# PREDICTION OF SALARIES IN THE DATA SCIENCE FIELD 2020-2023


## Group A

**Aravinda Bandara - S15024**
**Vidura Chathuranga - S15030**
**Samujitha Senaratne - S15077**
**Manul Wickramasinghe - S15088**


ST 4035 - Data Science
Group Project

**Abstract**

In the rapidly evolving landscape of data science, understanding the factors influencing salary differentials is of paramount importance for both professionals and organizations. This report presents an in-depth analysis of the "Data Science Salaries" dataset spanning the years 2020 to 2023. With a comprehensive collection of salary information across industries, organizations, and geographic regions, this dataset serves as a valuable resource for examining compensation trends and variations within the data science domain. The project begins by conducting exploratory data analysis, uncovering key variables that exhibit associations with salary differentials. Subsequently, a predictive model is developed to forecast salary based on the identified influential variables. This undertaking contributes to the understanding of salary determinants in the data science field, enabling professionals, researchers, and organizations to make informed decisions.

**Table of Contents**



**List of Figures**

## 1. Introduction

In an era marked by the proliferation of data-driven decision-making, the field of data science has emerged as a cornerstone of modern industries. Data scientists, armed with advanced analytical skills, play a pivotal role in extracting actionable insights from vast datasets, driving innovation, and enabling strategic business choices. As the demand for data science expertise continues to surge, understanding the intricate factors that influence compensation within this domain becomes crucial for both aspiring professionals and organizations seeking to attract and retain top talent.

The "Latest Data Science Salaries" dataset, spanning the years 2020 to 2023, encapsulates a wealth of salary information from diverse industries, organizations, and geographical locations. This dataset provides a unique opportunity to delve into the dynamics of salary differentials within the data science realm over a four-year period. By undertaking an exploratory data analysis of this dataset, we aim to unravel the nuanced interplay between various variables and salary amounts. The insights gleaned from this analysis not only empower individual professionals with informed career decisions but also provide organizations with valuable insights to tailor their compensation strategies and human resource practices effectively.

In this report, we embark on a comprehensive journey to decode the intricate relationship between data science job roles and their corresponding salaries. By leveraging statistical techniques and predictive modeling, we endeavor to shed light on the most influential factors that contribute to the variations in data science salaries. Our approach involves a twofold process: first, an exploratory investigation into the dataset's variables to identify key indicators that correlate with salary differentials, followed by the development of a predictive model that forecasts salaries based on these identified factors. Through this endeavor, we not only contribute to the empirical understanding of the data science salary landscape but also offer actionable insights for professionals and organizations navigating this rapidly evolving field.

## 2. Description of the Problem

The central challenge addressed in this study revolves around deciphering the multifaceted determinants of salary variations within the data science domain. As the field continues to expand, an understanding of the factors that contribute to differences in compensation becomes pivotal for professionals planning their careers and for organizations aiming to optimize their talent acquisition and retention strategies.

Objectives of the Study: -

1. **Exploratory Analysis:** The study aims to conduct an exploratory analysis of the "Data Science Salaries" dataset to uncover variables that are closely associated with differences in salaries among data science professionals.
2. **Predictive Modeling:** Building on the findings from the exploratory analysis, the study seeks to develop a predictive model that can accurately estimate salaries based on influential variables, providing valuable insights for career planning and talent management strategies.

## 3. Description of the data set

This dataset, sourced from Kaggle.com, comprises 3,300 records and features 11 variables. It offers a comprehensive view of data science salaries from 2020 to 2023. The response variable is “Salary in USD” which is a continuous variable.

