This project, "Decoding the Job Market for Data Professionals," has been a deep dive into understanding the dynamics of the data professional landscape. I undertook this study because I recognized the critical need to address challenges faced by job seekers in identifying in-demand skills and negotiating fair salaries, as well as the struggles educational institutions and organizations encounter in aligning curricula with industry needs and bridging skill gaps. The rapid digitization of businesses and the surge in data generation have positioned data as the "new oil" of the global economy , with the global data analytics market projected to reach $103 billion by 2025. This immense growth has led to a 650% increase in roles like data scientists and machine learning engineers since 2012, with average salaries around $176,213 annually. Understanding these trends is crucial for fostering economic growth and workforce satisfaction.

**My Research Objectives:**

My primary goal was to answer three critical 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?

**How I Conducted the Study:**

I followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a systematic and iterative methodology that guided me through six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

**Data Sources and Features:** I sourced my dataset from Glassdoor and Kaggle, which included features like job titles, company ratings, salary ranges, specific skills (e.g., Python, AWS, Excel), industries, and seniority levels.

**Data Pre-processing:**

* **Exploratory Data Analysis (EDA):** My initial inspection showed no missing values, and I found a mix of numerical and categorical variables. The average salary across job profiles was $100,626, with significant variability, ranging from $13,500 to $254,000. Histograms and box plots revealed slightly right-skewed distributions for salary variables, with

max\_salary showing the largest spread and most outliers.

* **Data Cleaning:** I didn't find any missing values, so no imputation was necessary. Binary skill indicators were also cleanly structured and consistent.
* **Data Transformation:** To address skewness and improve the symmetry of the salary distributions, I applied the Box-Cox Transformation to avg\_salary, min\_salary, and max\_salary.
* **Outlier Detection and Treatment:** I identified outliers in the salary variables using the interquartile range (IQR) method and capped extreme values to reduce their impact on my regression models.
* **Correlation Analysis:** My correlation heatmap showed strong positive relationships among the salary variables. Interestingly, individual skill indicators like Python, AWS, and Excel had weak linear correlations with salaries when viewed in isolation. However, I later found them to be significant predictors in combination with other features during regression analysis.

**My Analytical Methods/Models:**

* **Clustering Analysis:** I used both K-Means and Hierarchical Clustering to segment job profiles. The Elbow Method helped me determine that

*k*=4 was the optimal number of clusters for K-Means. Hierarchical Clustering also consistently identified 4 clusters.

* **Classification Analysis:** I employed Decision Tree and Naïve Bayes models to predict seniority levels.
* **Regression Analysis:** For predicting average salaries, I utilized Linear Regression and Lasso Regression. Salary variables underwent Box-Cox transformation, and numerical features were normalized using Z-score normalization.

**My Findings:**

**Clustering Analysis:** I identified four distinct job profile groups:

* **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. This was the largest cluster in my dataset.
* **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. This was the second largest cluster.

Hierarchical Clustering, with a Silhouette Score of 0.2068 and a Davies-Bouldin Index of 1.6132, slightly outperformed K-Means, which had a Silhouette Score of 0.1939 and a Davies-Bouldin Index of 1.7822. This indicated that Hierarchical Clustering produced better-defined and more compact clusters.

**Classification Analysis:**

* My Decision Tree model achieved perfect scores across all metrics, including an

**accuracy of 100%**. It showed flawless classification.

* Naïve Bayes also performed robustly, with an

**accuracy of 96%** and scores of 0.93 for Precision, Recall, and F1-Score.

* The Decision Tree revealed that

avg\_salary was the most significant predictor for seniority levels.

**Regression Analysis:**

* Linear Regression achieved an

R2 of **60%**, explaining 60% of the variation in average salary. Its Adjusted

R2 was 58%.

* Lasso Regression achieved an

R2 of 58% and an Adjusted R2 of 57%.

* Key predictors of average salary included

max\_salary, min\_salary, and technical skills like Python and AWS.

* The Linear Regression equation I derived was:

avg\_salary=0.0141×max\_salary−0.0052×min\_salary−0.0392×Rating+0.1335×python\_yn+0.0713×aws−0.0169×excel.

**My Insights and Recommendations:**

**Insights:**

* My analysis highlights patterns consistent with prior literature, such as salary increases with seniority. However, my study also emphasizes the crucial role of technical skills in modern job profiles, diverging from some previous findings.
* Roles requiring skills like Python and AWS are consistently linked to higher salaries and more senior positions.
* My findings underscore that average salary is a primary determinant when classifying seniority levels.
* It's clear that while individual skills might not show strong independent linear correlations with salaries, their interaction within multivariate models is critical for accurate predictions.

**Recommendations:**

* **For Job Seekers:** I highly recommend focusing on acquiring and refining in-demand technical skills, especially Python and AWS, as these are significant drivers of salary and career progression. Understanding how these skills influence seniority will aid in better career planning.
* **For Employers:** I advise creating competitive salary packages that are specifically tailored to roles requiring particular technical skills, as this will help attract and retain top talent.
* **For Educators:** Based on my findings, I urge educational institutions to design and update training programs that directly address identified skill gaps in technical areas, ensuring that curricula align with industry demands.
* **For Recruitment Platforms:** The clustering and regression insights from my study can be applied to improve how candidates are matched with suitable roles, streamlining the hiring process.
* **For Future Research:** My study encountered limitations, primarily due to the dataset lacking key features like geographic location, education level, and industry trends, which reduced the precision of clustering and regression results. Therefore, I strongly recommend incorporating a more diverse and detailed set of features in future studies. I also believe that exploring advanced techniques such as deep learning could significantly enhance clustering and regression performance. Finally, utilizing longitudinal data would provide more dynamic insights into evolving job market trends over time, and cross-validation with external datasets would improve generalizability and reduce bias