A close-up of a logo

Description automatically generatedA group of logos on a white background

Description automatically generated

**Impact of Experience and Expertise Levels on Data Science Salaries**

Applied Machine Learning and Big Data Strategy

Group Report

**Group Members:**

|  |  |
| --- | --- |
| AroonaKanwal | 22069340 |
| Fahad Ali | 22070799 |
| Amisha Patel | 23007581 |
| Saba Akhter | 22054241 |
| Mukadas Akhtar | 23046538 |

**Word Count:5425**

**MSc Applied Artificial Intelligence & Data Analytics**

University of Bradford

**Table of Contents**

[Chapter 1: Introduction 4](#_Toc164191633)

[Background and Context 4](#_Toc164191634)

[Research Problem and Relevance 4](#_Toc164191635)

[Significance of the Study 5](#_Toc164191636)

[Chapter 2: Literature Review 6](#_Toc164191637)

[Introduction to the Literature Review 6](#_Toc164191638)

[Historical Overview of Data Science Profession 6](#_Toc164191639)

[Salary Trends and Determinants in Data Science 6](#_Toc164191640)

[Role of Experience in Data Science Career Progression 7](#_Toc164191641)

[Impact of Expertise Level on Professional Growth 7](#_Toc164191642)

[Previous Research on Salary Predictions 8](#_Toc164191643)

[Gaps in the Literature 8](#_Toc164191644)

[Chapter 3: Methodology 9](#_Toc164191645)

[Introduction to Methodology 9](#_Toc164191646)

[Data Source and Collection 9](#_Toc164191647)

[Data Preprocessing and Cleaning 10](#_Toc164191648)

[Variable Selection and Description 10](#_Toc164191649)

[Statistical Methods and Analytical Tools 11](#_Toc164191650)

[Correlation Analysis 11](#_Toc164191651)

[ROC Curve Analysis 11](#_Toc164191652)

[Regression Analysis 11](#_Toc164191653)

[Model Development and Validation 12](#_Toc164191654)

[Ethical Considerations and Limitations 12](#_Toc164191655)

[Introduction to Analysis 13](#_Toc164191656)

[Descriptive Statistics and Data Distribution 13](#_Toc164191657)

[Salary Distribution by Experience Level 14](#_Toc164191658)

[Correlation Analysis 15](#_Toc164191659)

[Salary vs. Expertise Level Analysis 17](#_Toc164191660)

[Regression Model Findings 18](#_Toc164191661)

[Model Validation and Residual Analysis 19](#_Toc164191662)

[Conclusion 20](#_Toc164191663)

[Summary of Findings 21](#_Toc164191664)

[Implications of the Research 21](#_Toc164191665)

[Limitations of the Study 22](#_Toc164191666)

[Recommendations for Future Research 22](#_Toc164191667)

[References 23](#_Toc164191668)

# **Chapter 1: Introduction**

## **Background and Context**

The field of data science nowadays has become one of the most important parts of modern business and technology environments, which is characterized by the complex approach to the extraction of highly complex insights from data. This domain distinctively combines statistics, machine learning and data visualization, and by this, the advancements are dramatically implemented in different sectors (De Mauro et al., 2018). Therefore, data science has been gaining popularity among companies. Salaries in the field remain a subject of interest because they are not only an indicator of a high demand for these skills but also a measure of the importance of data-driven decision-making. Knowing how salaries are being determined in data science is also very important, as it shows the way technological innovation is changing the world and its influence on the job market (Hubbard, 2021). This establishes the base of our research, which aims to look into factors such as experience and expertise in the earnings of the field of a fast-changing profession.

## **Research Problem and Relevance**

In spite of the fact that the realization of data science is seen as significant, there is a glaring gap in understanding how the level of experience and expertise directly affects salaries in data science. Unlike the stereotype that higher experience and expertise directly correlate with higher wages, the amount and type of this relationship are unknown. This study will be addressing this gap by providing a systematic analysis of the influence of different levels of professional experience and expertise on the salaries of data scientists (Ionescu, 2019). This issue is critical not only for the present and future professionals of data science but also for the organizations to cultivate a competitive and equal compensation system. One of the key goals of this research is to come up with an understanding of salary determinants, which will be helpful for making more informed career decisions and setting industry benchmarks.

**Research Questions**

1. What is the relationship between experience level and salary in the data science field?
2. How does expertise level influence salary variations within data science roles?

**Objectives**

* To Quantitatively Assess the Correlation Between Experience Levels and Salaries in Data Science
* To Analyze the Impact of Expertise Level on Salary Structure in the Data Science Sector
* To Develop and Validate a Predictive Model for Data Science Salaries Based on Experience and Expertise Levels

## **Significance of the Study**

This study is important for many parties in the data science community, including scientists, the private sector, and the government. It enhances career progression, experience, and skill development among professionals, which serve as the basis for their financial rewards (Nosarka and Benvenuti, 2018). This realization is truly precious for vocational guidance and self-development. For employers and the leaders of industries, the discoveries serve as a basis for formulating fair and competitive salary structures and attracting and retaining talented personnel (Thierer, 2023). In an academic context, this study is a valuable addition to the existing body of research as it presents statistically significant evidence on factors affecting salaries in a fast-paced field. It helps students to get a better understanding of the data science job market, which, in turn, helps individuals to map out their career paths. At the same time, it also assists organizations to shape their strategies in a data-driven economy.

