Iuliia Bolgova

Final Report: Analysis of the influence of socio-economic indicators of countries on salary trends in the fields of Data Science

1. Introduction

1.1 Problem Statement

The Data Science field has expanded significantly in recent years, leading to changes in wages and working conditions. It is important to understand how data science salaries correlate with various socio-economic indicators internationally. This research will help identify the relationship between professionals' income levels, countries' economic conditions, and the quality of life of the population, aiding professionals and organizations in making informed career and salary decisions.

1.2 Research Objective

The objective of this study is to understand the factors influencing data science salaries and develop a model to predict these salaries.

1.3 Datasets

1.3.1 Latest Data Science Salaries Dataset:

- Provides insights into compensation trends and variations in data science from 2020 to 2024.
- Columns include Job Title, Employment Type, Experience Level, Salary, Company Location, Salary in USD, and more.

Data Sources:
https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023

1.3.2 Global Country Information Dataset 2023:

- Offers information on demographic statistics, economic indicators, healthcare metrics, education statistics, etc.

- Key attributes include Population Density, GDP, Life Expectancy, Minimum Wage, Unemployment Rate, etc.

Data Sources:

https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023

2. Data Wrangling

The data wrangling process for this project involved two primary datasets: "Latest Data Science Salaries" and "Global Country Information Dataset 2023." The objective was to clean and merge these datasets to enable comprehensive analysis of data science salaries in relation to global socio-economic indicators.

2.1 Data Loading and Initial Inspection

2.1.1 Latest Data Science Salaries Dataset:

- **Files:** The dataset was provided in seven separate parts.
- **Combined Dataset:** After concatenation, the dataset contained 5,702 records and 11 columns.
- **Columns Included:** Job Title, Employment Type, Experience Level, Expertise Level, Salary, Salary Currency, Company Location, Salary in USD, Employee Residence, Company Size, and Year.

2.1.2 Global Country Information Dataset 2023:

- **Initial Records and Columns:** The dataset included detailed socio-economic indicators for 195 countries, with 42 columns.
- **Columns Included:** Population density, GDP, life expectancy, education enrollment rates, minimum wage, unemployment rate, etc.

2.2 Data Cleaning

2.2.1 Cleaning the Global Country Information Dataset:

- **Handling Missing Values:**
    - Columns containing unwanted characters (e.g., commas, percentages, dollar signs) were cleaned.
    - Missing values for "Minimum wage" were filled using external data sources or the median value for the dataset.
- **Removing Duplicates:**

    - Identified and removed duplicates, resulting in a cleaner dataset.
- **Addressing Specific Missing Values:**
    - Filled missing values for columns like "Official language" and "Currency code" using known values or external sources.
    - Removed rows with missing values in critical columns like "Life expectancy," "Gross primary education enrollment (%)," and "Tax revenue (%)."
- **Post-Cleaning Dataset:** The cleaned dataset contained 195 records and 42 columns.

2.2.2 Cleaning the Salaries Dataset:

- **Removing Duplicates:** After inspection, no duplicates were found in the dataset.
- **Correcting Data Types:** Ensured consistency by correcting data types for columns like "Minimum wage," "Population," and "GDP."
- **Handling Missing Values:** Verified that no missing values remained in any columns post-cleaning.
- **Post-Cleaning Dataset:** The cleaned dataset contained 5,702 records and 27 columns.

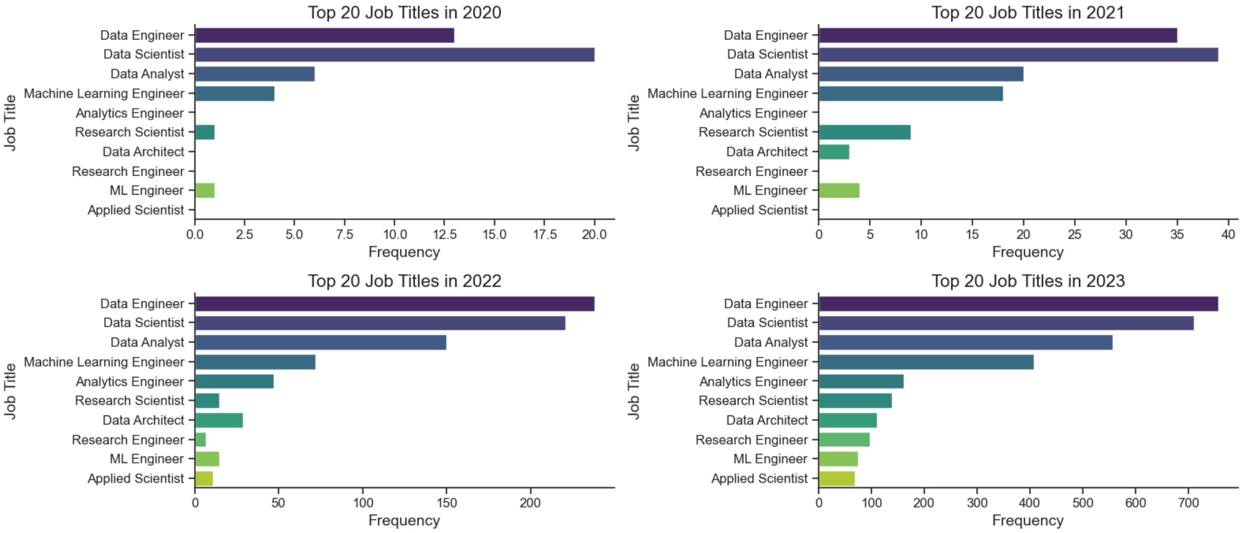
2.2.3 Merging the Datasets

- **Merging Process:** The cleaned "Latest Data Science Salaries" dataset was merged with the "Global Country Information Dataset 2023" using the company location to match countries.
- **Resulting Dataset:** The merged dataset contained 5,702 records and 47 columns, combining salary data with comprehensive socio-economic indicators.

3. Exploratory Data Analysis

3.1 Top 20 Job Titles by Year:

Top 20 job titles by frequency were plotted for the years 2020-2023. Note that 2024 was not included due to insufficient data.

