# **Abstract**

This study explored job market dynamics by performing clustering, classification, and regression analyses on job-related data. Clustering analysis, using K-Means and Hierarchical Clustering, revealed four distinct job profile groups based on salary ranges, seniority, and skills like Python and AWS. Classification analysis utilizing Decision Tree and Naïve Bayes models achieved 100% and 96% accuracy, respectively, with Decision Tree outperforming in distinguishing seniority levels. Regression analysis identified key predictors, including max\_salary and min\_salary, with Linear Regression explaining 60% (R² = 0.60) of the variation in average salary. The analysis highlights patterns consistent with prior literature, such as salary increases with seniority, but diverges by emphasizing the role of technical skills in modern job profiles. Implications include a data-driven approach to understanding salary structures and skill demands, aiding professionals and employers. Limitations include dataset bias and moderate predictive accuracy, suggesting further refinement of features for improved insights.

# Introduction

The rapid digitization of businesses, growing connectivity, and widespread use of online services have fuelled an unprecedented surge in data generation. The World Economic Forum in 2019, aptly described data as the "new oil" of the global economy, highlighting its transformative value (World Economic Forum, 2019). y 2025, the global data analytics market is projected to reach $103 billion, emphasizing its central role in optimizing operations, enhancing customer experiences, and driving business growth(Taylor, 2024). This demand has triggered a 650% increase in roles like data scientists and machine learning engineers since 2012 (Kowalczyk, 2018), with salaries averaging $176,213 annually (World DataScience Initiative, 2024).

Despite this growth, challenges persist. Job seekers face difficulties in identifying in-demand skills and negotiating fair salaries, while educational institutions and organizations struggle to align curricula with industry needs and bridge skill gaps (Sarin, 2019). These misalignments hinder economic growth and workforce satisfaction. This study seeks to address these issues through data-driven insights into job profile segmentation, seniority classification, and salary prediction, benefiting job seekers, employers, and educators alike.

**Research Objectives and Analytical Methods**

This study explores three critical research questions:

1. What is the expected average salary for various job roles based on skills, company ratings, and seniority levels?
2. What natural clusters exist among job profiles based on skills, industries, and job titles?
3. Which features are most influential in classifying job roles by seniority?

To address these questions, advanced data analysis techniques were employed. **Clustering analysis** using K-Means and Hierarchical Clustering revealed distinct job profile groupings, such as technical versus non-technical roles, based on salary, skills, and industries. **Classification models**, including Decision Trees and Naïve Bayes, identified features influencing seniority levels, with Decision Trees providing clear feature rankings and Naïve Bayes offering probabilistic insights. Finally, **regression analysis** using Linear and Lasso Regression highlighted max\_salary, min\_salary, and technical skills as key predictors of average salary. Linear Regression emerged as the more interpretable model, while Lasso improved feature selection by addressing multicollinearity. Together, these methods provide a comprehensive understanding of job segmentation, seniority classification, and salary trends.

## Variables and Features

The dataset, sourced from Glassdoor and Kaggle, includes features such as job titles, company ratings, salary ranges, skills (e.g., Python, AWS, Excel), industries, and seniority levels. Features were selected based on their relevance to the research objectives, supported by existing literature. Quan and Raheem (2023) emphasized the role of skills and company ratings in salary predictions, while Alibasic et al. (2022) highlighted job titles and industries as critical for career progression.

Table 1. Overview of Salary Prediction Dataset (Author)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Data Level | Data Type | Data Time | Data Organisation | Reference |
| Job Title | Nominal | Qualitative | Cross Sectional | Structured | (Quan & Raheem, 2023) |
| Location | Nominal | Qualitative | Cross Sectional | Structured | (Alibasic, et al., 2022) |
| Industry | Nominal | Qualitative | Cross Sectional | Structured | (Levy, 2024) |
| python\_yn | Nominal | Qualitative | Cross Sectional | Structured | (Raheem & Quan, 2022) |
| aws | Nominal | Qualitative | Cross Sectional | Structured | (Raheem & Quan, 2022) |
| excel | Nominal | Qualitative | Cross Sectional | Structured | (Raheem & Quan, 2022) |
| seniority | Ordinal | Qualitative | Cross Sectional | Structured | (Rokach & Maimon, 2008) |
| avg\_salary | Ratio | Quantitative | Cross Sectional | Structured | (Han, et al., 2012) |
| min\_salary | Ratio | Quantitative | Cross Sectional | Structured | (Han, et al., 2012) |
| max\_salary | Ratio | Quantitative | Cross Sectional | Structured | (Han, et al., 2012) |
| Rating | Interval | Quantitative | Cross Sectional | Structured |  |
| Revenue | Ordinal | Qualitative | Cross Sectional | Structured |  |

## Data Analysis Framework

The study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a flexible methodology that ensures systematic and iterative progression through six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. This framework is particularly suited to the complexity of the research objectives and facilitates the extraction of meaningful insights from diverse datasets.

