Analyzing In-Office Salary Trends in AI, Data, and Machine Learning Roles

Final Report (by André Jardim)

Introduction

The AI, data, and machine learning industries are experiencing significant growth as more organizations embrace data-driven strategies and automation. However, with a strong preference for in-office roles within these fields, understanding compensation trends is critical to attract and retain skilled talent. This project will analyze salary trends for in-office positions across AI, data, and machine learning roles, focusing on variations by job title, experience level, company size, company location, and changes over time from 2020 to 2024. With a majority of entries originating from the U.S., this analysis will explore how U.S. salary trends align with or diverge from international patterns. Additionally, the dataset lacks a primary key, resulting in duplicate entries that must be carefully managed to preserve data integrity.

Business Impact

By examining in-office salary trends in these fast-growing fields, the analysis will provide companies with benchmarks for competitive compensation strategies, aiding in talent acquisition and retention. This data-driven approach will also benefit industry professionals, offering clear benchmarks that support informed career decisions and negotiations. For policymakers and analysts, the findings can illuminate regional disparities in salary and the relative competitiveness of the U.S. against other regions.

General Dataset Information

File Name: salaries.csv

Description: AI, ML, Data Salaries 2020 – 2024

Dataset Details: 59,325 Rows / 11 Columns

Size: 3,304 KB (3.3 MB)

Source: AIJobs.net

Method of Collection: Direct download from AIJobs.net

Data Analysis & Computation

Access the full Jupyter Notebook for this project here.

Data Profiling Steps:

- Checked first 10 rows for initial structure and formatting and verified columns names and data types for consistency.
- Calculated key salary metrics, such as minimum, maximum, and average, to understand the data's range and typical values.
- Checked for zero or negative values in the *salary_in_usd* column.
- Confirmed no null values across columns.
- Examined distinct values in key columns such as *experience_level*, *company_size*, *company_location*, *job_title*, *employment_type*, *remote_ratio* and *work_year*.
- Verified ISO 3166 country codes conformity for *company_location* column.

- Ensured USD salaries match *salary_in_usd* column values and checked non-USD salaries for correct labeling and conversion consistency.
- Conducted salary distribution analyses for salary by experience level, company size, company location, job title, and work year.
- Analyzed remote ratio and employment type distribution.

Observations:

- Dataset covers AI, data, and machine learning roles across 258 unique job titles.
- Medium-sized companies are most represented, with 57,099 records.
- Salaries are standardized to USD, with original currency information available.
- Heavy U.S. focus: 53,809 records from the U.S., 1,925 records from Canada, and 1,688 records from United Kingdom. All other countries contribute fewer than 200 records each.
- Records are concentrated in recent years, particularly 2024 (48,851 records) and 2023 (8,522 records).
- Full-time employment type includes 59,094 records. Only full-time roles will be considered for analysis.
- In-office roles comprise 46,013 records. Remote roles (with remote ratio values of 50 or 100) will be excluded from analysis.

Data Wrangling Steps:

- **Dataset Structure, Data Types, and Missing Values:** Displayed the structure of the dataset, data types, and missing values for an initial understanding.
- **Unique Values in Categorical Columns:** Printed unique values for categorical columns to identify any inconsistencies or necessary adjustments.
- **Convert Country Codes to Full Names:** Converted ISO country codes in the *company_location* column to full country names using the pycountry library.
- **Map Columns to Descriptive Labels:** Mapped the *remote_ratio*, *company_size*, and *experience_level* columns to more descriptive labels.
- **Filter Dataset for Full-Time, In-Office Roles:** Filtered the dataset to include only full-time, in-office roles, removing rows that did not meet these criteria.
- **Check Duplicate Rows:** Checked for duplicates in the filtered dataset and printed the number of duplicate rows before and after filtering.
- **Company Location Counts:** Analyzed the frequency of different company locations in the filtered dataset.
- **Aggregate Salary Data:** Aggregated the salary data by calculating the median salary for each group of key columns, including *work_year*, *experience_level*, *company_size*, and others. Rows removed during aggregation were saved for reference.
- **Job Title and Company Location Counts After Aggregation:** Examined the frequency of job titles and company locations after aggregating salary data.
- **Salary Distribution Across Top 10 Countries:** Plotted the top 10 countries with the highest average salary.
- **Location Count Threshold Filtering:** Applied a threshold to filter out locations with fewer than a specified number of records, ensuring that the salary trends analyzed are statistically relevant. Additionally, a pie chart was plotted to visualize the distribution of company locations in the filtered dataset.
- **Plot Salary Distribution for Outlier Detection:** Created a plot to visualize the salary distribution and detect potential outliers.
- **Capping Outliers in Salary Data:** Capped salary values that fall outside the 5th and 95th percentiles for each combination of work year, experience level, company location, and company size. This helped mitigate the influence of extreme salary values, ensuring that the data is more representative of typical salary trends within each group.

