# S&DS625 Report

Zihan Wang

December 15th, 2023

## Abstract

As the data science job market burgeons, potential employees often grapple with gauging whether the salary offered aligns with market standards, due to the absence of a reference framework. Addressing this gap, This project delves into the analysis of data science-related job salaries, leveraging a dataset sourced from Kaggle. The dataset encompasses a range of attributes such as work year, experience level, employment type, job title, salary in USD, employee residence, remote ratio, and more. However, the focal point of the analysis rests on key predictors like experience level, job title, employee residence, and remote ratio to unveil their influence on salaries, measured in USD and adjusted for inflation based on the offer year. The project employs linear regression to incorporate predictors such as job title, country of employee residence, work mode (remote, hybrid, on-site), and experience level (from entry to senior levels), with the inflation-adjusted salary in USD serving as the outcome variable. The findings from linear model underscore a significant correlation between the stated predictors and data science job salaries, highlighting notable variations across different countries and experience levels. Subsequently, a comparative error analysis of various models led to the selection of the linear regression model with transformed salary for salary prediction. This choice was informed by comparing the root mean squared error of the models, where linear regression model demonstrated a smallest squared error. The final model can predict the estimated salary based on the employee's residence, job title, employee residence, and remote ratio etc. The culmination of this analysis not only provides valuable insights into the salary dynamics within the data science domain but also underpins a web application designed to offer employees a benchmark for evaluating job offers in this field. This application, fueled by the linear regression model's predictions, stands as a testament to the practical applicability of the project's findings, empowering data science professionals with a robust tool for informed salary negotiations.

## Section I: Introduction

The realm of data science is witnessing an unprecedented surge in job opportunities, reflecting the growing reliance on data-driven decision-making across various industries. With this surge comes a diverse range of salaries, influenced by multiple factors such as experience level, job title, geographical location, and working mode. This project is motivated by the need for a systematic approach to predict these salaries, providing data science professionals with a benchmark for evaluating job offers and helping them understand their market value.

The data underpinning this study is extracted from Kaggle, encompassing variables such as work year, experience level, employment type, job title, salary (both in local currency and USD), employee residence, remote ratio, company location, and company size. However, the core analysis focuses on predictors like experience level, job title, salary in USD, employee residence, and remote ratio. These factors are crucial in understanding the disparities in salaries across different conditions in the data science job market. The dataset is a comprehensive compilation of salary data across various countries, providing a global perspective on the data science job market.

Moving beyond the introduction, Section II delves into data exploration and visualization. This section reveals significant disparities in salaries when viewed across different variables such as experience level, job title, and geographic location. The analysis confirms that salaries in the data science field vary considerably under different conditions, highlighting the need for a predictive model.

In Section III, we develop a linear regression model to predict salaries based on the identified significant predictors. This model is used to estimate salaries under various conditions, reflecting the trends observed in Section II. The linear regression approach is chosen for its simplicity and interpretability, making it suitable for this kind of analysis.

Section IV discusses the results of the linear regression model, emphasizing its alignment with the trends observed in the data exploration phase. This section also provides a critical assessment of the model's performance, discussing its strengths and limitations in the context of salary prediction.

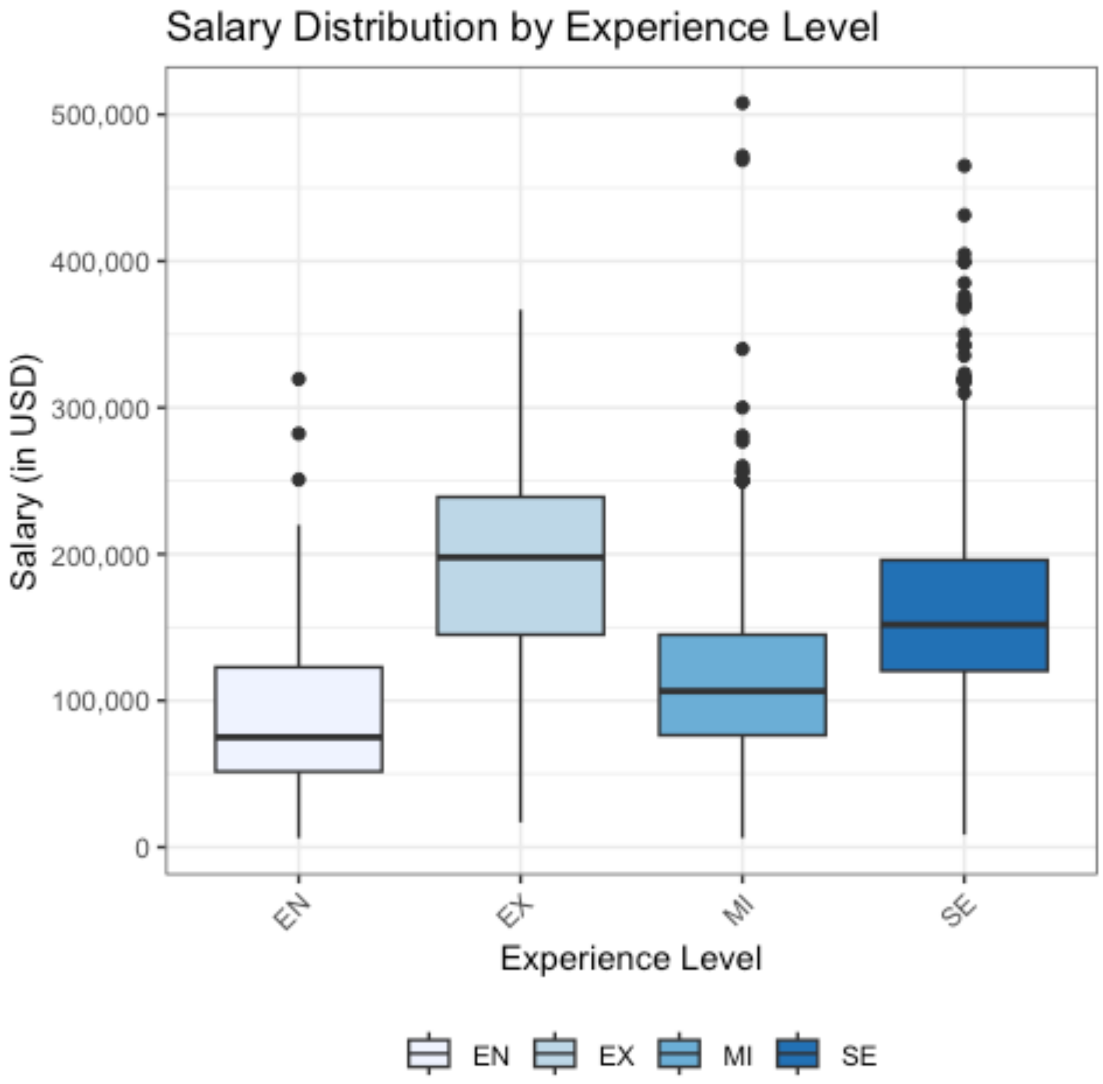
Finally, Section V concludes the report with a summary of findings, recommendations based on the analysis, and ideas for future research. This section will also touch on the potential implications of the study for data science professionals and employers alike, providing a broader perspective on the importance of salary prediction in the data science job market.

