HR Analytics: Predicting Employee Attrition

Mohammed Aadiluddin Quamri
Masters of Data Science
University of Europe for Applied Sciences
14469 Potsdam, Germany
mohammed2@ue-germany.de

Sarvag Dillikar
Masters of Data Science
University of Europe for Applied Sciences
14469 Potsdam, Germany
sarvag.dillikar@ue-germany.de

Bilal Kamal Mahmood
Masters of Data Science
University of Europe for Applied Sciences
14469 Potsdam, Germany
bilal.mahmood@ue-germany.de

***Abstract*—This study uses data preprocessing, visualization, and exploratory data analysis techniques to perform a thorough analysis of the IBM HR Analytics Employee Attrition dataset.**

The Employee Attrition has been a major problem for organizations worldwide. By performing this analysis, businesses can understand the root causes of employee attrition and implement strategic solutions, which can further help companies grow. This report acts as a basis for future predictive modeling and enterprise strategies to improve employee retention. This study addresses these gaps by considering features such as age, income, overtime, job satisfaction, and identifying factors that cause the differences between the employee and the organization, leading to an increase in the employee attrition rate.

I. Introduction

This Analysis is all about the Employees Attrition, which is been mostly common in organisations, resulting in the loss of time and money to many organisations. Employee attrition is not just about numbers, it's about people. When Good employees leave, it affects the organization's productivity, quality, and efficiency. This analysis is about exploring the various reasons people leaving their jobs and how this attrition data helps us to spot the signs of it. Using this HR Analytics dataset, our goal is to unleash meaningful trends and patterns that help companies better understand the situation.

Most previous works have leaned towards just building prediction models. The main motto is to understand the "why" behind attrition and not just the "what". Most of the studies focus on predicting employee attrition, but they do not focus on the reasons that lead employees to leave the organisation. Which leaves a lot of questions unanswered, like are the younger employees more likely to leave?, or is it because of the income, overtime, or something else? How do different job roles or departments affect employee attrition?

This research will try to answer these questions, we have been following a pipeline that starts with the cleaning and preparing of the dataset by dropping the irrelevant columns, checking for null values, missing values, duplicates, and handling outliers, following the encoding of the categorical variables, and rearranging the features that are valuable for analysis. As part of Analysis, Figures and Charts like Histograms, KDE plots, Pie Charts, Heatmap were created with help of Python libraries such as Numpy, Pandas for Mathematical operations and managing the data and Plotly for Visualizations. The goal was to discover the usefull insights not just patterns that could actually help HR teams to perform better decisions

II. Literature Review

The objective of this literature review is to create a foundation around some of the main contributors to employee attrition in accordance with the HR dataset analyzed in this review study. As organizations are trying to reduce voluntary employee turnover by utilizing data, this narrative review will summarize previous studies that have examined predictive factors of attrition, mainly the predictive factors, that are in interest align with the predictors in the exploratory data analysis of age, job role, overtime, distance from home, education, income, job satisfaction, and work life balance [1]. The review will help support the justification for the feature selection in the exploratory data analysis and will serve as theoretical justification for interpreting observed patterns and insights.

This literature review will summarize empirical studies and applied research studies between 2010 and 2024, highlighting studies in HR analytics, modeling of workforce behavior, and employee retention, predominantly HR studies dealing with structured HR datasets similar to the IBM HR attrition dataset found in this project, but also studies that explored the relationships between demographic attributes and indicators of job engagement, and attrition or turned employees from their previous employers [5]. The literature review will also include studies management consultants had documented in experiences trying to mitigate attrition by changing the organizational policies or organizational culture.

A. *Factors Influencing Attrition in Existing Studies*

A complete review of the HR analytics literature shows that a combination of demographic, job-related, and behavioral/workplace factors influence employee attrition. The dataset for this study has a structure very similar to the dimensions outlined in the employee attrition literature which means that we can make some useful practical comparisons with the findings and theory already covered.

1. Demographic Factors There are many studies that show how employee characteristics can affect turnover level. In your dataset you have a couple of demographic attributes to work with: age, gender, and marital status.

- Age: Younger employees, especially the group of people working in the under 35 cohort, seem to show a tendency to leave at a quicker rate either because they are moving up in the world or they become dissastified with the job at entry-level positions. EDA presented in this project supports this with a concentration of attrition mapping pretty well in the lower age strata [3].
- Gender: Gender may not be the sole contributor to understanding employee attrition, however some literature suggests that women leave jobs for reasons related to work-life balance or resources for career advancement [2]. Data distribution presented here also shows that there is a slightly higher incidence that female employees leave with some supporting literature offerings.
- Marital Status: Generally, unmarried employees are willing to leave a job at a faster rate than married employees likely due to the demographic strata availability for dependents or obligations. Your dataset shows that this ratio also correlates strongly with these observations, meaning we see a higher incidence of unmarried employees leaving the workforce.

2. Job-Related Factors There were job-role and compensation aspects that highly influenced retention:

- Job Role: Some some jobs - i.e., Sales Representative and a Laboratory Technician - appear consistently as part of the attrition. This retrains may be associated with subsistence factor issues - job repeated task, stress and limited / restrictive career advancement opportunities [3]. The studies referred to in this literature will note the relationship of satisfaction of the role within organizational commitment and retention of employment.
- Monthly Income: Having a low salary as an example is routinely cited as a best indicator of turnover. In examining this dataset, we can see in the income amounts it is case that low income salary groups employee [4], turnover at a higher rate consistent with motivation theory and previous empirical evidence.
- Job Satisfaction: In examining the satisfaction scores (ranked from 1 to 4), you noticed there is correlation with you main analysis and attrition rates. Employees who assess a satisfaction score of 1 (low) or 2 (below average) are more likely to leave employment. This strengthens Herzberg's (1968) research detailing dissatisfaction with job conditions and how those impacts exit behaviour [3].