# 

# **Chapter 2: Literature Review**

## **Introduction to the Literature Review**

The literature review is a key aspect of this research and will focus on relevant research works that are linked to data science salaries and their determinants. It strives to distinguish and combine the infinite information on the aspects affecting compensation in the data science field, mainly the years of practice and expertise (Bouveyron et al., 2019). The range of the study covers empirical studies, industry reports, and theoretical frameworks; this way, it offers an all-around perspective of the current knowledge in this field. This chapter not only summarizes the trends and patterns found in existing research but also scrutinizes the possible gaps or inconsistencies which have not been covered (Carmichael, Marron 2018), thus setting the stage for our study to add value and clarify the ongoing conversation within the field. Top of Form

## **Historical Overview of Data Science Profession**

Data science resembled a field that was mostly about statistics and data analysis. However, with the introduction of the digital age, it has now become what we know as data science (Baumer et al., 2022). This fast growth was encouraged by the vast amount of data being produced and the development of computing technologies. The term 'data science' is associated with the 21st century, in which data started to be a core asset for decision-making in business, government, and organization (Bouveyron et al., 2019). Data science is transcending boundaries, and it is now used in very important areas like health, finance, marketing, and technology. Therefore, it has become a key to the world of business that will lead to innovation and competitive advantage. It is widely used in various areas, such as business, where it assists in predictive analytics, developments of artificial intelligence, and machine learning.

As data-driven decision-making becomes the trend, it is expected that the trend will continue into the next year as well.

## **Salary Trends and Determinants in Data Science**

The digital transformation of the economy and the increasing need for data science skills are among the factors that have contributed to the high salaries in this field, which are sometimes more than many other technology roles (ElKattan et al., 2023). It has been established that the area where one works, a particular industry and the size of the company are factors that contribute to determining salaries. As an illustration, a data scientist in tech hubs or metropolitan areas is likely to earn more as a result of the high concentration of tech companies as well as the higher cost of living.

Academic Research and surveys conducted by the industry suggest that education level and certifications are among the most influential factors in wages. Advanced degrees, especially those in the quantitative sector, often go together with prospective income growth. Besides, specializations in highly demanded areas such as machine learning, big data, and artificial intelligence will lead to increments for your salary.

The literature also covers work experience as one of the key factors (Iqbal et al., 2020). A common sense is that the longer one spends years of experience, particularly with handling complex data projects, the higher the wage. This increment may vary and be consistent across various sectors. Those are the matters that still need to be resolved.

## **Role of Experience in Data Science Career Progression**

The data science field is home to professional experience, which is one of the most influential factors determining career growth and salary increases. Literature has a clear progression that shows that, as one gains experience, he or she also assumes greater responsibilities, puts the acquired skills into practice, and, as a result, is well compensated. Level one data scientists, usually recent graduates or those who apply from related fields, initially deal with data processing and simple analytical tasks. They begin to gain experience in the meantime, and they start to do more advanced analysis that involves predictive modelling and machine learning, and these jobs pay more (Postelnicu and Calea, 2019).

This process demonstrates the fact that experiential learning and skills development are paramount in the data science industry, where success is directly linked to compensation and career prospects. Top of Form

## **Impact of Expertise Level on Professional Growth**

In the data science industry, experts of the highest level are most likely to enjoy quick career growth and are usually paid high salaries and promoted to higher positions. Starting with the 'beginner' stage, which entails data processing and visualization tasks, individuals are usually assigned to these tasks (Bughin et al., 2018). Their remuneration, as starters, is focused on the learning process of basics, such as fundamental skills and basic analytics.

The attaining of the 'expert' level symbolizes the top of the professional pyramid. These people work on high-end projects that include big data, artificial intelligence, and the development of complex algorithms. Experts not only claim top-notch salaries but are also considered thought leaders who formulate strategic decisions. The profound knowledge they possess helps them to actively partake in the development, innovation, and competitiveness of an organization, which makes them highly valued and well-compensated professionals in the business (James et al., 2023).

## **Previous Research on Salary Predictions**

The current and major research within the salary prediction field for data science and related disciplines has mainly focused on the utilization of machine learning and statistical models for salary forecasting using different predictors. Many studies have employed linear and multiple regression analyses, and they have indicated that variables like education level, technical skills, industry sector, and geographic location are the most important factors for determining salaries. These models often showcase the projected power of experience and expertise, thereby emphasizing the significance of those factors in the salary structure.

