**Applied Probabilistic Machine Learning on HR Analytics Dataset**

**INTRODUCTION Problem Framing:**

* **The classification problem is to predict Attrition using various job and personal features.**
* **The regression problem is to estimate JobSatisfaction or MonthlyIncome as a function of predictors.**
* **The clustering problem is to discover hidden groupings in the employee base using Gaussian Mixture Models.**

**Existing Analyses vs. Our Approach:  
While several studies use standard machine learning models like decision trees or logistic regression, we emphasize:**

* **Uncertainty quantification**
* **Priors based on domain knowledge**
* **Bayesian model comparison**
* **Convergence diagnostics (e.g., R-hat, trace plots)**

**This adds interpretability and confidence bounds to the HR analytics process.**

**Motivation:** Employee attrition is a critical issue for organizations, impacting productivity, morale, and financial health. Understanding and addressing employee attrition factors is crucial for HR planning and retaining talent.

* Employee attrition is costly and often influenced by hidden variables like job satisfaction, performance scores, or work-life balance. Traditional models may provide point estimates, but probabilistic approaches offer a **distributional understanding** of outcomes, enabling organizations to quantify uncertainty in predictions. Identifying employees likely to leave can help organizations take proactive measures.

**Problem Statement:** Predict employee attrition, understand contributing factors, and identify latent groupings using probabilistic models.

Can we identify key drivers of employee attrition and satisfaction using Bayesian methods? Can probabilistic models offer interpretable and uncertainty-aware predictions for HR planning?

**Main Modelling Ideas:** We explore three main analysis tasks:

* + Classification: Predict if an employee will leave the company.
  + Regression: Predict Monthly Income/ Estimate employee satisfaction based on performance and work-related metrics.
  + Clustering: Discover latent employee groups based on satisfaction and performance using Bayesian Gaussian Mixture Models.

Or Discover latent segments within the employee base (e.g., dissatisfied vs. satisfied clusters).

These problems will be solved using **probabilistic machine learning techniques**, which are useful because they give us not just predictions, but also the uncertainty behind those predictions.

**Data Description & Problem Analysis**

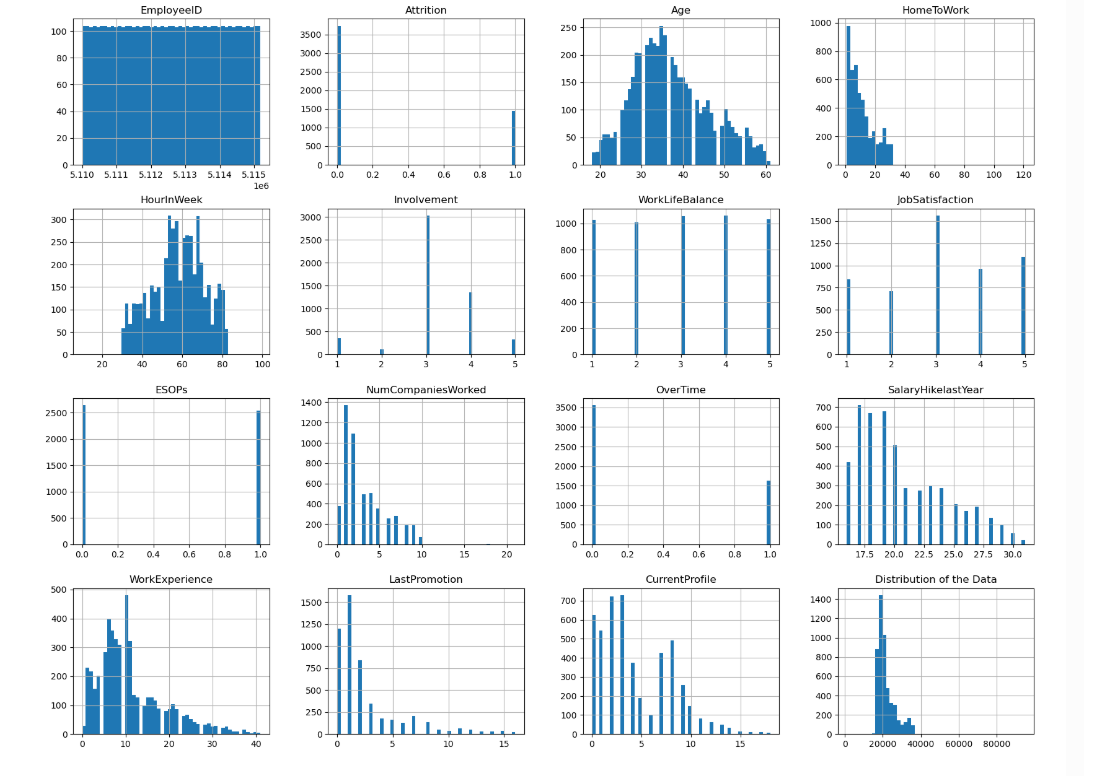
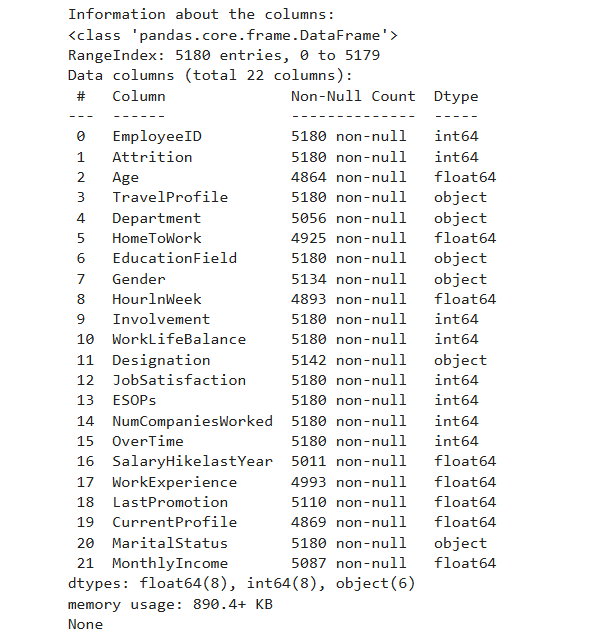
**2.1 Dataset Description:** The dataset collected in this study is a publicly available (on Kaggle) HR analytics dataset in .CSV format. It consists of **5,180 employee records** with **22 attributes**/ **features** encompassing both numerical and categorical variables. The dataset captures employee demographics, job role-related information, performance metrics, satisfaction metrics, and compensation details.