## Section II: Data exploration and visualization

This section presents a comprehensive overview of the key variables within the dataset, revealing insightful trends and relationships that inform our understanding of the factors influencing data science job salaries. Through meticulous data exploration and visualization, we unearth patterns that underscore the relevance of the chosen predictors in relation to the salaries reported.

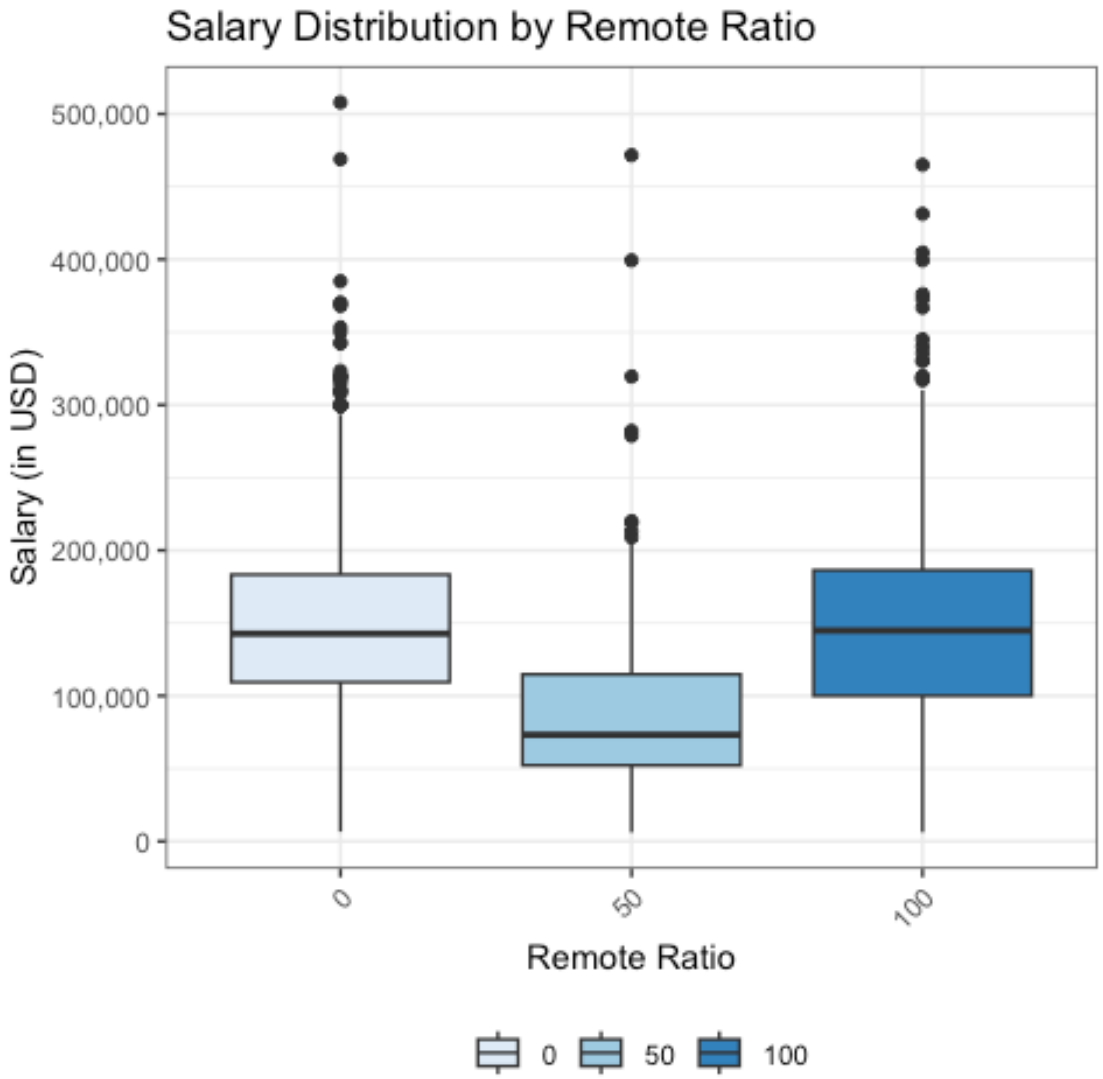
### *Mean Salary by Experience Level*

The first visualization showcases the mean salary delineated by experience level, revealing a clear stratification of earnings across different stages of professional development. Experienced (EX) professionals command the highest salaries, followed by senior (SE), mid-level (MI), and entry-level (EN) positions. This descending trend illustrates the premium placed on accumulated expertise within the data science field.