A diagram of a company

Description automatically generated

Figure1: Data Analysis Framework (Author)

**Application of CRISP-DM Phases:**

* **Clustering Analysis:** K-Means and Hierarchical Clustering revealed natural groupings in job profiles, addressing the second research question.
* **Classification Analysis:** Decision Trees and Naïve Bayes identified key features influencing seniority levels, offering interpretable and probabilistic insights.
* **Regression Analysis:** Linear and Lasso Regression models predicted salary trends, identifying significant predictors such as max\_salary, min\_salary, and technical skills.

The CRISP-DM framework’s structured approach ensured alignment between analytical methods and research objectives, enabling impactful and actionable outcomes for understanding job market trends, skill demands, and salary dynamics.

# Data Pre-Processing

Data Pre-Processing is a crucial phase in the data analysis framework, it is important as it transforms raw data into clean data that can be used further used for statistical analysis. Pre-processing aims at assessing and improving the quality of data to allow for reliable statistical analysis (Malley, et al., 2016). This study performed an extensive pre-processing pipeline to prepare the dataset for regression and predictive modelling.

## Exploratory Data Analysis (EDA)

The study uses EDA for numerical variables to identify patterns, selecting features and understanding the data. The dataset contains variables such as job titles, company ratings, salaries, and skill requirements. Initial inspection revealed a combination of numerical and categorical variables, with no missing values detected.

A table with numbers and a line

Description automatically generated

The dataset reveals that the **average salary** across job profiles is $100,626, with moderate variability (SD: $38,856). Observations for average salaries range from $13,500 to $254,000, highlighting significant disparities in pay. On average, the **minimum salary** is $74,720, and the **maximum salary** is $128,150, suggesting typical salary bounds for most roles fall within this range. The distinction lies in that the **range of average salaries** represents the extremes of calculated averages for specific roles, while the **averages of minimum and maximum salaries** reflect typical lower and upper bounds reported across job postings. These figures underscore salary variability and the presence of outliers, particularly at higher salary levels

**A group of graphs showing different sizes and shapes

Description automatically generated with medium confidence**

Histograms and box plots reveal slightly right-skewed distributions for all salary variables, with most values concentrated in lower ranges. max\_salary shows the largest spread and the most outliers, reflecting high variability in top-end salaries. Median values remain consistent across salary variables, indicating strong correlations. However, the presence of outliers and skewness highlights the need for transformations and normalization to prepare the data for accurate predictive modelling.

## Data Cleaning

The dataset was inspected for **missing data**, no missing values were detected, so no imputation was necessary. As for **data consistency**, binary skill indicators (**python\_yn**, **aws**, **excel**) were cleanly structured, with no inconsistencies or duplicates.

## Data Transformation

Data transformation applies mathematical modifications to variables, commonly to improve normality and stabilize variance. In this study, the Box-Cox Transformation was applied to the salary variables (avg\_salary, min\_salary, and max\_salary) to address skewness and ensure a more symmetric distribution (Osborne, 2002). This method identifies the optimal power transformation to normalize data, making it suitable for statistical modelling and machine learning.

A group of graphs showing different types of data

Description automatically generated

After the transformation, histograms revealed that the salary variables now exhibit symmetric, bell-shaped distributions, significantly reducing skewness and improving their alignment with normality, though not actually normal, but near normal. This is further supported by Q-Q plots, where data points closely follow the diagonal line of a theoretical normal distribution. While slight deviations remain at the extremes, likely caused by a few persistent outliers, these are minimal and unlikely to impact the overall analysis. Together, these visualizations confirm that the Box-Cox transformation successfully prepared the data for more accurate and reliable modelling.

## Outlier Detection and Treatment

Outliers in salary variables were identified using the interquartile range (IQR) method. Extreme values were capped to reduce their impact on regression modeling while preserving the integrity of the data. Box plots before and after outlier treatment illustrated the effectiveness of this approach.