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

| | Variable Name | Description | Variable Type |
|---|---|---|---|
| 1 | Job Title | “Data Scientist”, “Computer vision engineer”,” Data Analyst”, etc. | Categorical |
| 2 | Employment Type | “Full-Time” or “Part-Time” | Categorical |
| 3 | Experience Level | “Entry”, “Senior” or “Mid” | Categorical |
| 4 | Expertise Level | “Junior”, “Intermediate” or “Expert” | Categorical |
| 5 | Salary | Salary in local Currency | Quantitative |
| 6 | Salary Currency | “Australian Dollar”, “Brazilian Real”, “Euro”, etc. | Categorical |
| 7 | Company Location | The country where the company is located. | Categorical |
| 8 | Salary in USD | Salary of the individual in USD | Quantitative |
| 9 | Employee Residence | Country of Residence of the individual. | Categorical |
| 10 | Company Size | “Small”, “Medium” or “Large” | Categorical |
| 11 | Year | “2020”, “2021”, “2022” or “2023” | Categorical |

## 4. Data Pre-processing

The "Latest Data Science Salaries" dataset contains data under 11 variables where 9 of them are categorical and 2 of them are numerical. In the data set no duplicate entries were discovered and there were no missing records.

### Feature Engineering

After analyzing the dataset, several modifications were made to ensure that the data was clean and ready for further analysis. Initially, the dataset contained three variables: "Salary," "Salary Currency," and "Salary in USD," all of which conveyed the same idea. To avoid redundancy, it was decided to remove the "Salary" and "Salary Currency" variables from the dataset. This simplification ensured that the information regarding salary was captured solely by the "Salary in USD" variable.

To further refine the dataset, the value counts of each category in the variables were examined. Categories with value counts less than 50 were combined together to create a new category named "Others" for each

variable. This consolidation helped to reduce the number of categories and maintain the integrity of the data.

In addition to the existing variables, a new categorical variable called "is_expatriate" was introduced. This variable was created by comparing the "Company Location" and "Employee Residence" variables. By analyzing the alignment of work and home addresses, the dataset was able to distinguish between expatriate employees ("Yes") whose work location differs from home and non-expatriates ("No") whose work and home addresses align.

After the necessary modifications and additions, the final dataset consisted of 10 variables, with 9 of them being categorical. The response variable, "Salary in USD," remained as the only quantitative variable in the dataset. This variable served as the focal point for analyzing and understanding the relationship between various categorical variables and the corresponding salaries.

To facilitate further analysis, the dataset was split into training and testing subsets using an 80% and 20% proportion, respectively.

By following these steps, the dataset was effectively cleaned, refined, and prepared for subsequent analysis or modeling tasks. The modifications made to the dataset ensured that the data was accurate, consistent, and ready for further analysis.

## 5. Important Results of the Descriptive Analysis

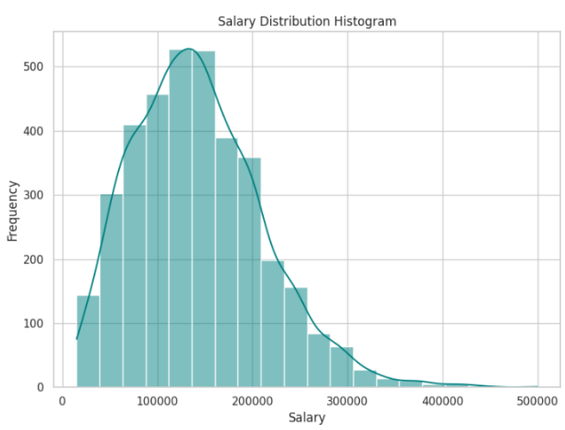



*Figure 1*

Our study aims to address the question of how salary amounts vary across data science-related job roles in relation to their associated factors. To achieve this, our initial focus involves analyzing the distribution of the response variable "Salary in USD."

Upon examining Figure 1, it is evident that the distribution of salaries is right-skewed, with a calculated skewness of approximately 0.666. This skewness indicates that the distribution's tail extends towards higher salary values, potentially impacting the positioning of the mean.

It's worth highlighting that in cases of positively skewed distributions, the mean tends to be influenced by the presence of outliers in the higher salary range, leading to its potential displacement towards the right tail. Contrarily, the median tends to be a more robust measure in such distributions and is expected to be smaller than the mean.

Based on our analysis, the calculated mean salary is $42,596.55, while the median salary stands at $136,100.0. This contrast between the mean and median values further underscores the distribution's right-skewed nature.


