Analyzing In-Office Salary Trends in AI, Data, and Machine Learning Roles (by André Jardim)

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 Profile

Initial Data Import and Setup:

- **Imported dataset into SQLite within Jupyter Notebook with 59,325 rows and 11 columns:**
  - **Columns:** *work_year, experience_level, employment_type, job_title, salary, salary_currency, salary_in_usd, employee_residence, remote_ratio, company_location, company_size.*

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.

Recorded Inconsistencies and Adjustments:

- **Remote Ratio, and Employment Type:** The *remote_ratio* column, currently stored as an integer, will be mapped to a categorical variable and then filtered to include only in-office roles. The *employment_type* column will be filtered to retain only full-time positions.
- **Company Location, Company Size, and Experience Level:** These columns will be expanded for clarity, including the use of full country names instead of ISO codes, as well as more explicit definitions for company size and experience level.
- **No Primary Key:** The dataset lacks a primary key, which could impact data integrity during merging or joining operations. Since many records with identical salary data are valid duplicates due to the survey nature, I will carefully review and manage duplicates during data wrangling to ensure consistency.
- **High Duplicate Count:** There are 31,212 duplicate records, mostly from respondents with identical salary information. These duplicates reflect real survey responses, and I will assess how to handle them (e.g., by averaging salaries or retention) during data wrangling to ensure accurate analysis and prevent overrepresentation.
- **Potential Outliers:** Some salary values fall outside expected ranges, indicating potential outliers. Further investigation will determine whether these outliers should be removed or adjusted using statistical methods like Z-scores or the Interquartile Range (IQR).

Data Wrangling

- **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.

Access the full Jupyter Notebook for this project here.