DQLab

Live Class

# Data Analysis for Employee Attrition Prediction

**Bootcamp Data Analyst with SQL & Python using Google Platform**

**Muammar Nurdin**

# Introduction

## Background

The high rate of employee attrition or turnover poses a significant challenge for the Human Resources division, as it affects productivity, recruitment costs, training expenses, and the loss of institutional knowledge. To understand the factors influencing employees' tendency to leave, a data-driven analytical approach is necessary.

In this context, analysis is conducted with the help of data visualization to intuitively illustrate patterns and relationships between variables. Employee characteristics such as age, education, job position, income, job satisfaction, and tenure are the main focus of this exploration.

DQLab

live Class

# Analysis Objectives

1. Presenting an overview of employee conditions based on various demographic and job-related aspects through data visualizations.
2. Identifying variables that have a significant correlation with attrition tendencies.
3. Providing data-driven visual recommendations to management and HR for designing more targeted employee retention strategies.

DQLab

live Class

# DATASET

## Dataset Description

This dataset is a collection of employee attrition predictions, comprising information from 10,000 individuals. It includes various demographic metrics, job characteristics, and employee performance indicators that can be used to analyze the factors influencing employee turnover rates. The dataset was obtained from a source on Kaggle (https://www.kaggle.com/datasets/ziya07/employee-attrition-prediction-dataset/data).

- Attrition
- Hourly Rate
- Attrition Rate %
- Overtime
- Absenteeism
- Job Involvement
- Years at Company
- Performance Rating
- Project Count
- Work Life Balance
- Job Satisfaction
- Training Hours Last Year
- Distance From Home
- Relationship with Manager
- Years in Current Role
- Average Hours Worked per Week
- Number of Companies Worked
- Years Since Last Promotion
- Work Environment Satisfaction

# Tools and Library

Google
Looker Studio

Google Collab

Python
Programming
Language

Pandas

Numpy

Matplotlip

Seaborn

# Data Cleaning

```python
# Tampilkan baris duplikat
duplicates = df[df.duplicated()]
print(duplicates)

# Tampilkan baris yang memiliki nilai kosong
missing_data = df[df.isnull().any(axis=1)]
print(missing_data)

# Melihat informasi pada data
df.info()
```

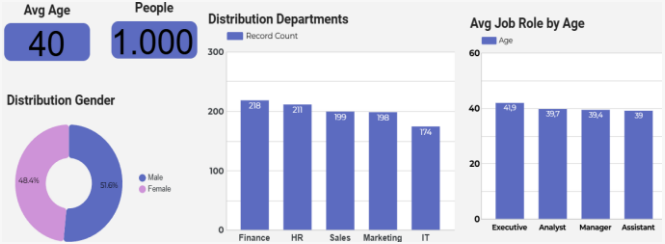
```
[0 rows x 26 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Employee_ID                     1000 non-null   int64
 1   Age                             1000 non-null   int64
 2   Gender                          1000 non-null   object
 3   Marital_Status                  1000 non-null   object
 4   Department                      1000 non-null   object
 5   Job_Role                        1000 non-null   object
 6   Job_Level                       1000 non-null   int64
 7   Monthly_Income                  1000 non-null   int64
 8   Hourly_Rate                     1000 non-null   int64
 9   Years_at_Company                1000 non-null   int64
 10  Years_in_Current_Role           1000 non-null   int64
 11  Years_Since_Last_Promotion      1000 non-null   int64
 12  Work_Life_Balance               1000 non-null   int64
 13  Job_Satisfaction                1000 non-null   int64
 14  Performance_Rating              1000 non-null   int64
 15  Training_Hours_Last_Year        1000 non-null   int64
 16  Overtime                        1000 non-null   object
 17  Project_Count                   1000 non-null   int64
 18  Average_Hours_Worked_Per_Week   1000 non-null   int64
 19  Absenteeism                     1000 non-null   int64
 20  Work_Environment_Satisfaction   1000 non-null   int64
 21  Relationship_with_Manager       1000 non-null   int64
 22  Job_Involvement                 1000 non-null   int64
 23  Distance_From_Home              1000 non-null   int64
 24  Number_of_Companies_Worked      1000 non-null   int64
 25  Attrition                       1000 non-null   int64
dtypes: int64(21), object(5)
memory usage: 203.3+ KB
```

Following a comprehensive data cleaning process, the dataset was confirmed to contain no missing values, duplicate records, or data type inconsistencies, consisting of 1,000 rows and 26 columns.