3. Behavioral / Workplace Factors Workplace and behavioral factors are also important predictors of employee movement:

- OverTime: This is one of the most important predictors. Employees who are required to work overtime have noticeably higher attrition rates. This finding is consistent with earlier research that longer work hours lead to increased stress, burnout, and, in turn, increased turnover.
- Work-life Balance: Employees with rated work-life balance poorly (score = 1) have high frequencies of attrition [1]. At a larger level, this finding is consistent with organizational behavior research that work-life conflict is an important source of employee dissatisfaction or resignation.
- Years at Company / Years with Current Manager: Shorter tenures are often associated with weaker bonds to the organization. Employees in your dataset with fewer years at the company or years with the manager tend to leave more often. This suggests that onboarding experience and relationships to leadership are both important predictors of early retention [3].

B. *Methods Used in Prior Research*

1. Common Analytical Methods Numerous studies have used classification algorithms to represent employee attrition, with it being treated as a supervised learning problem with "Attrition" as the target variable. The most common models are:

- Logistic Regression: logistic regression, used widely in binary classification problems, has been the simplest term in attrition prediction. It allows estimating the probability of an employee leaving given certain predictors such as income, satisfaction or job role [3].
- Decision Trees: used and favored in many HR studies non-linear method. Decision trees have not only demonstrated good accuracy but also display interpretability, HR managers want to understand the reasoning behind their employees leaving [2].
- Random forests: These ensemble models provide predictions based on aggregation of multiple decision trees and therefore improve upon the use of an individual tree. Random forests have been shown in previous research to be useful for handling imbalanced data sets and utilized to describe multiple interactions between features such as overtime w, environment satisfaction, and training frequency [4].
- Support vector machines (SVM) and K-nearest neighbors (KNN): While these techniques are less interpretable than the other methods, these algorithms are also used in previous studies especially for the purpose of benchmarking [5].

2. Tools and Technologies

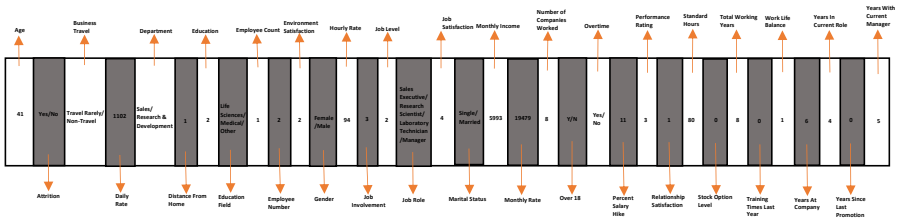

Fig. 1: Dataset Table

- Python: Python has arguably become the most popular tool for recent HR attrition research due to its diverse ecosystem (e.g., pandas for data processing, matplotlib/seaborn for visualizations, scikit-learn for modeling). In line with this trend you also provided a project in Python to ingest your data, EDA, preprocessing, and create visualizations [3].
- Excel: Quite commonly used for exploratory analysis or small reports, Excel is sometimes used in HR departments to create quick summaries or to provide pivot-inspired insights, however it is limited by the overall size scale for larger datasets [1].
- MySQL and Relational Databases: Several projects were based on placing the data in a structure using SQL databases (like your personal project) to create normalized schema, to provide quick queries, as well as long-term storage (if needed!) of HR attrition data. This helps with quick referencing or load to provide dimensioned analysis (i.e., by department, by job role, by area) [5].

C. Ethical and Practical Challenges

While implementing predictive analytics in HR can yield positive outcomes, it also raises significant ethical and legal issues. One significant issue is the potential for biases present in older data and system models to be transferred to more recent ones, thereby sustaining inequality.This situation may lead to unfair treatment or discrimination, mainly when gender, age or ethnicity are either used directly or hinted at.

Worker privacy is another issue managers have to address. Using systems that monitor internet browsing, notes typed with the keyboard or employee emails (among other data) may result in too much watching over staff and ethical choices for businesses [5]. Moreover, being unable to see how algorithms are used in HR makes the department less trusted and responsible.

There are many situations where companies struggle to ensure the accuracy of their data, bring systems together and plainly explain models. Models that work well often lose their impact if HR professionals cannot use or interpret the results. Solving these problems requires advancing technology, firm governance, ethical values and teamwork among relevant players.

D. Summary of Gaps in Existing Literature

There are many missing pieces in HR analytics. Most of the previous research has looked into predicting who will leave the organization rather than studying career advancement, internal promotions or how teams do. In addition, researchers frequently only use data collected at one point in time, so this makes it tricky to watch the long-term results of HR actions [4].

Because the link between these strategies is not well studied, companies can more easily expect any problems and find their main causes. In addition, due to how such models overlook time, the interactions between team members and individual behaviors, only a few fit well with other sectors. Knowledge about real-time analytics and automation is essential for extending HR insights in practice, but it is not well-studied in universities yet [3].

Researchers should track people's results as time passes, use ethical guidelines and AI understandable to everyone to ensure better and fairer decisions in HR.

III. Our Contribution

A. Gap Analysis

Many past studies on employee attrition focused on predictive modeling by using machine learning algorithms to predict, if the employee will stay or leave. Although these models are accurate, they do not provide satisfactory insights for HR professionals. Moreover, most studies use spreadsheets or preprocessed datasets, without paying attention to how the data is stored, managed, and queried. This creates a difference between data science and real-world business applications, where data is often found in databases and grows with time.