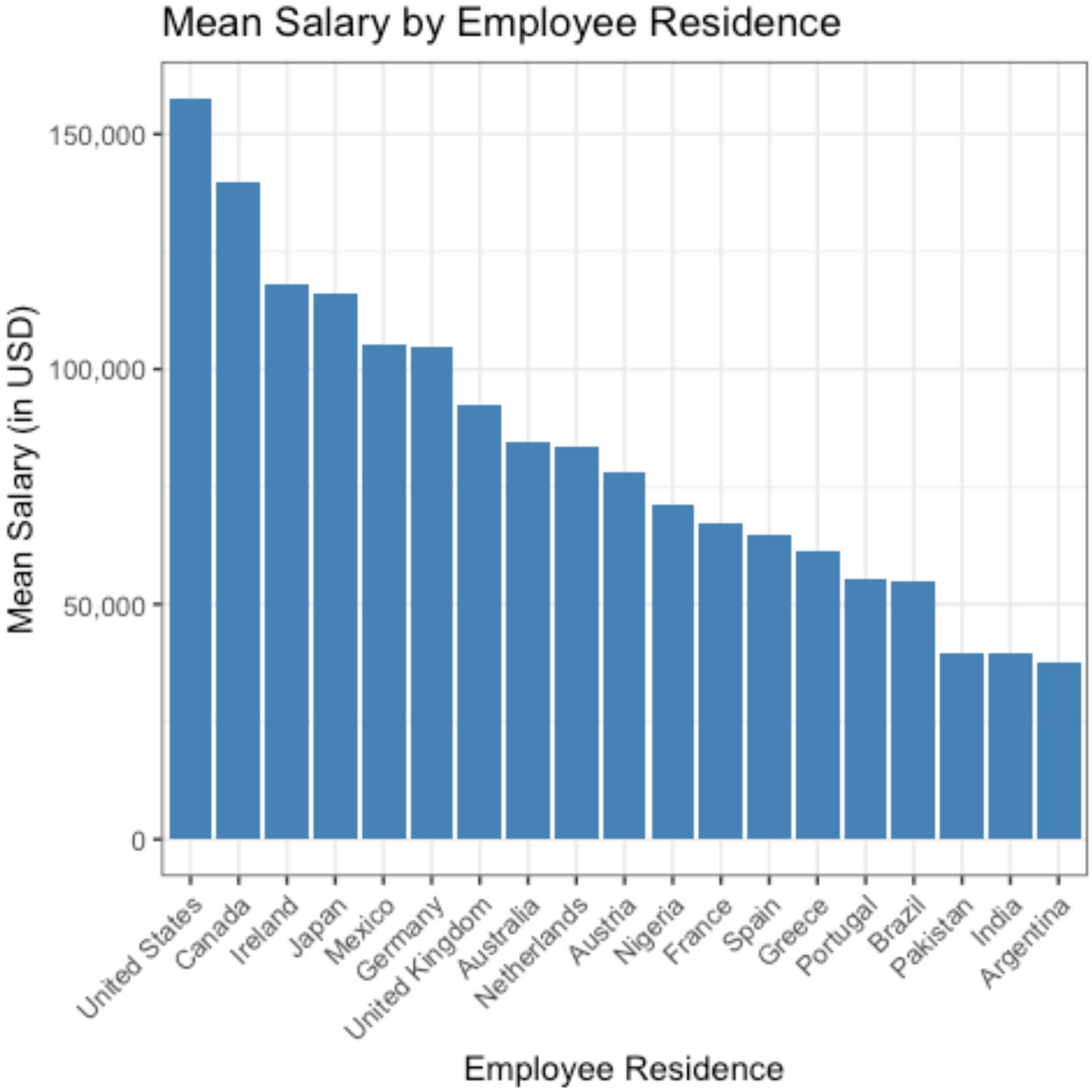
### *Mean Salary by Remote Ratio*

The second graph pivots to analyze the impact of remote working arrangements on salary. Here, a delineation is evident between fully remote (100), partial remote (50), and non-remote (0) roles. The findings indicate that full remote work does not necessarily translate to the highest salaries, prompting a deeper consideration of how workplace flexibility intersects with compensation.



