Human Resources Analyst Project

Aditi Shah

Psychological Sciences, Purdue University

December 27, 2023

Introduction

This project seeks to explore, aggregate, and visualize employee data retrieved from Kaggle, a prominent platform for datasets and data science resources. With a foundation in Python, my goal was to extract meaningful insights from a broad range of employee information, highlighting trends, correlations, and patterns that explain organizational dynamics.

Exploratory Data Analysis

The given dataset consists of 1470 records and 35 columns. Among these columns, 26 of them were stored as integer values like Age, Monthly Rate, Years at Company, and Number of Companies Worked, to name a few. Additionally, this dataset contains 9 non-numeric columns like Business Travel, Department, Gender, and Job Role. After utilizing specific Python pandas functions, I was able to conclude that all 1470 records were non-null. Immediately, the first column I chose to analyze was Age to ultimately understand the spread of each of these employees (Figure 1). Secondly, I was interested in learning the distribution of the different Departments within this dataset (Figure 2). In the two given figures below, I was able to see some of these basic distributions produced with Python.

Figure 1:

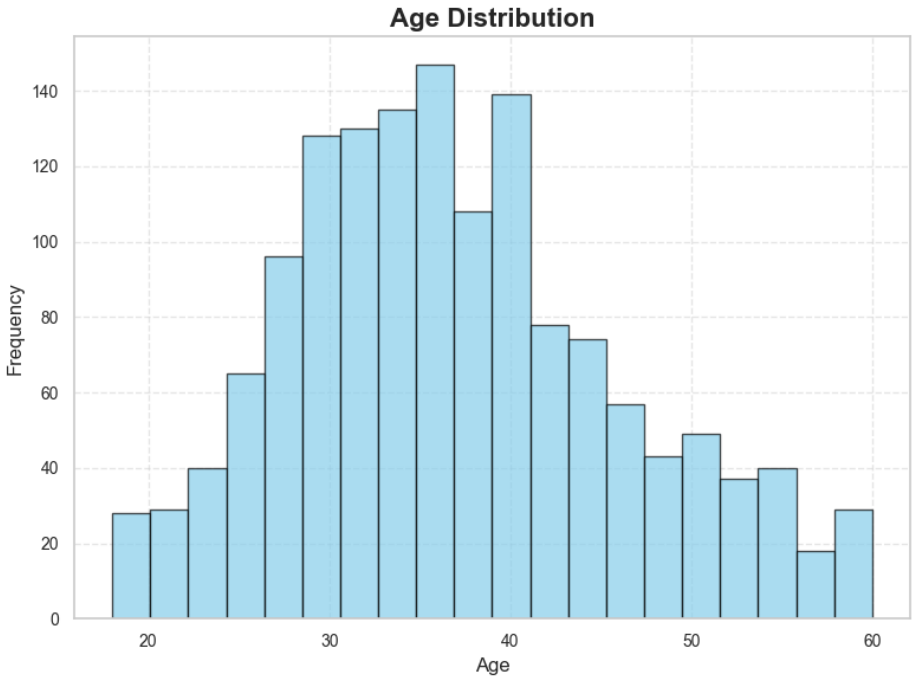
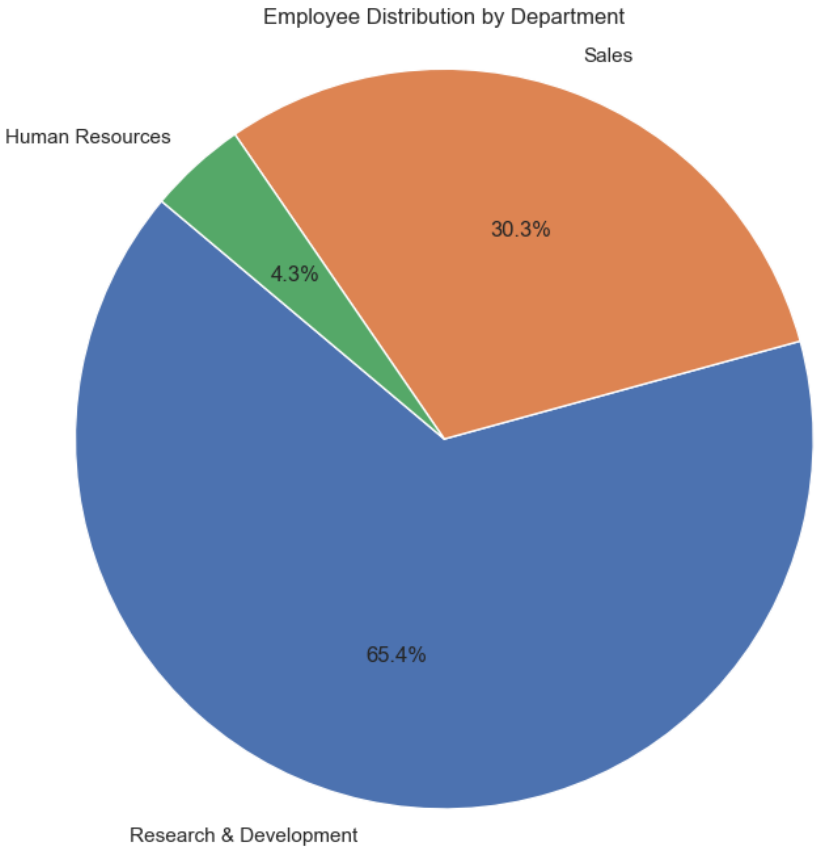