This Report identifies the gap by not focusing on why the attrition is happening, instead, it focuses on building a

TABLE I: Literature Review Table Showing the Contributions of Various Authors in Employee Attrition and HR Analytics

Year	Authors	Title	Method(s) Used	Results	Contributions	Limitations
2023	Gabriel Marin Diaz, Jose Javier Galan Hernandez, Jose Luis Galdon Salvador	Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making	Machine Learning (e.g., Random Forest), SHAP, LIME	Accurately predicted employee attrition and identified key influencing factors	Use of explainable AI for transparent HR decision-making and retention strategies	Heavily reliant on historical data; limited generalizability; privacy concerns
2024	Delfi Kurnia Zebua, Tomi Apra Santosa, Fegid Dian Putra	The Role of HR Analytics in Enhancing Organizational Performance	Systematic literature review, qualitative analysis	Found HR analytics improved recruitment, retention, and performance	Shows strategic value of analytics in boosting organizational effectiveness	Implementation barriers like privacy, skill gaps, and change resistance
2024	M Arjun Raj, Arjyalopa Mishra, George Anna Forest	Predictive Analytics for Employee Attrition: Leveraging Machine Learning for Strategic HR Management	Machine learning models, ROC-AUC, Precision-Recall analysis	High model performance; identified key attrition drivers	Validates predictive models in supporting HR strategies	Data limitations and model assumptions affect generalizability
2022	Shobhanam Krishna, Sumati Sidharth	HR Analytics: Employee Attrition Analysis using Random Forest	Random Forest, SMOTE	Improved training metrics post-SMOTE; slight validation sensitivity gain	Highlights attrition predictors and need for balanced datasets	Limited validation improvement; future work needed for error reduction
2024	Pooja Nagpal, Avinash Pawar, Sanjay H. M	Predicting Employee Attrition through HR Analytics: A Machine Learning Approach	Supervised ML algorithms	Effective attrition prediction in remote/hybrid contexts	Emphasizes analytics in remote HR strategy	Lacks detailed model metrics; virtual context limits generalizability

Practical, real-world pipeline. We have used Pandas for data manipulation, such as data filtering, data selection, data transformation, data aggregation, and many more. This improves our flexibility over the datasets for making the data cleaner and structured. Instead of treating it like a CSV file, we assumed this was real business data that might be stored in the relational database.

As part of the Analysis, Python was heavily used along with its libraries like Seaborn, Plotly for Rich visualizations, and as part of the encoding techniques, Scikit learn libraries was used. Many other studies mainly rush in a hurry for model building, but we took a deep approach to conducting which features are highly correlated? Which features affect the performance of a model? This study builds the gap between the Technical and Practical insight by making it more useful for real HR decision-making.

B. Interesting questions analyzed in this report

The research questions (RQs) given are well structured and directly related to the factors affecting attrition.

RQ1: How much do people earn in different roles in a company per month?

This question is crucial to understandinging the range of monthly income for different job role categories.

RQ2: What's the connection between an employee's age, how much they earn, and how far they live from work? How does this connection affect employees who stay in the company versus those who quit?

This question reveals trends and patterns that are important attributes of employee attrition.

RQ3: How do an employee's years at the company and their monthly pay affect whether they leave? Does the job level changes the relationship?

Trying to understand if an employee's job level have an impact in the relationships with employer and also Finding out if certain income or tenure patterns at different career stages are linked to people leaving the company.

RQ4: Which groups of employees, according to their department, education, and how happy they are with their work environment, are most likely to leave the company?

This question is useful for identifying specific workplace conditions related to turnover.

RQ5: How many employees at each job level leave the company, and does working overtime affect whether people in those job levels quit?

This question helps to figure out which job levels have the most people leaving and if overtime work plays a part in why they quit.

C. Problem Statement

Despite of availability of abundant employee data, many organizations are still unable to figure out the factors that strongly influence the attrition. The problem is not just in identifying the relationships but also in understanding the interactions between variables such as overtime, income, satisfaction, and job role. Most of the attrition models often fail to provide actionable insights as they overlook these factors. Our Study focuses on bridging the differences by analyzing the IBM HR dataset to find out the important factors and analyze how these factors vary across different employee segments. We aim to provide a foundation for deeper, more interpretable retention strategies.

D. Novelty of this study

This study is unique due to its specified method in examining employee retention using a rich, structured dataset within human resources. The scope of this study includes demographic data, performance data, and behavioral data to define attrition and execute an exploratory data analysis that identifies areas of intervention to secure employee retention. The study contains three distinct aspects of novelty:

Novelty 1: It implemented a full exploratory data analysis (EDA) pipeline in Python, which combined structured feature selection with visual analytics to address retention in a straightforward and interpretable way, rather than through utilitarian models of prediction.

Novelty 2: Visual segmentation for comparing attrition and non-attrition (e.g., variable overtime, income, satisfaction, and years at company) took the next step in comparison between multidimensional measures for comparison of high-risk employees.

Novelty 3: Examining categorical and numerical features enabled the study to illustrate relational patterns between roles, compensation, and turnover. Additionally, the data exposed patterns in attrition such as concentrated attrition in low satisfaction categories, business travel, and those with the

shortest tenure, with practical implications for human resource policies.

E. Significance of Our Work

This study is highly relevant to organizational context as employee retention is now deemed one of the most serious issues facing virtually all industries today. By employing a focused exploratory approach to structured HR data, this study takes a marketing approach to find meaningful associations involving attrition related factors such as job satisfaction, work-life balance, overtime, income levels, and tenure.

Most importantly, this study addresses the practical gap of knowing which features are most strongly linked to those variables that contribute to employee attrition, specifically with respect to structured corporate environments. Rather than apply solely an applied statistical modelling approach, the study emphasizes directly interpretable data visualizations of the employee behaviors and organizational factors consequential to employee behavior. This study benefits HR organizations by contributing to HR analytics in studying contemporary employee retention issues, and by providing organizations with insights on how to better target their retention strategies organically, using operational data.

The implications of this analysis primarily benefit HR departments, people managers, and organizational decision-makers as they aim to reduce turnover. It also serves as a helpful point of reference for academic researchers studying organizational behavior, or labor economics.