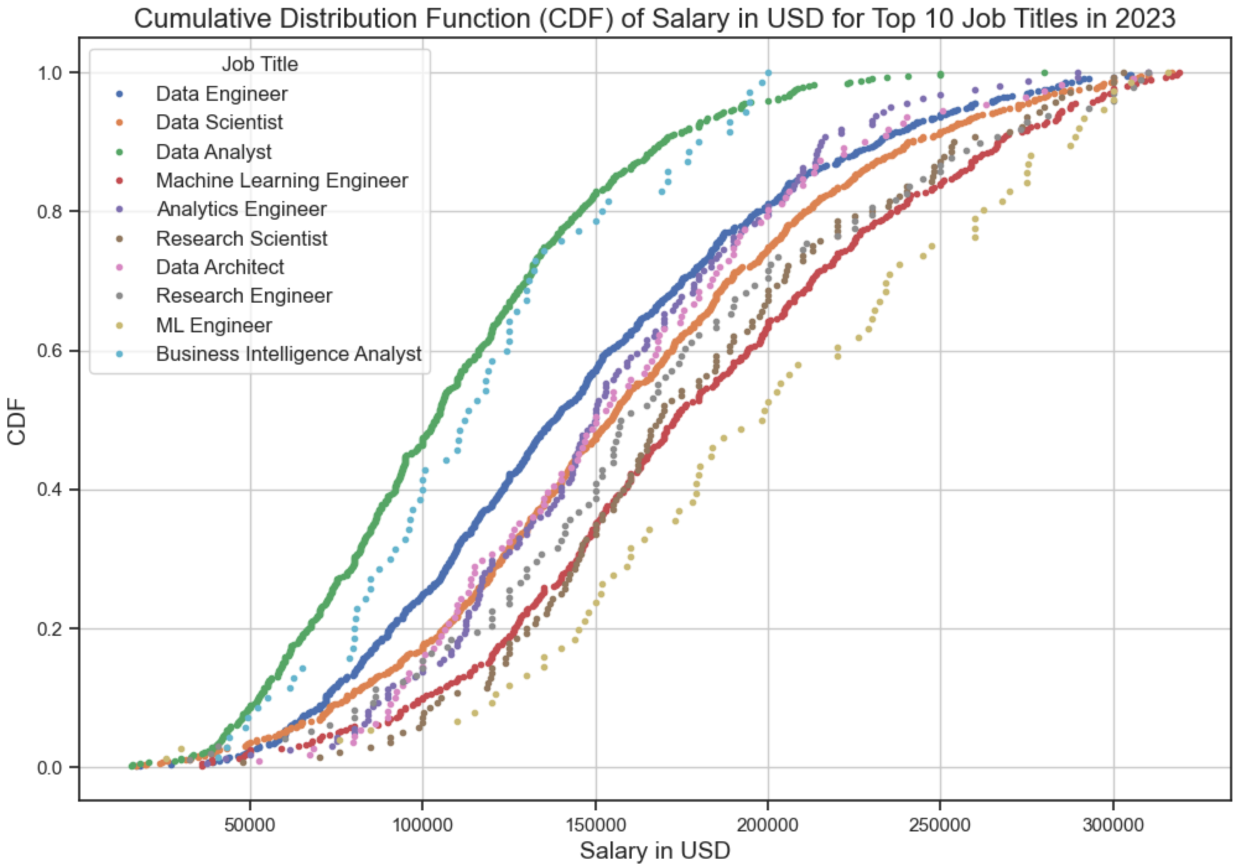


2020-2021: Data Engineer and Data Scientist consistently lead in the number of records. Increasing popularity of Data Analyst and Machine Learning Engineer.

2022-2023: The demand for Data Engineer and Data Scientist continues to grow. Significant growth in Data Analyst and Machine Learning Engineer positions. New popular job titles include Data Architect and Research Engineer.

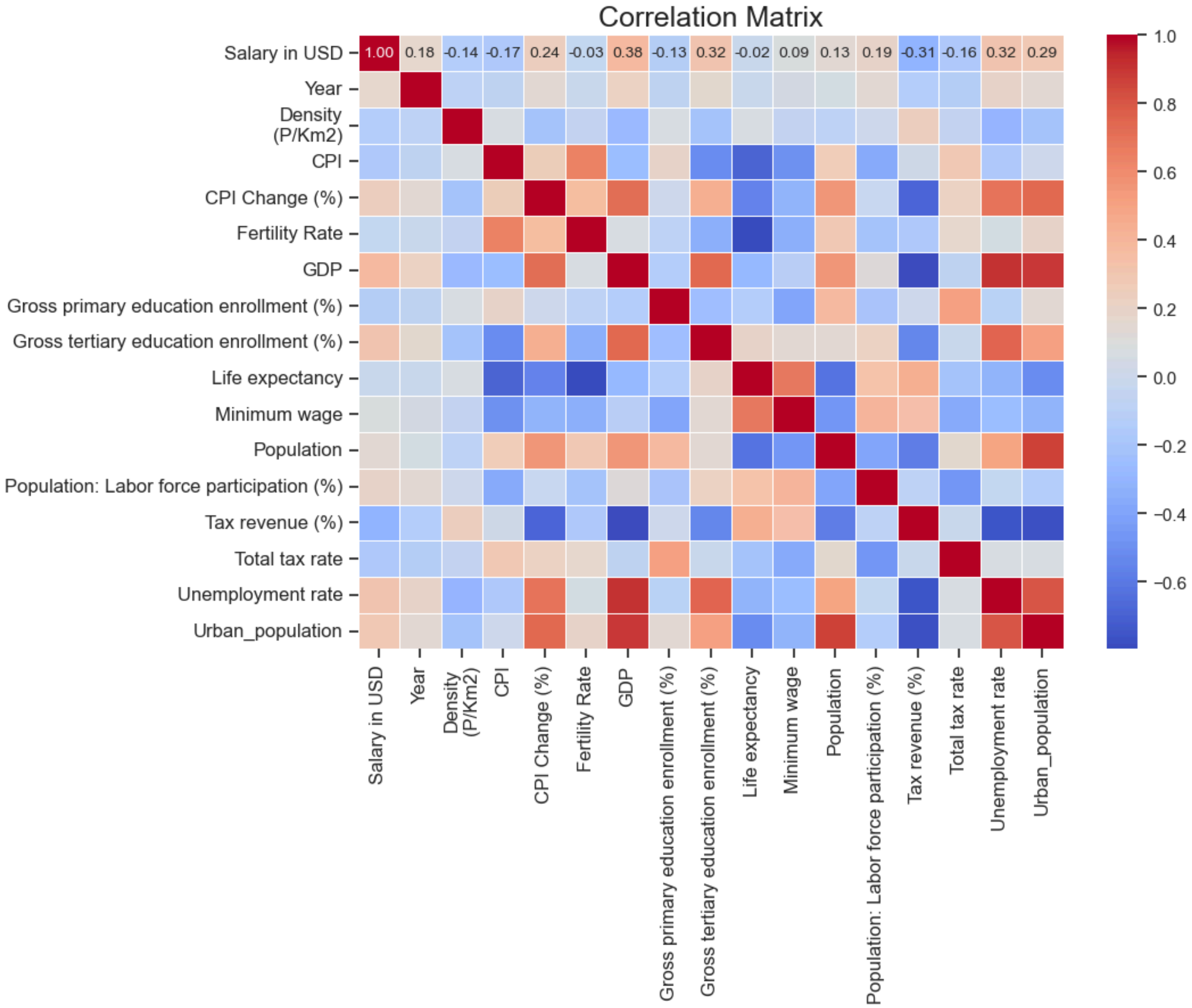
Conclusion: Data Engineer and Data Scientist remain the most in-demand job titles. Analytics and machine learning positions are also gaining popularity.

