Project Description: The Synthetic Employee Attrition Dataset is a simulated dataset designed for the analysis and prediction of employee attrition. It contains detailed information about various aspects of an employee's profile, including demographics, job-related features, and personal circumstances.

The dataset comprises samples, split into training and testing sets to facilitate model development and evaluation. Each record includes a unique Employee ID and features that influence employee attrition. The goal is to understand the factors contributing to attrition and develop predictive models to identify at-risk employees.

This dataset is ideal for HR analytics, machine learning model development, and demonstrating advanced data analysis techniques. It provides a comprehensive and realistic view of the factors affecting employee retention, making it a valuable resource for researchers and practitioners in the field of human resources and organizational development.

Project objectives

Project Objectives Summary:

1. Employee Performance Analysis

Key Columns: Performance Rating, Overtime, Job Satisfaction, Work-Life Balance.

2. Diversity & Inclusion Insights:

Key Columns: Gender, Education Level, Marital Status, Job Level

Focus: Gender/education distribution. Data Ready: Zero nulls in critical fields.

3. Compensation Trends:

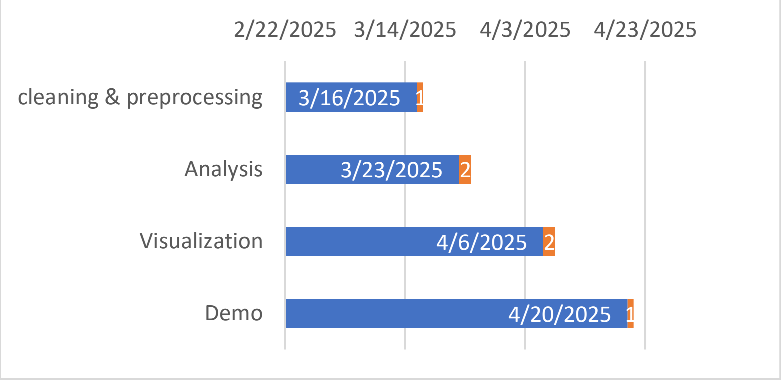
Key Columns: Monthly Income, Job Role, Number of Promotions

Focus: Salary fairness. Alert: 3k missing income, 3.5k missing promotions.

Milestones and deadlines:

- Data Cleaning & Preparation – Handling missing values, duplicates, and ensuring consistency.

- Exploratory Data Analysis (EDA) – Identifying patterns, trends, performance trends and anomalies.
- Data Visualization & Reporting – Creating dashboards and reports for actionable insights.
- Demo: This milestone will be for reviewing all we have done on this project.



Team leader: Sara Ahmed

Team Members and roles:

- Sara Ahmed
- Habiba Mohamed
- Salma Araby
- Amina Osama
- Marwa Ashraf
- Mariam Mohamed

We are following an Agile approach, where the entire team collaborates on the same milestone, this ensures flexibility and continuous improvement

Tools and technologies:

- Excel: it's used for storing data in structured ways (rows & columns) and for applying normalization on the dataset and breaking the table into number of tables.
- SQL: this tool is used in our project to analyze data and extract meaningful information through it. It helps to show the relationship and dependency between features of our dataset.
- Python: it is a tool for cleaning data and applying preprocessing on it using its useful libraries like: NumPy, pandas, matplotlib and seaborn.

- Tableau and Power BI: these tools are both used for visualizing the insights of data, making dashboards, helps in decision making and generate reports for stakeholders.

Key performance indicators KPIs

Data Cleaning & Processing:

Removed Duplicates:

- removed 3359 duplicate records to ensure data accuracy and prevent redundant calculations.

Handled Missing Values (Nulls):

- Years at Company → 2396 missing values.
- Monthly Income →2994 missing values.
- Number of Promotions →3598 missing values.
- Remote Work →2101missing values.

- Imputation: Fill missing numerical values (e.g., Monthly Income) with median/mean.
- Investigation: Determine if missing Remote Work values correlate with other fields (e.g., job roles).
- Drop or Flag: Consider removing rows with high missingness in critical columns (e.g., Number of Promotions).
- Formatting &Adjustments:

- Generated random Hire date that is compatible with age of employees
- Converted Hire Date to a standard date format (MM-DD-YYYY) for consistency in time-based analysis.

- Generated Years at Company column that is compatible with hire year of employee.
- Standardized categorical value of gender to avoid duplication due to inconsistent naming.

  Male →M
  Female →F

Scaling to Numerical Data:

- Normalized numerical values for monthly salary using Min-Max Scaling ensure consistency and improve analytical accuracy.