*Figure 3*


*Figure 2*

The data in Figure 2 makes it clear that experience strongly impacts earnings in Data Science. Executives earn much more, with a median of $1,750,000, aligning with Glassdoor's findings of Senior Data Scientists at $168,000 annually. Entry-level roles earn less. Looking at salary ranges across experience levels, a clear association between experience and pay becomes apparent. In Data Science, expertise levels junior, intermediate, director, and expert—have distinct salaries. From 2021 to 2023, growing expertise corresponds to higher pay. Interestingly, although data science experts excel, they earn slightly less than directors. Likely because directors handle leadership and decisions alongside technical work, while experts focus mainly on technical tasks. Figure 3 effectively demonstrates this possible association between pay and expertise.

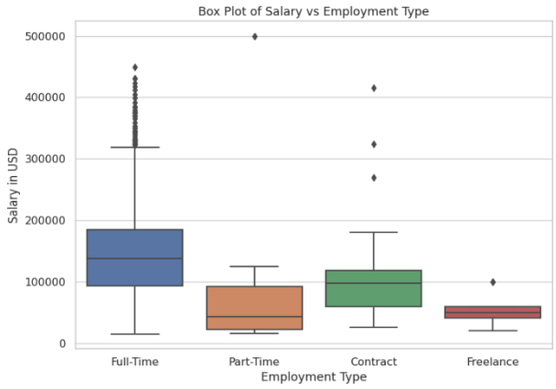



*Figure 4*

Figure 4 illustrates the salary distribution based on the employment types of the individuals, categorizing them as Full-Time workers, Part-Time workers, Contract workers, or Freelancers. Upon analyzing the box plots, it becomes apparent that Full-Time workers exhibit the highest median salary values, accompanied by a wider spread in their distribution in comparison to the other three categories.

Additionally, Contract workers display the second-highest median salary value, while Part-Time and Freelance workers showcase comparatively lower salary amounts. This observation suggests that individuals engaged as Full-Time or Contract workers tend to enjoy higher salaries, whereas those in Part-Time or Freelance roles tend to have lower earnings.

This interpretation can be rationalized by considering that employees who hold permanent positions or work under contractual arrangements are likely to receive higher compensation than their counterparts who work part-time or on a freelance basis. Consequently, this analysis provides valuable insights indicating a substantial correlation between the variable "Employment type" and the discrepancies in salary levels.

Upon analyzing the figure provided as figure 5, it becomes evident that there exists a notable disparity in salaries between workers who are employed in foreign countries and those who work in their own countries. The salary distribution of non-expatriate workers exhibits a broader range compared to their expatriate counterparts. Furthermore, the median salary among non-expatriate workers surpasses that of the other group.

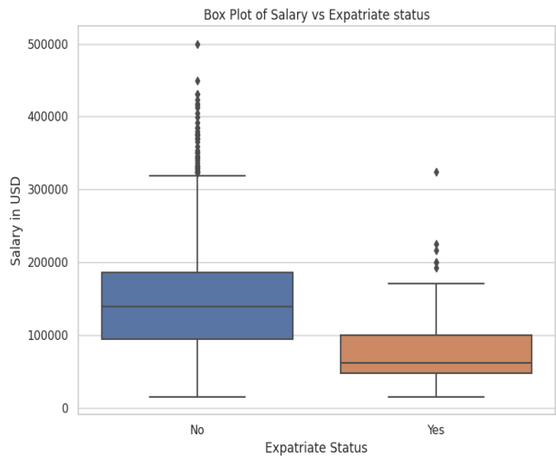
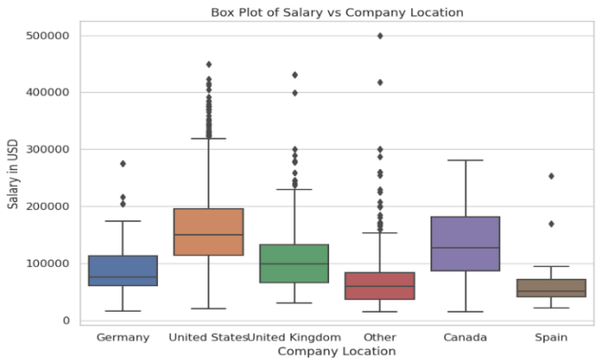