- **Salary Distribution Comparison:** Compared the salary distribution before and after capping the outliers. The distributions were not significantly different, indicating that the capping process did not distort the overall data significantly.
- **Drop Columns Not Required for the Analysis:** Dropped any columns that were not necessary for the analysis, streamlining the dataset for easier handling and interpretation.
- **Cleaned Dataset Overview:** Provided an overview of the cleaned dataset, highlighting the final structure and any changes made during the data wrangling process.
- **Cleaned Dataset Info:** Displayed the final cleaned dataset's information.
- **Summary of Removed and Adjusted Rows:** Summarized the rows removed or adjusted during the cleaning and wrangling process, providing clarity on the extent of data modifications made.
    - Rows removed due to non-full-time positions: 231 rows
    - Rows removed due to non-in-office positions: 13,219 rows
    - Rows removed during aggregation: 44,763 rows
    - Rows removed due to location count threshold: 408 rows
    - Rows capped as outliers: 149 rows
- **Dataset after filtering and capping:** 1,034 rows, 6 columns. The cleaned data has been saved to 'cleaned_data.csv' for further analysis.

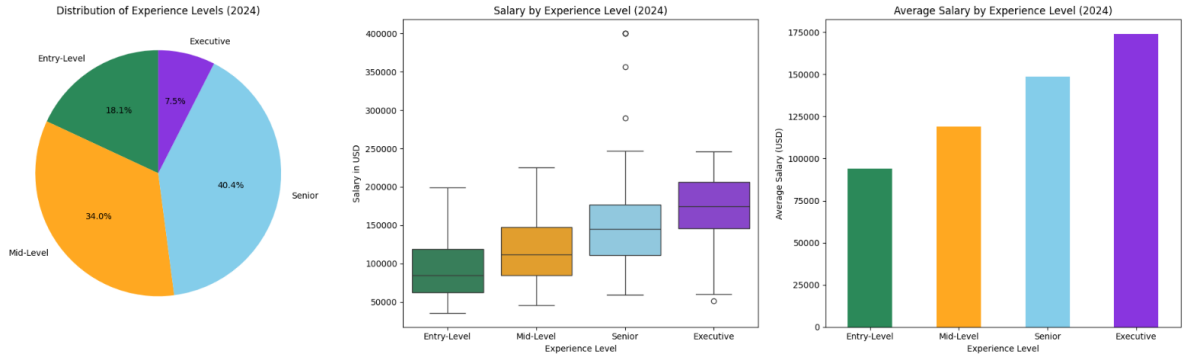
Cleaned Dataset Schema

Column Name	Data Type	Description
work_year	Integer	The year in which the salary data was reported. It represents the year of the survey or data entry.
experience_level	String	The experience level of the employee. This indicates the professional experience of the individual.
job_title	String	The job title of the employee. Represents the role held by the individual in the company.
company_location	String	The location of the company, represented by the country where the company is based.
company_size	String	The size of the company. This represents the scale of the organization based on employee count.
salary_in_usd	Float	The salary of the employee converted to USD. This is the standardized salary amount used for analysis, ensuring all salaries are in the same currency.

