**Alberto Gell**

**SNHU - DAT430**

**Objective**

The question we are addressing is: What factors are contributing to employee attrition within the organization, and how can we identify patterns in job dissatisfaction and lack of engagement to stem this attrition? By analyzing HR data, I aim to uncover metrics and patterns that help explain why employees are leaving and what factors might worsen attrition.

**Required Data**

We need to gather comprehensive HR data that includes key metrics such as:

* EmployeeNumber
* Job Satisfaction,
* Engagement,
* Department,
* YearsAtCompany,
* Attrition

These variables and other relevant factors will allow us to identify patterns related to attrition, employee dissatisfaction, and the impact of rapid company growth.

**Parameters:**

Appropriate parameters should be set by defining the relevant employee metrics (such as Attrition, JobSatisfaction, JobRoles,) We will also track EmployeeTenure(YearsAtCompany), WorkLifeBalance, and other engagement-related factors to identify trends and relationships between these variables and attrition rates.

**Data Sourcing and Preparation:**

The data was sourced from 15 CSV files, which include HR Training Data and HRData (1)–(14). And prepared by:

* Standardizing employee numbers for consistency across datasets.
* Filled in missing values for the training column
* Standardized the values in the gender column
* Merging all CSV files into a single dataset.
* Filtering and grouping employees by their respective JobRoles
* Saving the final dataset for further analysis.

**Preprocessing:**

The following preprocessing approaches were used:

**Handling missing data**:

The TrainingTimesLastYear column contained several missing values and the training column, while mostly complete, also had missing values. the missing values in TrainingTimesLastYear corresponded to non-missing values in the training column and both columns aligned with the EmployeeNumber, making it clear that these columns needed to be merged to provide a complete training record.

**Handling categorical data**:

Categorical data such as Department and JobRole was handled using string operations and grouping to ensure consistency across the datasets. Due to the inaccuracies in the 'Department' column, employees were grouped by their job roles. This decision was made to address the clear mismatches and to improve the accuracy for future analysis.

**Feature scaling**:

Scaling was not immediately necessary for this dataset but could be implemented if numeric features like salary or tenure require further comparison across employees. We can also utilize numeric features like MonthlyIncome or YearsAtCompany using normalization or standardization to achieve consistent comparisons across those factors.

**Reports**

By creating a dynamic reporting dashboard in PowerBI we can refresh the data source as new data is loaded. As the data changes, visualizations like bar charts, scatter plots, and heatmaps will automatically update to reflect the latest trends in employee attrition. When new data is added, the metrics and visualizations would update to show new patterns or shifts in employee engagement, job satisfaction, or attrition. This helps monitor ongoing trends in real time and identify areas needing immediate attention.

**Data Preparation for Visualizations in Power BI**

In Power BI, I performed additional data preparation steps to ensure the data was more interpretable for visualization purposes. Specifically, I duplicated several columns to create labeled versions of the numerical scales for better readability in charts and reports.

1. **Duplicated Columns**:

I duplicated the following columns: Job Involvement, Job Satisfaction, Performance Rating, and Work-Life Balance.

1. **Data Type Conversion**:

I converted the data type of the duplicated columns from numbers to text. This conversion allowed me to assign meaningful labels to each value in the scale, enhancing the clarity of the visualizations.

1. **Label Creation**:

For each duplicated column, I created labels for the numeric scales as follows:

**Job Involvement** (Scale: 1 to 4):

* + - * 1 = Very Low
      * 2 = Low
      * 3 = High
      * 4 = Very High

**Job Satisfaction** (Scale: 1 to 5):

* + - * 1 = Very Low
      * 2 = Low
      * 3 = Medium
      * 4 = High
      * 5 = Very High

**Performance Rating** (Scale: 1 to 5):

* + - * 1 = Very Low
      * 2 = Low
      * 3 = Medium
      * 4 = High
      * 5 = Very High

**Work-Life Balance** (Scale: 1 to 4):

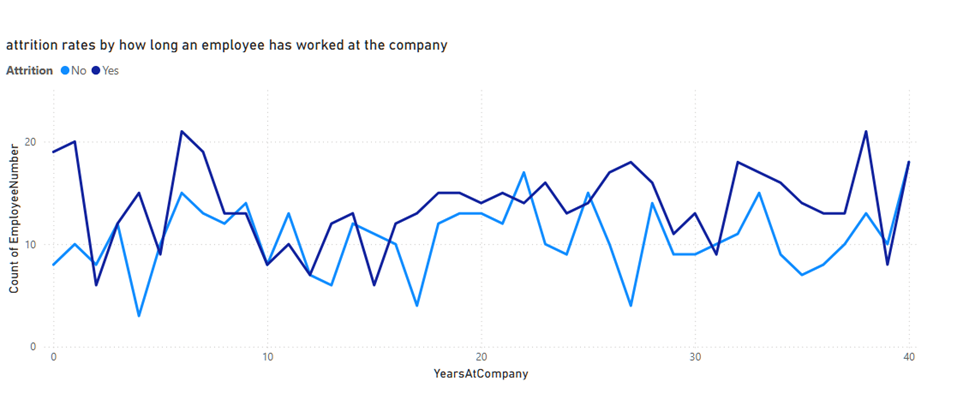
* + - * 1 = Poor
      * 2 = Fair
      * 3 = Good
      * 4 = Excellent

**Decision Making:**

Visual reports (such as bar charts showing Attrition Rates by Department or heatmaps of Job Satisfaction across Tenure) can help stakeholders quickly spot trends and problem areas. Visualizations provide a clearer understanding of the data and can reveal patterns that may not be evident in raw numbers alone. Using PowerBi to create interactive visualizations that allow users to filter by department, job role, and performance. Clear labeling and the use of intuitive visual elements will make these reports easy to interpret.