3.2 Cumulative Distribution Function (CDF) of Salary in USD for Top 10 Job Titles in 2023:



In 2023, salaries vary significantly across roles. Data Engineers and Data Scientists are the most frequent job titles, with most salaries clustering between $100,000 - $200,000 USD.

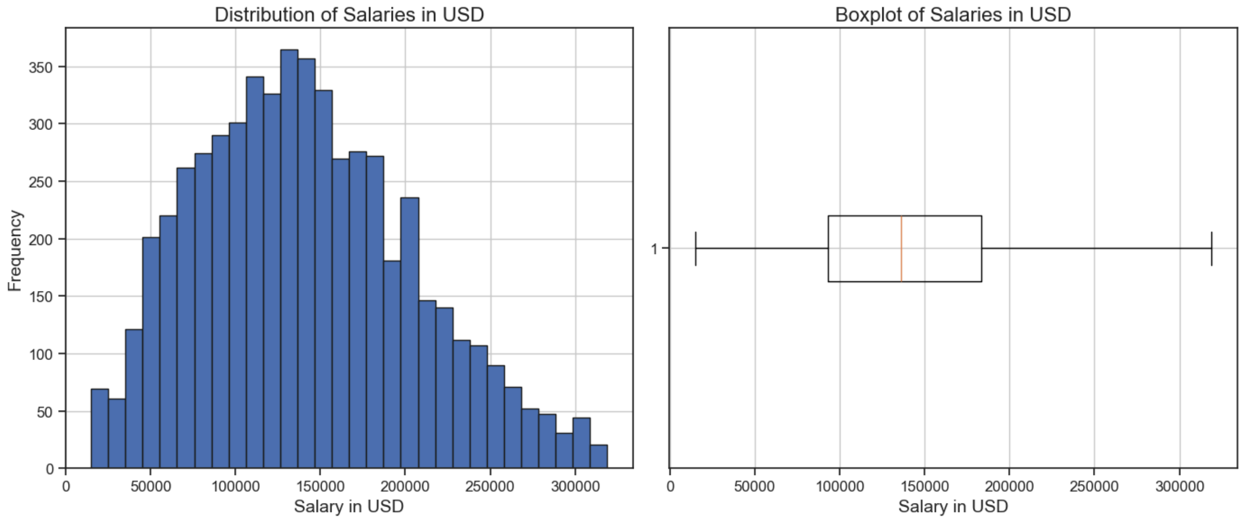
3.3 Correlation Analysis



The highest positive correlation with "Salary in USD" is with "Minimum wage" (0.35) and "Population: Labor force participation (%)" (0.28). Negative correlations include "Unemployment rate" (-0.15) and "CPI Change (%)" (-0.12). This suggests higher salaries are associated with higher minimum wages, greater labor force participation rates, lower unemployment rates, and more stable consumer price indexes.

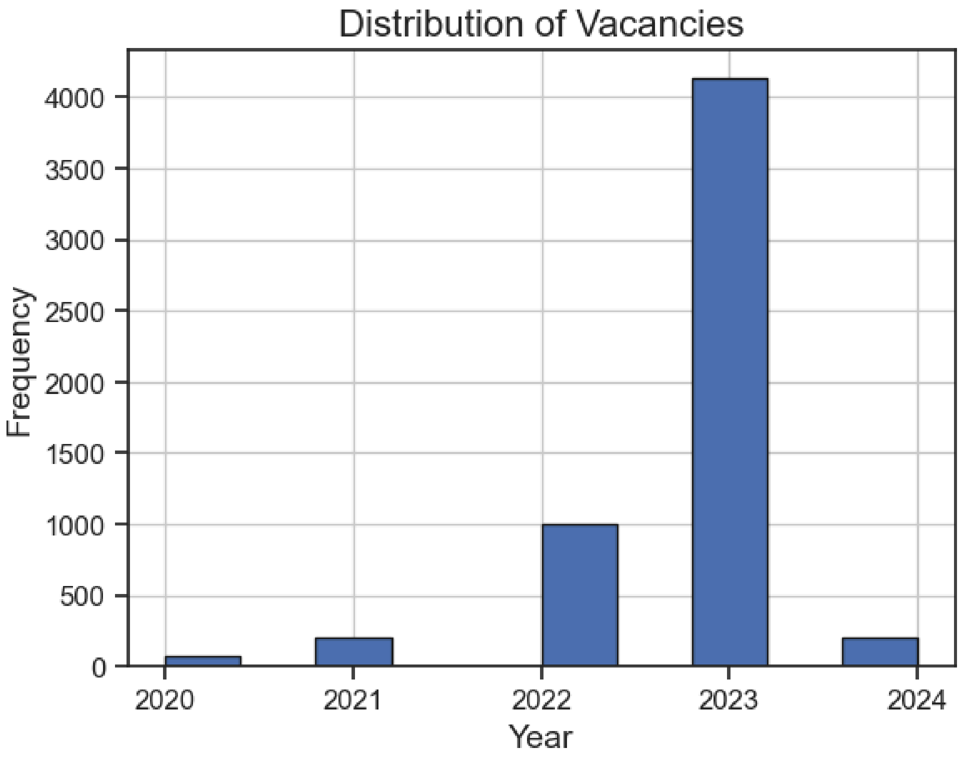
3.4 Salary Distribution

Histograms and Boxplots:



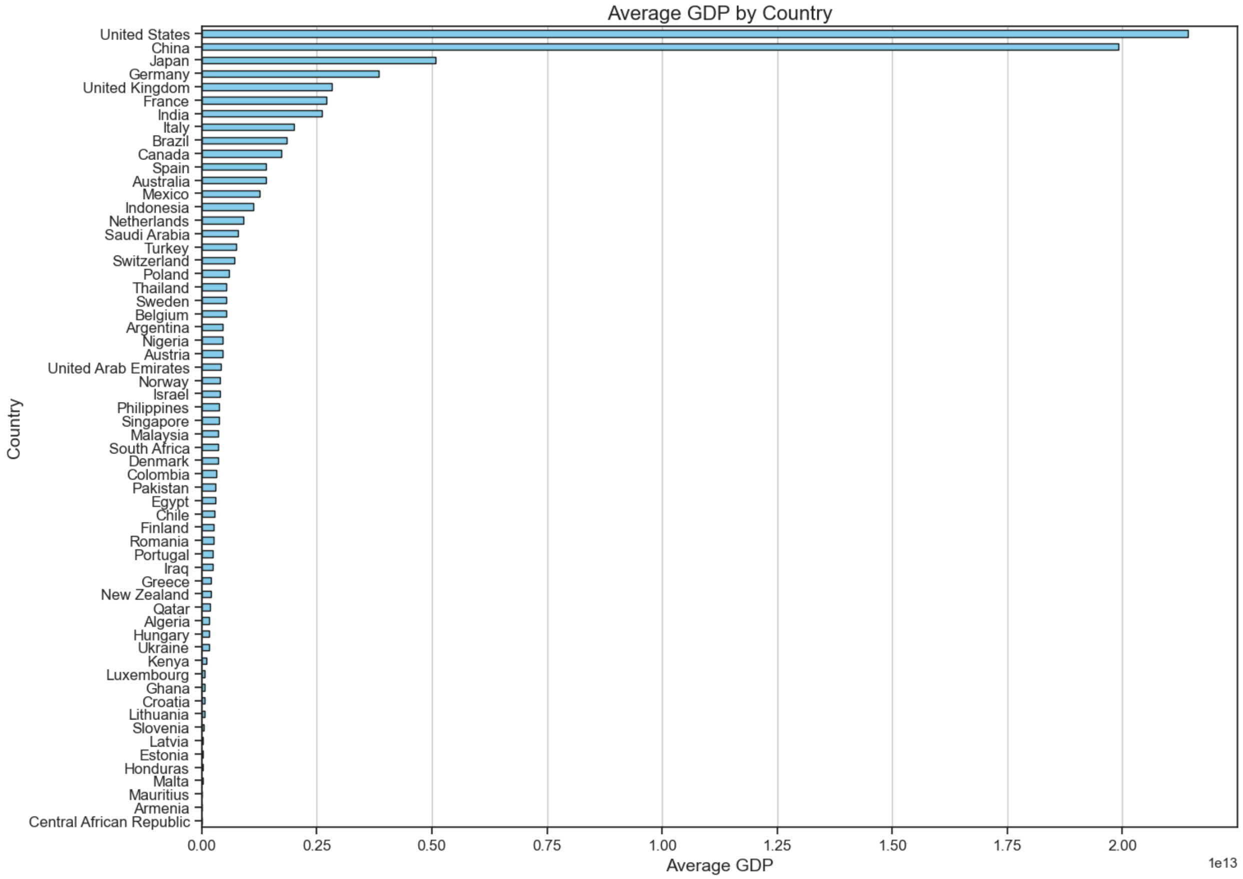
Most salaries are concentrated between $50,000 and $200,000, with a peak around $100,000 to $150,000. The distribution is right-skewed, indicating higher salaries are less common. The boxplot highlights outliers, extending up to $750,000.

3.5 Vacancy Distribution by Year



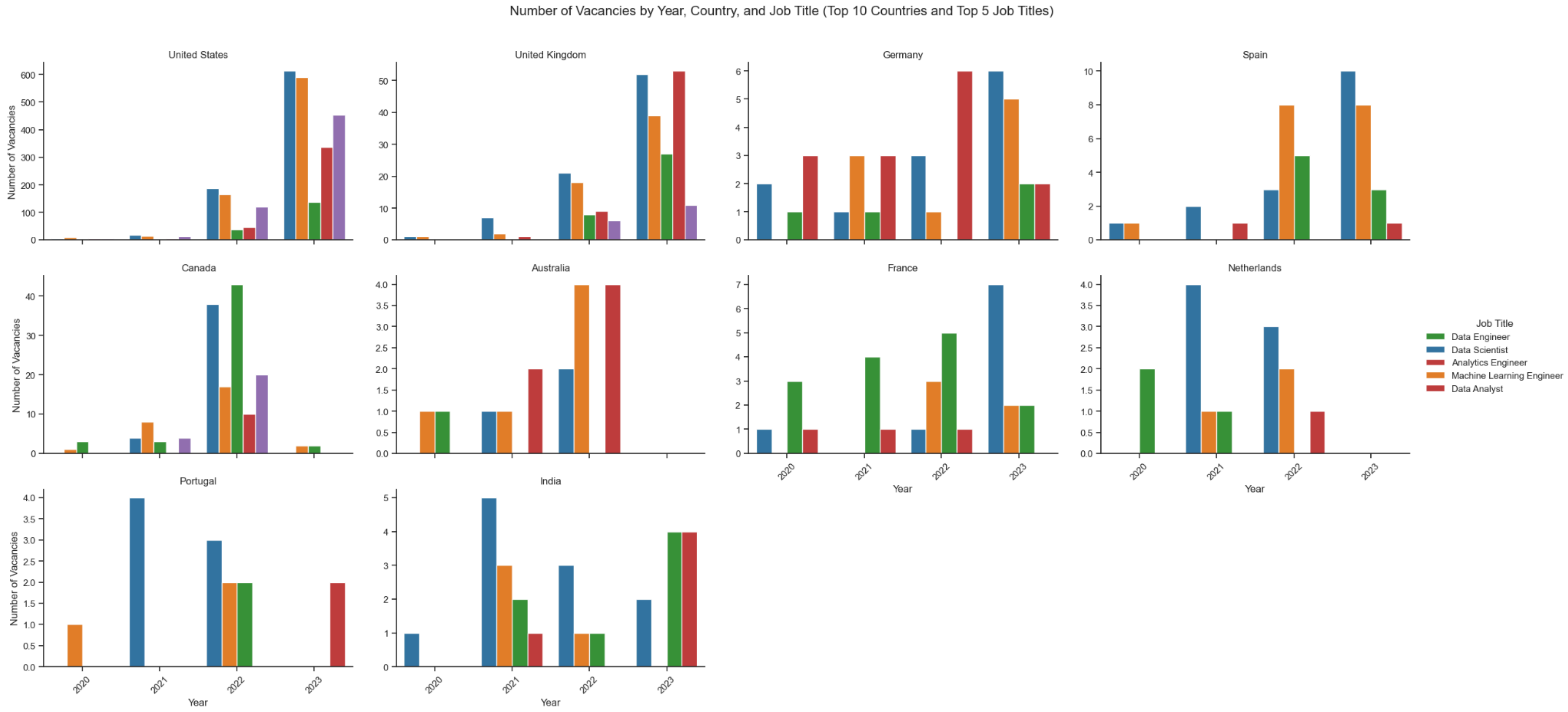
The majority of vacancies are concentrated in 2023. 2020, 2021, and 2024 show fewer vacancies, likely due to incomplete data for 2024.