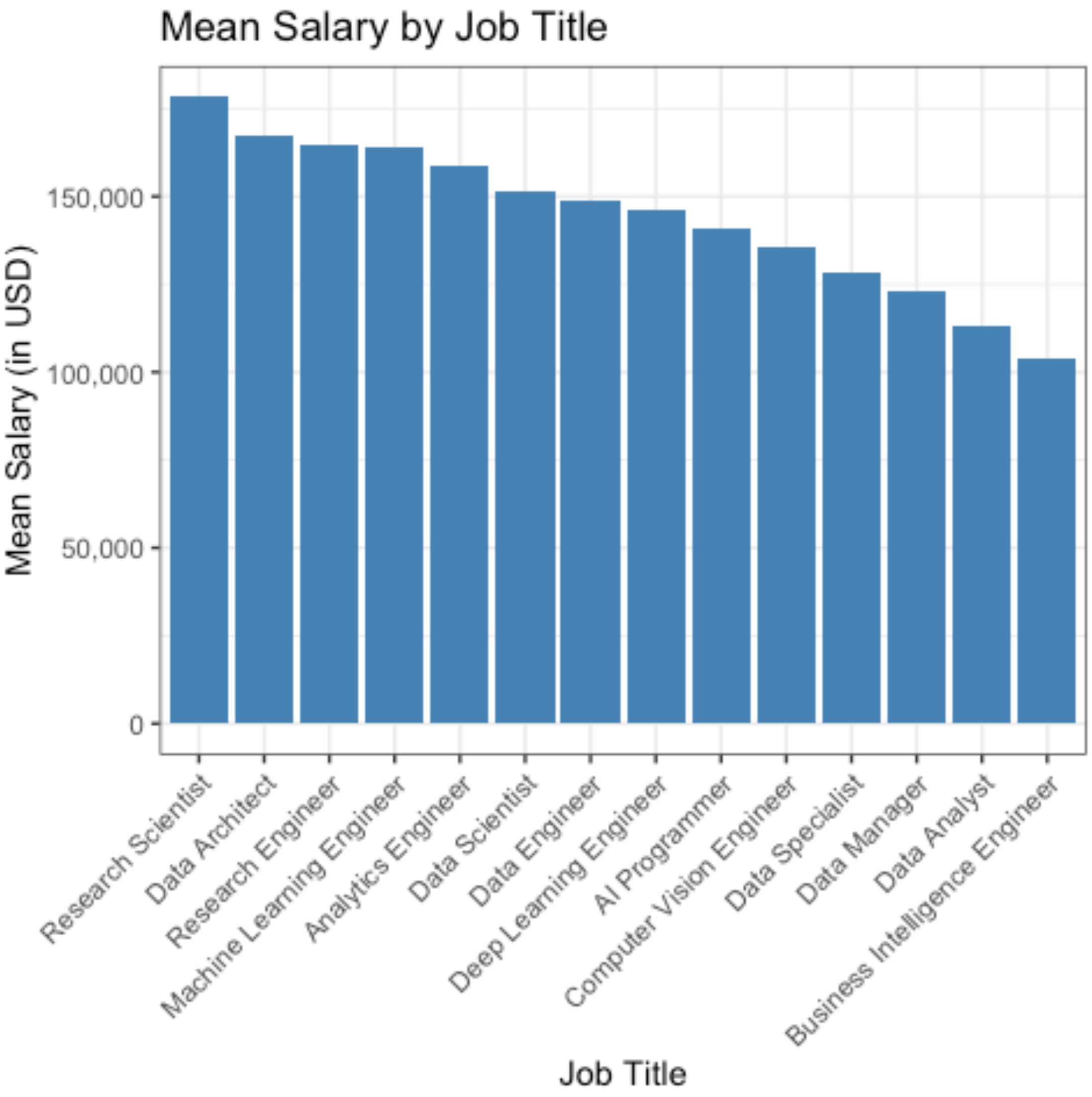
*Mean Salary by Employee Residence*

Delving into geographic influences, the third graph correlates mean salary with the employee's country of residence. The United States emerges as the top-paying country, followed closely by Canada and Ireland, descending through to Argentina. This gradient reflects the economic and industry-specific factors that shape salary benchmarks globally.



*Mean Salary by Job Title*

Lastly, we examine how different job titles within the data science umbrella correspond to salary variations. Research Scientists, Data Architects, and Machine Learning Engineers are among the top earners, with Business Intelligence Engineers and Data Analysts comparatively lower on the salary scale. This hierarchy reflects the market demand and perceived value of specialized skills.



These visualizations serve as a compelling prelude to the predictive modeling in Section III, where we construct and evaluate a linear regression model based on the highlighted predictors. The consistency of the trends observed here with the model's findings reinforces the conclusion that experience level, job title, employee residence, and remote ratio are significant determinants of salary in the data science domain. The visual evidence provided sets the stage for a nuanced discussion on the results of the model and their implications for data science professionals navigating the job market.

## Section III: Modeling/Analysis

In this section, we delve into the construction of a predictive model for data science job salaries. The primary tool of our analysis is a linear regression model, chosen for its simplicity and interpretability, making it a staple for exploratory analysis in quantitative fields.

