Final Report

Haoyu Yu

Introduction

As the AI and tech industry continues to grow, understanding salary distribution and the factors that influence it is crucial for both employers and employees. Several variables can impact salary, such as experience level, remote work, job title, company size, and geographical location. For organizations, knowing these factors can help in offering competitive salaries, while for professionals, it assists in career planning and salary negotiations. The dataset used in this analysis contains salary information for individuals working in the global AI and tech industry, spanning different roles, experience levels, employment types, and locations. The dataset includes data points such as salary (in both local currency and converted to USD), remote work ratios, company sizes, and the employee's country of residence. The following is an explanation of the important variables in the dataset to better understand the results.

Variable	Description
Work_year	The year the salary was paid
Experience_level	The experience level in the job during the year: - EN: Entry-level / Junior - MI: Mid-level / Intermediate - SE: Senior-level / Expert - EX: Executive-level / Director
Employment_type	The type of employment for the role: - PT: Part-time - FT: Full-time - CT: Contract - FL: Freelance
Remote_ratio	The amount of work done remotely: - 0: No remote work (<20%) - 50: Partially remote/hybrid - 100: Fully remote (>80%)
Company_size	The number of people working in the company: - S: Small (<50 employees)

Variable	Description
	- M: Medium (50-250 employees) - L: Large (>250 employees)

Based on this dataset, two questions worth exploring are proposed 1. What are the key factors that influence salary levels in the global AI and tech industry, and how do these factors interact to shape salary 2. Whether there are more interactions between different factors related to salary.

method

Data Acquisition

The dataset used in this analysis was obtained from the website AI Jobs Salaries, which provides detailed salary data for various roles in the AI and tech industry. The dataset includes variables such as work year, experience level, job titles, salaries in local currencies and converted to USD, remote work ratios, company size, and employee location.

Data Cleaning and Wrangling

Using `dim(salaries_data)` function to determine the dimensions of the dataset. The dataset contains 53,841 pieces of salary data and 11 variables. Using `str(salaries_data)` to determine the data type of each variable. Using `head(salaries_data)` and `tail(salaries_data)` to observe the first few rows and last few rows of data. And then, using `colSums()` and `is.na(salaries_data)` functions to view missing data in the dataset. There are no missing data in this dataset.Using `summary(salaries_data)` to check the statistical results of the dataset and look for any unreasonable data.

By observing the dataset, I found that the `job_title` variable is not convenient for statistical analysis because the title of each job is different. The `salary` variable and `salary_currency` variable can be replaced by `salary_in_usd` variable. Therefore, delete `job_title`, `salary` and `salary_currency` variables, making the data set more concise.

Since variables `experience_level`, `remote_ratio`,`company_size` and `employee_residence` are categorical variables, these variables are created as factors to facilitate subsequent calls.

Results

Factors that affect salary

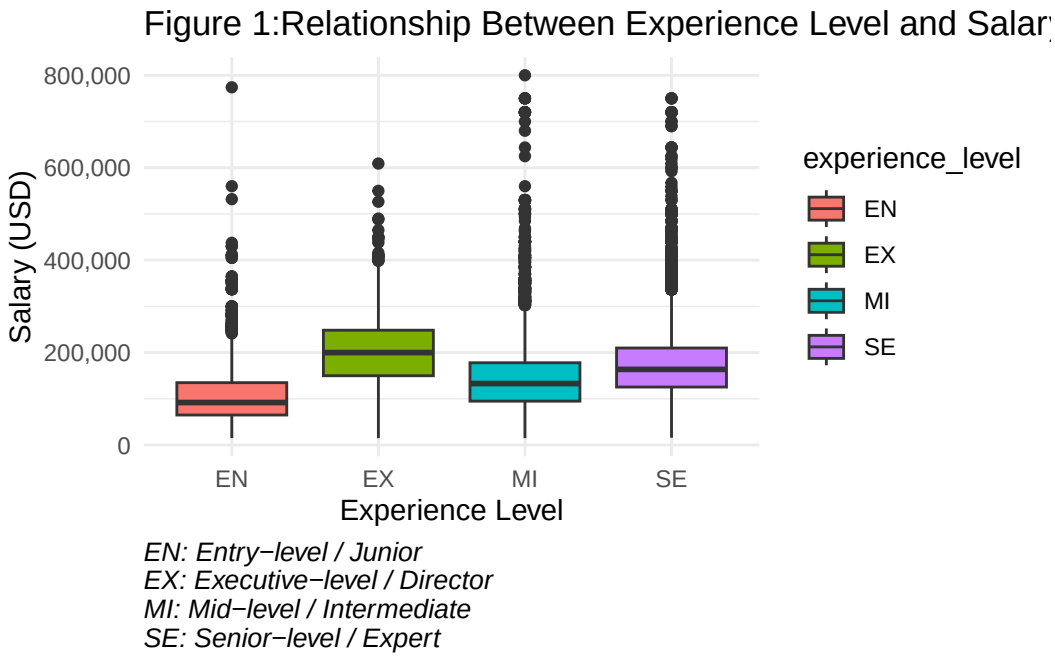


Figure 1: The box plot illustrates the relationship between experience levels and salary in the global AI and tech industry. Entry-level (EN) employees have the lowest median salary and the narrowest range, indicating a lower variation in earnings for junior positions. Executive-level (EX) employees exhibit the highest median salary and the largest salary range, reflecting a significant premium for leadership roles and strategic expertise. Mid-level (MI) salaries are slightly lower than Senior-level (SE) salaries but demonstrate a narrower range. Senior-level (SE) employees earn significantly more than entry-level (EN) positions and represent a bridge between managerial and technical expertise. The salary data aligns with industry norms where experience strongly correlates with potential salaries

