RMIT UNIVERSITY

School of Science, Engineering & Technology
Bachelor of Information Technology

Factors affecting Data Industry Salaries – Insights into the "Sexiest Field"

Subject Code:	COSC2968 — COSC3053
Subject Name:	Foundations of AI for STEM
Location & Campus:	Saigon South (SGS)
Student Name:	Kai Nguyen
Student Number:	s4126139
Lecturer/Tutor:	Dr. Nhat-Quang Tran
Assignment Due Date:	31 August 2025
Date of Submission:	30 August 2025
Word Count:	3059
Pages (incl. cover)	16

Declaration of Authorship

I declare that this assignment is my own work and that no part has been copied from any other source without appropriate acknowledgement.

August 30, 2025

Contents

1 Statement of the Problem

Objectives

Over a decade ago, Davenport and Patil (2012) emphasized the data science professional as "the sexiest job of the 21st century", highlighting its important role in the data-driven economy. Indeed, the demand for job positions related to data has increased due to the need of the labor market, along with the high salary level. This study aims to clarify and analyze the crucial factors that affect salaries (USD) by using the FOAI-assignment.csv dataset. Through data science methodology: data exploration, feature analysis, data preparation, and predictive modeling, the report attempts to provide a realistic understanding for estimating salary and optimizing recruitment's decision-making.

Research Questions

This study tries to shed light on 6 questions:

1. How can I exploit this dataset?
2. What are the factors affecting the salaries in the data field?
3. How can I explain the result of predictions?
4. Based on the results, what recommendations can be provided to the CEO in tackling the business's issue?

2 Data Exploration

2.1 Dataset Overview

The data originates from the 'In The Know' company, but no information about the author and creation date could be discovered in the file FoAI-assignment.csv. The dataset contains 11 columns and 1500 rows, recording the time from 2020 to 2023. Additionally, there are 2 types of data: variables and categories in the raw data set. Going through data quality briefly, it is clear that several records are missing in the salary and *salary_in_usd* columns, not only that hundreds of rows are duplicated.

2.2 Data Quality Checks

- **Missing Values**: The raw dataset contains 12 blank cells (6 cells in each salary and salary_in_usd column), which is equivalent to 0.4% of the full dataset. Because of the paucity of these missing values, I decided to drop their rows to continue analyzing the dataset.
- **Duplicate Rows**: After removing 6 rows, another problem arises in the dataset. There are 143 rows which has up to 241 copies. The majority is derived from 2022 and 2023, which concentrated in the US, medium-sized companies, and the senior level. I made a decision to remove these duplicates since they could have an adverse effect on the next analysis. Finally, the cleaned dataset has 11 columns and 1253 rows.

2.3 Univariate Analysis

In this part, I will analyze each column in the cleaned dataset to have a general perspective.

2.3.1 Target: salary_in_usd

The target variable of this study is *salary_in_usd*, representing the annual salary of employees in the dataset after dropping missing values. As shown in Figure 1, the mean salary of employees related to data fields is exactly 129, 141 USD, which is near the median 125, 000 USD, reflecting the balance of distribution. However, due to the existence of individuals with high salary levels from 300, 000 USD to a maximum 450, 000 USD, the distribution is right-skewed with a long tail. Additionally, the major part of income is normally distributed from 60, 000 USD to 200, 000 USD, indicating that the salary of jobs in the dataset is placed in middle-to-high income, while only a few with an outstanding salary could relate to experience, job title, currency, location, and company size, or could be considered as outliers.

Figure 1: Salary Distribution

2.3.2 Numeric variables

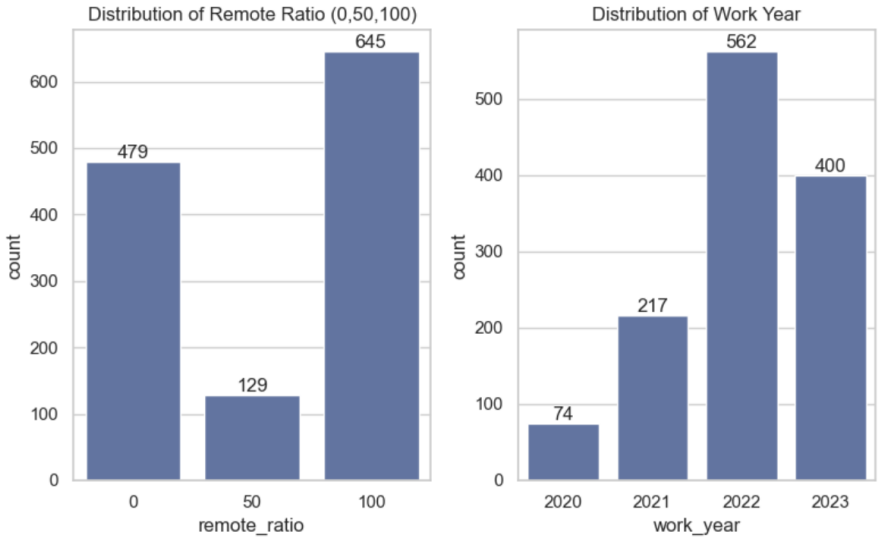

Figure 2: Remote Ratio and Work Year

- **Remote Ratio Distribution**
  The left bar chart of Figure 2 is divided into 3 three main groups, including: 0%(on-site), 50%(hybrid), and 100%(fully remote). It is clear that the number of fully remote

cases is highest, about 645 cases $\sim$ 51.48%, which describes the trend of working at home. However, hybrid mode accounts for only 129 cases, reflecting the difficulty and the minority in implementing this model.