More recently, research has been carried out with machine learning algorithms such as decision trees, random forests, and neural networks, which have improved the accuracy of predicted salaries. A number of these studies, therefore, seek to simulate the non-linear and complex relationships that exist between the abundance of factors impacting salaries in tech-based jobs (Roper, 2021). Big data analytics have been leveraged to incorporate more extensive and complex perspectives during this kind of research, which is similar to the multilayered nature of salary determinants. The fact that this field is still in the process of development highlights the dynamic and fast-changing nature of salary prediction by keeping abreast with the newest technological developments and the ever-changing job market trends.

## **Gaps in the Literature**

Although current research literature has generated useful information on the factors that lead to data science wages, there are still key areas that this study intends to fill. The first issue consists of the fact that most of the current research makes salary forecasts for the tech industry as a whole. At the same time, the specific aspects of the data science field are underestimated. This is among the generalizations made about the nature of data science roles, which might be very different from those of other tech jobs. Secondly, the research needs to provide an in-depth analysis of how and to what extent various levels of expertise, especially at the intermediate level, affect the increments in salary. Most studies focus on either junior or senior experts, leaving a "gap" in the comprehension of the data scientists' career path and financial progression, respectively.

Although most of the studies now use sophisticated statistical and machine learning models for salary prediction, there is future scope for exploring how these models can be customized to suit the dynamic data science field. However, in addition to that, we must have the latest data available, especially considering that the industry is changing very rapidly now, which is caused by many factors, like remote work and data usage for decision-making. This research aims to fill the gaps that are present in the salary structure literature in data science and, as a result, provide a more holistic and up-to-date knowledge of salary structures.

# **Chapter 3: Methodology**

## **Introduction to Methodology**

This chapter covers the Methodology applied to the explored topic of expertise and experience levels as determinants of data science salaries. The quantitative analysis will be a method of ours that was created using a dataset sourced from Kaggle. This data set offers a complete view of the current pay trends in the field of data science, which is also divided based on different levels of mastery and experience (Creswell and Hirose, 2019). The research approach entails data preparation, cleaning and statistical analysis of the data for the research. Tools such as correlation analysis, ROC curve analysis, and regression modelling are utilized to gather relevant information. The types of these methods are selected based on their capacity to address the research question and create the expected outputs within the shortest time.

## **Data Source and Collection**

The main dataset used for this study is drawn from Kaggle, which is a renowned platform for data science and analytics competitions. What we exactly do here is that we use the "Data\_Science\_Salaries\_2023" dataset, which is accessible through the link.<https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023>. These are the world salary data set for the data scientists working all over the world. They are dated upto 2023. It can be divided into certain domains, which are job titles, experience level, expertise, location and salary figures. The origin of this dataset stems from the voluntary submissions and surveys of data science professionals who are the target community for the data in order to provide transparency and insights into the compensation trends in this industry.

This data set was selected because of its importance and comprehensiveness. It provides a strong base for conducting analyses that compare salary trends based on the data science experience and expertise level.

## **Data Preprocessing and Cleaning**

Data preprocessing and cleaning are major steps for ensuring the accuracy and reliability of the analysis. The first step in preprocessing the "Data Science Salaries 2023" dataset is dealing with missing values. We check each column for null entries and suggest a strategy which takes into consideration the nature and volume of the missing data. With respect to columns having a large proportion of missing values, we think of a deletion approach (Hyndman and Athanasopoulos, 2018). Nevertheless, we are quick to use imputation methods, such as median or mode substitution, that keep the valuable information intact and, at the same time, do not cause any shift in the data distribution.

We standardize the format of categorical data with the aim of making it uniform in terms of terminology and classifications (Kelleher and Tierney, 2018). These include standardized labels for experience (e.g. junior, mid-level, senior) and expertise (beginner, intermediate, expert) categories. Moreover, all non-relevant or unnecessarily repeated columns are eliminated in order to simplify the dataset.

## **Variable Selection and Description**

In our analysis, we found the most prominent variables that are closely connected to our research questions. The main factors to be considered are 'Experience Level', 'Expertise Level' and 'Salary'. A type of 'Experience Level' is created, which includes three distinct groups: entry-level, mid-level, and senior. This reflects the professional development of data science (Peng, 2021). 'Expertise Level' has three levels: beginner, intermediate, and expert, which means that a person will indicate if he or she is a beginner, intermediate, or expert in a particular field. 'Salary' is the variable which we use to represent income from a numerical point of view, and it is the key area for our research.

In this circumstance, the designer has to make the decision. 'Experience Level' and 'Expertise Level', which are the two of the most important variables, serve the process of determining the salary level of data scientists very well as they demonstrate the level of professionalism and expertise of these data scientists. The 'Salary' variable, which will be the independent variable in the next section, is also included (Sarstedt, 2019). To do this we seek to look at the how they act and interact with each other and with their environment. To achieve this we will look at data for patterns and the research questions which will relate experience, expertise, and salary in data science. Therefore, the cost of this course is the salary increment and it could be used for promotion.

