Statistical Report
on
Simulated HR Data

BY : Ziad Asaad

April 2025

Employee Data Analysis Report

Dataset Description

The dataset used in this analysis is a synthetically generated dataset designed to simulate employee records. It is not sourced from any official organization or real-world company, and it does not reflect actual individuals or organizations. The data was created for educational and analytical purposes to demonstrate data cleaning, transformation, and visualization techniques.

It includes 20,000 records and contains the following information:

- *Full name*
- *Gender*
- *Date of birth*
- *Hire date*
- *Email*
- *Age*
- *Years of experience*
- *City and country of residence*
- *Job title*
- *Marital status*
- *Education level*
- *Salary details (amount and currency)*

This artificial dataset is valuable for practicing data preprocessing, exploratory data analysis, and visualization workflows.

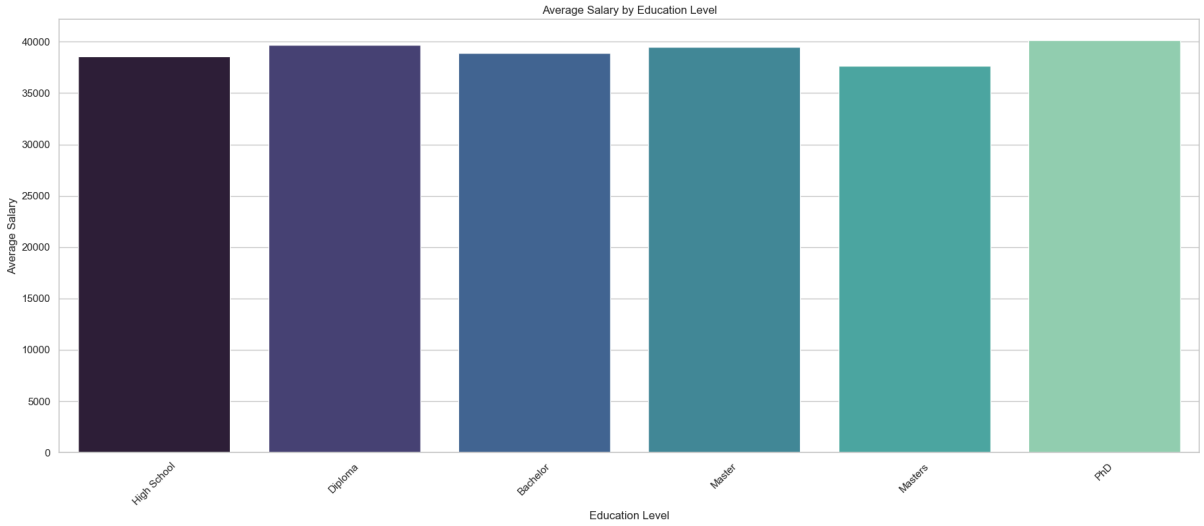
Introduction

This report presents a comprehensive analysis of employee data. The dataset includes information such as job titles, salaries, age, experience, education, city, country, and more. The purpose of this analysis is to explore patterns and insights related to salary, demographics, and job characteristics.

Data Cleaning & Preprocessing

- Removed unnecessary columns (`First_Name`, `Last_Name`)
- Standardized and converted the Date_of_Birth and Hire_Date columns to datetime format.
- Stripped whitespace and fixed capitalization in `Email`.
- Handled inconsistent date formats using `dateutil` parser.
- Filled missing `Age` values based on `Date_of_Birth` and converted to `int`.
- Extracted `City` and `Country` into separate columns.
- Cleaned `Job_Title` by:
- -Removing symbols and typos.
- -Standardizing names (e.g., `Teachr` → `Teacher`, `Sales Manger` → `Sales Manager`).
- Standardized `Marital_Status` values (`M`, `S`, etc.).
- Shortened `Education_Level` values (`Bachelor`, `Master`, etc.).
- Split `Salary` into:
- - `Salary_Amount` (as `int`)
- - `Salary_Currency`
- Filled missing salaries using the average for same `Job_Title` and `Country`.
- Standardized currency for countries where missing.
- Cleaned city and country names from special characters and typos.

