Project: Data Science Salaries Analysis

Objective

This project aims to conduct a comprehensive analysis of data science salaries to uncover key trends and insights within the data science job market. The analysis will focus on understanding salary progression over time, the impact of remote work, the influence of company size, and the distribution of salaries across various job titles and experience levels.

Dataset Description

The dataset used in this analysis contains detailed information about data science job postings and associated salaries. It includes the following key fields:

- **work_year**: The year in which the job information is recorded.
- **experience_level:** The professional experience level of the employee (e.g., Entry-Level, Mid-Level, Senior, Executive).
- **employment_type:** The type of employment (e.g., Full-time, Part-time).
- **job_title:** The specific job role (e.g., Data Scientist, Data Engineer, Machine Learning Engineer).
- **Salary:** The raw salary in its original currency (this column will be dropped after conversion).
- **salary_currency:** The currency of the raw salary.
- **salary_in_usd:** The salary converted to US Dollars, serving as the primary metric for analysis.
- **employee_residence:** The country where the employee resides.
- **remote_ratio**: The percentage of remote work for the position (0 for no remote, 50 for partially remote, 100 for fully remote).
- **company_location:** The country where the company is located.
- **company_size:** The size of the company (Small 'S', Medium 'M', Large 'L').

Project Instructions

1. Data Preparation

- Load the dataset into Jupyter NoteBook(Python) .
- Perform initial data inspection using df.info() to understand data types and non-null counts.
- Clean the dataset by dropping irrelevant columns such as 'Unnamed: 0' and 'salary'.

2. Exploratory Data Analysis (EDA) and Visualization

Conduct the following analyses and visualize the findings:

- **Workforce Dynamics & Salary Progression:**
  - Group data by **work_year** and calculate the average **salary_in_usd.**
  - Visualize the average salaries by year using a bar chart to show trends in salary progression over time.
- **Remote Work Trends:**
  - Analyze the distribution of **remote_ratio** to understand the prevalence of remote, partially remote, and no-remote positions.
  - Visualize this distribution using a bar chart.
- **Company Size Analysis:**
  - Examine the distribution of **company_size** categories ('S', 'M', 'L').
  - Map these categories to more descriptive labels (Small, Medium, Large).
  - Visualize the distribution of company sizes using a pie chart.
- **Job Title Analysis:**
  - Determine the frequency of different **job_title** entries.
  - Identify and visualize the top 5 most common job titles using a bar plot.
- **Salary Distribution by Company Size:**
  - Investigate how **salary_in_usd** varies across different **company_size** categories.
  - Calculate and visualize the average salary for each company size using a bar chart.
  - Plot the distribution of salaries for each company size using histograms (or KDE plots) to show salary disparities.
- **Experience Level Distribution:**
  - Analyze the distribution of **experience_level** categories.
  - Map abbreviations ('SE', 'MI', 'EN', 'Ex') to full descriptions (Senior, Middle, Entry Level, Executive).
  - Visualize this distribution using an appropriate chart (e.g., a pie chart or bar chart) to show the proportion of each experience level

Key Findings & Deliverables

- A fully executable Jupyter Notebook containing all data loading, cleaning, analysis, and visualization steps.
- Visualizations (bar charts, pie charts, histograms) illustrating key insights into:
  - Yearly average salary trends.
  - The prevalence of remote work.
  - The composition of companies by size.
  - The most common job titles in the data science field.
  - How salaries vary based on company size and experience level.
- Clear textual explanations within the notebook summarizing the insights derived from each analysis section.