Exercise 6.1- Sourcing Open Data

Data Source

Dataset was obtained via Kraggle and provides 2 million employee records from a Multi-National Corporation (MNC). The data includes personal identifiers, job-related attributes, performance, employment status, and salary information.

The purpose of this data is to provide analytics for a large-scale company. This data was selected for a variety of reasons. Primarily, the data requires analysis in python because it has 2 million rows. As a previous manager, I am interested in identifying how performance evaluations and modes of working correspond to salaries and employee attrition. Finally, the data is global, allowing for location-based analytics.

Data can be found here.

Data Profile

Data Cleaning and Understanding

The data is often used for practice and does not have many steps required for cleaning. Here are several steps taken to begin:

- Checked for duplicates and missingness
- Dropped unnecessary variables (one of which deidentifies the data)
- Reviewed Data Types for all the variables
    - Changed the date variable to the appropriate style
- Preparing the Data
    - Converted Indian Rupees Salary to USD
    - Split the Country and City location variable to two distinct variables
    - Created a binary 'Status2' variable that shows Active versus Inactive employees
- Created Summary Data Output
    - Both Quantitative and Qualitative summary outputs in Python
    - Salaries look to be a little low after conversion to USD- this may be a flaw in the data.

Data Limitations & Ethics

After reviewing the data, I did notice that the data is limited in a few instances. The data provides a static snapshot of employee records. It is unable to analyze how salary raises play a role in employee retention and attrition. Luckily, the data does incorporate a hire date which allows for some aspects of time to be reviewed.

Additionally, the employee salary itself is a limitation. The salary is originally in Indian Rupees and converting it to USD shows staff in high-ranking positions on very limited salaries. The average annual salary of $10,000 does not seem realistic. The cost of living in the US is 7 times greater and I will make this adjustment to the salaries so the average salary would sit around 70k which is more realistic when examining this company and its employees. This places the highest salary at over 210k. This does not seem problematic for someone in a CFO or CTO role.

The data ethically has few considerations since there is no evidence this corresponds to a real company or employees. The people are fictional and used only for the purpose of analysis and practice. To maintain proper ethical standards, I did deidentify the data after checking for duplicates. The analysis I am exploring does not require any individual employee's names. Additionally, each employee has an assigned employee ID which will function as the primary key in my analysis.

Questions to Explore:

1) Which countries have the largest concentration of employees?
    a. How does this relate to overall employee salary expenses?

2) Which countries hold the most leadership (CFOs, CTOs, etc.)?

3) How does the average salary compare to various departments, titles, and years of experience?

4) What is the work mode (on-site v. remote) distribution?
    a. Is there country variation in the allowance of remote work?
    b. How does work mode relate to employee attrition?
    c. What are the salary variations between on-site and remote employees

5) How has employee hiring shifted recently? Is there an emphasis on more remote work?

6) What is the correlation between performance evaluations and salary?

7) What department has the highest attrition rate? Which has the lowest?