- **Work Year**
  The right graph of Figure 2 describes the data collected over the years from 2020 to 2023. The gathered data significantly increased from 74(2020) to a peak of 562(2022), before falling to 400(2023). This could reflect the rapid growth in data related to remote work in the post-pandemic Covid period, followed by better stability in recent years.

2.3.3 Categorical variables

- **Job Title**

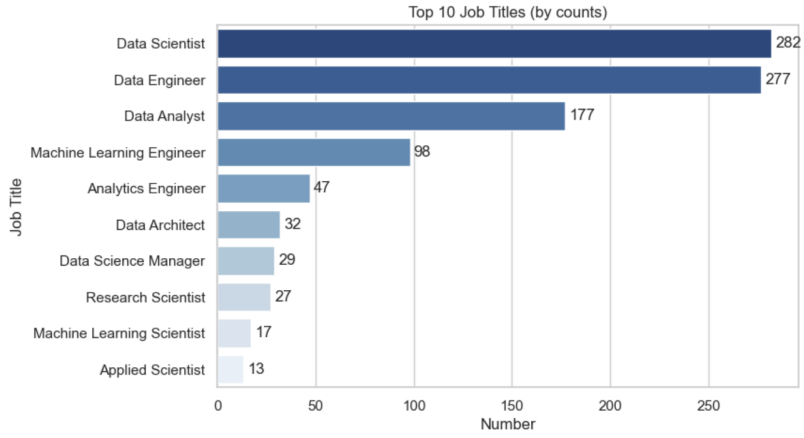

Figure 3: Salary Distribution

The dataset contains 69 different job titles. The majority are data scientists (nearly 22.5%), data engineers (22.1%), data analysts (14.13%), and machine learning engineers (7.82%) in the cleaned dataset (see Figure 3). However, there are 419 records for 65 job titles, which appear from 1 to 54 times, indicating the large distribution of the dataset with respect to job title. These rare records could be divided into 6 main fields, including Data Science & ML Research, Data Engineering/Infra/MLOps, Analytics& BI, Applied Data Science/Generalist, Management/Leadership, and others. This is also the challenge for data preparation in making a decision between dropping rare jobs and putting them into groups.

- **Location**
  With regard to geographic location, both companies and employees are predominantly based in the U.S., accounting for approximately 69.8% and 67.7% of the dataset, respectively (see Figure 4). These proportions are more than 10 times higher than those of the second-largest group (GB) in the raw dataset, indicating that the records were collected primarily from the U.S. Locations outside the U.S. can be grouped into four main regions: Europe, Asia, North America (excluding the U.S.), and Other (representing less common areas).

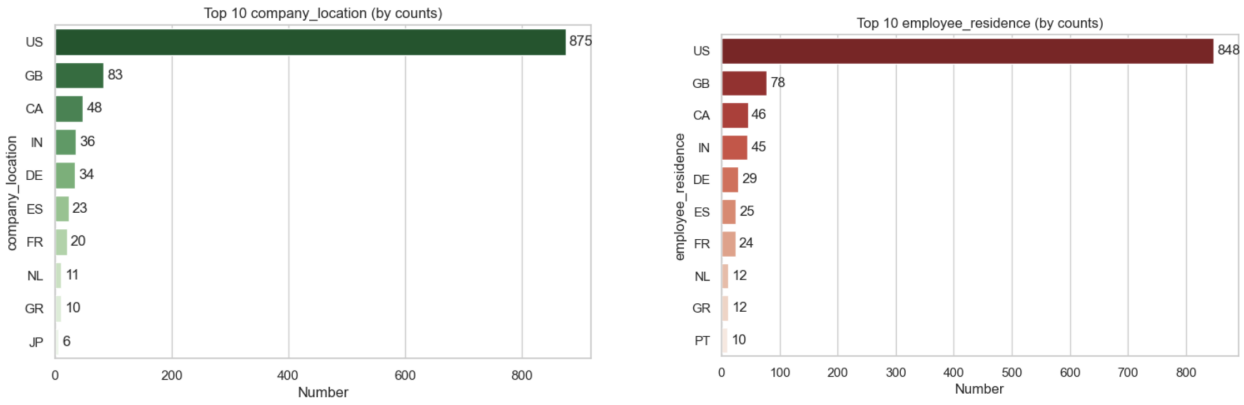

Figure 4: Location's Distribution of Companies and Employees

- **Experience Level**
  The experience level shows a heavy skew toward Senior level (713 records, ∼56.9%), followed by Mid-level (326), Entry-level (156), and Executive (58). This result may bias any subsequent analysis because the salary could depend on experience (Figure 5).
- **Company Size**
  Company size distribution extremely focuses on medium-sized organizations (859 observations, ∼ 68.6%), with small companies (106) and large companies (288). Similarly, these data could create an imbalance in the next analysis (Figure 5).
- **Employment Type**
  There is a contrast in the number of data between full-time (97.8%) and other types of employment. As a result, this feature can be ignored in the next steps of the process (Figure 5).