3.6 Average GDP by Country



The United States, China, and Japan have the highest GDP among the countries analyzed, indicating a potential link between a country's economic strength and the salaries offered for data science positions.

3.7 Vacancies by Country and Job Title



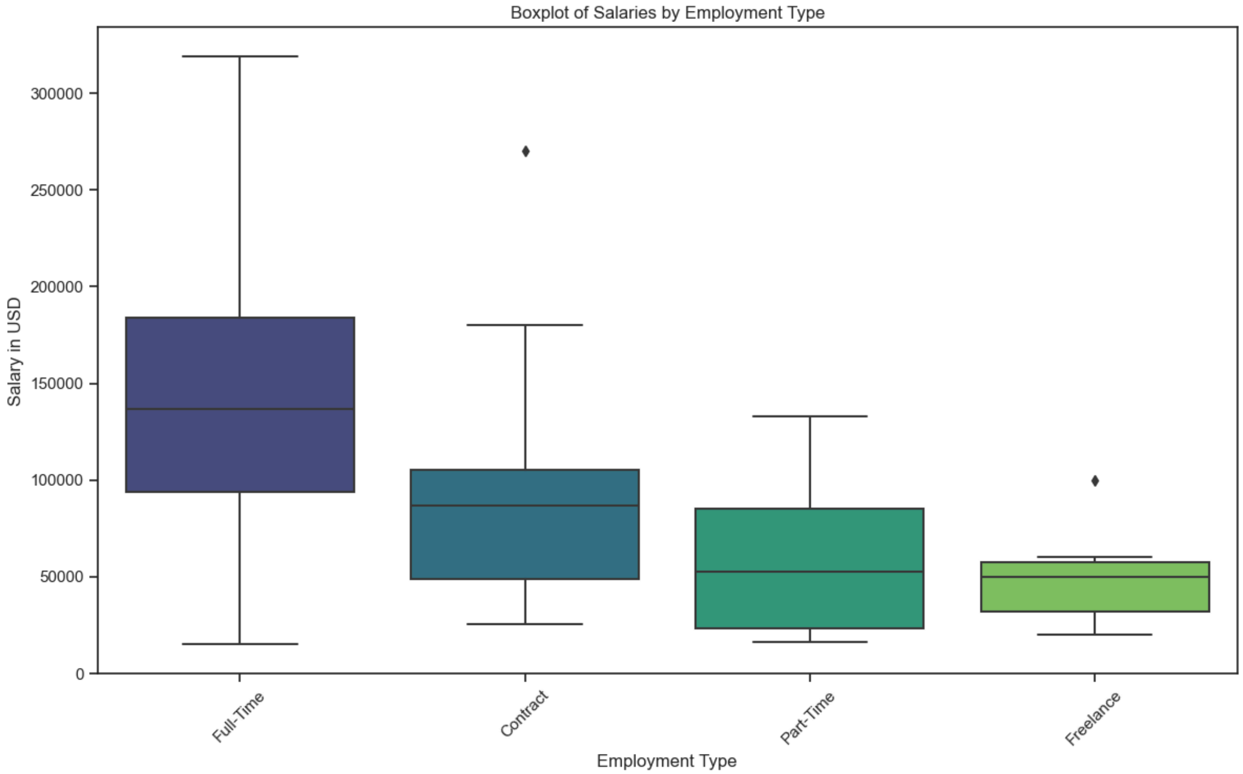
United States: Dominates the number of vacancies, especially in 2023. Most prominent roles are Data Engineer, Data Scientist, and Data Analyst.

United Kingdom, Germany, and Canada: Show a steady increase in vacancies over the years. The roles of Data Engineer and Data Scientist are consistently high.

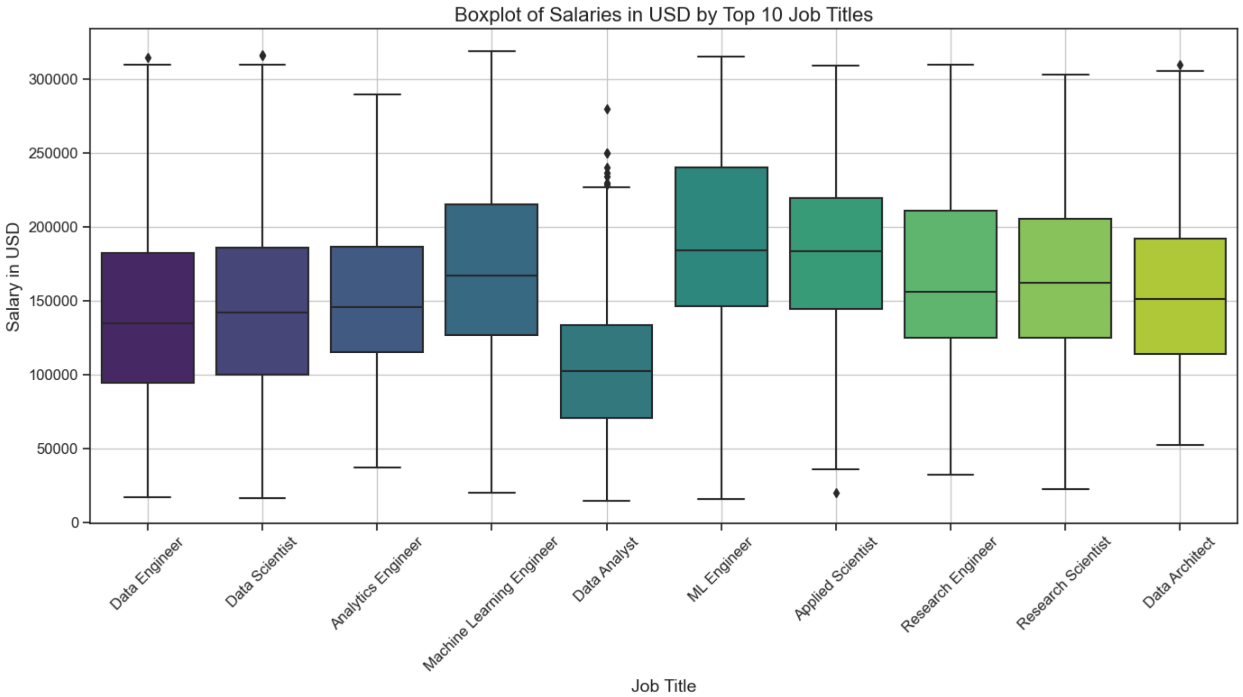
Other Countries: Countries like Spain, Australia, and India show a moderate number of vacancies, with a noticeable peak in 2023.