Data Analysis:

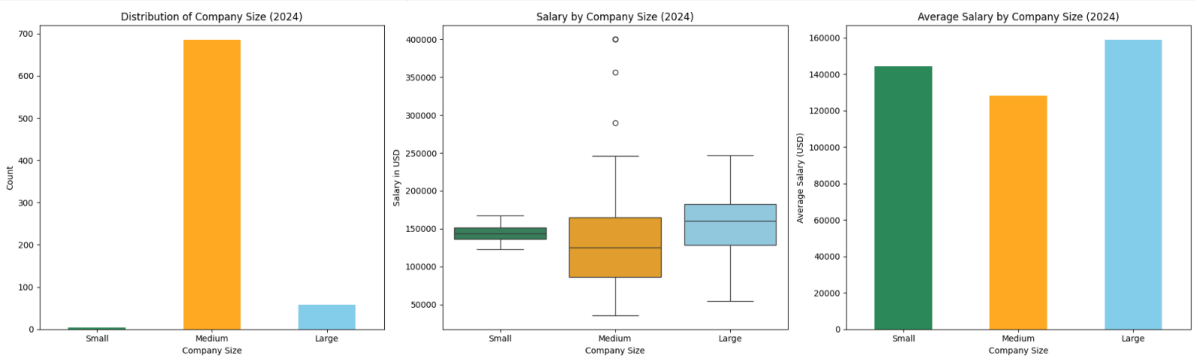
Salary Distribution by Experience Level in 2024:

- In 2024, salaries increase with experience level, with entry-level roles earning an average of $93,954, mid-level at $118,841, senior at $148,740, and executive at $173,866. The variability in salaries also rises across levels, with senior and executive roles exhibiting a wider range, reflecting diverse roles and responsibilities. Entry-level positions show lower salary variance, while senior and executive roles have significant spreads, indicating that higher positions often have larger discrepancies based on company size, industry, and role specialization. Overall, experience level strongly influences salary, with higher positions commanding substantially higher pay, though with increased variability.



Salary Distribution by Company Size in 2024:

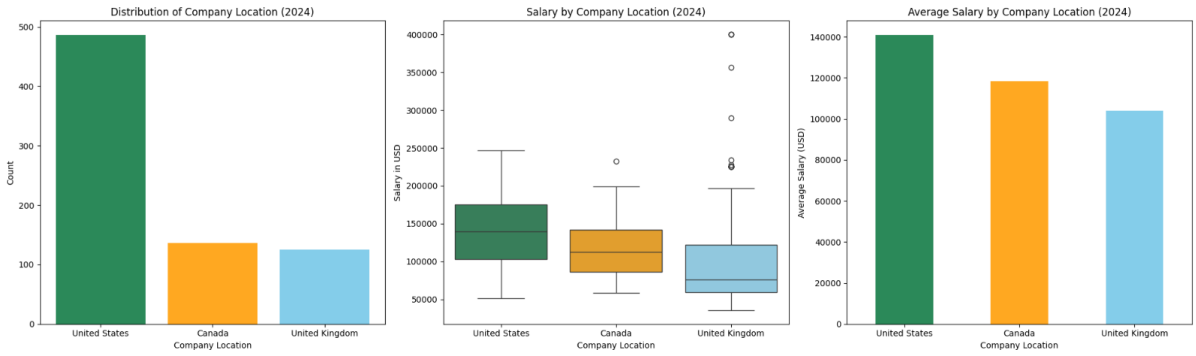
- In 2024, salaries are highest in large companies ($158,860), followed by small companies ($144,377) and medium companies ($128,079). The small company group is underrepresented in the dataset, with only 4 records remaining after data cleaning and transformation, which makes it harder to draw reliable conclusions from this group. Medium-sized companies show the most salary variability, with salaries ranging from $35,625 to $400,000. Large companies also have a broad salary range, from $54,438 to $247,065, but with fewer extreme values. Overall, salary tends to increase with company size, although medium-sized companies show the greatest variation in salary.



Salary Distribution by Company Location in 2024:

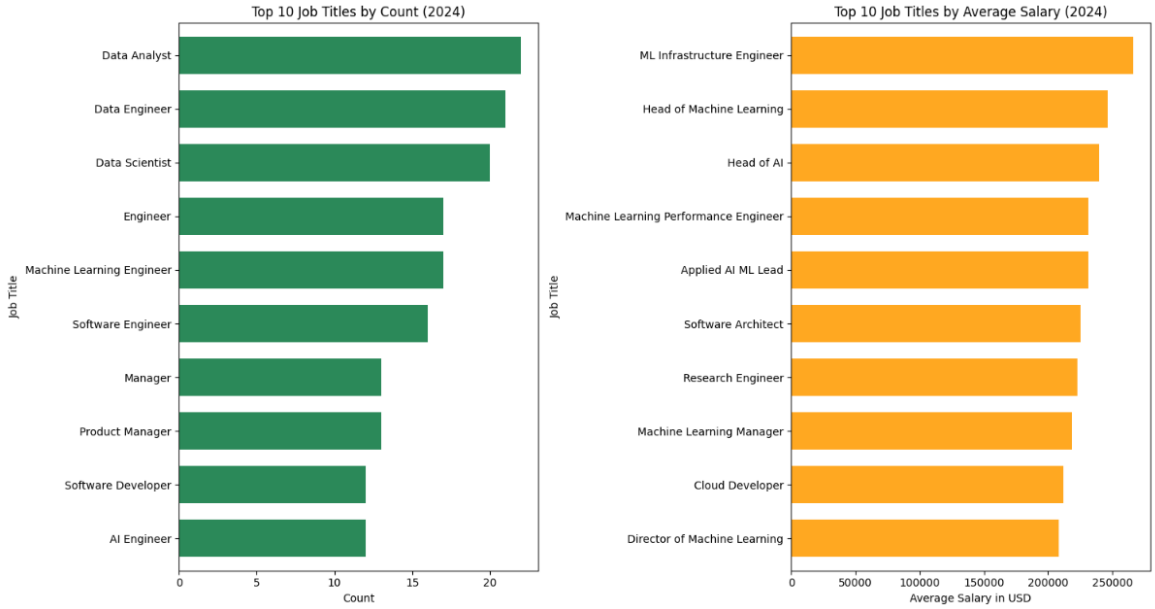
- In 2024, salary trends across different company locations exhibit notable variations. The United States stands out with the highest average salary of $140,832, followed by Canada at $118,240, and the United Kingdom at $104,003. The U.S. also shows the greatest salary range, ranging

from $51,050 to $247,065, indicating significant salary disparities within the region. In contrast, Canada has a narrower range, with salaries between $58,340 and $232,750, and a lower standard deviation, suggesting more consistency in salary distribution. The United Kingdom, while having the lowest average salary, displays a broader salary range from $35,625 to $400,000, with considerable variation, as reflected by the high standard deviation. These differences underscore the varying economic conditions, cost of living, and job market characteristics across these regions.



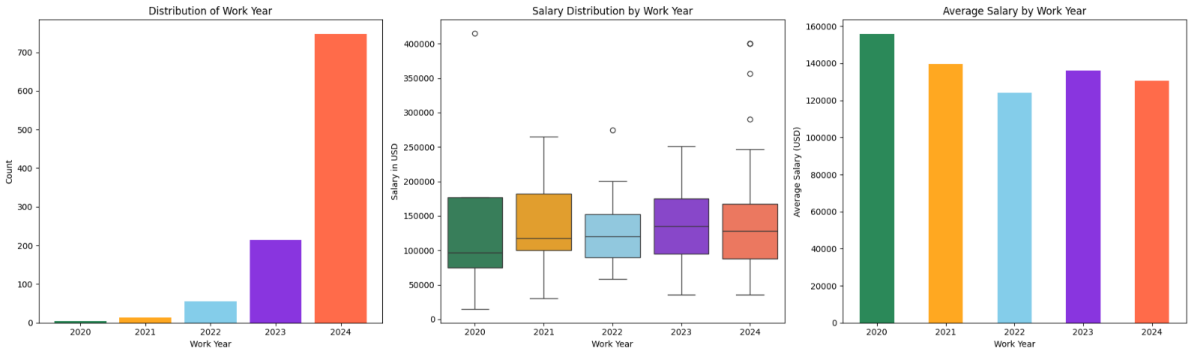
Salary Distribution by Job Title in 2024:

- The analysis of job titles and salary data for 2024 reveals key trends in both the prevalence and compensation of roles across the dataset. The top 10 job titles by count reflect the most commonly held positions in the industry, with Data Analyst, Data Engineer, and Data Scientist being particularly frequent. These roles, though widespread, show notable variability in salary, with positions like Machine Learning Engineer and Software Engineer offering higher average salaries.
- In contrast, the top 10 job titles by average salary are more specialized and managerial, with roles such as ML Infrastructure Engineer, Head of Machine Learning, and Head of AI commanding significantly higher compensation. These positions, however, tend to be less common, reflecting the demand for specialized skills and leadership in cutting-edge fields like machine learning and artificial intelligence. It is important to note that the presence of lower-frequency observations in the dataset—such as roles with only one or two instances—can skew the average salary for these job titles, potentially affecting the ranking of higher-paying positions.