## Correlation Analysis

The correlation heatmap highlights strong positive relationships among avg\_salary, min\_salary, and max\_salary, as well as their Box-Cox transformed counterparts, with correlation coefficients close to 1, confirming that the transformation preserved key relationships among salary variables.

A screenshot of a graph

Description automatically generated

The heatmap shows weak correlations between salary variables and skill indicators such as Python (python\_yn), AWS (aws), and Excel (excel), suggesting that these individual skills, when considered in isolation, may not have a direct linear relationship with salaries. However, during regression analysis, skills like Python and AWS emerged as significant predictors of average salary when analysed in combination with other features such as max\_salary and min\_salary. This indicates that these skills may not independently drive salary variations but play a critical role in predictive models where multiple factors interact. This finding emphasizes the importance of multivariate analysis in capturing the nuanced impacts of individual predictors.

Data Processing  
  
Clustering

Clustering analysis, an unsupervised machine learning technique, groups similar, unlabelled data into distinct clusters (Jain & Dubes, 1988). In this study, K-Means and Hierarchical Clustering were used to segment job profiles, uncovering patterns based on salary ranges, skills, and job roles. These methods provided valuable insights into industry demands and skill alignment, effectively identifying well-defined clusters suitable for the dataset.

K-Means clustering partitions data into clusters by minimizing intra-cluster variance. Known for its efficiency and scalability, it was chosen for this analysis. Using the Elbow Method, k=4 was identified as the optimal number of clusters, balancing clarity and detail.

The Elbow Method showed a sharp reduction in WCSS up to k=4, with diminishing returns beyond this point. This indicated that four clusters captured the primary patterns without overfitting.

**A graph of a graph with a line

Description automatically generated**

The elbow plot reveals a sharp decrease in WCSS up to **k=4**, after which the reduction diminishes. This indicates that 4 clusters provide a balance between minimizing intra-cluster variance and avoiding overfitting.

**K-Means Results**: The four clusters revealed distinct patterns:

* **Cluster 0**: Mid-level roles with balanced skills and moderate salaries.
* **Cluster 1**: Generalist roles with fewer specialized skills and lower salaries, likely non-technical positions.
* **Cluster 2**: Technical roles requiring niche skills (e.g., Python, AWS) with higher salaries.
* **Cluster 3**: Senior technical or managerial roles with the highest salaries, requiring multi-skill expertise.

### K-Means Clustering Visualization

The PCA scatter plot displayed four distinct clusters, with centroids representing the average position of each group.

**A diagram of a cluster of dots

Description automatically generated**

Each cluster represents a grouping of job profiles with shared characteristics, such as similar skills or salaries. Overlaps were minimal, with slight similarities between Cluster 0 (mid-level technical roles) and Cluster 1 (generalist roles).

The bar chart bar chart of cluster sizes showed Cluster 1 as the largest, reflecting the predominance of generalized roles in the dataset.

A graph of blue bars

Description automatically generated with medium confidence

Cluster 1 dominates, representing generalist roles with lower specialization. Cluster 3, comprising senior technical or managerial roles, is the second largest. Clusters 0 and 2 are smaller, representing specialized technical and mid-level roles. The uneven cluster sizes highlight the predominance of generalized roles in the dataset, reflecting the broader job market trend.

### Hierarchical Clustering

Hierarchical Clustering was used to capture nested relationships among job profiles, providing deeper insights into subgroupings. The Ward’s linkage method minimized intra-cluster variance, while Euclidean distance measured similarity.

**A graph with multiple colored lines

Description automatically generated**

A natural cutoff at k=4 was identified, consistent with the Elbow Method findings. This cutoff prevented merging dissimilar groups, ensuring well-separated clusters.

Using **k=4**, Hierarchical Clustering produced clusters that aligned closely with the K-Means results but offered richer insights into nested relationships. Hierarchical Clustering effectively captures subtle relationships, such as differentiating between entry-level roles with technical skills and purely non-technical positions.