Trend: The general trend across all countries is an increase in the number of vacancies, with a significant spike in 2023, indicating growing demand for data science professionals globally.

3.8 Boxplot of Salaries by Employment Type



Full-time positions offer the highest median salaries compared to contract, part-time, and freelance positions. Full-time roles also show a higher degree of salary variability. The distribution of vacancies by employment type shows a dominance of full-time positions across the dataset.



Median Salaries: Data Engineers and Data Scientists have similar median salaries, around $140,000. Machine Learning Engineers and Research Engineers have slightly higher median salaries compared to Data Analysts and Analytics Engineers.

Salary Range: All job titles show a wide range of salaries, indicating significant variation within each role.

Summary

Salary Distribution: The distribution of salaries shows a positive skew, with most salaries clustering around the lower end but with a long tail of higher salaries.

Correlation Analysis: Salary in USD has a moderate positive correlation with factors like CPI and minimum wage, and a negative correlation with unemployment rate and fertility rate.

Job Vacancies: There is a significant increase in job vacancies in 2023, especially in the United States, United Kingdom, Germany, and Canada. Data Engineer and Data Scientist are consistently high-demand roles across all countries.

GDP Analysis: The United States, China, and Japan have the highest average GDP, which could correlate with the higher salaries observed in these countries.

Future Outlook: The trend indicates a growing demand for data science roles globally, with increasing salaries and job vacancies. The data for 2024 is limited, and further analysis will be needed as more data becomes available.

3.9 Recommendations

1. **Focus on Full-Time Positions:** For higher salary prospects, prioritize full-time positions.
2. **Geographic Considerations:** Consider opportunities in countries with higher GDPs for potentially better salary packages.
3. **Specialization:** Specializing in roles like Data Engineer or Data Scientist may offer better salary prospects and job security.

4. Preprocessing and Training Data

4.1 Data Cleaning and Transformation

1. Dummy Encoding:
   - Categorical variables were converted into dummy variables using one-hot encoding, resulting in a dataset with 5,702 records and 278 columns.
2. Standardization:
   - Numeric features were standardized using StandardScaler to ensure uniformity in magnitude, aiding in improving the performance of machine learning models.
3. Data Check:
   - Verified that no object data types or NaN values were present in the dataset, ensuring data integrity and readiness for modeling.

4.2 Split into Testing and Training Datasets

To build and evaluate our machine learning models effectively, the dataset was split into training and testing sets. The training set is used to fit the models, while the testing set is used to evaluate their performance.

- **Features (X)**: All columns except the target variable (salary_in_usd).
- **Target Variable (y)**: salary_in_usd.
- **Data Split**: 80% training data, 20% testing data.

5. Applying the Machine Learning Models

5.1 Model Selection

This is a regression problem in supervised learning. The following regression models were used:

1. Random Forest
2. Ridge Regression
3. Lasso Regression
4. XGBoost
5. Gradient Boosting

5.2 Random Forest Model

5.2.1 Find Significant Features

Using RandomForestRegressor, the importance of features was determined, and insignificant features were removed. Features with an importance greater than an arbitrary threshold (e.g., 0.001) were retained.

Feature	Importance Level
salary_currency_United States Dollar	0.3139452537273851
job_title_Data Analyst	0.06574581243346314
experience_level_Senior	0.0627121487961952
expertise_level_Expert	0.06054022733389565
job_title_Machine Learning Engineer	0.03195802134271284
expertise_level_Director	0.02910916046114724
minimum_wage	0.027454919124997697
experience_level_Executive	0.020456180611366125
year_2023	0.014481510072970676
job_title_ML Engineer	0.014017558644162723
gross_tertiary_education_enrollment_pct	0.013886863547166547
year_2022	0.011557909377429374
job_title_Research Scientist	0.01111304413874286
expertise_level_Junior	0.010223633853746623
job_title_Applied Scientist	0.00963508109451584
company_size_Large	0.009329162678556203
experience_level_Entry	0.009264831947825342
company_size_Medium	0.009075740884434289

Feature	Importance Level
job_title_Data Manager	0.008685374245092971
employee_residence_United States	0.008459969657279636
job_title_Data Scientist	0.008269164405074065
gdp	0.007684598518009928

5.2.2 Hyperparametric Optimization

Hyperparameters were optimized using RandomizedSearchCV. This process involved splitting the dataset into training and validation sets and using 3-fold cross-validation to ensure a reliable performance assessment.

5.2.3 The Results

- **Removal of Insignificant Features**: Improved the model by focusing on significant features.
- **Cross-validation**: Provided a reliable assessment of the model's performance.
- **Hyperparametric Optimization**: Enhanced model performance by finding the best parameters.

Optimized Metrics for Random Forest:

- MAE: 0.6119
- MSE: 0.5943
- RMSE: 0.7710
- R^2: 0.4020
- Training Time: 87.39 seconds
- Prediction Time: 0.054 seconds

5.3 Gradient Boosting Regressor

Hyperparameters were optimized for the Gradient Boosting Regressor using a similar approach as Random Forest.

Optimized Metrics for Gradient Boosting:

- MAE: 0.6011
- MSE: 0.5716
- RMSE: 0.7554
- R^2: 0.4249

- Training Time: 34.31 seconds
- Prediction Time: 0.0035 seconds

5.4 XGBoost Regressor

Hyperparameters for XGBoost Regressor were selected and optimized.

Optimized Metrics for XGBoost:

- MAE: 0.5995
- MSE: 0.5711
- RMSE: 0.7557
- R^2: 0.4254
- Training Time: 5.19 seconds
- Prediction Time: 0.0067 seconds

5.5 Ridge Regression

Hyperparameters for Ridge Regression were optimized.

Optimized Metrics for Ridge Regression:

- MAE: 0.6060
- MSE: 0.5812
- RMSE: 0.7623
- R^2: 0.4152
- Training Time: 0.50 seconds
- Prediction Time: 0.0026 seconds

5.6 Lasso Regression

Hyperparameters for Lasso Regression were optimized.

Optimized Metrics for Lasso Regression:

- MAE: 0.8160
- MSE: 0.9940
- RMSE: 0.9970
- R^2: -0.0001
- Training Time: 0.40 seconds
- Prediction Time: 0.0019 seconds

5.7 Comparison of the Models

The performance of different ML models was compared using various metrics.