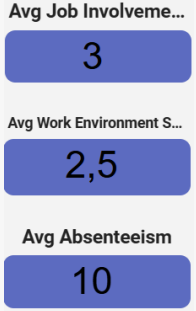
# Exploratory Data Analysis

## Exploratory Dashboard

**Demographic Metrics**

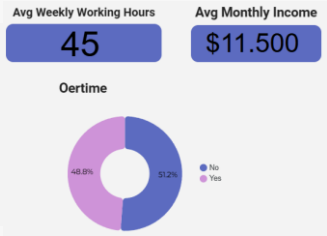



**Engagement & Satisfaction**

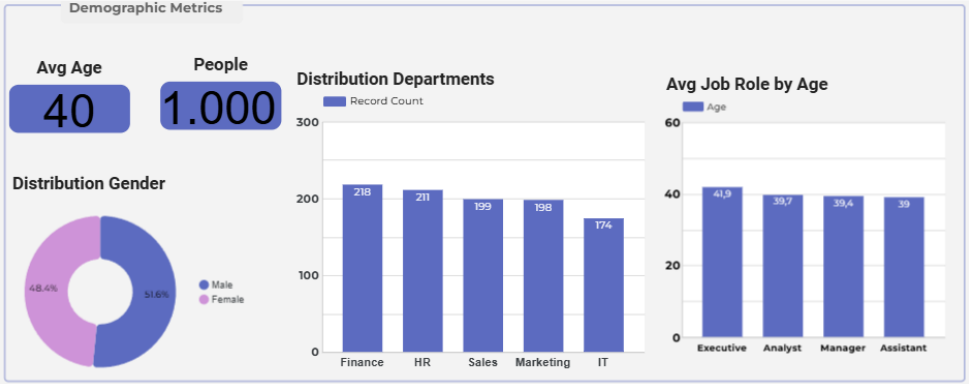



**Attrition**


**Work Performance**



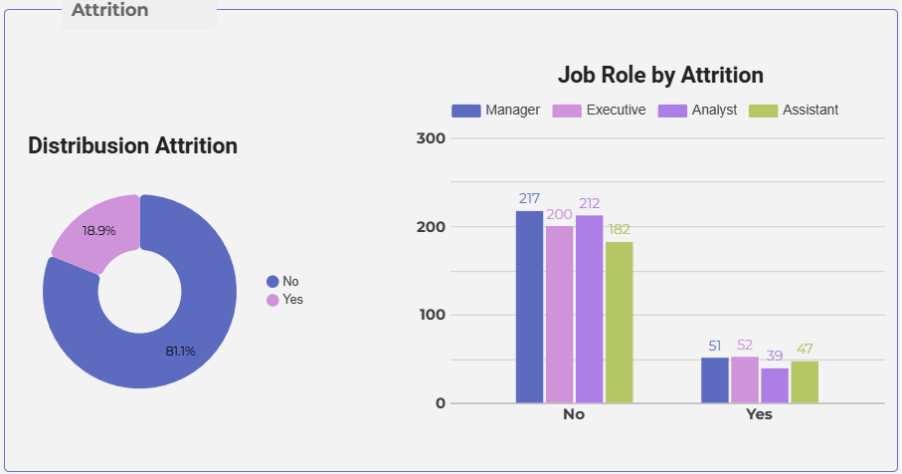
# Dashboard Demographic




## Insight

- The dataset records 1,000 individuals.
- The average age of 40 indicates that the workforce consists mostly of experienced adults.
- Gender distribution is nearly equal, suggesting balanced representation.
- Finance is the most represented department, while IT has the fewest employees indicating a potential imbalance across departments.
- Executives have the highest average age among job roles, implying seniority aligns with age.

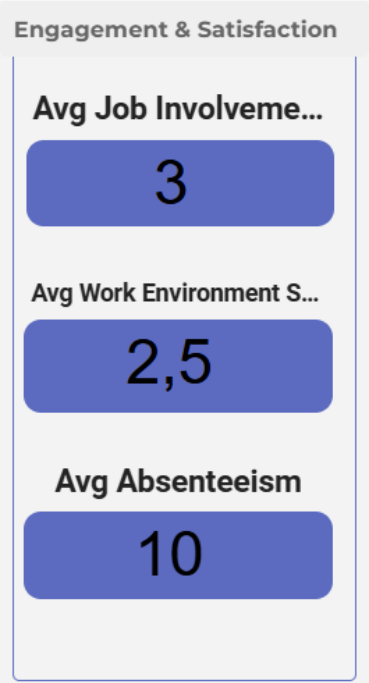
# Dashboard Attrition




**Insight**

- A total of 18.9% of employees in the dataset have left the company.
- Across all job roles, the number of employees who stayed is significantly higher than those who left.
- Among those who resigned, **Executive** and **Manager** positions have the highest attrition counts (52 and 51 respectively).
- The **Assistant** role, while having the lowest total headcount, also shows considerable attrition, with 47 people leaving.