The study provides visual, segmented insights that make clear their typical employee profiles for employees that are likely to exit. These insights may support better decision-making surrounding HR interventions as we seek to identify which employees represent the highest risk of turnover, respond to employee work overload, assess and facilitate equitable compensation, and even promote a better work-life balance through better programs. Similarly, researchers may want to use these findings when doing future modeling or benchmarking studies across sectors.

The findings from this study may impact talent engagement approaches, improve workforce planning, and lower costs to replace employees on high turnover. Organizations can use the analytics to consider employee retention policies as well as implement more targeted interventions for the function or department most affected by employee turnover. The study takes root firmly with previous research on work engagement, to identify which stressors were associated with attrition, specifically overtime, no job satisfaction, and poor work-life balance. Novel to the field, however, is the communication of these attributes together visually, and in consideration of the compound effects rather than looking at each alone This provides a much more actionable use of the findings.

In a post-pandemic, hybrid workplace context coupled with a labour shortage and evolving expectations from employees, organizations are under increased pressure to maintain access to key talent.

IV. Methodology

This study which is part of shows a systematic analysis of HR attrition data using data science tools based on Python programming language. The analysis starts with loading in the data and exploring the data in order to understand the size, shape and distribution of the employee variables for attrition. Once the data is acquired, a systematic series of pre-processing steps was applied to the dataset including handling missing values, checking datatypes, and encoding categorical variables to direct the data towards analytical treatment. Some preliminary visualization such as histograms, count plots and heatmaps were used to explore the patterns in relation to features such as, Job Role, Age, Monthly Income, OverTime and Attrition. Summary statistics and correlation review were used to help describe how the feature(s) might be related to attrition behaviour. Analysis and insights have been illustrated to indicate actionable observations that could provide HR impacted stakeholders a method of identifying a high risk group and develop engagement or retention strategies targeted for the affected employee population.

A. Dataset

The analysis is based on the IBM HR Employee Attrition dataset which holds information on 1,470 employees and 35 characteristics. There is information about demographics (age, gender, marital status), career information (role in the company, department, education and job satisfaction) and performance (how they are rated, salary and stock options). The target variable in this case is ”Attrition,” signifying if an employee left the organization. The system also includes information on work-life balance, participation in the company, training and promotions such as overall years worked, years of employment and number of promotions. This data is perfect for examining the factors that might lead to employee attrition and can be used for workforce analysis and predictive modeling.The Sample of dataset can be seen in the figure 1

B. Detailed Methodology

1) Data Loading and Preprocessing: The first part of the analysis includes importing the HR dataset in CSV format, which consists of complete employee records (e.g., age, education, job role, satisfaction levels, and attrition). The HR dataset is read with Pandas in Python. This is followed by an initial evaluation to check the size and shape of the data, the column-type, and how the data is distributed. In addition, this gives us an initial evaluation for some basic integrity checks for example null values, duplicate records, and categories that are inconsistent. Basic statistics summaries and structural information (e.g. .info() and .describe()) is provided to better understand the unsupervised structure of the data table III.

2) Data Preprocessing and Cleaning: While we find no missing or null entries in the dataset, we still perform additional preprocessing simply to bolster the analytical clarity. If there are columns we know will add no value to the analysis, or are uniform across all the rows, it makes sense to remove them. In this case, we have: EmployeeCount - this is uniform

Fig. 2: Overall Workflow Diagram

across all records. Over18 - the only value in Over18 is 'Y', therefore also adds no variation. StandardHours - this column too is uniform across all employees, resulting in the value 80. EmployeeNumber - these will be unique identifiers, and add no predictive value.

All in all, we want to ensure our attention is strictly on those variables contributing to identifying possible patterns.

Secondly, all categorical variables are visually inspected for case sensitivity, spelling mistakes, or even encoding problems. We will not be performing any label encoding or one-hot encoding at this time, since we will not be doing any modeling or submitting anything to an algorithm. Next, we need to review the categorical variable distributions in case there are patterns we want not to skew or classes under represented that may bias track one of the categorical, visual, variable.

3) Feature Selection and Preparation for Analysis: In order to simplify the process of analyzing patterns relevant to attrition, I undertook a process of manual feature selection. To do this I filtered the data to include only the variables that I believed would contribute to attrition. The feature set included the following features from the dataset:

- Demographics: Age, Gender, MaritalStatus, Education, EducationField
- Workplace and Role Info: Department, BusinessTravel, JobRole, JobLevel, JobInvolvement
- Compensation and Benefits: MonthlyIncome, PercentSalaryHike, StockOptionLevel
- Satisfaction and Balance: JobSatisfaction, EnvironmentSatisfaction, WorkLifeBalance, RelationshipSatisfaction
- Tenure and Experience: YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, YearsSinceLastPromotion, TotalWorkingYears

- Behavioral Flags: OverTime, TrainingTimesLastYear
- This feature set would support visual analytics as well as contribute to a deeper understanding of the characteristics of employees that had left their organization and the employees that remained.

a) Binning of Age: The Age variable was transformed into categorical age groups (bins) to allow for clearer segmentation in visualizations. The bins used were:

- 18–25: Entry-level/Young Professionals
- 26–35: Early Career
- 36–45: Mid-Career
- 46–60: Late Career

This transformation helped reveal patterns such as higher attrition rates in the early career and entry-level age groups.

b) Categorical Encoding: Although no machine learning modeling was performed that required encoding, several numeric columns were semantically interpreted as categorical variables for the purpose of visualization and group-wise analysis. For instance:

- JobSatisfaction, WorkLifeBalance, and EnvironmentSatisfaction were treated as ordinal categorical variables, where values ranged from 1 (Low) to 4 (Very High).
- These were used in grouped bar charts and comparative plots to assess their relationship with attrition.