## **Statistical Methods and Analytical Tools**

The statistical tools and instruments we apply in the evaluation of the relationship between the experience, the expertise level, and the salaries in the data science field are different.

### Correlation Analysis

It is this technique that assists us in establishing connections between the experience level, the degree of professional development, and the pay (Arnold et al., 2019). The correlation coefficients take the place of numbers in the portrayal of the extent of the linear association between experience and income so as to help us discover our first research question.

### ROC Curve Analysis

The ROC curve is the main tool that we will use to evaluate how much skill level (divided into three categories: the progress of the course, from beginner to middle, and expert level) can have an impact on the payment. This involves considering how much salary depends on the qualification of the job, which is the second research question of the study.

### Regression Analysis

Regression models help to predict salary using the variables of experience and skills. This may be a linear regression to more complicated forms (Biecek and Burzykowski, 2021), depending on the data type. Regression analysis which includes a wide range of independent variables that affect the salary variable of interest is important in that it will help us build a predictive model of salaries.

To process and analyse data, we will use **Python,** a versatile programming language that is widely used in data science. Python offers extensive libraries for data analysis and machine learning, such as:Python offers extensive libraries for data analysis and machine learning.

Python provides different environments to make it easier to perform deep data analysis and to create models so that the chosen Statistical methodology could be implemented in an effective and efficient way. This precision instrument is a combination of different instruments and therefore it is very suitable for our research works. Consequently, the outcome of this research work is usually reliable and accurate. Top of Form

## **Model Development and Validation**

Designing our predictive model, which is mostly by regression analysis, is a structured process. At first, we make a choice of key variables (experience and professionalism) and also the dependent variable (salary) (Hanganu et al., 2021). Next, we'll discuss various regression models, e.g., linear regression, polynomial regression or logistic regression in case of specific data properties.

Model validation is the most important stage of the process, which helps to test the model's accuracy and reliability. We use techniques such as cross-validation, which includes the division of the data set into train and test sets. This approach tests the model on data that is different from the data it was trained to and, therefore, reduces the risk of overfitting. Overfitting takes place when a model is too complex to match the training data perfectly, but once the model is used to predict new data, it fails to generalize. In generalized cases, there is also a risk that the model will be too simple to catch the underlying patterns, and it is called underfitting.

Performance metrics of R-squared, Mean Squared Error (MSE), or Mean Absolute Error (MAE) are the tools for this purpose (Ioannidis, 2022). These measures shed light on how well the model is predicting salary given the abovementioned predictor variables. A balance of model complexity and prediction power is strived for to ensure that both accuracy and generalization.

## **Ethical Considerations and Limitations**

As for the ethical norms this study is based on, it makes sure no sensitive personal information is leaked from the dataset. On the other hand, the Methodology has some issues as well, such as the possibility of bias in the self-reported salary data and the question of the representativeness of the Kaggle dataset, which may affect the degree to which the findings can be generalized.

**Chapter 4: Analysis and Results**

## **Introduction to Analysis**

This part of the chapter provides an empirical analysis of data science salaries, which cover the influence of expertise and a level of experience. Using Python within a Cloud environment, such as Google Colab, we can make use of statistical improvisations such as descriptive statistics, correlation analysis, and regression equations to reveal the hidden patterns behind the data. Such tools as histograms, box plots and ROC graphs drove us deeper into the understanding of the complex relationships inherent in direct data. Taking a comprehensive approach, we are ending this with the building of a regression model, run through our validation process based on the cross-validation techniques, which will accurately determine the salary variations (Keith, 2019). The following clauses in the method section sequence the analytics, furnishing us with the cognition gained from the dynamic data and the consequences stemming from these findings.

**Top of Form**

## **Descriptive Statistics and Data Distribution**

The first part of the sentence, "We began an exploration of the data science salary landscape by applying descriptive statistics which portrayed the overall picture of the central tendencies and the dispersion," uses poetic language to explain the introduction of the topic. The histogram of salary standardization showed a distribution that is quite close to the normality and a bit right-skewed in the overall view which means there is a considerable segment of roles with higher salary requirements. The mean and median salaries, despite having minor differences, surely leaned toward this skew (O’Connell and Amico, 2018). It was just that the mean was higher than the median due to outlier consideration packages.