**Example Of Key Elements in Visualizations:**

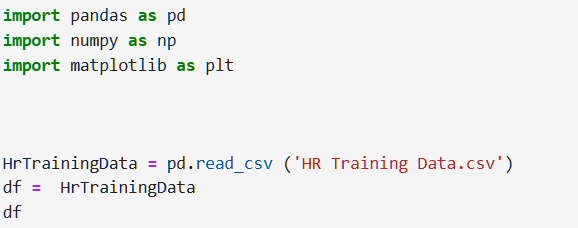
**Attrition by years worked at the company**

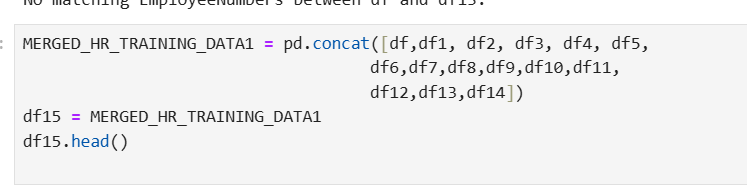
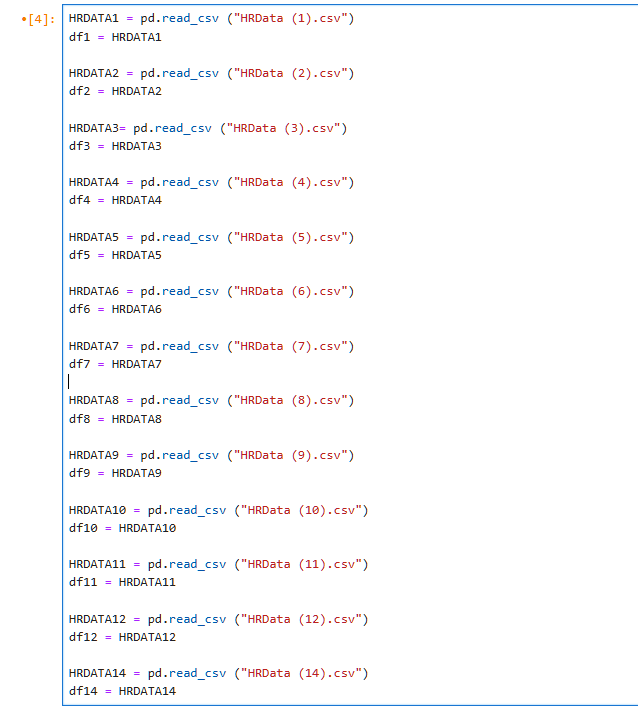


## Part 2

**Merging the Datasets:**

The first step involved merging the 15 datasets, including HR Training Data and HRData (1)–(14). This was achieved by reading each CSV file, standardizing the employee numbers in HRData (13), ensuring no duplicate employee numbers, and then concatenating the data into a single DataFrame.

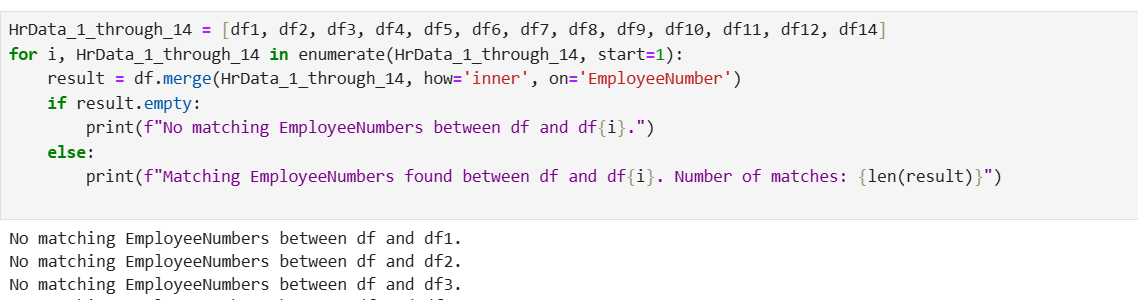




**Preprocessing the Data:**

During preprocessing, the data was cleaned and formatted. The csv file HRData(13) had its employee numbers standardized to a 9-digit format with leading zeros where necessary. I ensured that there were no duplicate employee numbers across any of the datasets.





**Standardizing The Gender Values**

The values for gender were inconsistent within the files, many values were illustrated as either "Male/Female" and "M/F" this was resolved by renaming the values such as, F for Female and M for Male.

A screenshot of a computer

Description automatically generated

**Grouping Employees by Job role:**

Due to the inaccuracies in the 'Department' column, I transitioned to categorizing employees by their job roles rather than departments

A screenshot of a computer

Description automatically generated

**Filling In The Missing Training Data**

the missing values in TrainingTimesLastYear corresponded to non-missing values in the training column and both columns aligned with the EmployeeNumber

A screenshot of a computer program

Description automatically generated

The merged data was saved as an exportable CSV file, and nine additional csv files corresponding to a specific job role which can now be used for further analysis.