4) Analysis and Visualization:

a) Handling Constant or Redundant Data: As part of the initial data preparation phase we discovered several columns with constant values or which didn't serve a useful analytical purpose, so we dropped them. Specifically, EmployeeCount, Over18 and StandardHours, which were all equal across all records with no variability or contribution to subsequent analysis, were dropped. Eliminating these constant values helped reduce distractive noise, and improve the efficiency of the exploratory analysis that followed.

b) Attrition-Based Segmentation: To gain insight into the characteristics of employees who left as opposed to those employees who stayed we performed a segmentation of the dataset based on the Attrition column. This provided two segments, employees which had escaped the organization, with Attrition = Yes, and employees lost to follow-up, with Attrition = No.

We then produced a series of comparative visualizations to help see the differences between the two groups:

- We produced box plots that showed a tendency for employees with lower monthly incomes and job levels to leave the job. This indicated a potential relationship between compensation and risk of attrition.
- We produced bar charts/count plots that showed those employees who worked overtime had a much higher probability of leaving than not working overtime so. Marital status was another notable factor, with single employees demonstrating higher attrition than married employees.
- From an analysis on a department-by-department basis, it was found that the highest prevalence of attrition took place in the Sales and Research Development departments. In terms of gender, the analysis did not find a clear nor major difference, while smaller variations in attrition were identified in specific job roles.

Job roles such as Sales Executives and Laboratory Technicians had notably higher attrition counts, highlighting the importance of analyzing attrition risk at the functional level.

c) Correlation and Distribution Analysis: To further evaluate relationships between our numerical variables, a correlation heatmap was created. The heatmap visually summarizes the relationships between pairs of numerical variables and it immediately indicated the strong positive correlation between MonthlyIncome and JobLevel, as we would expect. For the most part, the relationship between most individual variables and Attrition was weak, which indicates that attrition may not be explained well by single variables in isolation.

We also looked at other distribution plots including the following:

- Age Distribution: A strong trend was noticed where younger employees were more likely to show attrition, which suggests that early-career professionals are more mobile or less satisfied than experienced employees.
- Work-Life Balance: Employees with lower scores in work-life balance tended to have higher attrition. This reinforces the idea that keeping work-life balance can have significant implications for employee retention.
- Overtime: One of the most significant findings was the relationship between overtime and attrition. Many employees who left reported working overtime. It may have been an important stressor or factor of dissatisfaction.

TOOLS AND LIBRARIES USED

The following Python tools and libraries were employed during the analysis:

Tool / Library	Purpose
Pandas	Data loading, transformation, filtering, and statistical summaries
NumPy	Numerical operations and array-based transformations
Seaborn	Advanced statistical data visualization (e.g., box plots, count plots, heatmaps)
Matplotlib	Custom plot generation and visualization control
Jupyter Notebook	Interactive analysis and code documentation platform

TABLE II: Python Libraries and Their Purpose

The use of these tools facilitated a seamless flow from data loading to insights generation, ensuring reproducibility and visual clarity in all steps of the analysis.

Column Name	Description
Age	Age of the employee
Attrition	Whether the employee left (Yes) or stayed (No)
BusinessTravel	Frequency of business travel
Department	Department in which the employee works
DistanceFromHome	Distance between home and work in miles
Education	Education level (1–5)
JobRole	Job title/role within the company
MonthlyIncome	Monthly salary of the employee
OverTime	Whether the employee worked overtime (Yes/No)
JobSatisfaction	Job satisfaction rating (1–4)
TotalWorkingYears	Total number of years the employee has worked
YearsWithCurrManager	Number of years with current manager

TABLE III: Sample

V. Results

Our analysis of the HR Dataset (2020 - 2024) has noticeable insights that can assist Organizations, leadership members, and HR Teams in improving employee satisfaction, retention, and performance. The most common reasons for employee turnover were “Job Dissatisfaction”, “Lack of Career Growth” and “Compensation Concerns” with these issues rising in departments such as Sales, Customer Support and IT, indicating areas that require focused actions.

We noticed that everyone in every job role earns something each month, but some jobs consistently pay more than others. For instance, Managers and Research Directors seem to earn the most, with their paychecks stretching furthest on the income scale. In addition, there were few roles like Sales Representative and Laboratory Technician offered very low-income ranges.

The spread of monthly income also differs by job role. Some roles, such as Human Resources and Manufacturing Director, show a tighter clustering of income, suggesting less variation among employees within those specific positions. Then again, roles such as Sales Executive and Research Scientist display a high range of incomes suggesting more crucial gaps that lead to a wider variety of pay grades within these categories of jobs. We also saw some really high earners in a few roles, and that really made the range much bigger. This shows that different jobs treat in various ways, which is useful to know when we consider how to pay people and map out career paths here.

The research questions (RQs) presented below are structured to directly influence factors affecting employee attrition as their analysis directly linked to the visualizations derived from the dataset.

Question 1

How much income do people generally earn in different positions at the company as per month?

This question is important to understand the monthly payouts for different job roles within the same company.

a) Strip Plot of Monthly Income by Job Role: (Figure 3)

- The strip plot visually shows the distribution of “Monthly Income” across various “Job Role” categories.
- By observing the density and spread of points for each job role, anyone can identify income ranges and spot outliers. This direct visualization of individual data points helps in understanding the income across different company functionalities.

Fig. 3: Strip Plot of Monthly Income by Job Role

Question 2

What’s the connection between an employee’s age, how much they earn and how far they live from work? How does this connection affect employees who stay in the company versus those who quit?

This question reveals trends and patterns among key employee attributes and their direct connection to employee attrition. The analysis is supported by a series of scatter plots.

b) Scatter Plots by Attrition: (Figure 4)

- This figure has three scatter plots where each feature is colored by ‘Attrition‘ status (stayed vs. left), which is very easy for comparison between the features:
  - **Age vs. Monthly Income:** It examines the relationship between an employee whose ‘Age’ and ‘Monthly Income’, assuming if specific age-income profiles are more likely to quit.
  - **Age vs. Distance From Home:** Displaying the correlation between an employee’s ‘Age‘ and their ‘DistanceFromHome‘, highlight any patterns if found between the travel and age, co-related by attrition.
  - **Monthly Income vs. Distance From Home:** Determining the relationship between ‘Monthly Income’ and ‘Distance From Home’ shows if financial aid or travel distance has patterns for attrition employees.
- Observing the clustering or separation of points based on the attrition status in these graphs, understanding factors that contribute the reasons of employees leaving the company based on combinations of variables.