Some key variables include:

* **Attrition** *(target for classification)*: Binary variable indicating if the employee has left the job (1 = Yes, 0 = No).
* **MonthlyIncome** *(target for regression)*: Continuous variable representing employee’s salary.
* **WorkLifeBalance**, **Involvement**, **JobSatisfaction**: Scales measuring subjective employee experience.
* **OverTime, MaritalStatus, Department –** These are categories that describe the employee’s work schedule, personal life, and job department.

Out of the 22 columns:

* **16 are numerical** (like age, salary, experience).
* **6 are categorical** (like gender, department, marital status).



**Problem Framing:**

* The **classification problem** is to predict Attrition using various job and personal features.
* The **regression problem** is to estimate JobSatisfaction or MonthlyIncome as a function of predictors.
* The **clustering problem** is to discover hidden groupings in the employee base using Gaussian Mixture Models.

**Existing Analyses vs. Our Approach:**  
While several studies use standard machine learning models like decision trees or logistic regression, we emphasize:

* **Uncertainty quantification**
* **Priors based on domain knowledge**
* **Bayesian model comparison**
* **Convergence diagnostics (e.g., R-hat, trace plots)**

This adds interpretability and confidence bounds to the HR analytics process.

**OR Difference from existing work:** Most analyses use frequentist models; we’re applying Bayesian techniques (PyMC, Bambi, etc.), offering uncertainty quantification.

**Exploratory Data Analysis**

Some columns have missing values. For example, not all employees have their **age** or **salary hike last year** recorded. We'll take care of these missing values during data preprocessing.

A graph of blue bars

AI-generated content may be incorrect.

**Data Preprocessing:**

**Initial Cleaning**

* **Dropped empty rows**: Used dropna(how='all') to remove rows where all fields were missing.
* **Inspected missing values** in columns like EmployeeID, HomeToWork, Gender, Age, Department, MonthlyIncome, etc.

**Missing Value Imputation**

* HomeToWork: Missing values replaced with the median.
* Age: Filled missing values with the median.
* Gender:
  + Imputed with 'N' (unknown).
  + Standardized: 'Male' → 'M', 'Female' → 'F'.
* MonthlyIncome: Filled missing values with the mean.
* HourlnWeek: Filled with the median.
* LastPromotion: Replaced missing values with 0.
* CurrentProfile: Imputed with a value 2.5 (likely representing an average/median).
* Categorical fields like Designation, Department: Filled missing values with 'Unknown'.

**Standardization**

* Marital Status: 'M' renamed to 'Married'

**Feature Engineering**

Add the correlation graph here.

**Model Descriptions**

* **Model 1 (Classification - PyMC):** Bayesian Logistic Regression to predict Attrition. Use informative priors based on domain knowledge or literature.
* **Model 2 (Regression - Bambi or PyMC):** Predict Monthly Income using Bayesian Linear Regression.
* **Model 3 (Clustering - Probabilistic GMM):** Use scikit-learn’s BayesianGaussianMixture model to discover groups of employees.