### *Linear Regression Model*

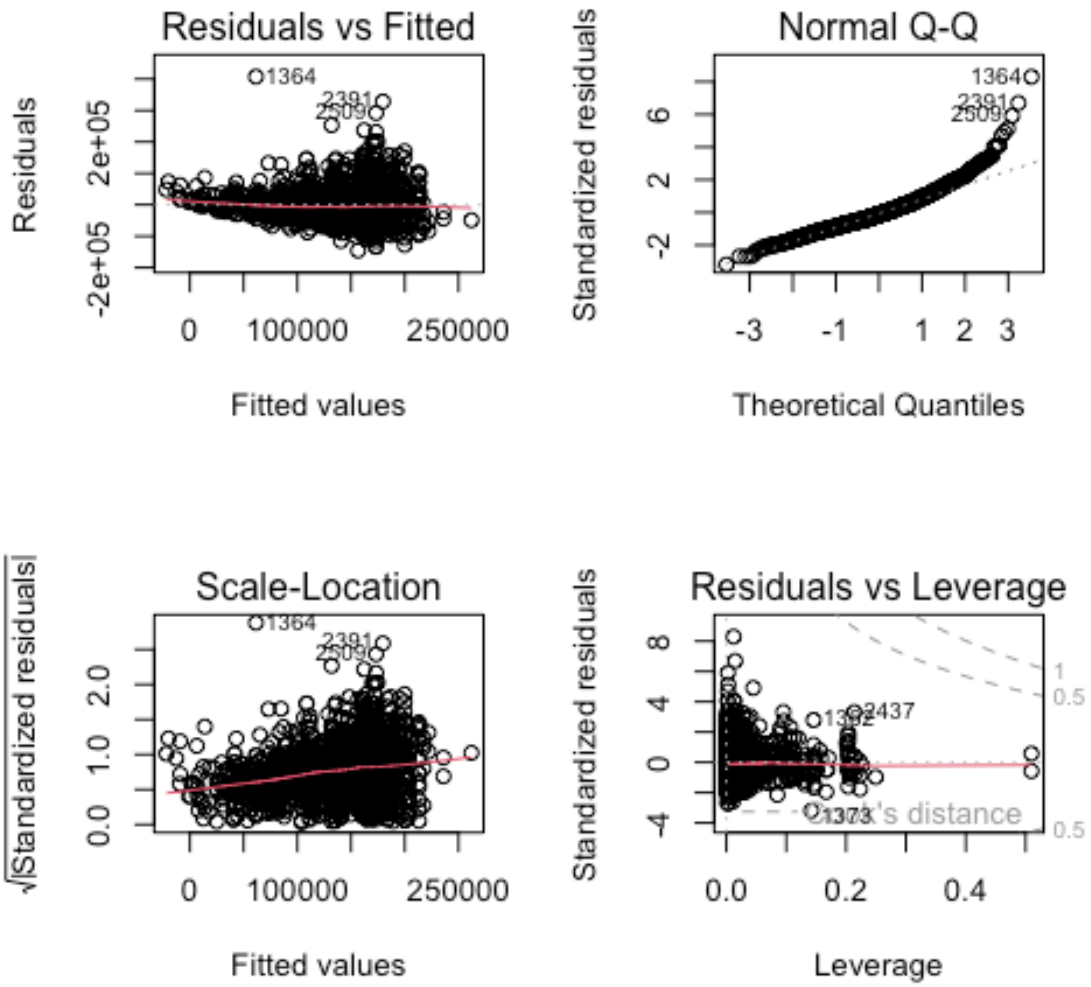
The linear regression model assumes a linear relationship between the independent variables (or predictors) and the dependent variable (or outcome). The predictors in the linear model are:

- Experience Level ($X_{i_1}$): Level of seniority
- Remote Ratio ($X_{i_2}$): Level of remoteness in the Job, 0 represents fully on-site, 50 represents hybrid, 100 represents fully remote.
- Job Title ($X_{i_3}$): Title of the job
- Employee residence ($X_{i_4}$): The residence of the country that the work is located.

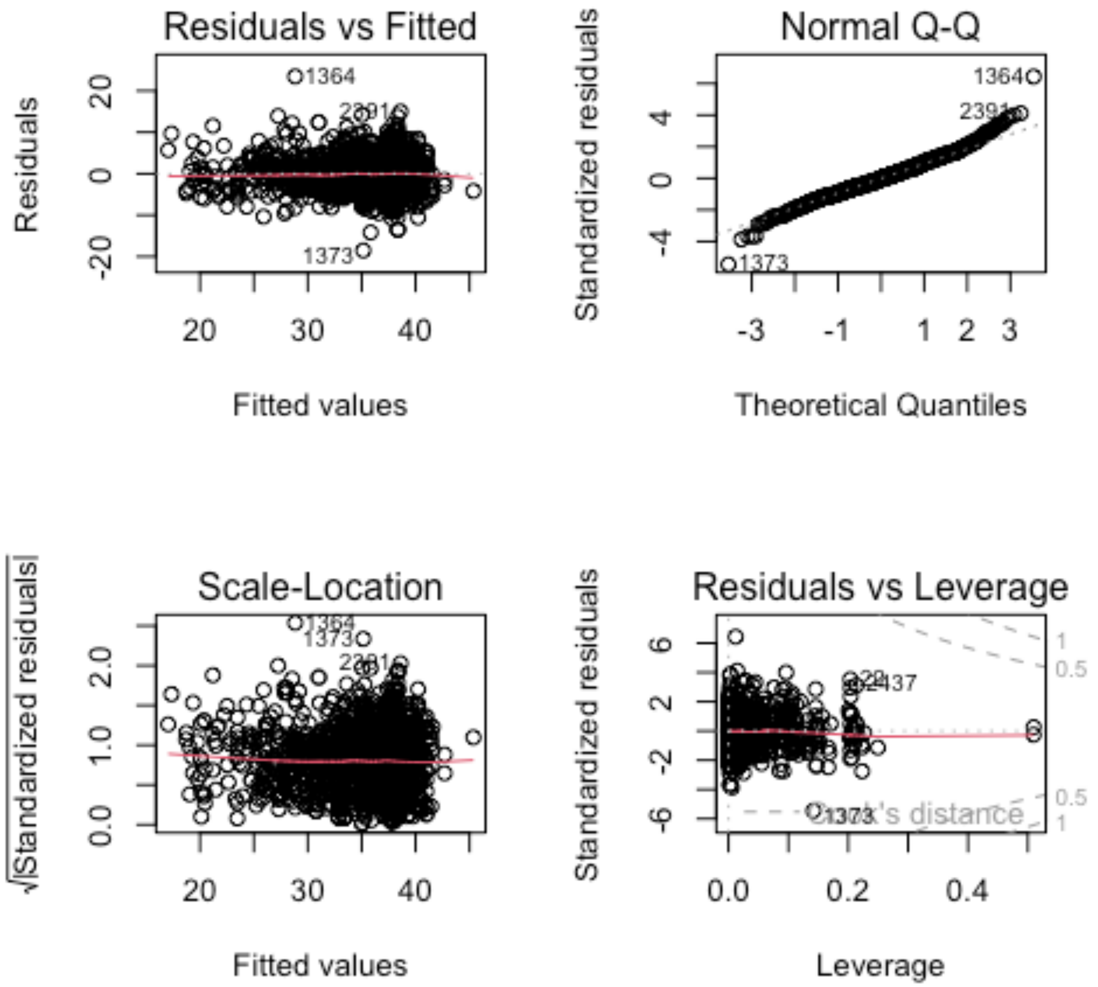
Our model takes the form: $Y_i = \beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i_2} + \beta_3 X_{i_3} + \beta_4 X_{i_4} + \beta_5 X_{i_1} X_{i_2} + \epsilon$, where $Y_i$ represents the salary in USD, $\beta_0$ is the y-intercept, $X_{i_1}$ through $X_{i_4}$ represent the four predictors, $\beta_1$ through $\beta_4$ represent the regression coefficients for the predictor variables, $\beta_5$ represents the regression coefficients for the interactive terms of experience level and remote ratio, and $\epsilon$ is the error term.