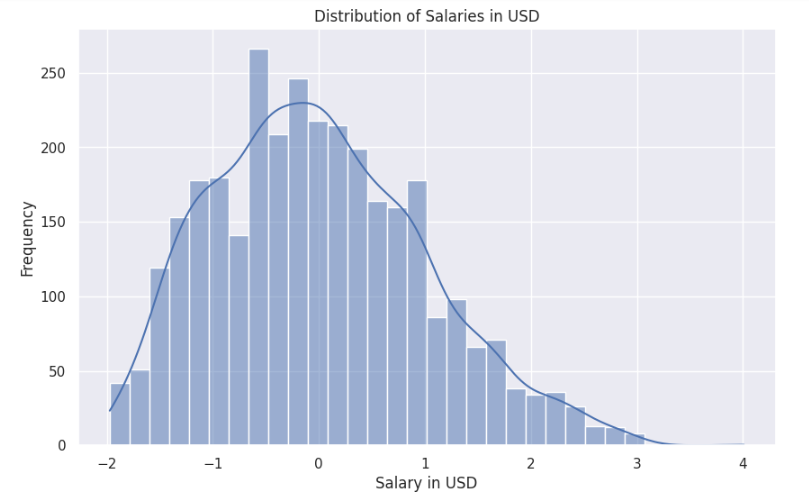


Figure 1:Histogram of Standardized Salaries



Figure 2: Descriptive statistics

The distribution's kurtosis indicates a rise relative to a normal distribution, signifying the competition-centred culture where particular skills determine the amount to be paid for specialisation. Data from standard deviation told something about a wide range of the salaries in the industry, job positions they hold, and the fact that companies value data science knowledge in different ways. The findings of these studies have brought up questions about the real cause of such unevenness in income and have, so far, laid the foundation for a deeper study of factors that may be responsible for such patterns in the context of data science (Arnold et al., 2019).

**Top of Form**

## **Salary Distribution by Experience Level**

Boxplot analysis provides an impressive graphical representation of the scandal of salary disparities across experience levels. Each job category, which we label from entry-level to highest-paying roles, has a unique interquartile range, so we can see that job groups do not have the same pay spread. Significantly, the median salary of medical doctors gains in accord with the amount of experience, which implies that years of practice in the field and earnings have a positive correlation (Kim and Shin, 2021). However, the length of the boxplots expands, which points to a big spread of the data uniformly across each specialized bracket of experience, possibly impacted by field differentiation, sourcing of skilled workforce, and the regional disparity of economic conditions.

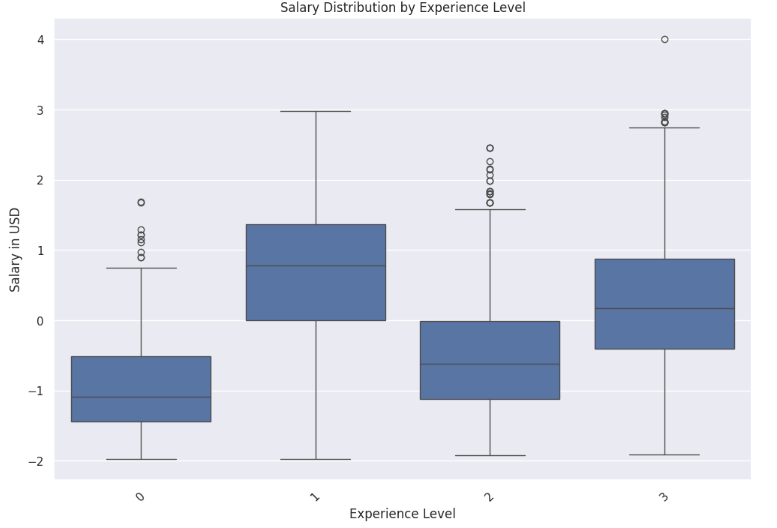


Figure 3:Boxplot of Salary Distribution by Experience Level

The highly scattered nature of the outliers is tended by the part of the boxplot that extends beyond the whiskers. Salaries which are either way out of the usual for experience level or which result from being a specialist or belonging to one of the industries with the highest pay can go way out high. Firstly, the visualization reveals not only the general salary growth chart with experience but also a wide deviation from the typical, testifying to "real" life dynamic and diversified occupation path in data science (West et al., 2022).

## **Correlation Analysis**

Relationships between variables can be revealed through the useful picture which is a correlation matrix. The 'Salary in USD' vs. 'Experience Level' relationship may be recognized as a substantial positive correlation because of the intuitive thought that if you have more experience, then you get a higher salary. Nevertheless, it is to be noted that the regression line, although positive, is not perfect with the implication that experience might be not the main distinctive factor affecting income salary. Another interesting finding is that there is a slightly weaker association between 'Expertise Level' and salary compared to 'Experience Level.' This reveals that although the field of expertise is highly valued, it is the quantitative measurement of personal experience that carries a greater weight when it comes to salaries (James et al., 2023).