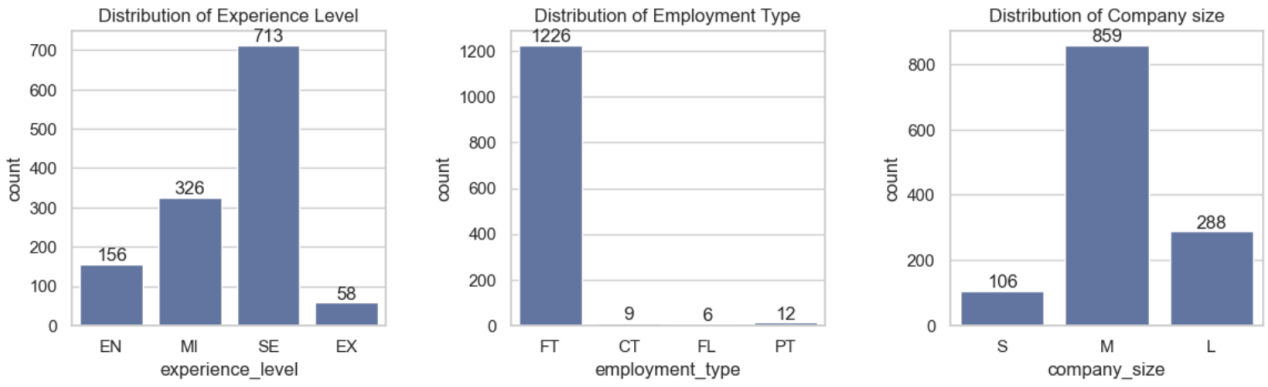

Figure 5: Salary Distribution

2.4 Correlation Analysis

In this part, I will attempt to analyze the relationship between salary_in_usd and other features, and also check currency conversion.

2.4.1 Salary_in_usd vs other features

- **Salary_in_usd vs Experience Level**

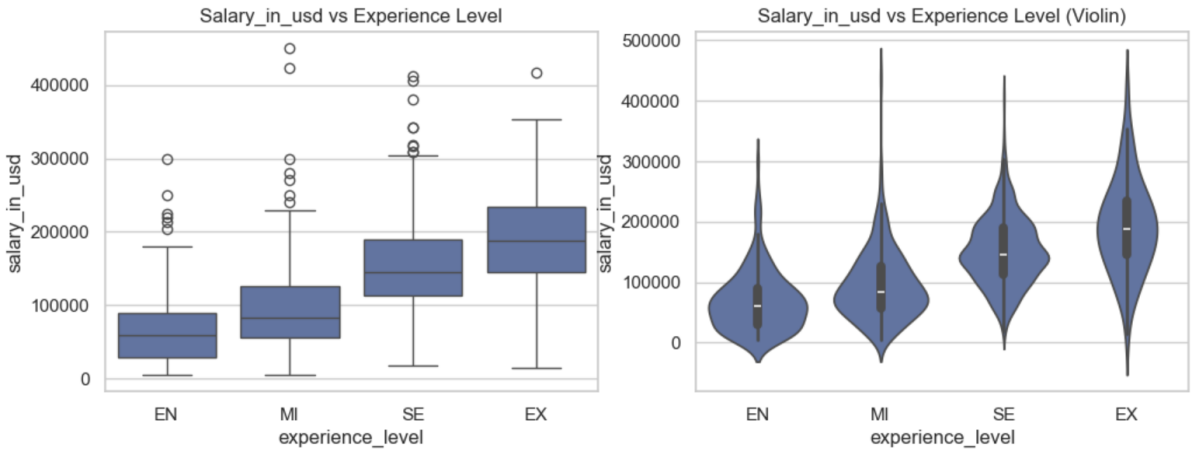

Figure 6: Salary vs Experience Level

Figure 6 "Salary vs Experience Level" uses the box plot and violin plot to describe the relationship between salary and experience. In the box plot, it is noticed that the increase of the median relies on the experience level, and outliers appear in each level, reflecting the differences in salary within the same group. The violin plot adds a different view of salary distribution, indicating that the entry level is more converged, while the executive level is largely distributed and skewed toward higher salary levels. In short, both plots show the trend of salary and highlight the distribution of salary in terms of experience level.

