## **Data Science Project: HR Data Analysis – Who's Likely to Quit?**

**Case Study Overview:** Employee retention is a critical factor for organizational success. High employee turnover leads to increased hiring costs, loss of institutional knowledge, and decreased team productivity. Understanding which employees are at risk of quitting (attrition) enables HR departments to implement proactive engagement and retention strategies, thereby preserving talent and maintaining workforce stability.

**Objective:** The objective of this project is to analyze HR data to identify patterns and factors that correlate with employee attrition. This will help HR prioritize retention strategies, reduce hiring costs, and maintain workforce stability by predicting who is likely to quit.

### **Part 1: Data Generation (Python)**

Let's generate a synthetic dataset simulating HR employee data, including factors that might influence attrition.

import pandas as pd

import numpy as np

import random

# Set a seed for reproducibility

np.random.seed(50)

random.seed(50)

# Define parameters for data generation

num\_employees = 2000

departments = ['Sales', 'HR', 'IT', 'Marketing', 'Finance', 'Operations', 'R&D']

job\_roles = {

'Sales': ['Sales Executive', 'Sales Manager'],

'HR': ['HR Generalist', 'HR Manager'],

'IT': ['Software Engineer', 'DevOps Engineer', 'Data Scientist', 'IT Manager'],

'Marketing': ['Marketing Specialist', 'Content Creator'],

'Finance': ['Accountant', 'Financial Analyst'],

'Operations': ['Operations Associate', 'Logistics Manager'],

'R&D': ['Research Scientist', 'Product Developer']

}

education\_levels = ['High School', 'Bachelors', 'Masters', 'PhD']

gender = ['Male', 'Female', 'Other']

performance\_ratings = [1, 2, 3, 4, 5] # 1: Poor, 5: Excellent

data = []

for i in range(num\_employees):

employee\_id = f'EMP{i:04d}'

age = random.randint(22, 60)

dept = random.choice(departments)

role = random.choice(job\_roles[dept])

edu = random.choice(education\_levels)

gen = random.choice(gender)

# Tenure in years

tenure\_years = random.randint(0, age - 20) # Ensure tenure is less than age

# Monthly Income - base + role/dept factor

base\_income = random.uniform(30000, 150000)

if dept == 'IT': base\_income \*= 1.5

if role in ['Manager', 'Director']: base\_income \*= 1.8

monthly\_income = int(base\_income + random.uniform(-10000, 10000))

# Job Satisfaction (1-5, 5 highest)

job\_satisfaction = random.randint(1, 5)

# Work-Life Balance (1-5, 5 best)

work\_life\_balance = random.randint(1, 5)

# Overtime (hours per month)

overtime\_hours = random.randint(0, 40)

if random.random() < 0.2: # 20% chance of higher overtime

overtime\_hours = random.randint(40, 80)

# Performance Rating

performance = random.choice(performance\_ratings)

# Attrition (True/False) - introduce correlations

attrition = False

# Factors increasing attrition likelihood:

if job\_satisfaction < 3:

if random.random() < 0.6: attrition = True # High chance if low satisfaction

if work\_life\_balance < 3:

if random.random() < 0.5: attrition = True # High chance if poor WLB

if overtime\_hours > 40:

if random.random() < 0.4: attrition = True # Increased chance with high overtime

if performance < 3:

if random.random() < 0.3: attrition = True # Some chance if low performance

if monthly\_income < 50000 and tenure\_years < 2:

if random.random() < 0.2: attrition = True # Junior, low pay

# Factors decreasing attrition likelihood:

if job\_satisfaction > 4 and work\_life\_balance > 4:

if random.random() < 0.1: attrition = False # Override if very happy

if monthly\_income > 150000 and tenure\_years > 5:

if random.random() < 0.1: attrition = False # Override if senior, high pay

data.append([

employee\_id, age, gen, dept, role, edu, tenure\_years,

monthly\_income, job\_satisfaction, work\_life\_balance,

overtime\_hours, performance, attrition

])

df = pd.DataFrame(data, columns=[

'employee\_id', 'age', 'gender', 'department', 'job\_role', 'education\_level', 'tenure\_years',

'monthly\_income', 'job\_satisfaction', 'work\_life\_balance',

'overtime\_hours', 'performance\_rating', 'attrition'

])

# Display basic info and head

print("Generated Data Info:")

print(df.info())

print("\nGenerated Data Head:")

print(df.head())

# Save the dataset to a CSV file

df.to\_csv('hr\_data.csv', index=False)

print("\nDataset 'hr\_data.csv' generated successfully!")