Model	**MAE**	**MSE**	**RMSE**	**R^2**	**Training Time (s)**	**Prediction Time (s)**	**Complexity**	**Scalability**	**Maintenance Costs**
Random Forest	0.6119	0.5943	0.7710	0.4020	87.39	0.054	High	Good	Medium
Ridge Regression	0.6060	0.5812	0.7623	0.4152	0.50	0.0026	Low	Excellent	Low
Lasso Regression	0.8160	0.9940	0.9970	-0.0001	0.40	0.0019	Low	Excellent	Low
XGBoost	0.5995	0.5711	**0.7557**	0.4254	5.19	0.0067	High	Excellent	High
Gradient Boosting	0.6011	0.5716	**0.7554**	0.4249	34.31	0.0035	High	Good	High

5.8 Conclusion

1. **XGBoost and Gradient Boosting**: Showed the best results in terms of prediction accuracy.
2. **Random Forest**: Also performed well and is recommended depending on training and prediction time requirements.
3. **Ridge Regression**: Provided a good balance between simplicity, scalability, and performance.
4. **Lasso Regression**: Performed the worst and is not suitable for this dataset.

5.9 Recommendations

1. **Model Selection**: Random Forest and XGBoost are the most suitable models for salary prediction.
2. **Considerations**: Ridge Regression may be preferred for its simplicity and scalability.
3. **Exclusions**: Lasso Regression is not recommended due to its poor performance.

6. Applying the Machine Learning Models

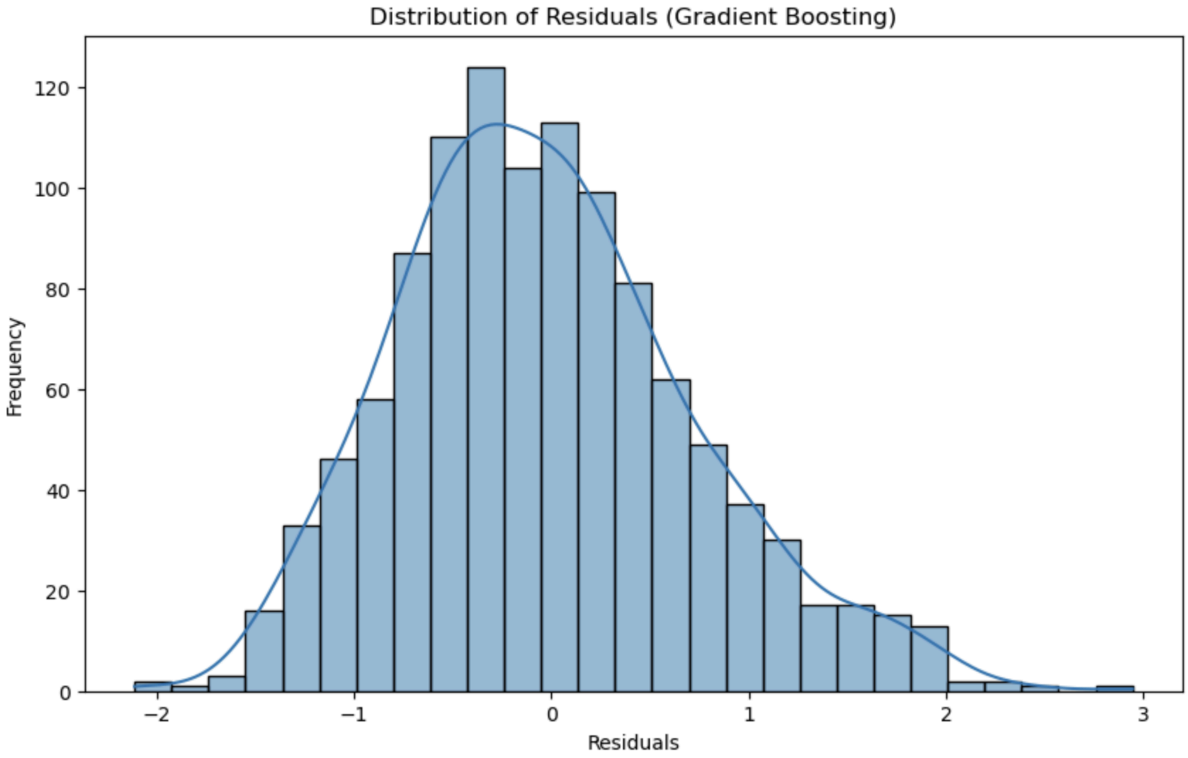
6.1 Gradient Boosting Regressor (Selected Model)

Based on model performance and evaluation metrics, Gradient Boosting was chosen as the preferred model for further analysis and predictions.

6.2 Residuals Distribution

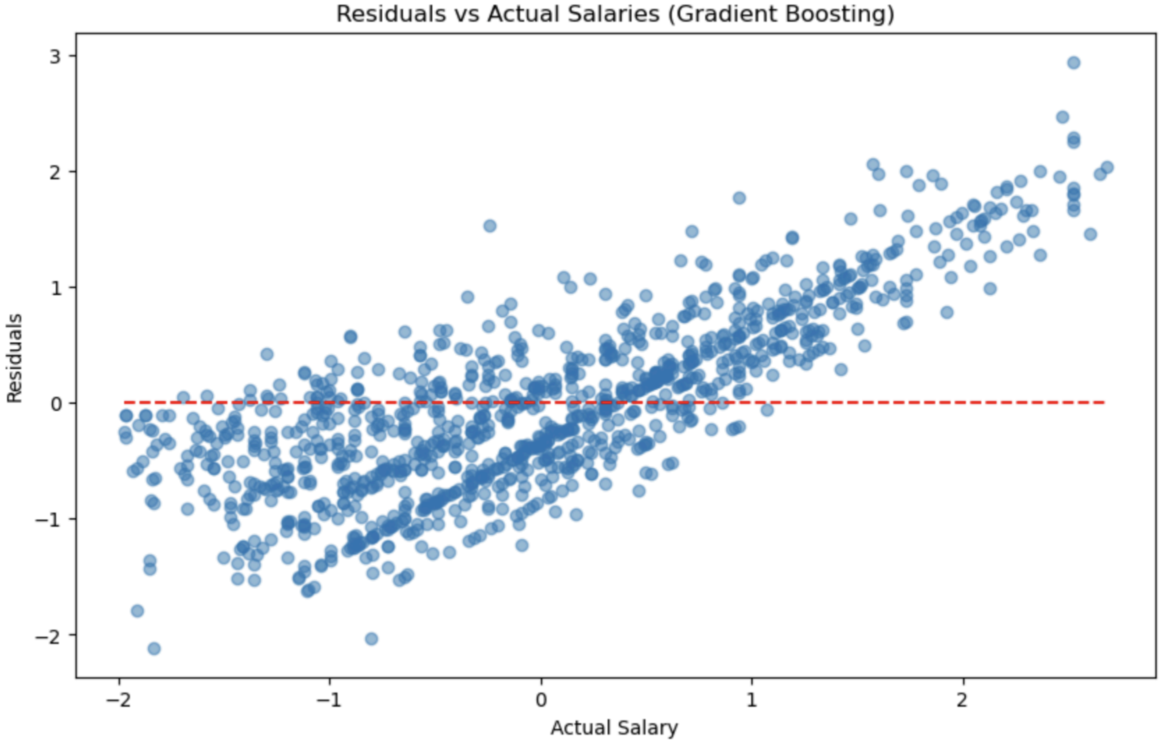
The distribution of residuals provides insights into the performance of the Gradient Boosting model. Residuals are the differences between the actual and predicted salaries.