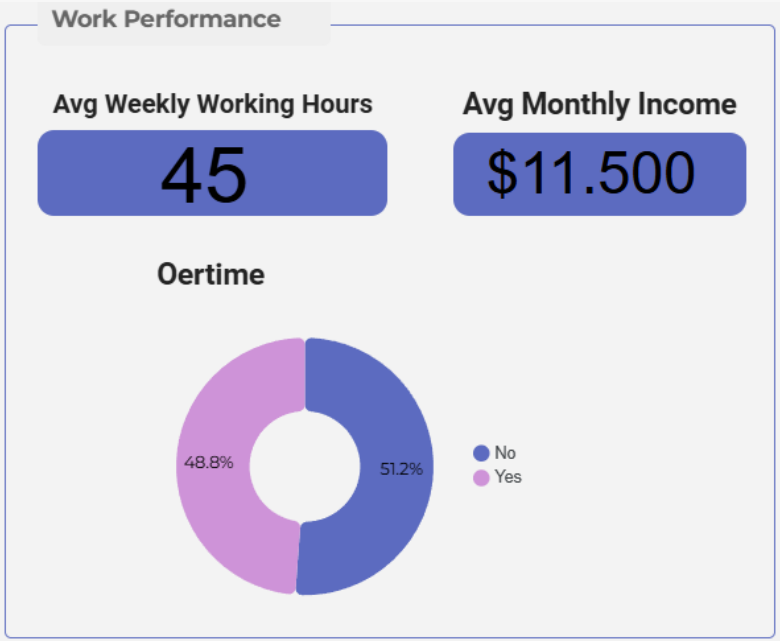
# Dashboard Engagement & Satisfaction




**Insight**

- The **average job involvement score** is 3, which indicates a moderate level of employee engagement with their work responsibilities.
- The **average work environment satisfaction** is 2.5, suggesting that employees tend to have a neutral to slightly low perception of their working conditions.
- The **average absenteeism** is 10, meaning employees are, on average, absent for 10 days. This could signal disengagement, dissatisfaction, or other external factors affecting attendance.

# Dashboard Work Performance




**Insight**

- The **average weekly working hours** is **45 hours**, which indicates that employees are working slightly above the standard 40-hour workweek.
- The **average monthly income** is **$11,500**, suggesting that the organization provides relatively high compensation.
- **Overtime distribution** is nearly balanced, with **48.8% of employees working overtime** and **51.2% not working overtime**, showing that workload pressure is shared fairly equally among employees.

DQLab

live Class

# Correlation Analysis

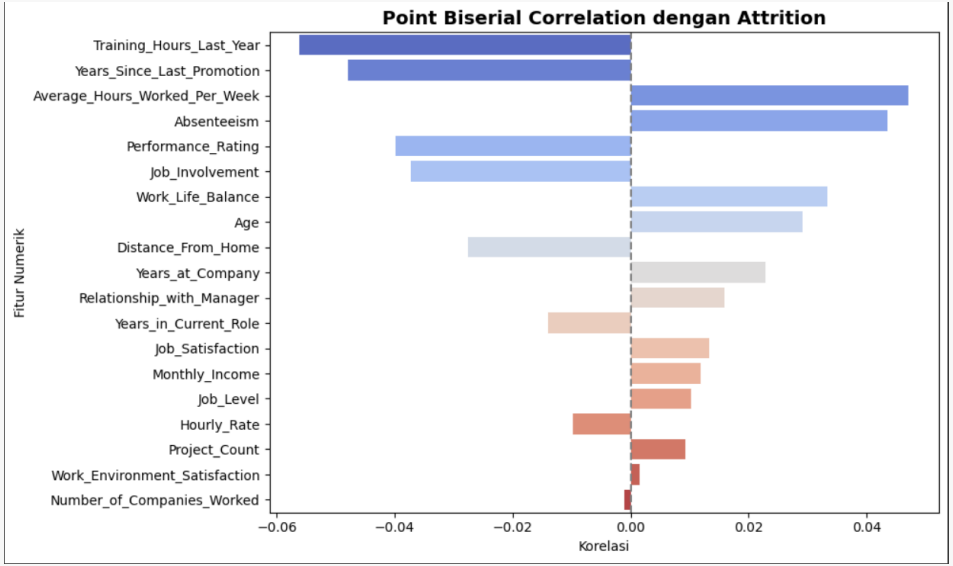
## Point Biserial Correlation

The Point Biserial Correlation is used to measure the relationship between a numerical variable and a binary variable (with two categories). In this context, it is applied to examine the relationship between numerical features and the target variable Attrition (whether an employee leaves the company or not).