### **Part 2: Data Science Tasks for Students**

Students will use the generated hr\_data.csv file to perform the following tasks using Python (Pandas, NumPy, Matplotlib, Seaborn).

**Task 1: Data Loading & Initial Exploration**

* Load the hr\_data.csv into a Pandas DataFrame.
* Display the first 5 rows and check info().
* Identify the number of rows and columns.
* Get describe() statistics for numerical columns.
* Check for any missing values.

**Task 2: Data Cleaning & Preparation**

* **Handle Missing Values:** Address any NaN values (if present in the generated data).
* **Data Type Conversion:** Ensure all columns have appropriate data types (e.g., attrition as boolean, numerical columns as int/float, categorical as object/category).
* **Outlier Detection:** Explore distributions and identify outliers in numerical fields like age, tenure\_years, monthly\_income, overtime\_hours. Discuss how you would handle them.
* **Categorical Variable Handling:** Ensure categorical variables (gender, department, job\_role, education\_level, performance\_rating, job\_satisfaction, work\_life\_balance) are correctly identified and ready for analysis (e.g., consider converting to 'category' dtype).

**Task 3: Feature Engineering**

* Create new features if needed:
  + tenure\_buckets: Categorize tenure\_years into bins (e.g., '0-2 years', '3-5 years', '6-10 years', '10+ years').
  + income\_bins: Categorize monthly\_income into bins (e.g., 'Low', 'Medium', 'High').
  + satisfaction\_level: Group job\_satisfaction into 'Low', 'Medium', 'High'.
  + overtime\_status: Categorize overtime\_hours into 'None', 'Low', 'Medium', 'High'.

**Task 4: Exploratory Data Analysis (EDA) & Visualization**

* **Overall Attrition Rate:**
  + Calculate the overall attrition rate for the company.
* **Attrition by Demographics:**
  + Visualize attrition rates by gender, age (e.g., age groups), and education\_level using bar charts.
* **Attrition by Department & Role:**
  + Calculate and visualize attrition rates by department and job\_role. Which departments/roles show higher attrition rates?
* **Attrition by Work Factors:**
  + Analyze and visualize attrition rates based on job\_satisfaction, work\_life\_balance, and overtime\_hours. Do low satisfaction or high overtime correlate with leaving?
  + Use bar charts or boxplots to show these relationships.
* **Attrition by Performance & Income:**
  + Analyze and visualize attrition rates by performance\_rating and monthly\_income (or income\_bins). Is there a relationship between salary/performance and attrition?
* **Correlation Analysis (Optional):**
  + Create a heatmap to visualize correlations between numerical features (including a numerical representation of attrition if you map True/False to 1/0) and other relevant variables.

**Task 5: Key Findings & Business Recommendations**

* Present at least **three key insights** related to attrition risks based on your EDA. For example:
  + "Employees with low job satisfaction (ratings 1-2) and high overtime (40+ hours/month) are twice as likely to quit compared to their counterparts."
  + "Department X and Department Y have attrition rates 30% higher than the company average, indicating specific issues within these teams."
  + "There's a noticeable correlation between lower monthly income within the first 2 years of tenure and higher attrition, suggesting early career compensation might be a factor."
* Based on your findings, provide actionable recommendations for HR:
  + Improve work-life balance initiatives for employees consistently working high overtime.
  + Target retention programs and leadership training in high-risk departments.
  + Conduct salary reviews or implement performance-based incentives where income correlates with attrition risk, especially for early-career employees.
  + Implement regular job satisfaction surveys to proactively identify at-risk employees.
* Summarize your key takeaways and explain how these predictive insights can support strategic HR decisions to reduce turnover and preserve talent.

### **Part 3: Optional Visualization using Power BI**

For students interested in Business Intelligence tools, you can also use Power BI to create an interactive dashboard based on the hr\_data.csv dataset.

**Suggested Power BI Tasks:**

* Import the hr\_data.csv into Power BI.
* Create calculated columns and measures (using DAX) such as Attrition Rate, Total Employees, Employees Attrited.
* Design an interactive dashboard with the following visualizations:
  + KPIs: Total Employees, Total Attrited Employees, Overall Attrition Rate.
  + Bar chart: Attrition Rate by Department.
  + Bar chart: Attrition Rate by Job Role.
  + Bar chart: Attrition Rate by Job Satisfaction Level.
  + Bar chart: Attrition Rate by Overtime Hours (or categories).
  + Table/Matrix: Average Monthly Income and Attrition Rate by Tenure Bucket.
  + Slicers for department, job\_role, gender, education\_level, and performance\_rating.