- **Salary_in_usd vs Job_Group**

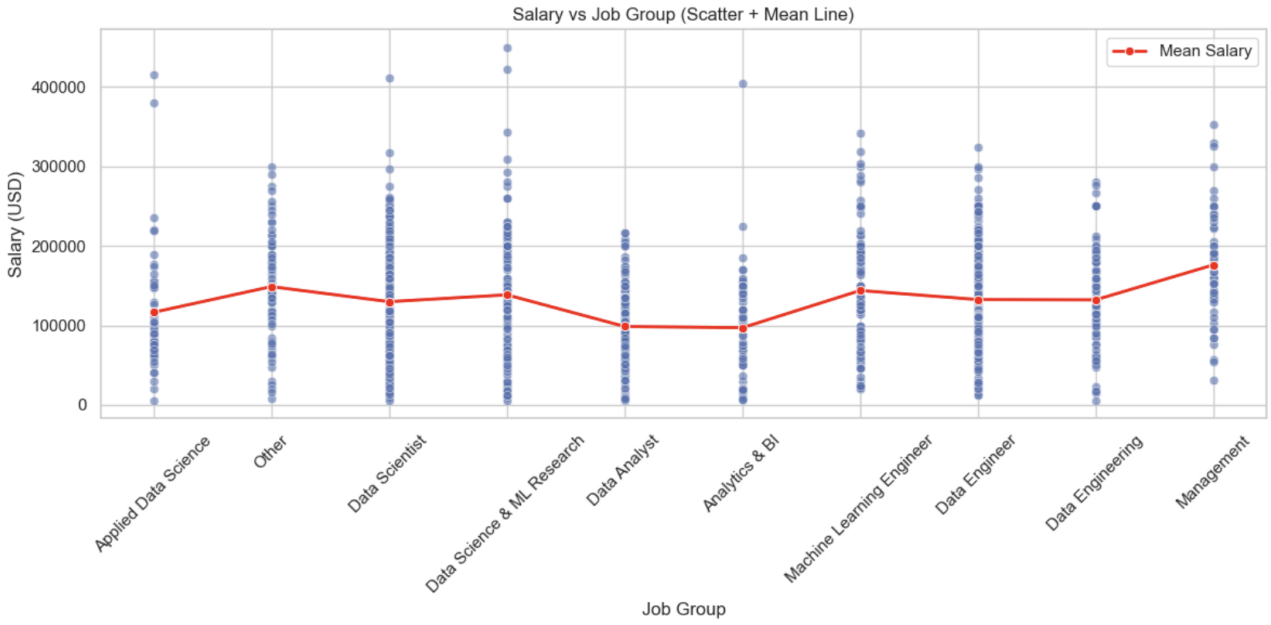

Figure 7: Salary vs Job

As I mentioned in the job title analysis, there are a variety of job titles. Therefore, I grouped them into 10 major categories. Fortunately, the salary distribution within each job group is not extremely divergent. From an overall perspective, the mean salary does not differ significantly between the groups and ranges from 100,000 USD to 175,000 USD.

It is evident that the management group has the highest mean due to the presence of managers and senior-level roles. Additionally, the Analytics field, including data analysts and BI professionals, has the lowest mean. Data engineers and data scientists, who account for the highest number of records in the dataset, have a mean salary in the middle range. In this view, it can be observed that there are numerous outliers in each job group because of the rare jobs in the job group.

- **Salary vs Region**

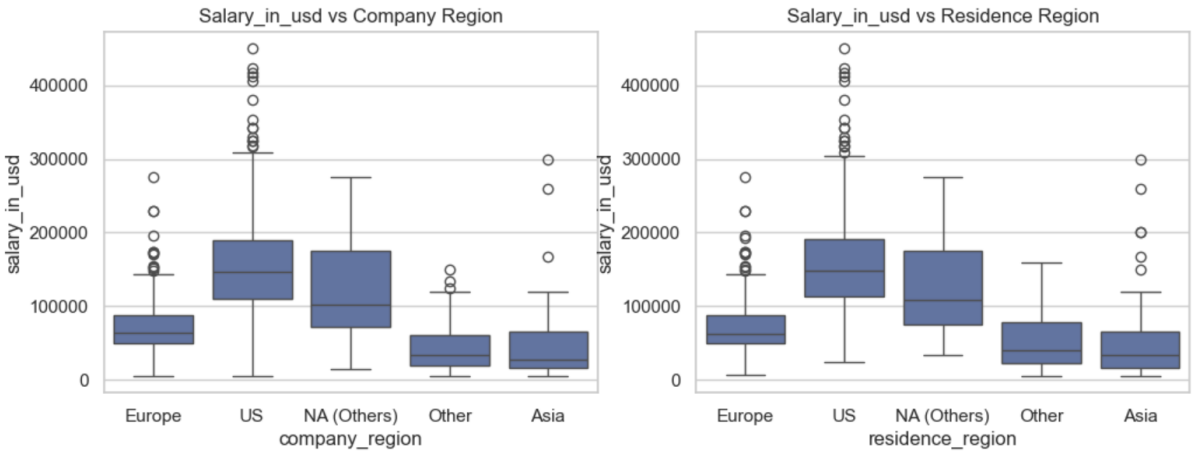

Figure 8: Salary vs Region

Figure 8 "Salary vs Region" illustrates the equivalence between the salary distribution of companies' & employees' region. It is clear that America has the highest salary level and a large distribution in both company and residence region. A reverse pattern could be observed in the data of others. Furthermore, the US has the highest number of outliers far from the normal salary, which could create an imbalance in the target distribution.