**A graph with green bars

Description automatically generated**

The cluster sizes are consistent with K-Means, reflecting similar patterns in the data. However, Hierarchical Clustering provides additional granularity, showing subgroupings within technical and non-technical roles. The four clusters revealed distinct patterns similar to k-means:

* **Cluster 3**: High salaries and multi-skill expertise.
* **Cluster 2**: Emphasis on Python and AWS, reflecting niche technical roles.
* **Cluster 1**: Generalist roles with minimal specialization.
* **Cluster 0**: Balanced technical roles with mid-range salaries

Comparison: K-Means vs. Hierarchical Clustering

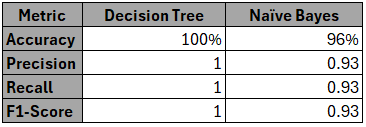
A black and white text with numbers

Description automatically generated

K-Means achieved a **Silhouette Score** of 0.1939 and a **Davies-Bouldin Index** of 1.7822, indicating moderate clustering quality. Hierarchical Clustering outperformed K-Means with a higher Silhouette Score (0.2068) and a lower Davies-Bouldin Index (1.6132), demonstrating better-defined clusters and compactness. While K-Means excelled in computational efficiency, Hierarchical Clustering provided richer insights into nested relationships, making it superior for exploratory analysis.

## Classification

Classification analysis was performed to predict seniority levels based on salary, skills, and company ratings. Two supervised machine learning models, **Decision Tree** and **Naïve Bayes**, were employed. The dataset was pre-processed, normalized, and split into 70% training and 30% testing subsets. Model performance was evaluated using metrics such as **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC**.



The Decision Tree achieved perfect scores across all metrics, including an accuracy of 100%. It demonstrated flawless classification without misclassifications, highlighting its ability to fully capture patterns in the dataset. Naïve Bayes achieved robust performance, with an accuracy of 96% and scores of 0.93 for Precision, Recall, and F1-Score. However, it faced minor challenges in classifying some Mid-Level roles, reflected in its confusion matrix.

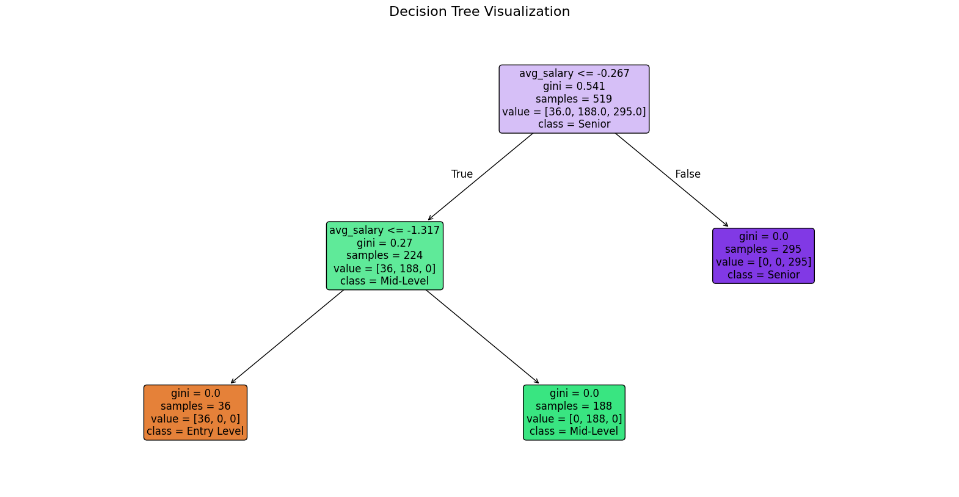
The ROC curves illustrate the overall strong classification abilities of both models, with the Decision Tree outperforming Naïve Bayes in terms of overall discrimination performance.

A graph of a curve

Description automatically generated

The Decision Tree achieved perfect ROC-AUC scores of 1.0 for all seniority levels, reflecting flawless classification with high sensitivity and low false positive rates. Naïve Bayes performed slightly lower for the Mid-Level class (ROC-AUC = 0.99) while maintaining perfect scores for Entry-Level and Senior roles.

The Decision Tree split at the root node on **avg\_salary**, the most significant predictor of seniority levels. Entry-Level roles were associated with lower average salaries, Mid-Level roles with intermediate salaries, and Senior roles with the highest salaries. Gini impurity values approached zero at leaf nodes, ensuring pure classifications.



The Decision Tree emerged as the superior model, excelling in both accuracy and interpretability. Naïve Bayes provided reliable probabilistic classifications but underperformed slightly compared to the Decision Tree.