Analysis & insights:

- **Key Analytical Questions**

  Based on the HR dataset and business objectives, the following critical questions were explored:

- **Basic Exploratory Questions:**

1. What is gender distribution across the entire company?
   → *Measures workforce diversity.*

2. **Percentage of Employees with Dependents by Gender**
   → *Supports family support policies and benefits planning.*
3. **What is the average monthly income by job role?**
   → *Measures market competitiveness across roles.*
4. **How many employees have a job satisfaction score below 3?**
   → *Measures potential disengagement.*
5. **What percentage of employees work overtime?**
   → *Identifies work-life balance issues.*
6. **What is the average number of promotions per job level?**
   → *Measures career growth opportunities.*
7. **What is the distribution of employees by education level?**
   → *Maps educational background of workforce.*
8. **How is leadership opportunity distributed by job rule?**
   → *Tracks inclusivity and leadership access.*
9. **What is the average company reputation rating across job roles?**
   → *Gauges internal perception of external brand.*

- **Medium Questions:**

10. **How does job satisfaction vary by job role and gender?**
    → *Measures engagement and inclusivity.*
11. **What is Recognition Level by Performance Rating**
    → *Evaluates fairness and alignment in employee recognition.*
12. **What is the Attrition rate among employees with low performance ratings?**
    → *Measures performance-impact on attrition.*
13. **What's the average tenure of employees at each job level?**
    → *Measures retention by seniority.*
14. **Do employees with remote work have different average satisfaction scores?**
    → *Measures effectiveness of hybrid/remote policies.*
15. **What is the median salary for each education level?**
    → *Measures fairness and competitiveness.*
16. **Correlation between Years of Experience and Monthly Income**
    → *Tests alignment between experience and compensation.*
17. **Do employees with more Years of Experience have higher performance ratings?**
    → *Measures effectiveness of experience on performance*
18. **What is the overtime distribution across job roles?**
    → *Measures workload balance.*
19. **Is there a link between distance from home and job satisfaction?**
    → *Measures impact of commuting on engagement.*

- **Advanced Questions:**

20. **What is the Percentage of high performers who have leadership opportunities**
    → *Measures whether leadership roles are being given to top talent.*
21. **What is the Percentage of high performers with innovation opportunities**
    → *Measures whether creative talent is being engaged effectively.*
22. **Do higher-paid employees report higher job satisfaction and Work-Life Balance?**
    → *Correlates well-being and compensation.*
23. **Are women equally represented in higher job levels across departments?**
    → *Measures gender equality in advancement.*

24. **What is the average Job Satisfaction of employees with vs. without dependents**
→ Supports family support policies and flexible work structures.
25. **Which job role has the highest turnover (Attrition) among top performers?**
→ Identifies critical retention risks.

- **Insights**

The following insights reveal how various factors relate to one another and their impact on performance, satisfaction, and retention.

Employee ID → Unique identifier (primary key) used to track employees.

Education Level & Job Role → Helps understand how academic background impacts job level, income, and promotion frequency.

Years of experience & Company Tenure → Reveal patterns in retention, career progression, and promotion cycles.

Job Level & Job Role → Provide structure for analyzing career stages and seniority across departments.

Performance Rating → Core measure of employee effectiveness.

Work-Life Balance & Overtime → Quantify workload intensity and help assess stress, satisfaction, and potential burnout.

Number of Promotions → Measures career mobility, employee motivation, and resignation risk.

Job Satisfaction, Work-Life Balance, Manager Support (not present), Employee Recognition, Leadership Opportunities → Behavioral KPIs tied to retention, engagement, and performance.

Remote Work → Used to assess modern work preferences, hybrid work effectiveness, and its effect on performance and resignation rates.

Distance from Home → Helps analyze the impact of commute time on job satisfaction, attendance, and attrition.

Attrition → Used to evaluate how different features (satisfaction, income, training, etc.) contribute to turnover risk.

Company Reputation & Innovation Opportunities → Help assess how external and internal perception of the company influences employee satisfaction and retention.

Company Size → Allows segmentation by business scale to explore its influence on promotion rates, satisfaction, and engagement.

Gender, Marital Status, and Number of Dependents → Enable demographic-based segmentation to uncover patterns in attrition, satisfaction, and financial strain.

- **Analysis to be performed**

- Correlation between Employee Satisfaction & Resignation
- Workload vs. Job Satisfaction Regression

- Work-Life Balance & Remote Work Regression

***Visualization & Reporting*:**

- Visualizations help in decision-making
  - transforming raw data into meaningful insights.
  - allow HR teams to identify trends, detect patterns, and make data-driven decisions.
- Creating interactive dashboards
  - provide a real-time view of key workforce metrics such as attrition rates, salary distributions, performance trends, and employee satisfaction.
  - help stakeholders quickly analyze data, compare metrics, and take proactive actions to improve workforce management.

***Forecasting and predicting*:**

We will apply machine learning models on our project to predict values in future based on given information by splitting data into training set and test set then apply the model on the training set to enable the model to predict values on test set

First, Regression Analysis for Salary Prediction

Target Variable (Dependent):

- Monthly Income

Available Predictors (Independent Variables):

1. Years at Company

2. Education Level

3. Overtime

4. Job Satisfaction

5. Number of Promotions

6. Job Level

7. Job Role

8. Performance Rating

Data Preparation Required:

1. Impute missing values in:

   - Monthly Income (29.9% missing)

   - Number of Promotions (35.7% missing)

   - Years at Company (24.0% missing)

2. Encode categorical variables:

   - Job Role, Education Level, Marital Status

Final document & presentation:

To be continued.