- **Salary vs Work Year**

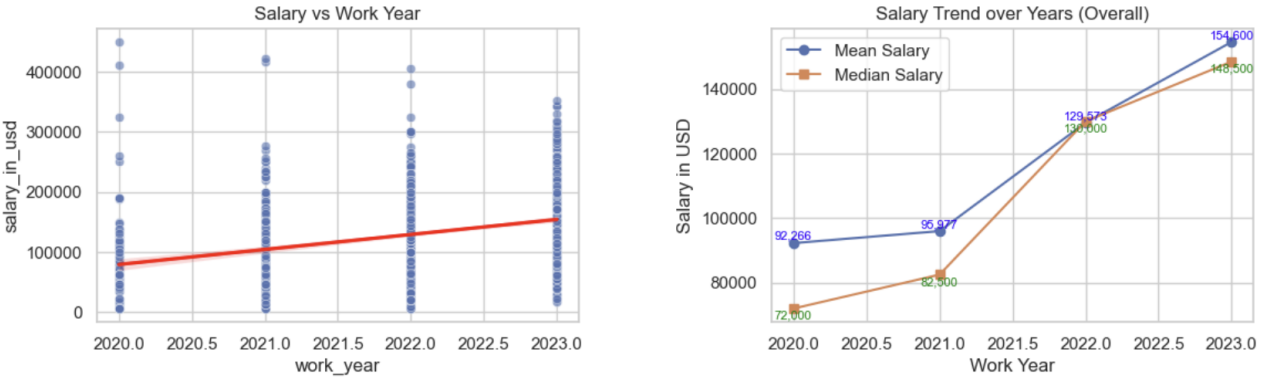

Figure 9: Salary vs Work Year

As shown in Figure 9, the average salary steadily increased throughout the period, rising from 92,266 USD in 2020 to 154,600 USD in 2023. A particularly significant jump occurred between 2021 and 2022, with median salaries increasing from 82,500 USD to 148,500 USD. This trend suggests a surge in demand for data-related jobs in the post-COVID-19 era.

- **Salary vs insignificant features (Hypothesis)**
  Regarding company size, it is notable that the mean of the salary distribution is highest at medium-sized companies instead of large ones. This is because the data from medium-scale companies outweighs the others. Similarly, the remote_ratio also shares the same distribution pattern with company size. Hence, I assume that these features are not important in the training model. Moreover, both features also contain some outliers, which we can flag for consideration during data preparation.

2.4.2 Currency & Currency Conversion Check

The dataset includes salary values in several currencies. I conducted an investigation on the conversion rates between USD and other currencies to ensure consistency and comparability. Specifically, I divided the salaries in other currencies by the salaries in USD and used statistical measures on these quotients, such as mean, median, and standard deviation, which could evaluate the conversion accuracy. These values were then compared against official ISO currency conversion rates for validation. As a result, all the records can be accepted.

2.5 Key Insights

After analyzing both the univariable and the correlation, there are four key sights:

- **Experience**: It is a fact that the salary increases with experience level.
- **Location**: Residents in North America earn significantly higher salaries compared to other regions.
- **Job Title**: It is trustworthy that the manager has a higher income, which aligns with expectations, while there is no significant difference in the salary level of others.
- **Work Year**: Salaries show a consistent upward trend over the years due to the demand and potential of employment related to data.

3 Data Preparation and Model Training

3.1 Feature Selection

As I analyzed in the data exploration section, the result of the target could be directly affected by four main factors in the training model. Therefore, five special features, including *experience_level*, *residence_region*, *company_region*, *job_group*, and *work_year*, should be chosen to train the model. For other features, such as *salary_currency*, *company_size*, and *remote_ratio*, I tried dropping each feature first to see the performance of the model, but the performance is worse than the one that I put all of them in to train the model. For example, if I dropped the *salar_currency* feature, the model would increase nearly 3000 USD in root mean square error(RMSE) compared to the model that contains all features. This can be explained that these features could support the model in understanding the unusual salary values. Finally, I dropped the salary column since it's a data leak.

3.2 Handling Data Issues

3.2.1 Sparse categorical features

As I mentioned and explained in Data Exploration, I dropped missing values and duplicates before analyzing the data because they adversely affect the training model. Although I deleted

247 rows $\sim$ 16, 46% information of the raw dataset, another issue emerged about the rare records in some features, such as *job_title, employee_residence,* and *company_location.* For example, Thailand, Malta, and other nations appear only once in both *employee_residence* and *company_location.* Therefore, we have to find a way to deal with these categories. The first method is dropping rows that contain the rare values, which occur less than 7 times $\sim 0, 5\%$ information of the dataset. Unfortunately, a host of rows could be deleted, which led to the loss of information and the low performance of the model because the size of the dataset was originally small. Another more effective method is to first analyze the relationships between categories in each feature and then merge them into a new category. This transformation supports simplifying the data, reducing noise, and making the features more useful in model training. However, I kept originality for *salary_currency* because we cannot create a new currency.

