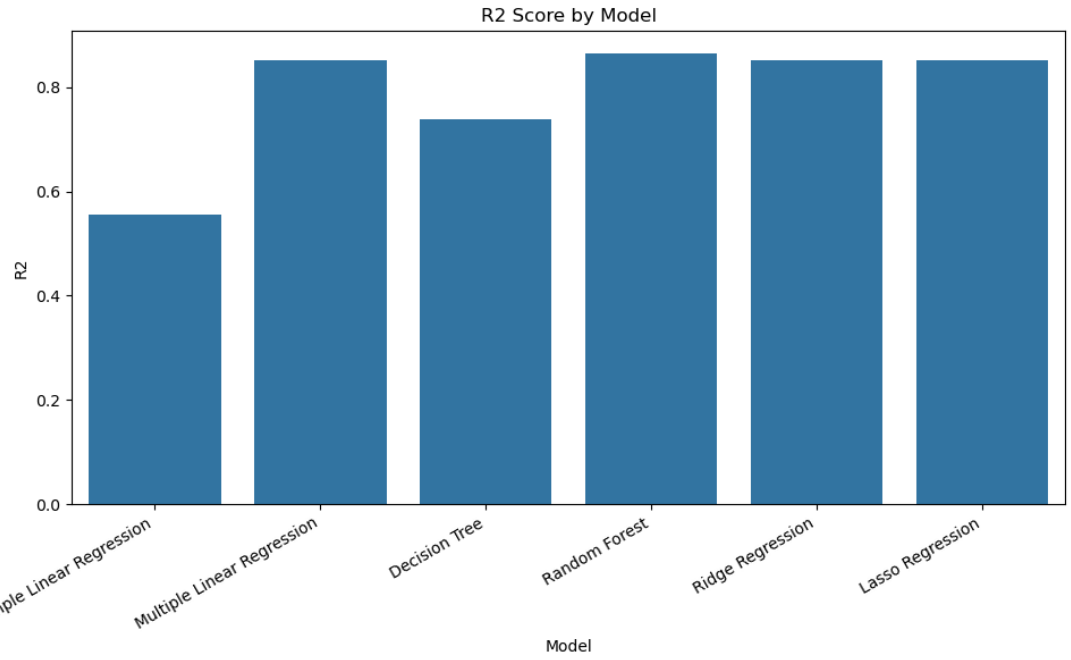
- **Linear Regression** – to establish a baseline model and interpret coefficient significance.
- **Ridge and Lasso Regression** – to handle multicollinearity and perform regularization.
- **Decision Tree Regressor** – to capture non-linear interactions and segment-based predictions.
- **Random Forest Regressor** – an ensemble method that improves accuracy and generalization by averaging multiple decision trees.

5.1 Model Performance Evaluation

Each model was evaluated on both training and testing datasets using standard regression metrics:

- **R^2 (coefficient of determination)** – proportion of variance explained,
- **MAE (Mean Absolute Error)** – average absolute prediction error,
- **MSE (Mean Squared Error)** and **RMSE (Root Mean Squared Error)** – penalize larger errors more strongly.

Figure 5.1: Model performance comparison (R^2, MAE, MSE, RMSE)