| | |
|---|---|
| 0.9 – 1 | Very strong |
| 0.7 – 0.9 | Strong |
| 0.4 – 0.6 | Moderate |
| 0.1 – 0.3 | Weak |

A positive correlation indicates that the higher the value of the numerical feature, the greater the likelihood of the employee leaving.
A negative correlation indicates that the higher the value of the numerical feature, the lower the likelihood of the employee leaving.

# Point Biserial Correlation for Numeric Columns




## Insight

- A moderate positive correlation between **Average Hours Worked Per Week** and **attrition** suggests that employees who work longer hours are more likely to leave the company.
- A moderate positive correlation between **Absenteeism** and **attrition** may indicate that frequent absences are an early signal of disengagement or dissatisfaction.
- A moderate negative correlation between **Training Hours Last Year** and **attrition** suggests that employees who receive more training are less likely to leave.
- A moderate negative correlation between **Years Since Last Promotion** and **attrition** implies that lack of career advancement is associated with higher turnover.
- A weak negative correlation between **Job Satisfaction** and **attrition** reflects that lower job satisfaction may slightly contribute to employee departure

DQLab

Live Class

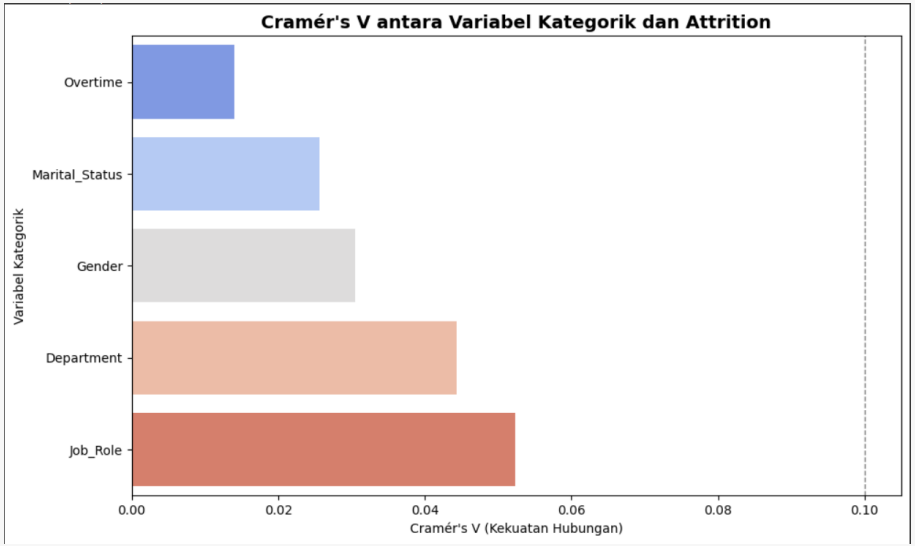
# Categorical Correlation Analysis

## Cramér's V

Cramér's V is used to measure the strength of the association between two categorical (nominal) variables. In this context, it is used to assess the relationship between categorical features and the target variable Attrition (whether an employee leaves the company or not).

Values closer to 0 indicate a very weak or no association between the variables, while values closer to 1 indicate a stronger relationship.

# Cramér's V for Categorical Columns




**Insight**

- **Job Role** shows the strongest association compared to other variables, although the value is still very weak (around 0.05).
- **Department** has a slightly higher association strength than Gender, Marital Status, and Overtime.
- **Overtime**, **Marital Status**, and **Gender** have Cramér's V values close to zero, indicating almost no association between these variables and the likelihood of employees leaving.
- Overall, these results suggest that the categorical factors in this dataset are not strong predictors of Attrition.

DQLab

live Class

# Research Findings

This data analysis aims to examine the influence of demographic factors, job engagement, and employee performance on attrition likelihood.