3.2.2 Handling outliers

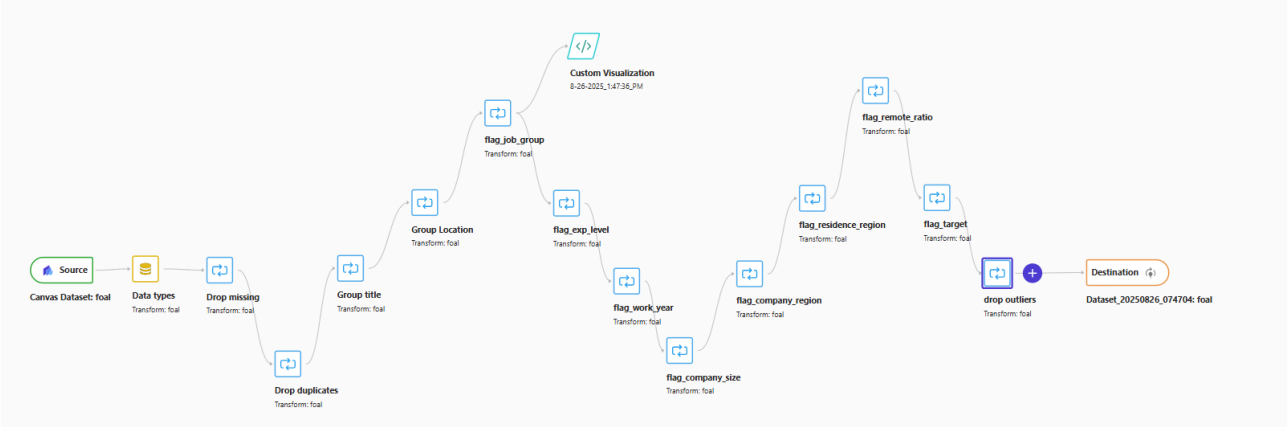

Figure 10: Data Preparation process

The Interquartile Range (IQR) method was applied to detect and handle outliers. This method identifies values that are far from the normal range of the conditional distribution of salary given particular features. After detecting these values, I could create new columns to flag outliers with 1 and normal values with 0 (see Figure 10). The next step is to analyze all rows (60 rows) containing outliers to decide whether they should be removed. Finally, I decided to drop all these rows because they were discrete and distributed quite equally in each feature. Not only that these rows equivalent to 4, 78% information in the dataset, rather than deleting hundreds of rows without combination.

3.3 Model Training and Performance

3.3.1 Train Dataset Information

The final dataset contains 1193 rows and 22 columns, including implied_rate to check currency conversion, grouped columns, and flag columns. Regarding to training dataset, it also contains 80% of 1193 samples for training, 20% rest for testing, and 8 features, including *experience_level, job_group, residence_region, company_region, work_year, company_size, remote_ratio,* and *salary_currency.* This dataset provides the best performance for the predictive model, surpassing other datasets that I tried to handle. Figure 10 illustrates the cleaning process: drop missing values, drop duplicates, flag outliers by the IQR method, before drop them.

3.3.2 Predictions

I made 6 predictions to clarify how *experience_level* and *work_year* affect *salary_in_usd*.

- **Experience Level**: I just changed *experience_level* and kept the same values of other factors (Figure 11). The salary increase aligned with the experience level, so it matched the correlation analysis.

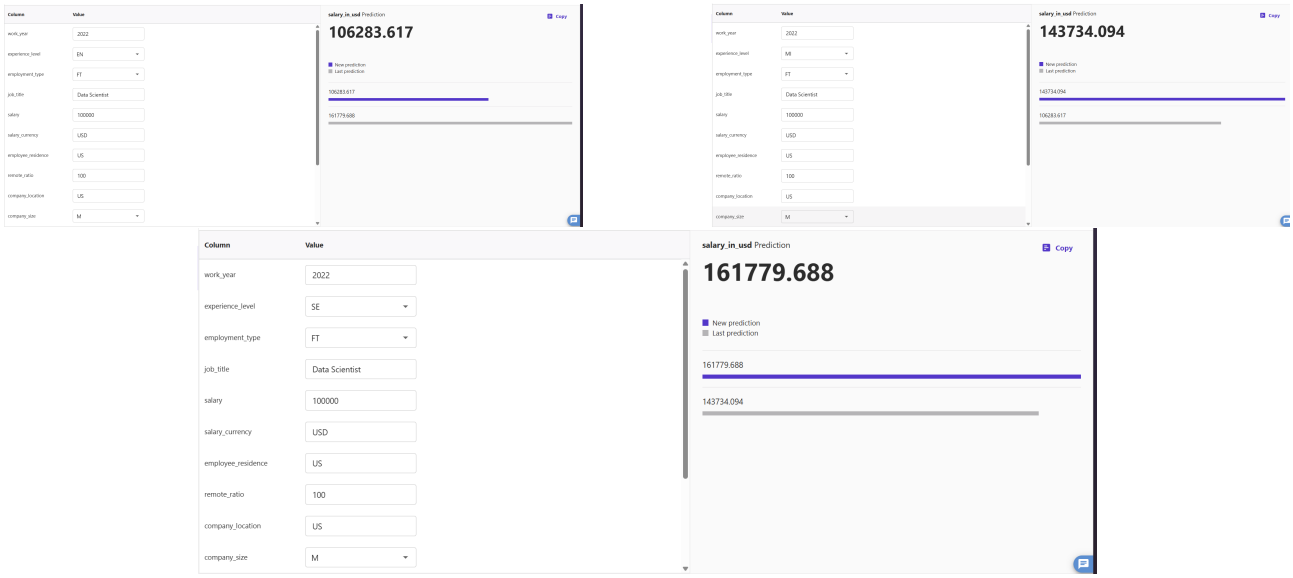

Figure 11: Change Work Year

- **Work Year** Similarly, I adjusted *work_year* and kept the same values of other factors (Figure 12). The predictive salary increased steadily year by year; thus, it also matched the correlation analysis.