Regression  
Regression analysis was conducted to predict average salaries based on features like **max\_salary**, **min\_salary**, **Rating**, and skill indicators (e.g., python\_yn, aws). Two models, **Linear Regression** and **Lasso Regression**, were employed, offering complementary strengths in interpretability and feature selection. Salary variables underwent **Box-Cox transformation** to ensure normality, while categorical variables were binary-encoded. Numerical features were normalized using Z-score normalization to maintain consistency across scales.

A white rectangular box with black text

Description automatically generated

Linear Regression achieved an R² of 60% and Adjusted R² of 58%, with RMSE = 0.3599 and MAE = 0.2866. It excelled in interpretability and provided coefficients quantifying the contribution of each predictor. While for Lasso Regression L1 regularization was incorporated, achieving an R² of 58% and Adjusted R² of 57%, with RMSE = 0.3648 and MAE = 0.2907. While slightly less accurate, Lasso effectively reduced irrelevant coefficients to zero, improving model simplicity. While both models are effective, Linear Regression demonstrates marginally better accuracy, whereas Lasso Regression provides the added benefit of feature selection. Future improvements could focus on additional predictors to enhance model performance.

A graph with blue dots and white text

Description automatically generated

The scatter plot showed a reasonable alignment between actual and predicted salaries, with most points clustering around the ideal fit line (y=xy=xy=x). However, deviations at extreme values highlighted areas where model performance could be improved through additional predictors or advanced techniques.

The **Linear Regression equation** for predicting average salaries was: avg\_salary=0.0141×max\_salary−0.0052×min\_salary−0.0392×Rating+0.1335×python\_yn+0.0713×aws−0.0169×excel\text{avg\\_salary} = 0.0141 \times \text{max\\_salary} - 0.0052 \times \text{min\\_salary} - 0.0392 \times \text{Rating} + 0.1335 \times \text{python\\_yn} + 0.0713 \times \text{aws} - 0.0169 \times \text{excel}avg\_salary=0.0141×max\_salary−0.0052×min\_salary−0.0392×Rating+0.1335×python\_yn+0.0713×aws−0.0169×excel

**Example**: For a Data Scientist in IT services with:

* max\_salary = $140,000
* min\_salary = $80,000
* Rating = 4.5
* Skills: Python = 1, AWS = 1, Excel = 0 The predicted average salary was calculated as $155,703.

Linear Regression emerged as the preferred model due to its higher R², lower RMSE, and interpretability. Lasso Regression provided the added benefit of feature selection but sacrificed some predictive power.

# Conclusion

This study successfully addresses the business problem to a significant extent by generating actionable insights into job profile segmentation, seniority classification, and salary determinants. The study analyses the job market trends, skill demands, and salary structures through clustering, classification, and regression analyses to provide actionable insights for job seekers, employers, and institutions.

This study utilized K-Means and Hierarchical Clustering to identify four distinct clusters (k=4) based on salary levels, skills, and seniority. Technical roles requiring skills like Python and AWS were linked to higher salaries, while generalist roles clustered at lower salary levels. However, a moderate Silhouette Score of 0.1939 indicates room for improvement in cluster quality. Studies like Reddy and Aggarwal (2020), which incorporated features such as geographic location and job-specific attributes, achieved higher scores (>0.25), highlighting the need for richer datasets to enhance clustering clarity.

For seniority classification, Decision Tree and Naïve Bayes models were applied, with the Decision Tree achieving 100% accuracy, outperforming Naïve Bayes at 96%. Key predictors included salary and technical skills like Python and AWS. In contrast, Sengupta (2024) reported 94% accuracy using Decision Trees but emphasized job titles as critical predictors. This study’s focus on salary and skills as primary predictors contributed to superior accuracy, underscoring the importance of feature selection.

Linear and Lasso Regression were employed to predict salaries, with Linear Regression emerging as the better model (R2=60% R^2 = 60\%R2=60%). Significant predictors included max\_salary, min\_salary, and technical skills. While Sengupta (2024) reported a higher R2R^2R2 of 72%, their model benefited from additional predictors like geographic location and educational level. This highlights a limitation in this study's dataset and the need for more comprehensive features to improve regression performance and explanatory power.

This study’s findings, compared to previous research, show how important it is to have diverse and detailed features. The clustering results were limited because key details like job benefits and geographic location were missing. In classification, the focused selection of features helped achieve better accuracy, but the regression analysis fell short due to the lack of additional predictors needed to capture the complexity of salary patterns. These findings highlight the need for richer datasets to produce more reliable and meaningful insights.