1. Average Salary by Education Level

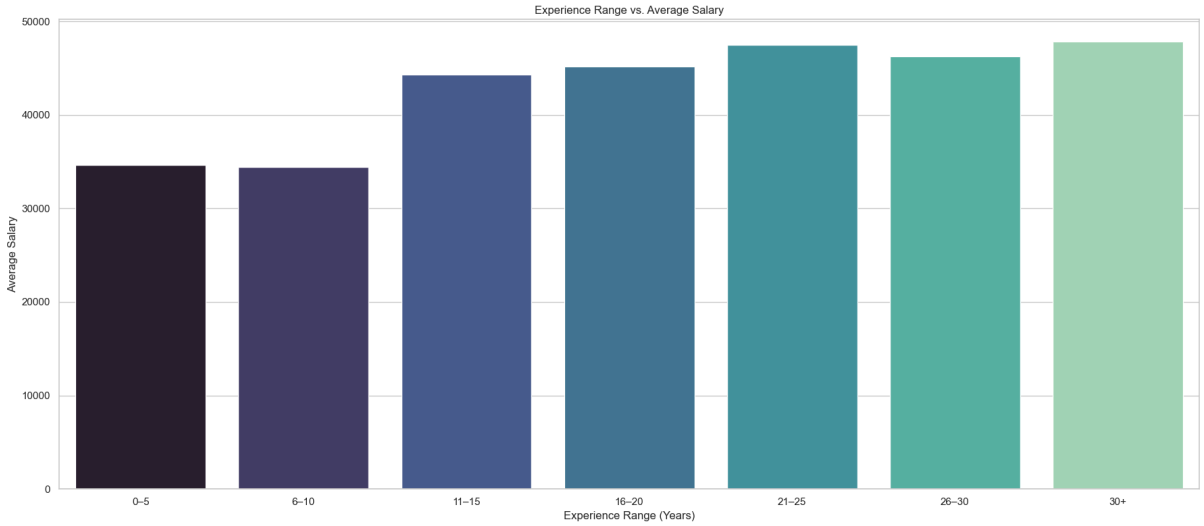


Description: This barplot visualizes the average salary in USD for each education level across the dataset.

Insight: Higher education levels such as `PhD` and `M.Sc.` are associated with higher average salaries. Conversely, individuals with a `High School` or `Diploma` education tend to have lower salaries.

---

2. Experience vs Salary (Grouped)

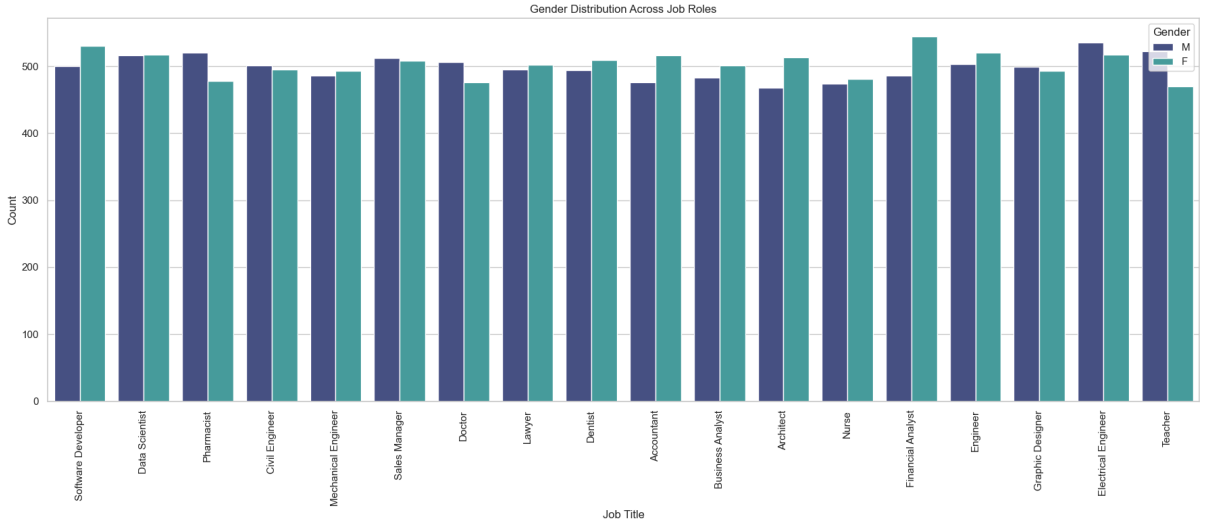


Description: A barplot that groups experience years into bins (e.g., 0-5, 6-10, etc.) and shows average salary per bin.