### *Assumptions and Transformations*

The preliminary assumption checks revealed deviations from normality, linearity, and heteroscedasticity. The normal Q-Q plot tests the assumptions of normality, The Residuals vs. Fitted plot check linearity linearity between predictors and outcome variable, the Scale-Location plot checks a heteroscedasticity of the model. In the original multiple linear regression model, the normal Q-Q plot indicated a departure from normal distribution. The Residuals vs. Fitted plot shows a flat linear line for the residuals, indicating a linear relationship. In the Scale-Location plot, the linear line is not completely horizontal with increase in squared root of standardized residuals as fitted values increase, but the points are relatively spread out.



To address non-normality of the original model, a Box-Cox transformation was applied to the outcome variable, which resulted in a more normally distributed outcome and improved linearity, as evidenced by the transformed residuals plot. The transformation exponent was determined to be 0.3030303, resulting in a transformed salary variable for our multiple regression model. From the plots below, we can see normality, linearity, and heteroscedasticity are better addressed, especially from the Normal Q-Q plot, where the points are more in the linear line with few outliers. Therefore, the final model linear regression model with transformed outcome variable: $Y_i^d = \beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i_2} + \beta_3 X_{i_3} + \beta_4 X_{i_4} + \epsilon$, where $d = 0.3030303$ is the Box-Cox transformation value.



*Model Summary*

The regression analysis was performed on the transformed salary variable, yielding the following significant results. The intercept $\beta_0$ and coefficients $\beta_1$ through $\beta_4$ reflect the relationship between each predictor and the salary, adjusted for the other variables in the model. For instance, the coefficients for the experience level categories—Entry (EN), Mid (MI), and Senior (SE), Executive (EX)—are positive and the values of coefficients are increasing, indicating that higher experience levels have higher salaries.

The job title and employee residence predictors also show varying levels of influence on the salary, with some categories being significant and others not. For example, the negative coefficients for various job titles (eg. Analytics Engineer, Data Analyst) suggest that when compared to the baseline category (AI Programmer), these roles tend to have lower salaries. Conversely, positive coefficients for countries like the United States and Canada indicate higher salaries compared to the baseline residence (Argentina).

The model's performance was evaluated using the Residual Standard Error (RSE), Multiple R-squared, and F-statistic. The Multiple R-squared of 0.5525817 suggests that approximately 55% of the variability in the transformed salary is explained by the model. The F-statistic is highly significant, indicating that the model fits the data better than a model with no predictors.

```
[1] "Multiple R-squared of the transformed linear model is 0.523402811569115"

[1] "Adjusted Multiple R-squared of the transformed linear model is
0.515288874838634"
```