Figure 2:



Methods

Primarily, with the extraneous number of employees and variables included, it was difficult to process which variables would be related to one another. Therefore, I calculated a correlation matrix through the use of the pandas DataFrame functions. Considering there were multiple variables to be considered in this matrix, I narrowed down my results by only including those columns that met a minimum threshold of a 0.5 correlation coefficient. In the figure shown below (Figure 3), I was able to find pairs of variables that met this threshold to conduct further analysis.

Figure 3: Variables with Moderate-Strong Correlation

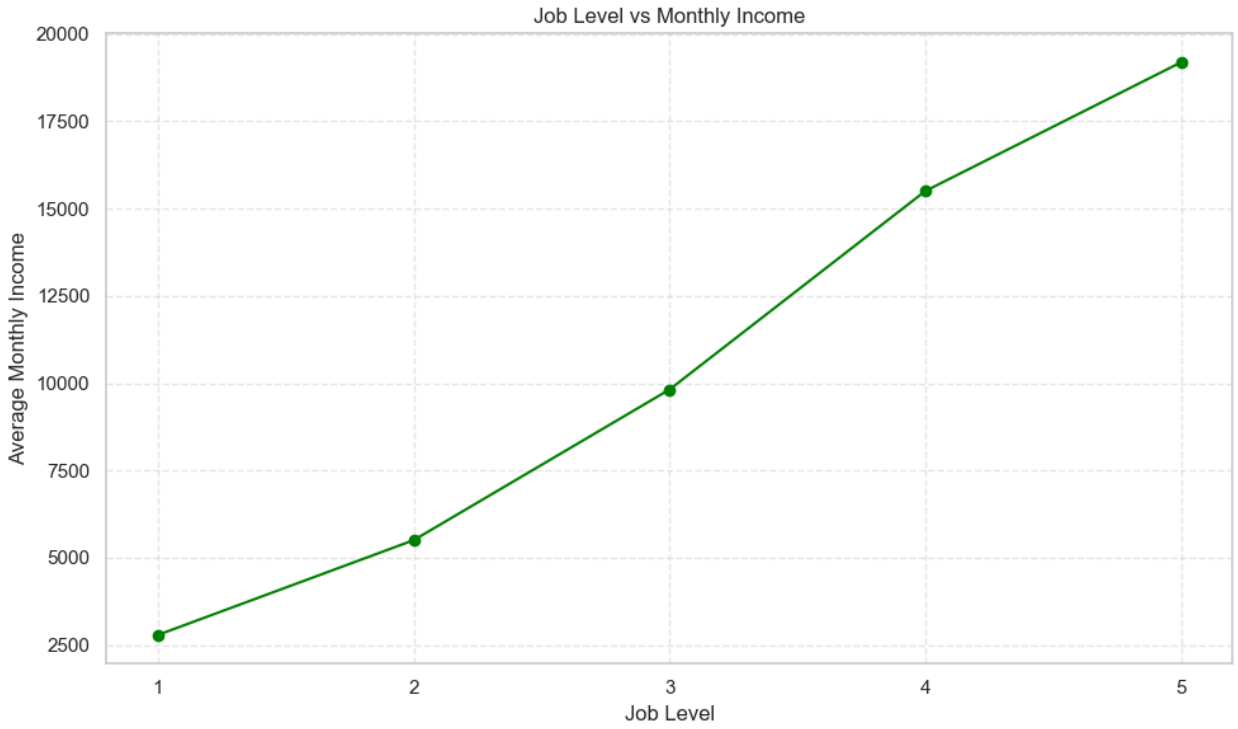
```
[('JobLevel', 'MonthlyIncome'),
 ('JobLevel', 'TotalWorkingYears'),
 ('MonthlyIncome', 'TotalWorkingYears'),
 ('PercentSalaryHike', 'PerformanceRating'),
 ('YearsAtCompany', 'YearsInCurrentRole'),
 ('YearsAtCompany', 'YearsWithCurrManager'),
 ('YearsInCurrentRole', 'YearsWithCurrManager')]
```

After determining these pairs of variables, I went ahead and started creating visualizations with pandas and Matplotlib to identify trends in the dataset.