**Demographic & Attrition** →

This analysis examines the influence of demographic characteristics including age, gender, marital status, and education level on employee turnover.
It investigates whether specific demographic groups are more prone to attrition; however, the findings reveal only a very weak correlation.

**Engagement & Satisfaction** →

How Job Engagement, Work Environment, and Absenteeism Influence Employees' Decision to Stay or Leave.
Findings indicate that high absenteeism, low job satisfaction, and a lack of training and promotion opportunities are key factors that increase the likelihood of attrition.

**Work Performance** →

The Relationship Between Weekly Working Hours, Monthly Income, and Overtime Habits with Employee Attrition Decisions.
High working hours and income do not guarantee retention if engagement and growth are lacking.

# Research Findings

## Correlation Analysis Findings:

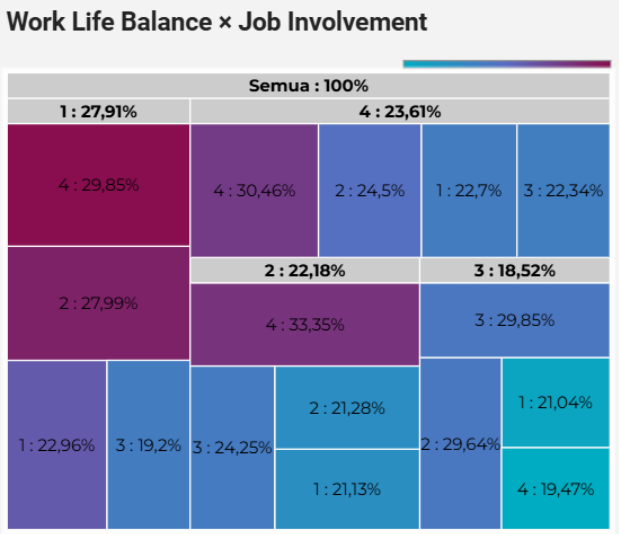
### Numerical Variables (Point Biserial Correlation)

No strong correlations were found, but weak patterns still provide valuable insights for shaping employee retention strategies.

### Categorical Variables (Cramér's V)

Despite generally low correlation strengths, factors such as overtime habits and job roles remain key considerations in retention management.

# Satisfaction and Engagement - 1
# Work Life Balance × Job Involvement

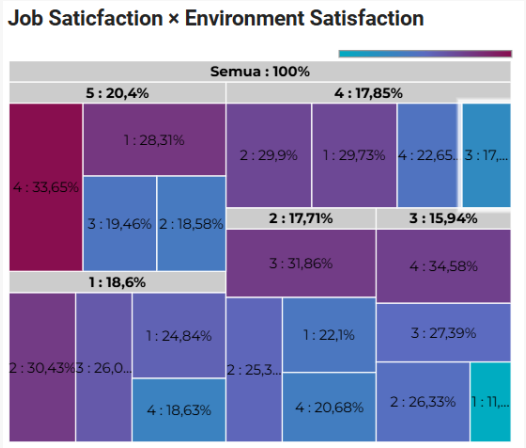



| | Work_Life_Balance | Job_Involvement | Attrition Rate (%) |
|---|---|---|---|
| 1. | 4 | 1 | 0,28 |
| 2. | 2 | 1 | 0,26 |
| 3. | 4 | 2 | 0,24 |
| 4. | 4 | 4 | 0,24 |
| 5. | 1 | 1 | 0,21 |
| 6. | 2 | 4 | 0,19 |

1 - 16 / 16

Low **Work Life Balance** and low **Job Involvement** show higher **attrition** (~28%). High balance and involvement reduce attrition (~19%).

## Satisfaction and Engagement - 2
# Job Satisfaction × Environment Satisfaction




| | Job Satisfaction | Environment Satisfaction | Attrition Rate (%) |
|---|---|---|---|
| 1. | 5 | 4 | 0,28 |
| 2. | 5 | 1 | 0,24 |
| 3. | 1 | 2 | 0,23 |
| 4. | 2 | 3 | 0,23 |
| 5. | 3 | 4 | 0,23 |
| 6. | 4 | 2 | 0,22 |

1 - 20 / 20

Very high **Job Satisfaction** with high **Environment Satisfaction** shows the highest **attrition** (28%). Moderate–high satisfaction combinations have lower attrition (~20–23%).

# Strategic Recommendations for HR

**Strategic Overtime Management >**

Insight : Overtime shows the strongest correlation with attrition (Cramér's V).
Recommendation : Tighten overtime monitoring. Ensure work-life balance through leave, compensation, or fair policies.