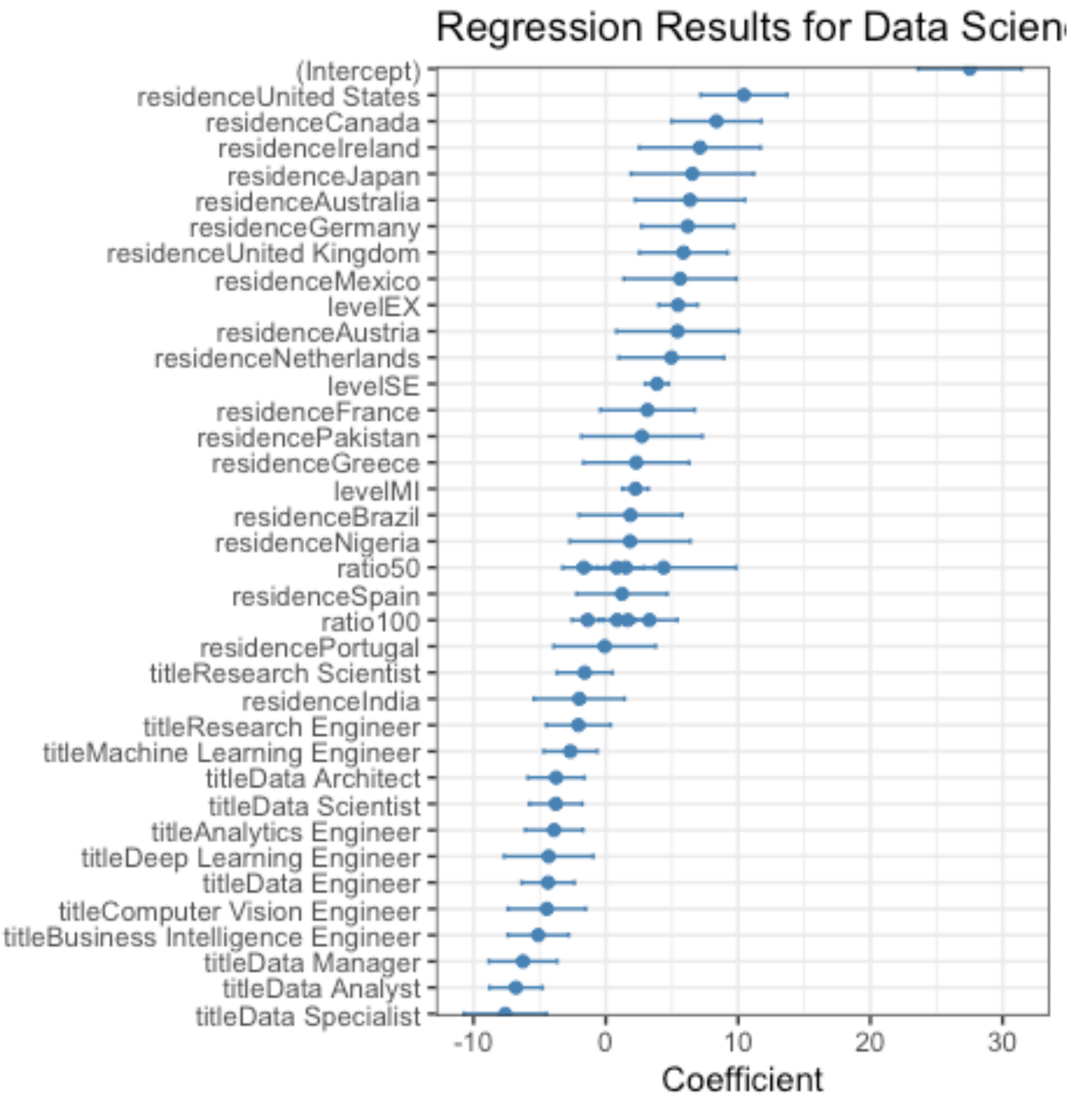
The model appears appropriate for the data, given the improvements post-transformation and the significant predictors. However, there is still unexplained variance, suggesting the potential for model enhancement or the existence of non-linear relationships not captured by this model.

In conclusion, the linear regression model, with the applied Box-Cox transformation, provides a robust framework for understanding the determinants of data science job salaries. It reveals the crucial role of experience, job title, and geographic location in shaping salary outcomes, offering valuable insights for job seekers and employers alike.

## Section IV: Visualization and interpretation of the results

The insights drawn from our regression analysis provide substantial implications for the data science job market. Our model's results, depicted in the regression coefficient plot below, offer a quantified look into how various factors affect data science salaries.

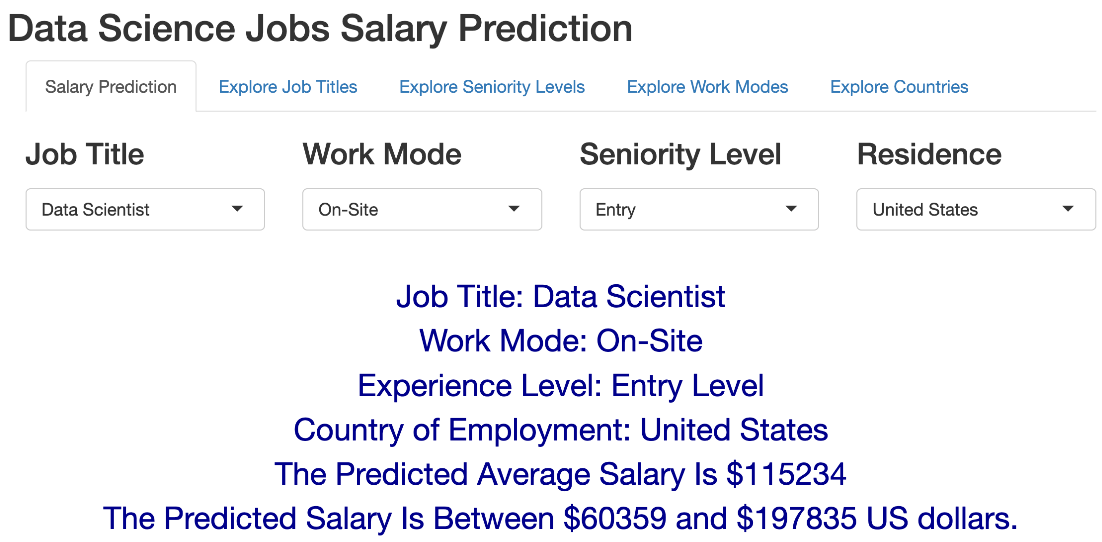
The coefficient plot indicates that geographical location has a significant impact on salaries, with 'United States' and 'Canada' showing the most substantial positive effect. Conversely, job titles such as 'Data Analyst' and 'Data Specialist' are associated with lower salaries, suggesting that certain roles within the data science field command premium pay. The positive coefficients for experience levels—particularly 'Executive' (EX)—highlight the premium on experience in the industry.



These findings have direct implications for data science professionals assessing job offers and career paths. The model underscores the importance of location and role specificity in salary negotiations. For instance, a Data Scientist in the United States can expect a significantly different salary from their counterpart in Spain, emphasizing the economic variances across regions.

### *Interactive Salary Prediction Tool*

An interactive R Shiny web application is also created to allow users to predict the range of salaries based on their job title, work mode, seniority level, and country of residence. There are 5 tabs of the application. The application's first tab is shown as the first picture below, which showcases the predicted salary range for selected inputs. When people select inputs of Job Title, Work Mode, Seniority Level, and Employment Residence, the application will provide an immediate, personalized salary range estimate that reflects the model's outputs.