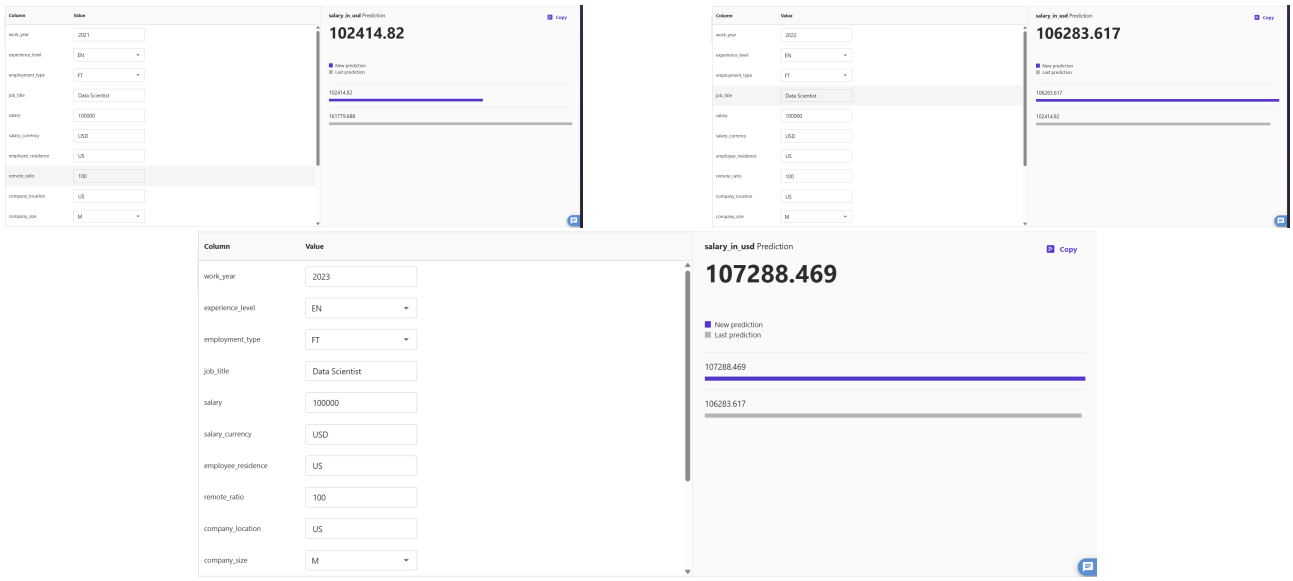

Figure 12: Change Experience Level

3.3.3 Evaluation Metrics

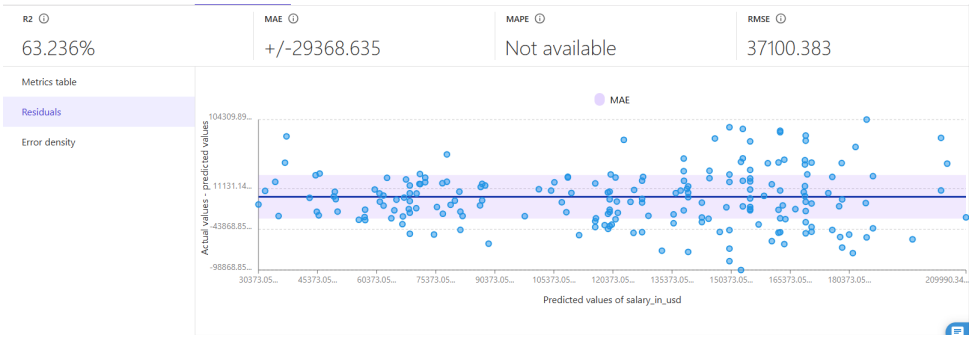

Figure 13: Model Output

Figure 13 describes the difference between actual salary values and predicted ones. MAE $\pm 29,385$ USD "measures the average absolute difference between the predicted and actual" salary (Harris, 2025), while RMSE is about $37,100$ USD, which Hodson (2022) analyzed that RMSE is often more sensitive to large errors than MAE. However, the salary ranged from $10,000$ USD to nearly $290,000$ USD (fluctuating about $280,000$ USD), MAE and RMSE just account for 10.5% and 13.25% of this fluctuation, respectively, pointing out that the relative errors could be acceptable in the large value range.