6.2.1 Histogram of Residuals



- ○ The histogram shows the distribution of residuals, which ideally should be centered around zero and follow a normal distribution if the model has captured the patterns well.
- ○ The provided histogram shows that most residuals are close to zero, indicating that the model predictions are generally accurate. However, there is some spread, suggesting areas where the model could be improved.

6.2.2 Residuals vs. Actual Salaries Plot



This scatter plot displays residuals against actual salaries. In an ideal model, residuals would be randomly distributed without any clear pattern.

The plot shows that most residuals are close to zero, but there is some dispersion, especially for higher and lower actual salary values. This indicates that the model might underperform for extreme values.

6.3 Example of Predicted vs Actual Salaries

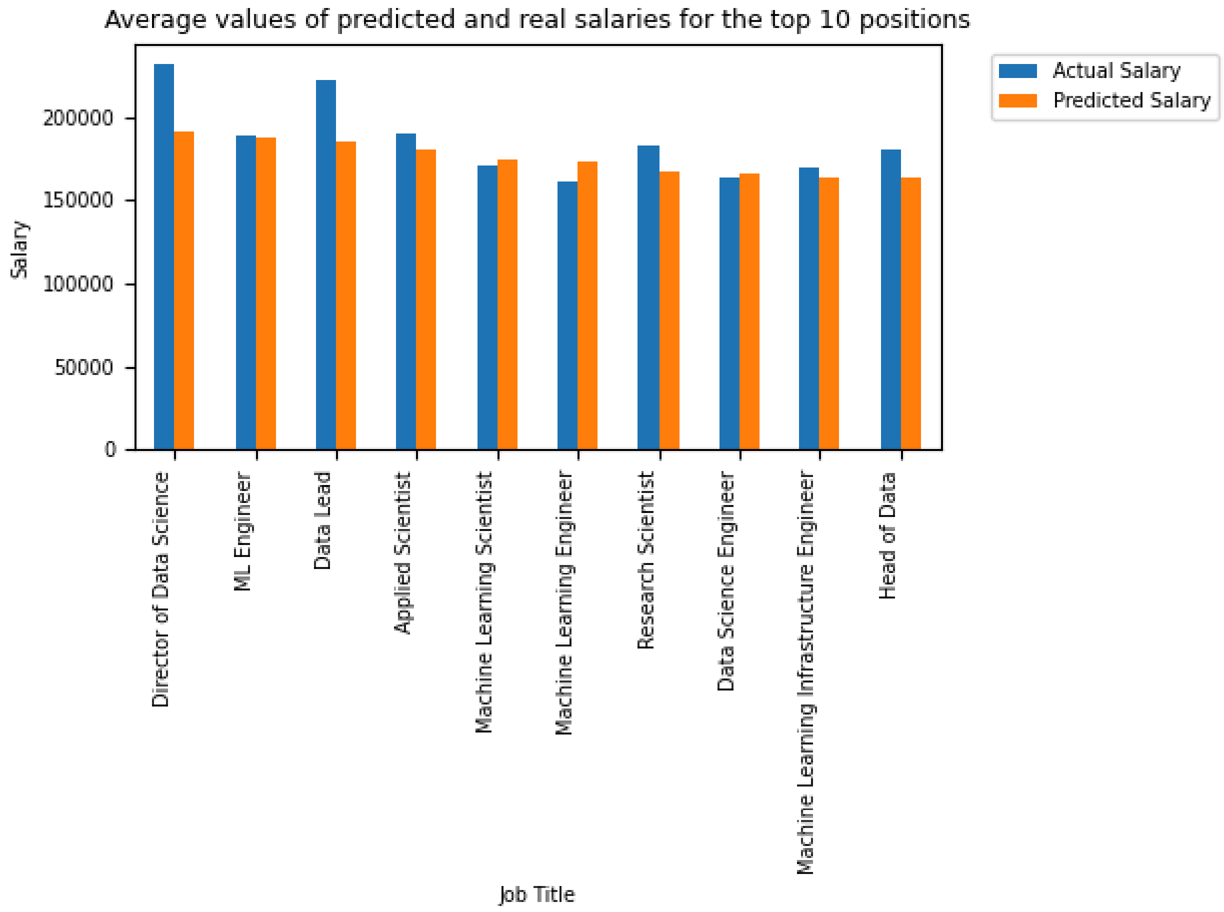
To illustrate the model's performance, a few examples of the predicted salaries compared to the actual salaries are presented below:

6.3 Example of Predicted vs Actual Salaries

To illustrate the model's performance, a few examples of the predicted salaries compared to the actual salaries are presented below:

Job Title	**Actual Salary**	**Predicted Salary**
Director of Data Science	232038.12	190985.24

Job Title	Actual Salary	Predicted Salary
ML Engineer	188916.85	188216.80
Data Lead	222007.22	185289.82
Applied Scientist	189668.81	180643.97
Machine Learning Scientist	170703.32	173878.11
Machine Learning Engineer	161281.71	173493.83
Research Scientist	183121.68	167145.79
Data Science Engineer	163335.33	166077.19
Machine Learning Infrastructure Engineer	169813.91	163147.69
Head of Data	180587.78	163138.39



The model is accurate for technical roles but tends to underestimate salaries for senior and leadership positions. Further refinement is needed to better capture the dynamics of high-level roles.

6.4 Summary

In general, although the model copes well with most wage values, it has some difficulties in predicting extreme values, which is a common problem in regression problems. Additional adjustment or more complex models may be required to improve predictions for these emissions.

7. Future Directions

To further improve the model's accuracy and reliability, the following steps are suggested:

1. Advanced Feature Engineering:

Add interaction and polynomial features to capture complex relationships.

Introduce domain-specific features relevant to data science.

2. Hyperparameter Tuning:

   Use Bayesian optimization and grid search to fine-tune model parameters.

3. Incorporating External Data:

   Include additional economic indicators, market trends, and regional salary data to improve model accuracy.

8. Conclusion

The XGBoost and Random Forest models accurately predicted data science salaries, with minor underestimations for higher-level positions. The model provides valuable insights into salary trends, aiding in informed decision-making. Future improvements, such as advanced feature engineering and the inclusion of external data, can further enhance model performance.