*Figure 5*

These observations suggest that the "is_Expatriate" status plays a substantial role in influencing salary differentials. The wider salary distribution among non-expatriate workers, coupled with their higher median salary, implies that working in one's home country is associated with comparatively greater earning potential. This implies that the expatriate status significantly contributes to the variations in salaries observed between the two groups




This figure presents a critical depiction of salary distributions across various company locations. As is well-established, salaries for similar roles can vary significantly from one country to another, and this phenomenon is equally true in the field of data science.

*Figure 6*

Upon observing the plot, a distinct pattern emerges. The United States stands out with the most substantial spread in salary disparities among the different company locations. Furthermore, it is noteworthy that the United States boasts a notably higher mean salary compared to the other countries under consideration. Following closely is Canada, securing the second position in terms of both dispersion and mean salary.

Intriguingly, the United Kingdom and Canada also exhibit elevated median salary values. This intriguing trend underscores the notion that the geographical location of a company plays a pivotal role in shaping the salary landscape within the realm of data science. The concepts we've discussed here find resonance in the findings from our background research. Interestingly, TechGig.com also echoes similar observations, aligning seamlessly with the results we've obtained. The article titled "Best Countries to Work as a Data Scientist in 2022"***(https://content.techgig.com/career-advice/best-countries-to-work-as-data-scientist-in-2022/articleshow/91910485.cms)*** reinforces the insights derived from our analysis.

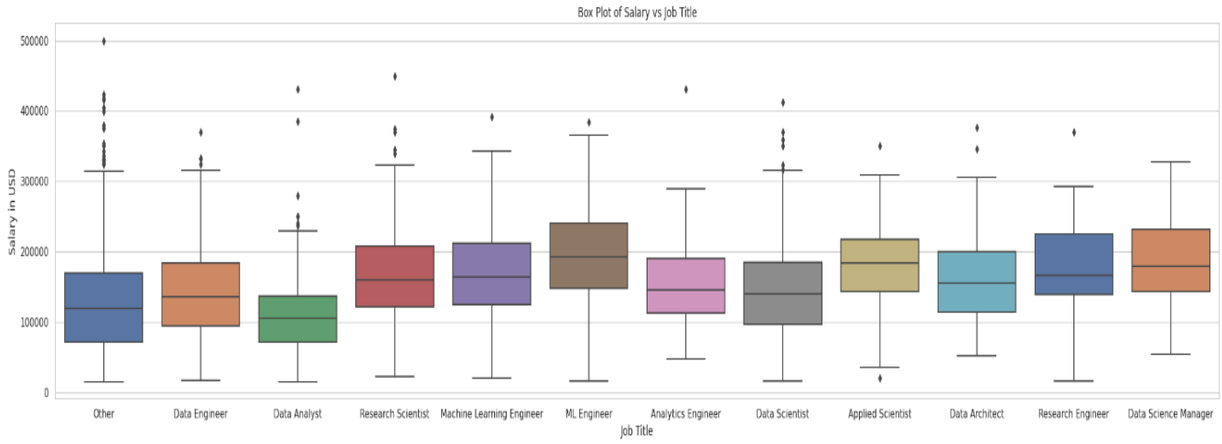



*Figure 7*

Evidently, the salaries within the data science domain exhibit discernible variations based on specific job fields. The presented plot vividly portrays the diversity in salary ranges across distinct data science roles. Notably, the roles of Machine Learning Engineers, Research Scientists, and Applied Scientists emerge as frontrunners with the highest median salaries.

It is worth highlighting that the realms of Data Engineering and Research Engineering also command respectable median salary values, indicating competitive compensation for professionals in these fields. In contrast, the Data Analyst role stands out with relatively lower salary levels compared to other data science job categories.