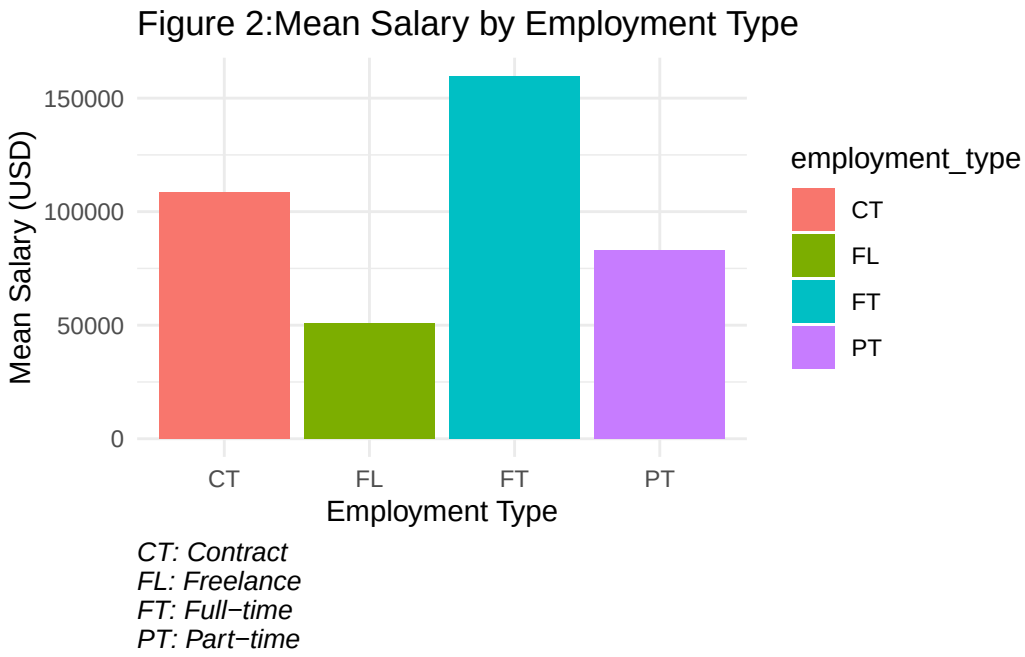
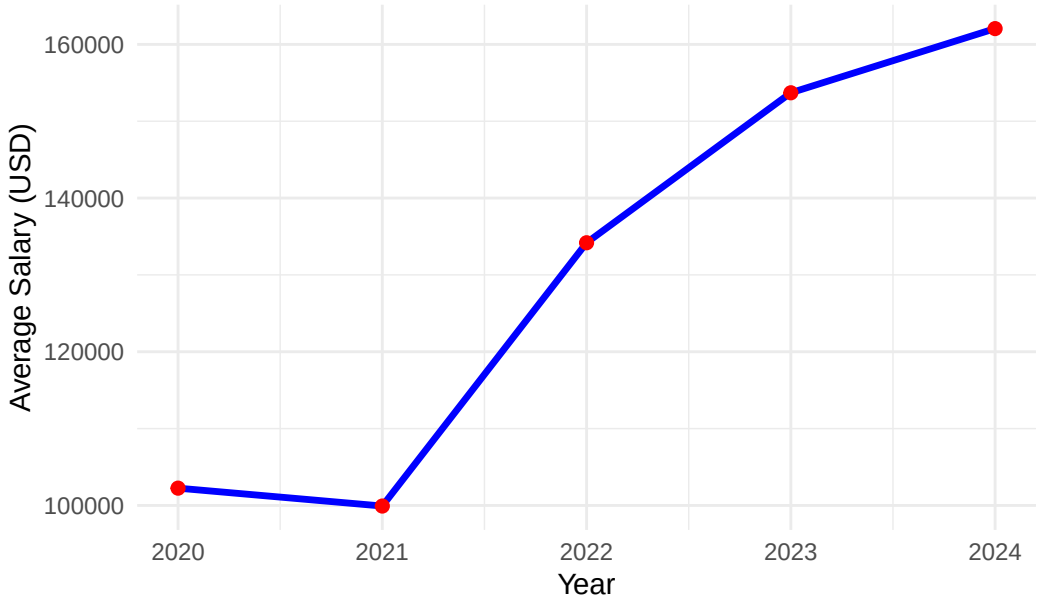


Figure 2: Full-time employees earn the highest mean salary among all employment types. This is likely due to the stability and benefits associated with full-time roles. Contract workers have the second-highest mean salary. This reflects the premium often paid for specialized or short-term expertise. Freelancers earn the lowest mean salary, while part-time workers earn slightly more than freelancers but significantly less than full-time or contract workers. This could be due to fewer hours worked or the lack of long-term stability. The data suggests that the type of employment significantly impacts salary levels, with more stable employment types commanding higher pay.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

Figure 3:Average Salary by Year



The average salary shows a steady increase from 2021 to 2024, with a particularly sharp rise between 2021 and 2022. The slower growth from 2023 to 2024 indicates a potential stabilization in salary increments. However, the slight decrease in 2021 compared to 2020 reflects the changing dynamics of the industry, which may be the impact of covid 19. The continuous increase in average salaries reflects the growing demand for skilled professionals in the AI and tech industry.