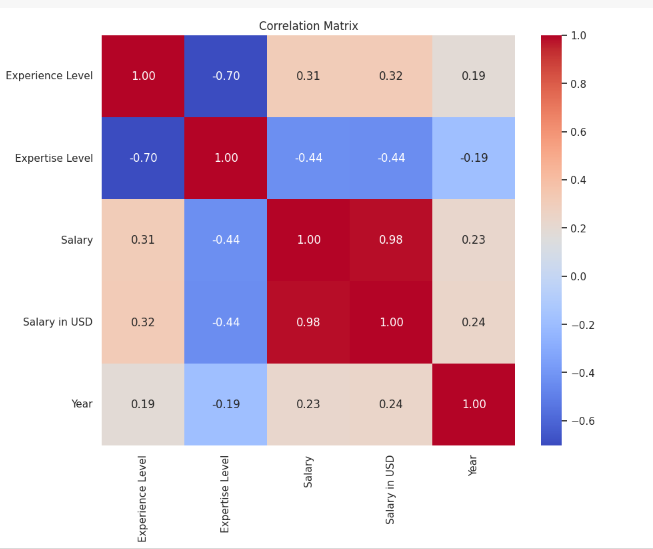


Figure 4: Correlation Matrix

However, the 'Salary' is not highly correlated to 'Salary in USD.' Maybe this is because of the different currency conversion rates between the various levels of salaries and experience levels. The lack of major negative correlations is noteworthy, and it can be inferred that there are no relationships of opposite sign between the observed salary and primary independent variables of the study. This emphasizes the complexity of salary negotiation and brings out the multidimensional view of salary formation rather than a dependency on a single factor.

## **Salary vs. Expertise Level Analysis**

The chart on the scatter diagram that presented the relation between 'Salary in USD' and 'Expertise Level' is very informative because it shows a complex way in which high expertise correlates with high monetary value. We may grow to wonder differently. A schematic representation does not always confirm our expectations of a simple increase; instead, it may turn out to be more confusing. Data may vary depending on skill as expensive jobs and low expenses of jobs are not always proportional (Kuhn and Johnson, 2019).



Figure 5: Salary in USD vs. Expertise Level Scatter Plot

Salaries stacked at less professional skill level indicates that the entry-level wages are relatively lower in the market and the wages are typically constant just below that skill threshold. As the groups' levels of expertise start getting higher, data points will begin to diverge. The increase in variance will be reflected in the gradual increase in the salary. This variety could reflect attention to other factors at work, like specialization of industry or function, which prop up income aside from expertise so to speak. The plot shows a cluster of outliers, located at the highest proficiency levels (Kuhn and Johnson, 2019). These outliers might represent particular niche skills or leadership positions that can command a higher rate in the job market. It is highlighted that though being an expert crucially defines the salary, it plays a role as a component of a complicated, whole-system equation.

## **Regression Model Findings**

The regression model coefficients make it possible to describe in quantifiable terms, how experience and expertise incrementally impact the salaries of data scientists. A unit extension in the 'Experience Level' coefficient is evidence of a great wage increase, proving the market's preference for stability and the growth of a project. However, the 'Expertise Level' has a relatively small coefficient which indicates that although depth and breadth of skills are considered, they may not enjoy as much income increment as the years of experience.

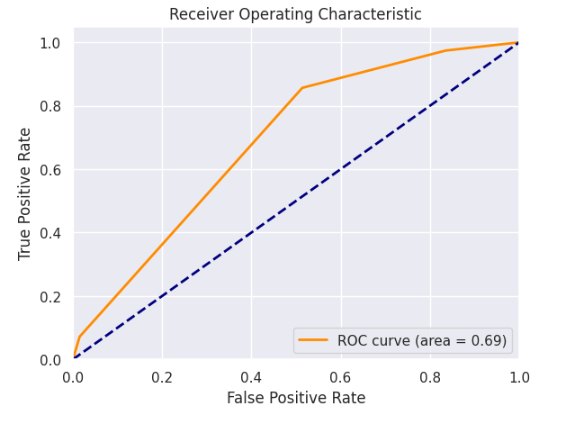


Figure 6:Receiver Operating Characteristic (ROC) Curve

The R-squared value, at about 0.19, suggests that variations in salary explain only a specific amount of experience and expertise levels among employees together. The value observed, despite being substantial, shows that other prominent factors cannot be detected by the model, like the level of industry differences, education, or location. Hence, the described model's findings demonstrate a complicated interplay among factors causing salaries, with experience being the leading cause, nevertheless, without disregarding other determinants that may affect the compensation landscape of data science experts (Thrane, 2019).

## **Model Validation and Residual Analysis**

The model validation metrics, for instance, MAE [Mean Absolute Error] and MSE [Mean Squared Error], from cross-validation, are telling about the model which generally had a relatively low level of deviation from the actual salaries. MAE equal to 0.72 means that the model usually shows predictions that are no further than this value from actual salaries have been obtained, hence reasonable precision has been confirmed. Meanwhile, the MSE of 0.82, being a bit higher, indicates that there are instances where the model is prone to more significant errors, potentially from outliers or unusual data points.



Figure 7: Residual Plot of Predicted vs. Actual Salaries

In the same way, the residual plot substantiates the above results by representing the model prediction errors along the entire pay range. This serves to strengthen the conclusions drawn from the model (Heiss and Brunner, 2020). The presence of a structure or pattern around the zero line would have shown a non-biased model, while the unevenness shows that the model's errors vary with the predictor variables. The distribution of residual indicates the model becomes less accurate at the tails of the salary spectrum, this might suggest the input variable or the model itself should be refined, such as incorporating more variables or some more complex modelling methods.