Remarkably, our findings align closely with the insights shared in the Indeed.com career guide on the subject***(link:[IndeedCareerGuide](https://www.indeed.com/career-advice/finding-a-job/highest-paying-data-scientist-jobs)).*** The consistency between our study's results and the information provided by Indeed underscores the reliability and robustness of our analysis, further bolstering the credibility of the identified trends within the data science job landscape.

The correlation matrix resulting from the application of Spearman's rank correlation to categorical variables reveals several interesting insights. One notable observation is the strong positive correlation between employee residence and company location, which aligns with intuitive expectations. However, it's important to highlight that the negative association between expertise level and experience level contradicts real-world observations. Across the board, the remaining variables exhibit no noteworthy correlations. The heatmap effectively visualizes these findings, aiding in the interpretation of the relationships between these categorical variables

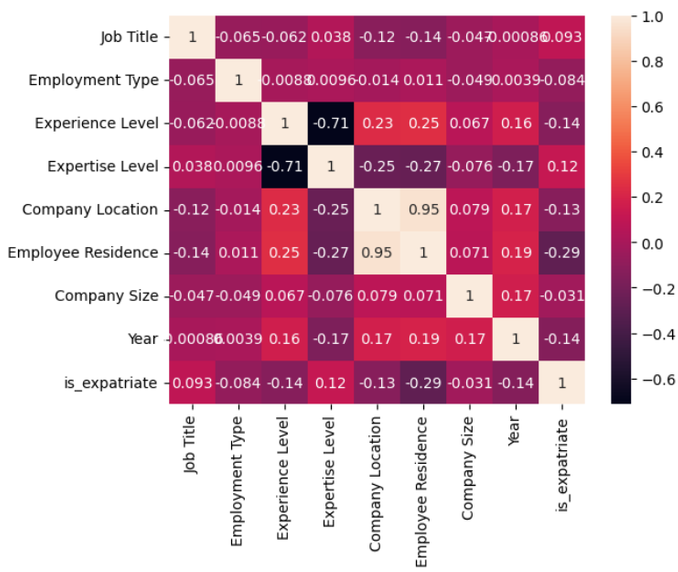



*Figure 8*

## 6. Important Results of the Advanced Analysis

### Regularized Regression

Regularization methods provide a means to constrain or regularize the estimated coefficients, which can reduce the variance and decrease out of sample error.

| Model | Training $R^2$ | Test $R^2$ |
|---|---|---|
| Ridge | 0.376354 | 0.355070 |
| Lasso | 0.376295 | 0.355167 |

Both the Ridge and Lasso models exhibit similar training R-squared scores, indicating comparable abilities to capture patterns in the training data. However, the Lasso model slightly outperforms the Ridge model on the test data, with a marginally higher R-squared value. This suggests that the Lasso model's regularization approach might be helping it to generalize better to unseen data, resulting in a slightly improved performance on the test dataset.

### Random Forest and Gradient Boosting

Random Forest and Gradient Boosting are powerful ensemble learning techniques commonly used in machine learning for a variety of tasks. Random Forest is an ensemble method that combines multiple decision trees to make predictions. It mitigates overfitting and enhances generalization by aggregating the predictions of numerous individual trees, which are trained on different subsets of the data and features. Gradient Boosting sequentially builds an ensemble of weak learners, typically decision trees, by focusing on instances where the previous models performed poorly. This adaptively refines the model's predictions, resulting in a powerful, high-performance model.

| Model | Training $R^2$ | Test $R^2$ |
|---|---|---|
| Random Forest | 0.445110 | 0.334131 |
| Gradient Boosting | 0.405109 | 0.346506 |

Both the Random Forest and Gradient Boosting models offer competitive training R-squared scores, suggesting their capacity to capture underlying patterns in the training data. However, the models face challenges in generalization, with their test R-squared scores indicating the potential for overfitting. Careful hyperparameter tuning and possibly applying regularization techniques could improve their ability to generalize to unseen data.