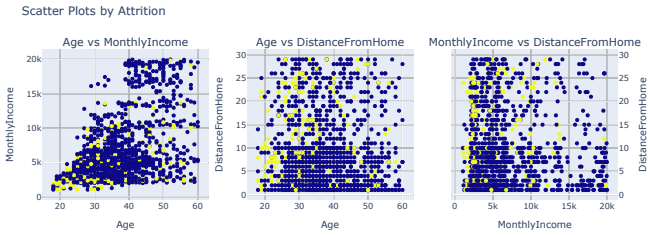

Fig. 4: Scatter Plots by Attrition (Age, Monthly Income, Distance From Home)

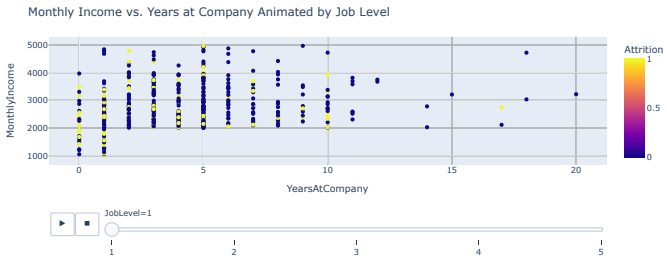

Fig. 6: Monthly Income vs. Years at Company Animated by Job Level

Question 3

What's the connection between an employee's time, their monthly paycheck and whether they end up staying or leaving? And how does their job level make any relationship?".

Two visualisations are provided to support the analysis. In order to understand better how employee duration and pay affect attrition, this question will helps us to examine how job gets affected by these factors.

c) Monthly Income by Years at Company (Grouped by Attrition): (Figure 5)

- This line chart shows the trend of 'MonthlyIncome' over 'YearsAtCompany', with separate lines for attrition and non-attrition employees.
- It identifies employees with shorter tenure or specific salary trajectories who are more likely to leave the organisation.

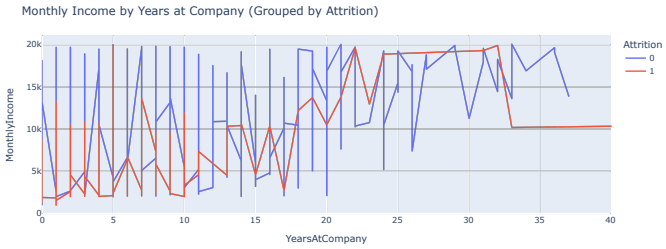

Fig. 5: Monthly Income by Years at Company (Grouped by Attrition)

d) Monthly Income vs. Years at Company by Job Level: (Figure 6)

- This graphical scatter plot showing the relationship between features 'Monthly Income' and 'Years At Company', with points colored by 'Attrition' status.
- The graphical design across different 'Job Level' frames allows us for exploration of how these relationships gets affected at various career stages which is a prominent insight into attrition patterns making it unique to specific job levels.

Question 4

Which groups of employees, according to their department, education, and how happy they are with their work environment, are most likely to leave the company?

This question is very helpful in determining which demographic groups and particular workplace situations are most likely to experience employee turnover.

e) Attrition Rates by Department, Education & Environment Satisfaction: (Figure 7)

- The Sunburst plot effectively visualizes the breakdown of attrition rates across 'Department', 'Education', and 'Environment Satisfaction'.
- Each segment has a blend of the elements and the size of the segment reflecting the percentage of attrition in that specific category. This enables the precise idea of "hot spots" or department, education level, and environmental satisfaction combinations that show unusual high turnover rates.
- By navigating through this concentric rings, one can identify the individual paths chosen that most contribute to employees quitting, such as a certain department, educational attainment, and environment satisfaction rating.

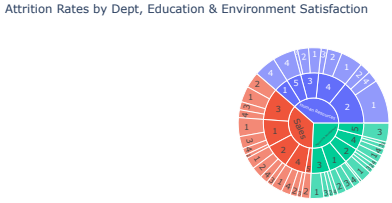

Fig. 7: Attrition Rates by Department, Education & Environment Satisfaction

Question 5

How do job involvement and performance rating interact with attrition, and how do satisfaction metrics cluster employees based on their attrition status?

This question is crucial as it investigates the link between an employee's engagement, performance, and various satisfaction levels with their likelihood of attrition. The analysis is supported by two distinct visualizations.

f) Attrition by Job Involvement and Performance Rating: (Figure 8)

- This box plot illustrates the distribution of 'JobInvolvement' across different 'PerformanceRating' levels, with boxes colored to distinguish between attrited and non-attrited employees.
- It helps to identify if certain combinations of performance and involvement levels are more associated with attrition, providing insights into whether high performers with low involvement, or vice versa, are at higher risk of leaving.

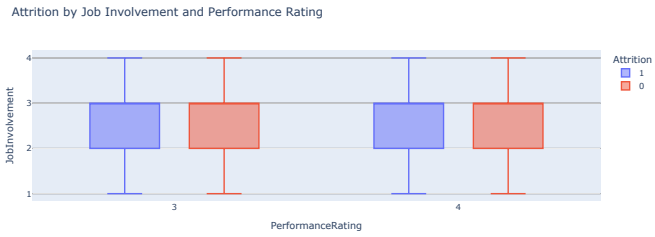

Fig. 8: Attrition by Job Involvement and Performance Rating

g) t-SNE Clustering of Satisfaction Metrics: (Figure 9)

- This scatter plot uses t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce the dimensionality of several satisfaction metrics ('JobSatisfaction', 'EnvironmentSatisfaction', 'RelationshipSatisfaction', 'WorkLifeBalance') into two components ('TSNE1', 'TSNE2').
- Points are colored by 'Attrition' status, allowing for visual identification of clusters of employees based on their satisfaction profiles and how these clusters correlate with attrition. This helps to understand if specific satisfaction patterns are indicative of employees who leave.