**Prioritize Balance and Engagement >**

Insight : Low Work-Life Balance + Low Job Involvement increases attrition risk by up to 28%. High levels of both reduce attrition by up to 19%.
Recommendation : Enhance employee engagement through self-development programs, regular feedback, and flexible work arrangements to support work-life balance.

**Job Satisfaction Pattern Analysis >**

Insight : Satisfied employees may feel confident to switch jobs—not due to discomfort, but because they feel ready to level up.
Recommendation : Establish a clear career path and accelerate internal promotions to retain high-satisfaction employees.

DQLab

Live Class

**Focus on Retaining Senior & Experienced Talent >**

Insight : Variables such as Years at Company, Total Working Years, and Monthly Income show a negative correlation with attrition (Point Biserial Correlation).
Recommendation : Tighten overtime monitoring. Ensure work-life balance through leave, compensation, or fair policies.

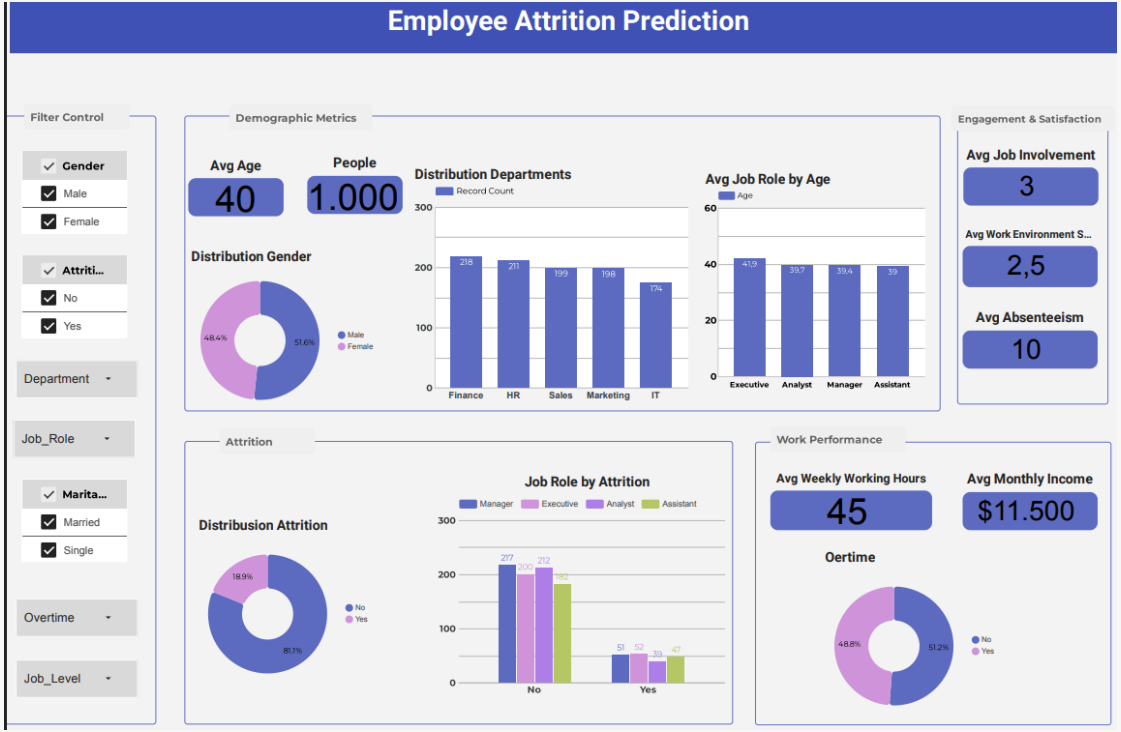
**Impact of Demographic and Commute Factors>**

Insight : Distance from home and frequent business travel contribute to higher attrition rates.
Recommendation : Consider hybrid work options or satellite office placements for employees with long commutes or frequent business trips.

**Targeted Interventions Based on Job Role >**

Insight : Certain roles, such as Sales Executive and Laboratory Technician, show higher-than-average attrition rates.
Recommendation : Implement personalized approaches for high-turnover roles—such as mentoring, job rotation, or workload adjustments.

Attrition is not always driven by dissatisfaction; it can also result from low engagement, overtime-related burnout, or poor work-life balance. Effective retention strategies require a holistic and personalized approach — focusing not only on compensation, but also on employees' daily work experiences.




Appendix Dashboard

DQLab

Live Class

# Thank you!


muammarnurdin28@gmail.com

linkedin.muammarnurdin