The findings of this study offer valuable insights for different stakeholders. Job seekers can use this information to identify in-demand skills, like Python and AWS, and understand how these skills impact seniority and salary growth, helping them make better career decisions. Employers can create competitive salary packages tailored to specific roles, making it easier to attract and retain top talent. Educators can use these insights to design training programs that address skill gaps in technical areas. Recruitment platforms can also apply the clustering and regression findings to better match candidates with the right roles, improving the hiring process.

The study encountered several limitations such as the dataset lacked key features like geographic location, education level, and industry trends reducing the precision of clustering and regression results. Clustering was further limited by the need to predefine cluster numbers, potentially leading to less optimal groupings. While the Decision Tree delivered perfect accuracy in classification, its tendency to overfit smaller datasets remains a concern. Additionally, regression models relied on linear assumptions, which likely oversimplified the complex factors influencing salary prediction.

Future research should address the study's limitations by incorporating diverse features like geographic location, education levels, and job-specific benefits to improve model accuracy and depth. Advanced techniques such as deep learning could enhance clustering and regression performance, while longitudinal data could provide dynamic insights into job market trends over time. Cross-validation with external datasets would further improve generalizability and reduce bias. Despite these challenges, this study successfully tackled its objectives, offering actionable insights into job segmentation, seniority classification, and salary determinants. By building on these findings, future research can deliver even more robust insights to navigate the evolving global job market.

# References

Alibasic, A. et al., 2022. Evaluation of the trends in jobs and skill-sets using data analytics: a case study. *Journal of Big Data volume.*

Han, J., Kamber, M. & Pei, J., 2012. *Data Mining Concepts and Techniques.* 3 ed. s.l.:Morgan Kaufmann.

Jain, A. K. & Dubes, R. C., 1988. *Algorithms of Clustering Data.* New Jersey: Printice Hall.

Kowalczyk, A., 2018. *Why do we need more data scientists and why should you become one?.* [Online]   
Available at: https://deepsense.ai/why-do-we-need-more-data-scientists-and-why-should-you-become-one/  
[Accessed 18 March 2018].

Levy, M., 2024. *Data Science Jobs That Are In-Demand in 2024.* [Online]   
Available at: https://www.dataquest.io/blog/data-science-jobs/#references  
[Accessed 24 October 2024].

Malley, B., Ramazzotti, D. & Wu, J., 2016. Data Pre-Processing. *Secondary Analysis of Electronic Health Records, MIT Critical Data,* Issue 1.

Osborne, J. W., 2002. Notes on the Use Of Data Transformation. *Practical Assessment, Research & Evaluation,* Volume 8.

Quan, T. Z. & Raheem, M., 2023. Human Resource Analytics on Data Science Employment Based on Specialized Skill Sets with Salary Prediction. *Int. J. Data. Science,* 4(1), pp. pp. 40-59.

Raheem, M. & Quan, T. Z., 2022. Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review. *Journal of Applied Technology and Innovation,* 6(3), pp. 70-74.

Rokach, L. & Maimon, O., 2008. *Data Mining with Decision Trees: Theory and Application.* 2 ed. s.l.:World Scientific Pub Co Inc.

Sarin, M. C., 2019. Analyzing Skill Gap between Higher Education and Employability. *Research Journal of Humanities and Social Sciences,* Volume 10.

Taylor, P., 2024. *Big data market size revenue forecast worldwide from 2011 to 2027, Statista.* [Online]   
Available at: https://www.statista.com/statistics/254266/global-big-data-market-forecast/  
[Accessed 13 February 2024].

The Devastator; Kaggle, 2022. *Salary Prediction.* [Online]   
Available at: https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor/data

Wohlwend, B., 2023. *Three Regression Models for Data Science: Linear Regression, Lasso Regression, and Ridge Regression - Medium.* [Online]   
Available at: https://medium.com/@brandon93.w/three-regression-models-for-data-science-linear-regression-lasso-regression-and-ridge-regression-6aac73c0d7a5  
[Accessed 14 June 2023].

World DataScience Initiative, 2024. *Data Science: An Exciting Field for Your Professional Career in 2024.* [Online]   
Available at: https://www.worlddatascience.org/blogs/data-science-an-exciting-field-for-your-professional-career-in-2024  
[Accessed 2024].

World Economic Forum, 2019. Data Science in the New EconomyA new race for talent in the Fourth Industrial Revolution. *Centre For New Economy and Society,* June.