Salary Distribution by Work Year:

- The salary data across different work years reveals some notable trends. In 2020, the average salary was the highest at $155,775, but with a very high standard deviation ($176,924), indicating a wide range of salaries, including extreme values. In 2021, the average salary dropped to $139,793, with a more consistent distribution (lower standard deviation of $70,035). The trend continued in 2022, where the average salary fell further to $124,198, and the distribution tightened. 2023 saw a slight recovery with an average of $136,222, while 2024's average salary was $130,556.
- It's important to note that years with fewer data points, such as 2020 (4 observations) and 2021 (14 observations), may not be fully representative of the broader trends and could have skewed the overall averages. The variability in these years highlights the influence of low sample counts on salary averages.



Challenges & Solutions

High Duplicate Count: Aggregated data carefully to maintain representation without overemphasizing recurring entries.

Skewed Data: U.S.-centric entries required sensitivity in global trend analyses.

Outliers: Applied robust statistical methods to cap extreme values without compromising data integrity.

Data Mapping: Converted abstract column values (e.g., ISO codes) to user-friendly formats for better interpretation.

Limited Work Year Representation: Earlier years (2020–2023) had fewer entries, making trends less reliable. The analysis focused primarily on 2024, where data was most robust, ensuring accurate insights into current trends.

Limited Company Location Diversity: With a significant concentration of data from the U.S. and a few other countries, some locations lacked sufficient representation. Locations with fewer than a minimum threshold of entries were excluded to maintain statistical relevance.

Description of Dashboard

Overview

An interactive dashboard was developed to explore in-office salary trends, allowing users to filter by company location and analyze key insights through various visualizations.

Key Features

- **Filters:**
  - Users can filter data by company location, enabling a focused analysis of geographic salary trends.
- **Visualizations:**
  - **Bar Chart 1:** Average salary by company location, showcasing geographic disparities.
  - **Bar Chart 2:** Salary breakdown by experience level, highlighting variations based on expertise.
  - **Bar Chart 3:** The most common job titles, providing insights into role distribution.
  - **Bar Chart 4:** Top 10 highest salaries by job title, identifying the most lucrative positions.
- **Scorecards:**
  - Display the average salary for each company location, offering a quick summary of geographic salary benchmarks.

Use Case

- **The dashboard is designed for:**
  - **Employers:** To benchmark competitive compensation by location, experience level, and job title.
  - **Job Seekers:** To compare salary expectations based on geographic trends and job roles.
  - **Analysts:** To identify top-paying roles and understand salary dynamics across various regions and experience levels.

Conclusion & Recommendations

Conclusion:

In conclusion, the analysis of in-office salary trends in AI, data, and machine learning roles reveals that experience, company size, and geographic location are key factors influencing compensation. Higher experience levels correspond with higher salaries, particularly for senior and executive positions, which exhibit greater variability. Larger companies tend to offer higher salaries, while medium-sized companies show the greatest salary range. The United States leads in average salary, with Canada and the United Kingdom presenting more consistent but lower salaries.

The dataset, focused primarily on 2024, also highlights the diversity of job titles within the industry, with a significant number of different roles contributing to salary variability. Specialized roles, such as ML Infrastructure Engineer and Head of Machine Learning, command higher salaries, while more common roles like Data Analyst and Data Scientist show lower average salaries but are more widely represented. These insights are valuable for companies to optimize their compensation strategies and for individuals seeking clarity on in-office salary expectations in the evolving AI and machine learning fields.

Recommendations:

Based on the findings, it is recommended that future studies expand the dataset to include more diverse company locations, remote work roles, and a broader range of job titles to provide a more comprehensive view of salary trends. Companies should also consider factors like company size and location when designing competitive compensation packages, while professionals may benefit from focusing on specialized roles to maximize earning potential.

References & Acknowledgements

Data Source: AIJobs.net

Tools: Python libraries, Jupyter Notebook, SQLite, Tableau.

Acknowledgements: Thanks to AIJobs.net for providing the dataset and to contributors to open-source libraries.

A special thanks to **Correlation One** for their invaluable knowledge and support in helping develop the technical and analytical skills necessary to complete this project.

Links

Project Repository

Tableau Dashboard

Jupyter Notebook