Analysis Findings

The first visualization I developed was a line chart to plot the upwards trend of an employee's given job level (1-5) and the average monthly income. Of course, correlation does not imply causation, however the given graph shows an almost proportional relation between these two variables (Figure 4).

Figure 4:

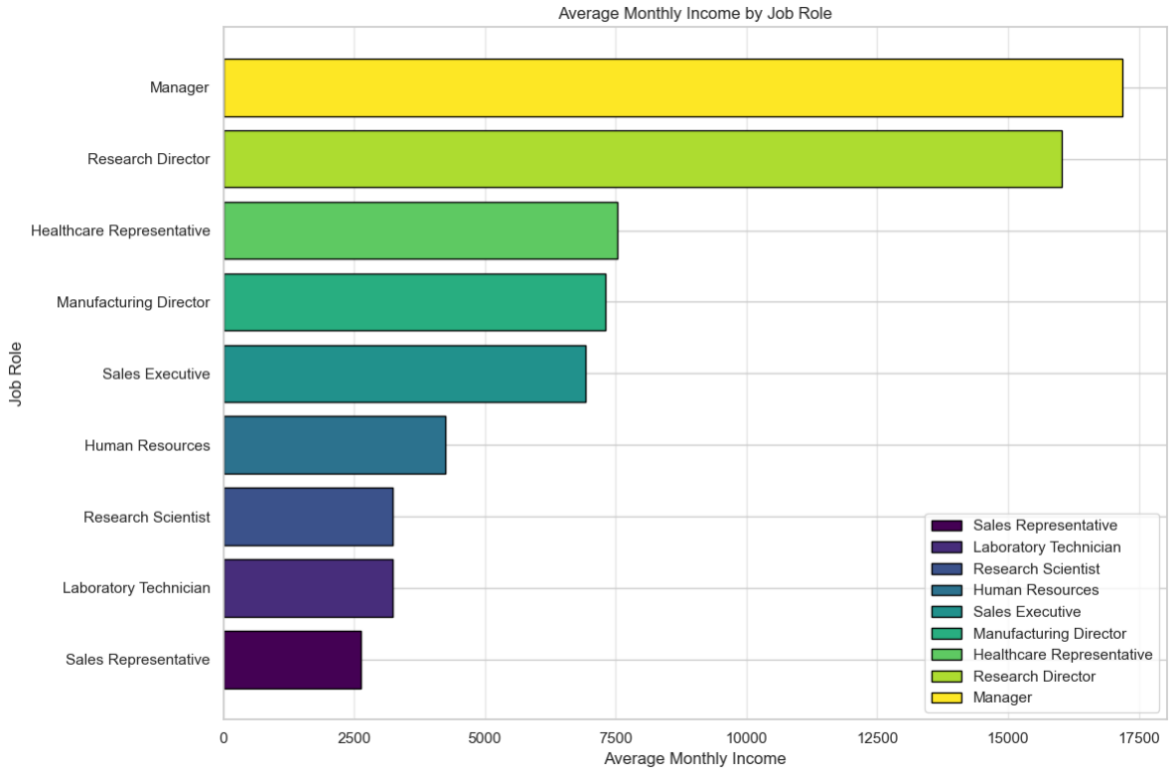


The next study I wanted to understand was the relation between the years an employee has been at the company and the years he or she has been in their current role. To do this, I utilized linear regression to predict our dependent variable, years in current role. However, this statistical model seemed to have scattered data points, indicating a large variance and moderate correlation. The line of best fit showed a positive trend between the two variables (Figure 5).

Figure 5:

Finally, I decided to stray away from the correlations and understand which job functions tend to make more of an average monthly income. For this research study, I utilized Matplotlib to create a horizontal bar chart that lists each of the job roles in the dataset and reflect it with its average monthly income (Figure 6). The given visualization below can be useful for a young adult interested in learning on job outlooks for a given career.

Figure 6:



Discussion

In conclusion, this introductory project allowed me to properly understand the different aspects a Human Resource specialist normally encounters. Understanding employee characteristics can drive better business decisions and allow organizations to cater more employees for their overall well-being in career growth. However, I was able to infer some limitations in this dataset as some values for a given column, take “Department” for example, had an uneven number of records for each of its unique values. Therefore, some of these aggregations created, for example “Average Monthly Income,” were not always as expected. Additionally, one column that could have been useful in this dataset is “State” or “City.” Knowing either of these would have allowed us to make more sense on aggregates like “Average Monthly Income” for a given region in the country.