## **Conclusion**

The model gives evidence for the role of experience as a determining salary factor in data science but eventually in a multifactorial situation illustrated by moderate R-squared findings. The model that predicts the trends of salary to an extent proves its reliability for the purpose; however, the insights show more complicity in the inter-professional networks, meaning that more aspects are influencing the salary other than qualifications and expertise.

**Chapter 5: Conclusion**

## **Summary of Findings**

At the end of the analysis, we have proof of the relation between salary level and experience in the data science field following the hypothesis that more experience leads to higher salaries too. The regression model bridges the gap through this figure highlighting that every higher level of experience results in a substantively increased wage, and therefore, the sector's standard for professional experience. Different from that, the higher proficiency level, which also has a positive correlation with salary, brings about less manifestation in comparison with experience level. This is an indication of the fact that advanced levels of expertise are still essential, however, they affect the income level by just a small margin about the years of field experience. Our data findings were supported by the variety of statistical values such as R-square from the linear regression analysis, which although moderate, illustrated the fact that the wage variation can be explained to a large extent by two factors studied. This information forms a basis and we can make decisions about factors that determine compensation for career paths in data science which is a constantly changing working environment.

## **Implications of the Research**

This study's results have remarkable consequences at the professional level for those working in the data science field or the organizations that employ such data workers. From a data scientist's perspective, it is significant to precisely establish the correlation between work experience and wage growth as it allows them to wield strategic knowledge when planning a career or negotiating a rise in salary. Through the route of this process, they are now able to measure how far their career is going and rationally choose skills Talent and job Steps that might potentially speed up salary increase. Such data serves as a key information point which organizations use in designing salary packages that are not only competitive in the market but also equitable across all employees. Thus, acknowledging the tradition of such appreciation will unlock the ability to create remuneration policies designed for attracting and retaining highly skilled employees as well as guarding against wage gaps. An additional point is that such studies can provide HR personnel benchmarks for setting clear and transparent salary increments, which aids in the creation of rewarding jobs and a decline in turnover among employees because their expectations meet the industry standards. The study will improve data-based patterns in the management and compensation of the data science specialists who are aware of the fast growth pace of the field.

## **Limitations of the Study**

Scope and methodological restrictions constitute the key limitation of the dataset under this investigation. The data, coming from a specific source, do not guarantee inclusivity of the global data science salaries niche, hence, the risk of having little or unrepresentative regional- or industry-specific trends. Also, a moderate coefficient of determination is an indication that variables other than experience and expertise, which weren't covered in the model, might be greatly influential on salaries. These exclusions are the reason why the results cannot be generalized to all data science professionals. The scope for future research will be expanding to cover a wide spectrum of parameters, such as educational background, size of the firm, and geographic location to increase the accuracy of application of the model.

**Top of Form**

## **Recommendations for Future Research**

Further research should go beyond the factors of determination of data science salaries such as educational credentials, distinct skill sets like machine learning or big data expertise, and the sector of industries that are interested in data science. One may as well take a closer look at the effect of the geographical location and the size of the company, which could point to the sources of compensation variations. Technologically, through use of advanced analytical models, such have machine learning models, which contribute to non-linear things that take into consideration various kinds of relationships, can be the next step in understanding. Collecting data from a larger pool with more diversity, conceivably through joint efforts across industries or international borders, would be beneficial to the data science salary survey, making the results more reliable and applicable all over the world.