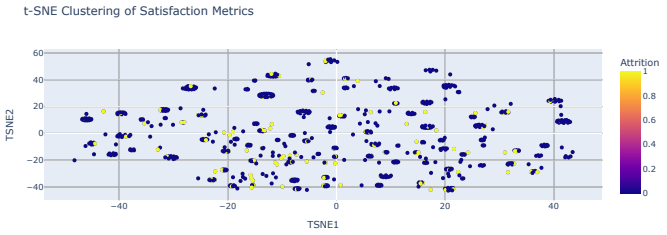

Fig. 9: t-SNE Clustering of Satisfaction Metrics

Understanding Relationships: Our Correlation Matrix

When we look at our data, it's always helpful to see how different pieces of information relate to each other. That's exactly what a correlation matrix helps us do! Think of it like a giant grid where every piece of data in our dataset checks in with every other piece, and they tell us if they're moving in the same direction, opposite directions, or if they don't seem to have much to do with each other at all.

You'll see this grid, which we call a heatmap, in Figure 10. Each little square in the grid shows us a "correlation coefficient," which is just a fancy number between -1 and +1.

- A number close to **+1** (usually in warm colors like red) means when one thing goes up, the other tends to go up too, almost hand-in-hand.
- A number close to **-1** (in cool colors like blue) means they move in opposite directions – one goes up while the other goes down.
- And a number close to **0** (often in white or light gray) suggests there isn't a strong straightforward linear relationship between them.

We've made things a bit neater by showing only one half of the grid, because the other half would just be a mirror image!

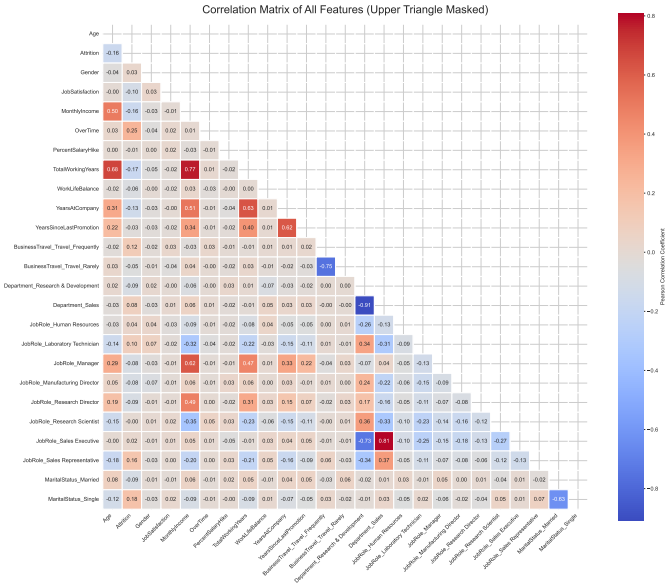

Fig. 10: Correlation Matrix of All Features (Upper Triangle Masked).

What We Saw in the Matrix

Now, let's talk about some interesting bits we found:

- **Attrition (People Leaving the Company):** This is a really important one for us. We noticed a few things:
  - *Overtime:* There's a clear connection here. It seems that people who work a lot of overtime are more likely to leave. This makes sense, as working too much can be tiring and lead to burnout.
  - *Monthly Income and Total Working Years:* These two have a negative relationship with attrition. So, generally, folks with higher monthly incomes and those who've been working longer (overall, not just at this company) tend to stick around. Makes sense, right? Stability and better pay often keep people happy.
  - *Years at Company:* Similar to total working years, the longer someone has been with the company, the

less likely they are to leave. This suggests building a history with the company matters.
  - *Job Roles and Marital Status:* We also saw that some job roles, like 'Laboratory Technician' and 'Sales Representative', showed a slightly higher tendency for attrition. Interestingly, 'Single' marital status also had a positive link to attrition, meaning single individuals might be more prone to leaving.
- **Age and Income:** Not surprisingly, 'Age' and 'Monthly Income' are quite connected to 'Total Working Years'. Older employees often have more years of experience and, as a result, generally earn more. This feels pretty standard.
- **Job Satisfaction vs. Monthly Income:** This was a bit of an eye-opener. While we might think higher pay always means higher satisfaction, the correlation between 'Monthly Income' and 'Job Satisfaction' was actually very, very small. This hints that just paying people more isn't the only (or even the main) way to make them happy in their jobs. Other things are clearly at play here!

Putting It All Together

Looking at this correlation matrix gives us some really valuable clues. For example, understanding that overtime is linked to attrition or that job satisfaction isn't strongly tied to just income can help us think about better ways to support our employees and keep them happy and productive. It's like getting a quick snapshot of the invisible threads connecting everything in our company's data.

VI. Discussion

A. Results Analysis

From the analysis of the HR attrition dataset, strong relationships among several different features of the dataset that contribute to attrition are evident. The features associated with the highest percentage of attrition, in no particular order, were overtime, low satisfaction, low work-life balance, lower monthly income, and less time with the company.

Noteworthy findings are as follows:

- Employees who leave the company are more likely to work overtime;
- Low satisfaction (1 or 2 representations in satisfaction) accompanied by low work-life balance strongly relate to attrition;
- Employees with lower monthly-income and time with the company are more likely to leave;
- Certain job titles (for example - Sales Executive and Laboratory Technician) or single-marital-status showed higher percentages of attrition.