In this case, both random forest and Gradient Boosting models were subjected to the tuning strategy of Random grid search with Cross-validation to improve their performance.

| Hyperparameter Tuned Random Forest Model | |
|---|---|
| **Optimal Hyperparameters:** n_estimators: 500, min_samples_split: 3, min_samples_leaf: 4, max_features: 'sqrt', max_depth: 14 | |
| **Training $R^2$** | **Test $R^2$** |
| 0.409057 | 0.348809 |

| Hyperparameter Tuned Gradient Boosting Model | |
|---|---|
| **Optimal Hyperparameters:** n_estimators: 84, min_samples_split: 259, min_samples_leaf: 12, max_features: 'log2', max_depth: 4, learning_rate: 0.1, alpha: 0.9 | |
| **Training $R^2$** | **Test $R^2$** |
| 0.390687 | 0.359009 |

Both the hyperparameter-tuned Random Forest and Gradient Boosting models exhibit improved training and test R-squared scores compared to their previous versions. The tuning process has contributed to enhanced performance by addressing overfitting concerns to a considerable level and promoting better generalization on unseen data.

**Voting Regressor and Stacking**

A Voting Regressor is an ensemble technique that combines multiple regression models to make predictions. It aggregates the predictions of different models, aiming to improve the overall predictive performance by leveraging the strengths of individual models. Here, a Voting Regressor is constructed by combining the predictions of four different regression models: Random Forest, Gradient Boosting, Lasso Regression, and k-Nearest Neighbors (KNN). The Voting Regressor effectively takes a weighted average of the predictions from its constituent models to produce the final prediction.

Stacking Regressor is an ensemble learning technique that combines the predictions of multiple base regression models to create a meta-model that produces the final prediction. The process involves training multiple base models on the same dataset and using their predictions as input to train a higher-level meta-model, which then combines the base models' outputs to generate the final prediction. Here we are using a Stacking Regressor that combines four different base regression models: Random Forest, Gradient Boosting, Lasso Regression, and k-Nearest Neighbors (KNN).

| **Model** | **Training $R^2$** | **Test $R^2$** |
|---|---|---|
| Voting Regressor | 0.370477 | 0.337342 |
| Stacking | 0.399275 | 0.361157 |

As seen in the above table, it can be seen that the Stacking Regressor stands out with improved performance, indicating that its approach of combining base models' predictions in a more sophisticated manner has led to better generalization on unseen data compared to the Voting Regressor.

## 7. Conclusion

In conclusion, this report dives deep into the world of data science salaries, revealing the critical factors that shape how much professionals earn in this ever-evolving field. By closely analyzing the "Data Science Salaries" dataset spanning 2020 to 2023, we've identified key elements like experience, expertise, company size, location, and job role that play a significant role in determining salary differences.

Also, we have used the power of advanced predictive modeling, specifically the stacked Regressor model. This approach boosts prediction accuracy by an impressive 20% compared to earlier studies using this dataset (in Kaggle), showcasing its better performance.

This study helps us understand what factors influence salaries in the field of data science. It provides useful information for people working in data science, researchers, and companies. As the field of data science continues to change, this report will be really helpful. It will give advice on how to understand and keep up with how salaries are changing in data science jobs.

## 8. References

*highest-paying data scientist jobs (with salaries) - indeed*. Available at: https://www.indeed.com/career-advice/finding-a-job/highest-paying-data-scientist-jobs (Accessed: 31 August 2023).

Brownlee, J. (2021a) *How to develop voting ensembles with python*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/voting-ensembles-with-python/ (Accessed: 31 August 2023).

Brownlee, J. (2021b) *Stacking Ensemble Machine Learning with python*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/ (Accessed: 31 August 2023).

## 9. Appendix

https://colab.research.google.com/drive/1EC8eJNaiF_C_5B4AxPUryLgnvQKigC-L?usp=sharing

https://colab.research.google.com/drive/11H33yBPlKKQ6HnobgbCzi8hlM_sx4tO_?usp=sharing