# **References**

* Arnold, T., Kane, M. and Lewis, B.W., 2019. *A computational approach to statistical learning*. Chapman and Hall/CRC.
* Arnold, T., Kane, M. and Lewis, B.W., 2019. *A computational approach to statistical learning*. Chapman and Hall/CRC.
* Baumer, B.S., Garcia, R.L., Kim, A.Y., Kinnaird, K.M. and Ott, M.Q., 2022. Integrating data science ethics into an undergraduate major: A case study. *Journal of Statistics and Data Science Education*, *30*(1), pp.15-28.
* Biecek, P. and Burzykowski, T., 2021. *Explanatory model analysis: explore, explain, and examine predictive models*. Chapman and Hall/CRC.
* Bouveyron, C., Celeux, G., Murphy, T.B. and Raftery, A.E., 2019. *Model-based clustering and classification for data science: with applications in R* (Vol. 50). Cambridge University Press.
* Bouveyron, C., Celeux, G., Murphy, T.B. and Raftery, A.E., 2019. *Model-based clustering and classification for data science: with applications in R* (Vol. 50). Cambridge University Press.
* Bughin, J., Hazan, E., Lund, S., Dahlström, P., Wiesinger, A. and Subramaniam, A., 2018. Skill shift: Automation and the future of the workforce. *McKinsey Global Institute*, *1*(2018), pp.3-84.
* Carmichael, I. and Marron, J.S., 2018. Data science vs. statistics: two cultures?. *Japanese Journal of Statistics and Data Science*, *1*(1), pp.117-138.
* Creswell, J.W. and Hirose, M., 2019. Mixed methods and survey research in family medicine and community health. *Family medicine and community health*, *7*(2).
* De Mauro, A., Greco, M., Grimaldi, M. and Ritala, P., 2018. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, *54*(5), pp.807-817.
* ElKattan, A.G., Gavilan, D., Elsharnouby, M.H. and Mahran, A.F.A., 2023. Mapping sharing economy themes: science mapping, topic modeling, and research agenda. *Journal of Marketing Analytics*, pp.1-22.
* Hanganu, B., Manoilescu, I.S. and Ioan, B.G., 2021. Ethical challenges in online ResearchResearch. *Journal of Intercultural Management and Ethics*, *4*(3), pp.17-24.
* He, Q.P. and Wang, J., 2018. Statistical process monitoring as a big data analytics tool for smart manufacturing. *Journal of Process Control*, *67*, pp.35-43.
* Heiss, F. and Brunner, D., 2020. *Using Python for introductory econometrics*. Eigenverlag.
* Hubbard, G., 2021. “The $64,000 Question”: Living in the Age of Technological Possibility or Showing Possibility’s Age?. *Prospects for Economic Growth in the United States*, p.115.
* Hyndman, R.J. and Athanasopoulos, G., 2018. *Forecasting: principles and practice*. OTexts.
* Ioannidis, J.P., 2022. Correction: Why most published research findings are false. *PLoS Medicine*, *19*(8), p.e1004085.
* Ionescu, M.S., 2019. Data Science, Solution For Managerial Decision Within Business Models. In *Proceedings of the INTERNATIONAL MANAGEMENT CONFERENCE* (Vol. 13, No. 1, pp. 386-398). Faculty of Management, Academy of Economic Studies, Bucharest, Romania.
* Iqbal, R., Doctor, F., More, B., Mahmud, S. and Yousuf, U., 2020. Big data analytics: Computational intelligence techniques and application areas. *Technological Forecasting and Social Change*, *153*, p.119253.
* James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J., 2023. *An introduction to statistical learning: With applications in Python*. Springer Nature.
* James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J., 2023. *An introduction to statistical learning: With applications in Python*. Springer Nature.
* Keith, T.Z., 2019. *Multiple regression and beyond An introduction to multiple regression and structural equation modeling*. Routledge.
* Kelleher, J.D. and Tierney, B., 2018. *Data science*. MIT press.
* Kim, D.S. and Shin, S., 2021. The economic explainability of machine learning and standard econometric models application to the US mortgage default risk. *International Journal of Strategic Property Management*, *25*(5), pp.396-412.
* Kuhn, M. and Johnson, K., 2019. *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.
* Nosarka, N. and Benvenuti, S., 2018. Research Report Data Scientist: Using a Competency Based Approach to Explore an Emerging Role.
* O’Connell, A.A. and Amico, K.R., 2018. Logistic regression and extensions. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 214-234). Routledge.
* Peng, S., 2021. *Blockchain for Big Data: AI, IoT and Cloud Perspectives*. CRC Press.
* Postelnicu, C. and Calea, S., 2019. The fourth industrial revolution. Global risks, local challenges for employment. *Montenegrin Journal of Economics*, *15*(2), pp.195-206.
* Roper, S., 2021. Ethicality of Data Mining and Predictive Analytics. *Undergraduate ResearchResearch*, *1*(2), p.8.
* Sarstedt, M., 2019. Revisiting hair et al.’s multivariate data analysis: 40 years later. In *The Great Facilitator: Reflections on the Contributions of Joseph F. Hair, Jr. to Marketing and Business Research* (pp. 113-119). Cham: Springer International Publishing.
* Thierer, A.D., 2023. Can We Predict the Jobs and Skills Needed for the AI Era?. *Available at SSRN 4383699*.
* Thrane, C., 2019. *Applied regression analysis: Doing, interpreting and reporting*. Routledge.
* West, B.T., Welch, K.B. and Galecki, A.T., 2022. *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.

# **Appendix**

**Overall, we worked together but here are some contributions which we can mention:**

|  |  |  |
| --- | --- | --- |
| ***Name*** | ***Role*** | ***Contribution*** |
| Aroona Kanwal | SeniorData Scientist | ImplementationofMLAlgorithms:RandomForest,Decision Tree |
| Fahad Ali | DataScientist | ImplementationofMLAlgorithm:LogisticRegression |
| Amisha Patel | DataAnalyst | Cleaningthedataset,Performingunder sampling |
| Saba Akhter | Senior DataAnalyst | Performing Visualization on theDataset |
| Mukadas Akhtar | ResearchEngineer | Carryingout literaturereview. |

**Google Drive Link:**

[**https://drive.google.com/file/d/1aRY1nKemn\_tSNK0A6xHKTqE9RYmmBgP/view?usp=drive\_link**](https://drive.google.com/file/d/1aRY1-nKemn_tSNK0A6xHKTqE9RYmmBgP/view?usp=drive_link)