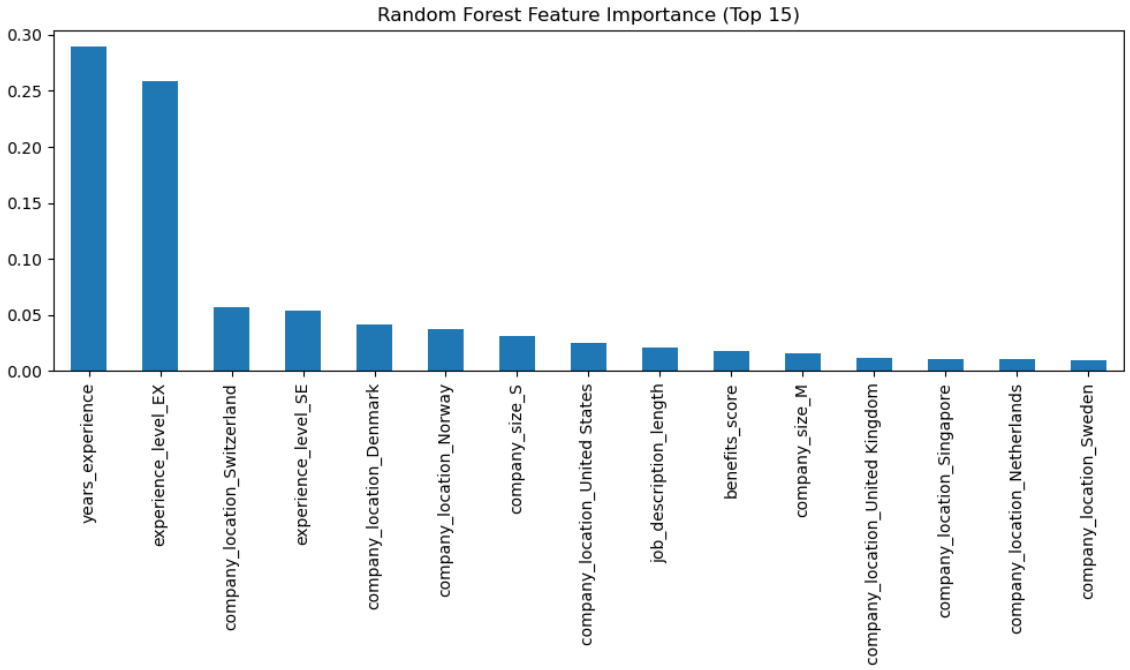


The Random Forest model achieved the highest R^2 and the lowest RMSE, making it the best-performing model overall. Linear regression provided interpretable coefficients, but underfit the data compared to non-linear models.

5.2 Feature Importance

Tree-based models such as Random Forest allowed for the extraction of feature importance scores, which quantify the contribution of each input variable to the predictive outcome.

Figure 5.2: Feature importance (Random Forest)



Interpretation: Variables such as job_title, experience_level, and country were among the most important predictors of salary.

6. Python vs Julia Results Comparison

To evaluate the versatility and performance of different programming environments for regression analysis, the project was implemented in both **Python** and **Julia**.

Python enabled a complete end-to-end workflow, including data preprocessing, exploratory data analysis, statistical verification, and modeling with both linear and ensemble methods. The availability of advanced libraries such as scikit-learn, seaborn, and statsmodels allowed for in-depth evaluation, interpretability (e.g., regression coefficients), and feature importance extraction from tree-based models.

Julia, on the other hand, was used to replicate key parts of the analysis, focusing primarily on linear regression. While the GLM.jl and DecisionTree.jl libraries provided a solid foundation for modeling, Julia lacked built-in support for feature importance in ensemble models. Additionally, statistical testing (e.g., Shapiro-Wilk) and some advanced visualizations were not directly available or required significant additional setup.

Despite these limitations, Julia successfully demonstrated the core relationships between features and salary, and confirmed key insights found in Python. However, for comprehensive and production-level regression workflows, **Python remains the more mature and flexible ecosystem**.

7. Conclusions

This regression analysis has identified several key factors that significantly influence salary levels in the global AI job market.

Based on both linear and tree-based models, the most impactful predictors include:

- **Job title** — specialist roles such as Machine Learning Engineer and AI Researcher consistently received higher compensation.
- **Experience level** — salaries increase sharply with professional seniority, particularly at the executive level.
- **Country of employment** — geographic disparities in salary were evident, with the United States, Switzerland, and similar markets leading.
- **Company-related factors** — such as benefits score and employment type also showed moderate effects on salary.

The **Random Forest Regressor** outperformed other models in terms of predictive accuracy (highest R^2, lowest RMSE), confirming its suitability for capturing complex interactions and nonlinear patterns. Linear regression, while less accurate, provided valuable interpretability and baseline metrics.

A comparative implementation in **Julia** confirmed the main insights but revealed limitations in functionality, especially regarding model diagnostics and feature importance extraction. **Python proved to be a more comprehensive environment** for end-to-end regression analysis.

In summary, the project successfully answered the research question:

"Which factors most significantly influence AI job salaries worldwide?"

These findings have practical implications for:

- AI professionals benchmarking salaries,
- recruiters setting competitive compensation ranges,
- policymakers analyzing international labor trends in the tech sector.

8. Repository and Resources

The full project, including all code (Python and Julia), cleaned datasets, Jupyter notebooks, visualizations, and the final report, is available on GitHub:

Project repository:

https://github.com/ftokarek/Analysis-of-Factors-That-Influence-Salaries-in-the-Global-AI-Industry

Final report:

Available as both .docx and .pdf in the /reports directory of the repository.

Cleaned dataset used in the analysis:

Stored as data/ai_job_dataset_clean.csv within the repository.

Original dataset source:

Global AI Job Market and Salary Trends 2025 – Kaggle