Insight: There is a clear upward trend: employees with more years of experience generally earn higher salaries, peaking in the 16+ years range.

---

3. Gender Distribution per Job Role

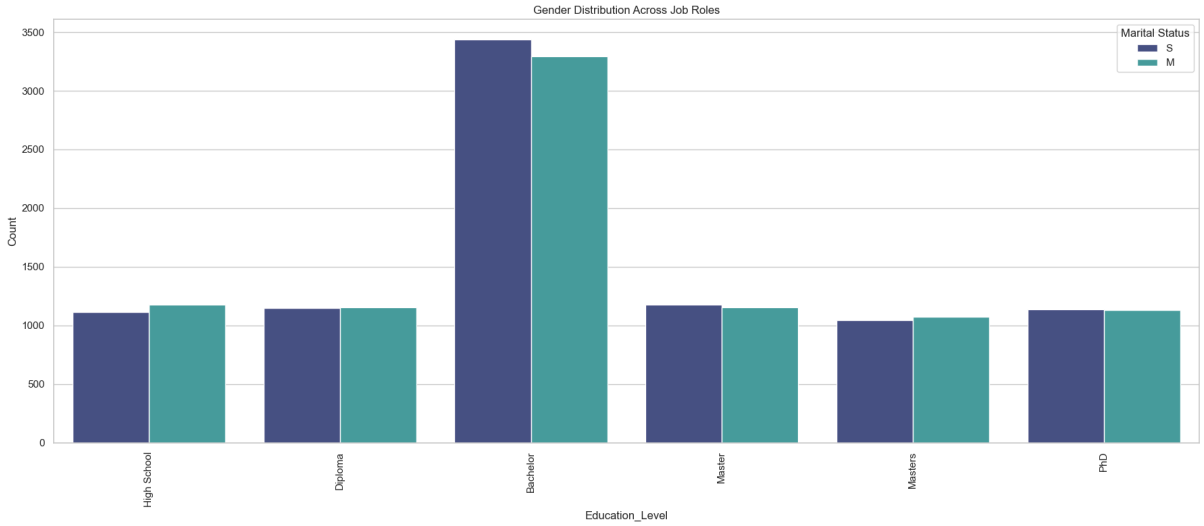


Description: A countplot showing the number of males and females across various job titles.

Insight: Some job roles like `Nurse` or `Teacher` have a higher number of females, while `Software Developer` and `Engineer` roles are male-dominated. This may reflect real-world industry trends.