4 Findings and Recommendations

4.1 Findings

Figure 14: Features' Impact

As Figure 13 shows, the model's R2 is medium, just 63.24%, which can be explained that the model could understand $63,24\%$ the relationship between features and the target (Amazon Web Services, n.d.). The *residence_region* play a most important role in the output ($36,52\%$), following by *experience_level* and *salary_currency* about $19,05\%$ and $16,837\%$, respectively (Figure 14). A reverse pattern could be seen in *company_size*, *company_region*, and *remote_ratio*. These results also imply that the order of factors affects salary. The graph in Figure 14 also indicates that the US strongly contributes to the result of the model, and a small range in impact on prediction points out the high accuracy of prediction when tackling US data.

4.2 Recommendation

1. **Optimizing salary policies using the result of the model**
    - **Recommendation**: It is advisable to build and utilize a dynamic salary framework based on several factors, such as employee nationality, experience, currency, and job title, shown in the analysis section.
    - **Reason**: Research shows that companies with national wage policies often pay similar money for the same role across different locations, but amend a little bit for local costs differentials and labor market conditions to remain competitive and fair within the global market. Hazell, J., Patterson, C., Sarsons, H., & Taska, B. (2021) demonstrate that a dynamic framework could bring various enormous benefits, such as:
        - Guaranteeing equality by adjusting salaries to experience and job group.
        - Maintaining market competitiveness due to currency and regional cost of living.
        - Adapting in real time through algorithmic adjustments as market standards and exchange rates change.
2. **International recruitment**
    - **Recommendation**: The company should recruit employees from nations outside of the US, especially Asia, or open more offices for working remotely to save the company's budget.
    - **Reason**: Both analysis and model output point out that employees from US requires a higher salary level as a whole (e.g., New York 80,000 USD–180,000 USD, San Francisco 100,000 USD–200,000) USD, while those from Asia has a lower salary average(e.g., Shanghai 40,000 USD–90,000 USD, Pune 18,000 USD–58,000 USD) (Horizons, 2024). Leaders of the company can take it into account to promote competition between job-seekers and find talent around the world with compatible salaries.
3. **Identifying anomalies and lack of transparency**
    - **Recommendation**: The authorities of companies should implement internal audit frequently to check the salaries that significantly deviate from the predicted salary.
    - **Reason**: Some outliers could be evident for injustice, mistakes, or the failure in the recruitment strategy. CPE Trainer(2024) indicated that regular anomaly detection ensures pay equity by discovering outliers for review and supports compliance with internal and external compensation standards.
4. **Establishing and developing data infrastructure**
    - **Recommendation**: It is recommended that building data centers and forming data teams should be prioritized to improve the quality and accuracy of models.
    - **Reason**: A modern data infrastructure—including automated pipelines, versioned datasets, and real-time processing—which reinforces the reliability of ML models. Improving data systems could bring benefits for firms, such as faster iteration, higher model accuracy, and lower operational costs. Sajid (2025) highlighted 4 key merits of data centers, including:

- Accessing high-quality data: centralized storage and governance could help firms ensure full rights to the data
- Scalability: Infrastructure expands with data volume, which helps with model upgrades continuously.
- Cost efficiency: Automated workflows optimize labor sources and resource waste.
- Competitive advantage: Exploiting data could help firms attract talent from opponents.

5 Conclusion

In conclusion, although the quality of the dataset is not good enough, key insights into the salary trend can be interpreted through analysis methods. Specifically, salary is strongly affected by residence nationality, experience, currency, and job title. Exploiting the result of prediction, firms could implement strategies, such as salary policies, international recruitment, and regular audits, to ensure sustainable development. To improve the performance of models in the near future, firms should enhance data quality by upgrading the data ecosystem. Ultimately, data plays an indispensable role in this era, so which firms that can utilize and grasp data will create a competitive advantage.

References

Amazon Web Services, Inc. (n.d.). Metrics reference - Amazon SageMaker. Amazon. https://docs.aws.amazon.com/sagemaker/latest/dg/canvas-metrics.html

CPE Trainer. (2024, November). *Predictive auditing: Revolutionizing financial oversight* [Webinar PDF]. https://on-demand.cpetrainer.com/wp-content/uploads/2024/11/Predictive-Auditing-Revolutionizing-Financial-Oversight-short-live-webinar.pdf

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*(10), 70–76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Hazell, J., Patterson, C., Sarsons, H., & Taska, B. (2021, November). *National wage setting* [Working paper]. Centre for Macroeconomics, London School of Economics. https://www.lse.ac.uk/CFM/assets/pdf/Christina-Patterson-National-Wages.pdf

Hodson, T. O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev., 15, 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022,2022.

Horizons. (2024, February). *The cost of global recruitment in 2024*. https://joinhorizons.com/wp-content/uploads/2024/02/The-Cost-Of-Global-Recruitment-in-2024.pdf

Michael Harris. (2025, February 3). Understanding RMSE, MSE, and MAE — Stats with R. Stats with R. https://www.statswithr.com/foundational-statistics/understanding-rmse-mse-and-mae

Sajid, H. (2025, July 3). *Why your AI data infrastructure is the real competitive advantage.* Encord. https://encord.com/blog/data-infrastructure-competitive-advantage/

A Appendix: Full Code

My code: https://github.com/kainguyen25/Analysis-and-Prediction-Salary.git
My original report: https://www.overleaf.com/read/drqtrpjfxcbh#96dcc2