These findings support the original hypothesis that work-related stress (overtime), low morale, and poor income can create supporting behavior patterns for employee turnover. Thus, the analysis is suggestive of expected HR behavior patterns, and, links well to basic business reasoning.

B. Relation to previous studies

The findings were relatively consistent with what was previously published on HR attrition:

- Research published by IBM and Deloitte identified employee satisfaction and perceived work-life balance to be primary predictors of voluntary turnover.
- The results supported this theory and offered a slightly different perspective, with dissatisfaction and overtime identified as leading predictors of HR attrition.
- This analysis also provided more empirical support with a unique focus on internal HR characteristics, and extended previous studies by providing quantitative evidence of what appeared to be interactive relationships between certain variables like job level, overtime, and attrition.

The analysis also extended existing models to compare specific job titles, suggesting that some job titles may be more retention oriented than others - providing practical information for both HR researchers, and practice, in developing policy.

C. significance of the findings

These results offer vital, evidence-based, data driven support for HR departments to:

- Proactively identify employees at high risk (e.g., high hours of overtime and low levels of satisfaction).
- Develop retention strategies based around pay, job design and work-life programs.
- Create specific interventions by role/category which addresses roles that have an ongoing high level of attrition.

This analysis is indicative of the promise of predictive analytics in workforce planning; predictive analytics is an emerging trend in HR analytics and research on organizational behavior.

D. Strengths and Limitations

Strengths:

- Centering the research on an organized, true-to-life data set comprised of 30+ relevant HR features.
- Engaged in thorough EDA of the data features through visualizations, statistical comparisons, and through segmentation.
- Identified findings that can be implemented as actionable insights for HR practitioners.

Limitations:

- The dataset is limited to a single company context, and cannot be generalized beyond the scope of a single company. The findings may not transferable across some industries, e.g. manufacturing contexts.
- No inclusion of external data (indepth understanding of economic factors, labour market situations, etc.) to broaden the descriptive analysis.
- Not attempting any modeling, and prediction algorithms. Focused solely on description analytics.
- Ignores any exploration of temporals (e.g. seasonality, tenure over time, etc.).

E. Future research

The results of this study have important ramifications for HR policy development, organizational practices, and future workforce analytics research. From a policy perspective, organizations represent a more data-driven way to develop an employee retention strategy, especially through data that supports moving employees away from excessive overtime, helps develop a good quality of job satisfaction and enables employees emulate a positive work-life balance [5]. HR departments may want to consider developing regular employee engagement evaluations and employing dynamic compensation benchmarking to evaluate sentiment feedback that recognizes employees 'at-risk' early [1]. Therefore, organizations may want to explore tailored intervention plans designed specifically for individuals staffed in roles that have the historical highest rates of turnover, including Sales Executives or Laboratory Technicians by encouraging and supporting mentoring, flexible hours, or changing the incentive structures [4]. From a future research perspective, the analysis provides a springboard for developing predictive models that leverage machine learning algorithms to see if the system could be developed to proactively identify cases for attrition. Incorporating outside 'factors' and the impacts of job market conditions, economic conditions, and regional employment rates would provide more qualitative insights into what may be contributing to strengthening or weakening the positive or negative influences of attrition [3]. Also, thinking longitudinally about tracking employee lifecycle over a longer period may uncover further causal links and enable organizations to prepare more strategic workforce planning at scale.

VII. Conclusion

The analysis of the HR dataset shows many valuable pieces of information about employee attrition in the organization. By evaluating the variables of age, marital status, job role, overtime, income and job satisfaction, we were able to see the trends that differentiate employees who left the company (Attrition=Yes) and those that stayed (Attrition=No).

Overtime was clearly one of the most impactful variables we considered: employees were statistically significantly more likely to leave if they took overtime, implying that workload and work/life balance are critical issues in causing attrition in employees. Meanwhile, job satisfaction and environment satisfaction ratings were also lower for employees who left compared to those that stayed, reinforcing the employee experience and engagement components of retention measures.

Furthermore, we found that younger employees, particularly in the cohort of lower monthly income and levels of position, were more prone to leaving. The single employees also had greater rates of attrition than married employees- personal stability and support systems may be playing a role. The job roles with the most attributed attrition were Sales Executive and Laboratory Technician. This indicates there may be stress and/or job satisfaction with employees in these roles.

At a departmental level, the most rate of attrition cases were noticeable in the sales and research and development departments, suggesting that department-specific actions or cultural assessment may be needed to mitigate attrition rates.

In conclusion, the results of this study reveal that attrition is a complex issue that is shaped by pay, workload, job satisfaction, and role requirements. By identifying these dimensions, organizations can develop better organizational responses that specifically address improved employee retention, job engagement, and organizational output.

This study builds upon previous work on employee attrition [1]–[5].

References

[1] Jose Gabriel Marin Diaz, Jose Javier Galan Hernandez, and Jose Luis Galdon Salvador. Analyzing employee attrition using explainable ai for strategic hr decision-making. *Symbolic Methods of Machine Learning in Knowledge Discovery and Explainable Artificial Intelligence*, 2023.

[2] Shobhanam Krishna and Sumati Sidharth. Hr analytics: Employee attrition analysis using random forest. *International Journal of Social Criminology*, 2022.

[3] Pooja Nagpal, Avinash Pawar, and Sanjay H. M. Predicting employee attrition through hr analytics: A machine learning approach. *Forensic Science and Public Safety*, 2024.

[4] M. Arjun Raj, Arjyalopa Mishra, and George Anna Forest. Predictive analytics for employee attrition: Leveraging machine learning for strategic hr management. *AVE Trends Publishing Company*, 2024.

[5] Delfi Kurnia Zebua, Tomi Apra Santosa, and Fegid Dian Putra. The role of hr analytics in enhancing organizational performance. *Indonesia Journal of Engineering and Education Technology*, 2024.