---

4. Marital Status vs Education Level

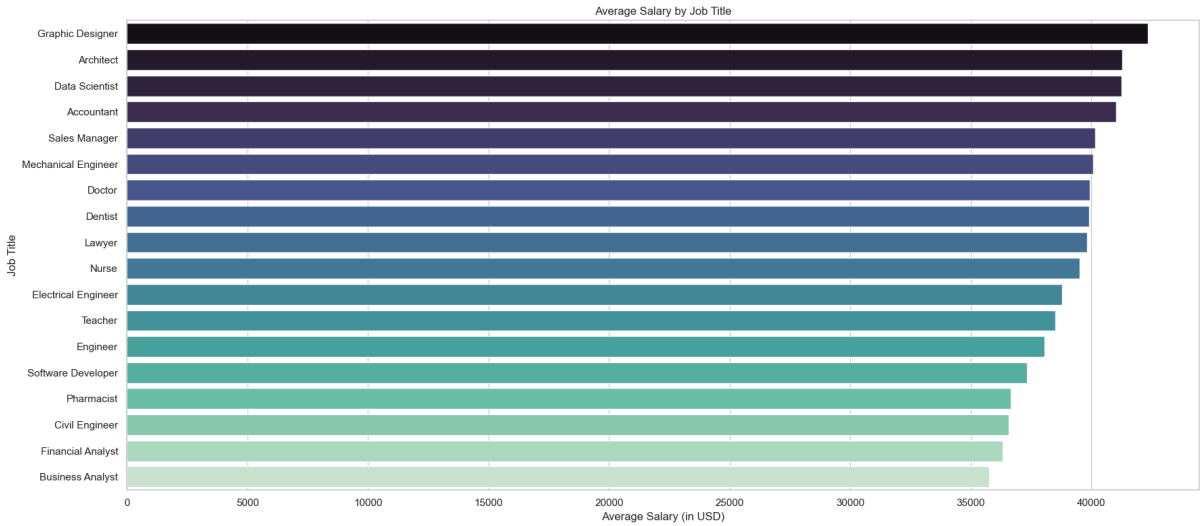


Description: A stacked barplot or grouped bar chart showing counts of marital status split by education level.

Insight: Married individuals tend to have higher education levels (Bachelor's and above), while single individuals are more prevalent among high school or diploma holders.

---

5. Average Salary by Job Title

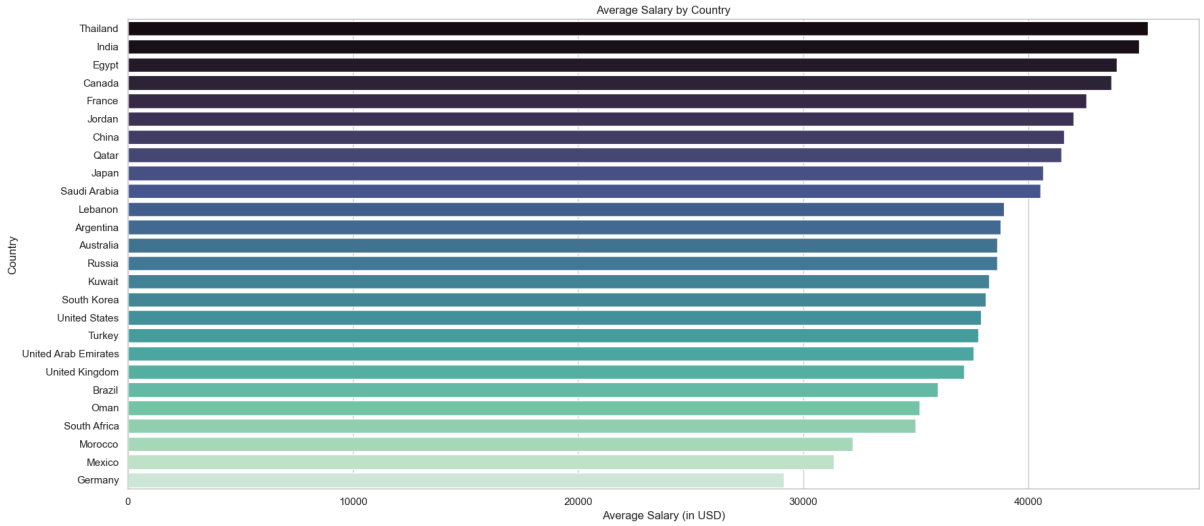


Description:
This horizontal barplot visualizes the average salary (in USD) for each job title in the dataset. The salaries have been converted to a standard USD value for accurate comparison.

Insight:
Surprisingly, creative and architecture-related roles like **Graphic Designer** and **Architect** top the salary chart, followed closely by **Data Scientist**. On the other hand, traditional business roles like **Business Analyst** and **Financial Analyst** appear at the lower end. This distribution may reflect sample bias or simulation logic, since it's generated data.

---

6. Average Salary by Country



Description:
This horizontal barplot visualizes the average salary (in USD) across different countries. The countries are listed along the y-axis, while the x-axis represents the salary range from 0 to 40,000.

Insight:
Salaries vary significantly by country, with Germany, the United States, and the United Kingdom appearing among the highest-paying nations. In contrast, countries like Thailand, India, and Egypt show lower average salaries. This disparity likely reflects differences in economic development, cost of living, and industry demand across regions.

Notably, some Gulf nations (e.g., United Arab Emirates, Saudi Arabia) also rank highly, possibly due to lucrative sectors like oil and finance. The data may be influenced by sample bias or currency conversion methods.

Conclusion

The data cleaning and analysis provided clear trends related to salary, job titles, experience, and location. These insights can help HR departments make informed decisions regarding compensation and recruitment.