Data Science Jobs Salary Prediction

Job Title: Data Scientist

Work Mode: On-Site

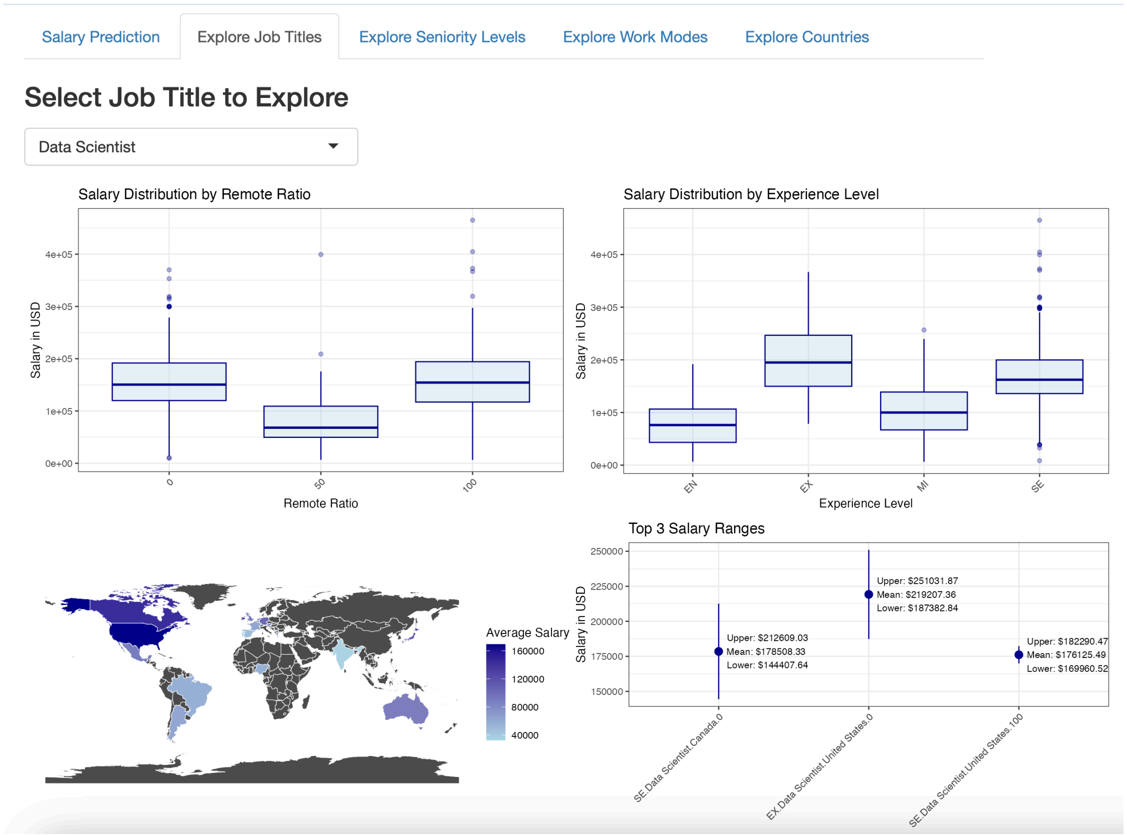
Seniority Level: Entry

Residence: United States

Job Title: Data Scientist
Work Mode: On-Site
Experience Level: Entry Level
Country of Employment: United States
The Predicted Average Salary Is $115234
The Predicted Salary Is Between $60359 and $197835 US dollars.

The subsequent 4 tabs offer an exploratory interface for users to understand salary distributions across different categories. For example, as shown in the second picture of the web application below, the second tab allows for exploring different data science related jobs. Selecting a job title reveals the distribution of salaries by remote ratio, experience level, and a world map visualizing the average salary by country. It also visualizes the top 3 salary combination of remote ratio, experience level, and employee residence based on the title selected. This interactive exploration not only adds depth to the user’s understanding of the market but also visualizes the model’s interpretability across different demographics.

The tabs focusing on Seniority Level, Work Mode, and Country (tabs 3 to 5, not shown here) further demonstrate the multifaceted nature of the data science salary landscape with similar features as the second tab. Users can examine the variances in salary distributions within each category and identify top 3 salary combinations, facilitating an informed decision-making process for career development.



Select Job Title to Explore

Data Scientist

## Section V: Conclusions and Recommendations

The analysis conducted in this project provides valuable insights into the factors influencing data science salaries across different regions and job titles. The regression results elucidate the impact of experience level, location, and job specificity on potential earnings within the data science field. Our findings confirm that geographic location is a significant salary determinant, with positions in the United States and Canada commanding higher wages. Additionally, the model highlights the premium on professional experience, particularly at the executive level.

The interactive web application developed as part of this study offers a practical tool for individuals to predict salary outcomes based on their professional profile. This can empower current and aspiring data science professionals with data-driven guidance when negotiating salaries or considering career advancements. Employers, too, can utilize this tool to ensure competitive compensation packages are offered to attract and retain top talent.

For future work, we recommend expanding the dataset to include more granular details on job titles and additional factors such as education level, certifications, and industry sectors. This could enhance the model's accuracy and provide a more comprehensive salary prediction. Incorporating time series analysis could also capture the trend of data science salaries over time, considering the rapid evolution of the technology sector.

Moreover, further research could explore the implications of emerging technologies and methodologies in data science, potentially affecting job demand and salary structures. Another avenue for development is the refinement of the web application to include more interactive features, such as real-time data updates and personalized reports, thus increasing its utility and user engagement.

In conclusion, while the current model offers a solid foundation for salary prediction within the data science realm, continuous improvement and updates will ensure its relevance and accuracy in a fast-paced and ever-changing industry.