Angela Hyson  
DATA 4390  
May 2, 2025

**Case Study into People Analytics**

### **Introduction**

In today’s business environment, organizations strive to maintain efficiency and effectiveness by embracing data-driven decision-making. Human Resource (HR) departments play a crucial role in sustaining organizational health and optimizing workforce performance. They achieve this by tracking key performance indicators (KPIs), developing predictive models to forecast employee performance and turnover, and assessing the impact of new policies, training programs, and workshops. People analytics, which involves the use of data and statistical methods to understand and improve employee performance and organizational health, has emerged as a critical tool for HR professionals.

This project serves as a case study in People Analytics, focusing on an HR department's efforts to analyze organizational health. Using statistical analysis and predictive modeling, I seek to identify trends in employee retention and termination, uncover patterns influencing workforce stability, and provide actionable insights. Additionally, I will develop an interactive dashboard to effectively communicate these findings to HR professionals and stakeholders, simulating the experience of presenting key business insights.

### **Objectives**

The primary objective of this project is to analyze employee data to improve organizational health through data science techniques. Specifically, the study aims to evaluate workforce performance and satisfaction, develop a predictive model for employee termination, and present insights through a dashboard. The end goal is to enable HR professionals to make proactive and informed decisions that enhance retention and employee well-being.

The study focuses on four major objectives. First, it involves analyzing organizational health using key HR metrics and statistical techniques. Second, a predictive model is developed to forecast which employees are at risk of termination. Third, an interactive dashboard is created to visualize findings and support exploration of the data. Finally, the study seeks to deliver strategic recommendations that HR leaders can use to foster a productive and equitable workplace.

### **Methodology**

#### Dataset Overview

The dataset used for this analysis originates from the publicly available Kaggle HR Analytics dataset (<https://www.kaggle.com/datasets/rhuebner/human-resources-data-set/data>). It consists of 311 employee records and 35 columns representing a wide range of human resource information. The data includes demographic details such as gender, race, and date of birth; job-related information like department, position, and manager ID; performance metrics such as performance scores and absences; engagement metrics including satisfaction and survey results; compensation details like salary and recruitment sources; and temporal data reflecting hiring and termination dates.

This rich dataset offers a wide view of the workforce and provides a foundation for meaningful statistical analysis and predictive modeling.

#### Data Cleaning and Preparation

To ensure data integrity and modeling accuracy, extensive cleaning and preprocessing were performed. Missing values were addressed by mapping existing data, particularly for Manager ID fields. The termination date was retained as NaN for active employees.

Data consistency was ensured by correcting capitalization, removing duplicate entries, and trimming leading spaces. The calculation of employee age was standardized using a reference date of December 31, 2019. Future birth dates were adjusted by assuming a century correction.

#### Feature Engineering

To enhance the predictive capability of the dataset, several new variables were engineered. A diversity status indicator was introduced to track diversity in hiring and evaluate the impact of initiatives such as diversity job fairs. Compensation benchmarks were calculated, including the department average salary and a salary-to-department ratio, which enabled the identification of pay equity issues.

Additional features included departmental performance benchmarking, derived from average performance scores and performance standards. The number of days since the last performance review was also computed to flag overdue evaluations. Termination reasons were simplified into grouped categories to make the predictive modeling and interpretation more manageable. Tenure was also calculated using either the termination date or the reference date, depending on employment status.

Through these steps, the dataset grew from 35 to 44 columns, improving the information available for analysis.

#### Data Preprocessing

To further prepare the data for statistical analysis and modeling, specific preprocessing techniques were applied. Winsorization was used to reduce the influence of outliers in numeric columns such as salary, age, and engagement scores while preserving the data shape. Log transformations were applied to skewed variables such as absences and salary, ensuring a more normal distribution.

Because the final models employed tree-based algorithms like Random Forest, standardization and normalization were not applied. This approach preserved the interpretability of original distributions, which was particularly important for exploratory and inferential statistics.

#### Statistical Analysis

Multiple statistical tests were conducted to explore relationships between employee characteristics and organizational metrics. A chi-squared test revealed a strong and statistically significant relationship between employee engagement and performance (*p* < 0.001). Employees who were highly satisfied tended to have higher performance ratings, emphasizing the importance of engagement strategies in performance management.

Conversely, an analysis of manager influence using a chi-squared test showed no significant differences in performance across managers (*p* = 0.2494). This suggests that employee performance is driven more by systemic factors, such as role clarity or training programs, than by individual managers.

Pay equity was assessed using the Mann-Whitney U test. The results confirmed that there were no significant differences in salaries across genders (*p* = 0.198), and similar patterns were observed across race and diversity status. These findings affirm that current compensation practices within the company are fair and unbiased.

#### Predictive Modeling

Initial modeling began with logistic regression to establish a benchmark. While the model was interpretable and highlighted some relevant features, it struggled to accurately identify the positive class (terminated employees). Insights from this baseline included elevated turnover risks among employees hired through certain sources, under specific managers, and within low-performing departments/roles.

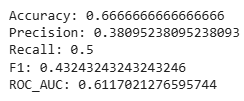


Figure 1: Results from Baseline Model (Logistic Regression)

To address these limitations, Random Forest classifiers were developed and tuned. The first Random Forest model reduced the number of features from 109 to 16 by removing descriptive and redundant variables, opting for using the ID features over the string descriptive version of the features. Class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique). This model did exceptionally well, however results were dominated by a dominant feature (DaysSinceLastReview) and may have been overfitting.

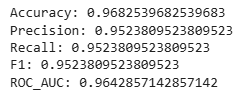


Figure 2: Results from 1st iteration of Random Forest Model

The second iteration of the model expanded to 29 features and reintroduced variables such as Manager ID, Performance Standard, and Recruitment Source based on the key features that were presented in the baseline model. Additionally, the dominant feature, DaysSinceLastReview was removed.

The final model achieved an Area Under the Curve (AUC) of 90.5%, with a precision of 71.4% and recall of 93.8%. These metrics indicate strong classification performance, especially in identifying true positives. While the model does produce some false positives, it effectively flags the majority of high-risk cases for HR intervention.

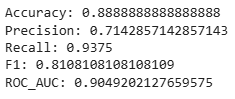


Figure 3: Results from 2st iteration of Random Forest Model

#### Risk Scoring and Distribution

Employees in the test set were assigned risk scores ranging from 0 to 1, representing the model’s predicted probability of termination. These scores were categorized into three groups: low risk (below 0.3), medium risk (0.3 to 0.7), and high risk (above 0.7).



Figure 4: Risk Distribution Flag Model

The resulting distribution revealed that the majority of employees fall within the low-risk category, reflecting a stable overall workforce. A smaller group was identified within the medium to high-risk ranges identifying these as at-risk employees, warranting targeted retention efforts such as stay interviews, career development planning, and tailored communication.

#### Interactive Dashboard

To support data exploration and decision-making, an interactive dashboard in Tableau was developed. This allows HR users to view key performance indicators such as turnover rates, engagement scores, and compensation equity. Features such as department filters and demographic segmentation enable stakeholders to drill down into specific groups and uncover trends.

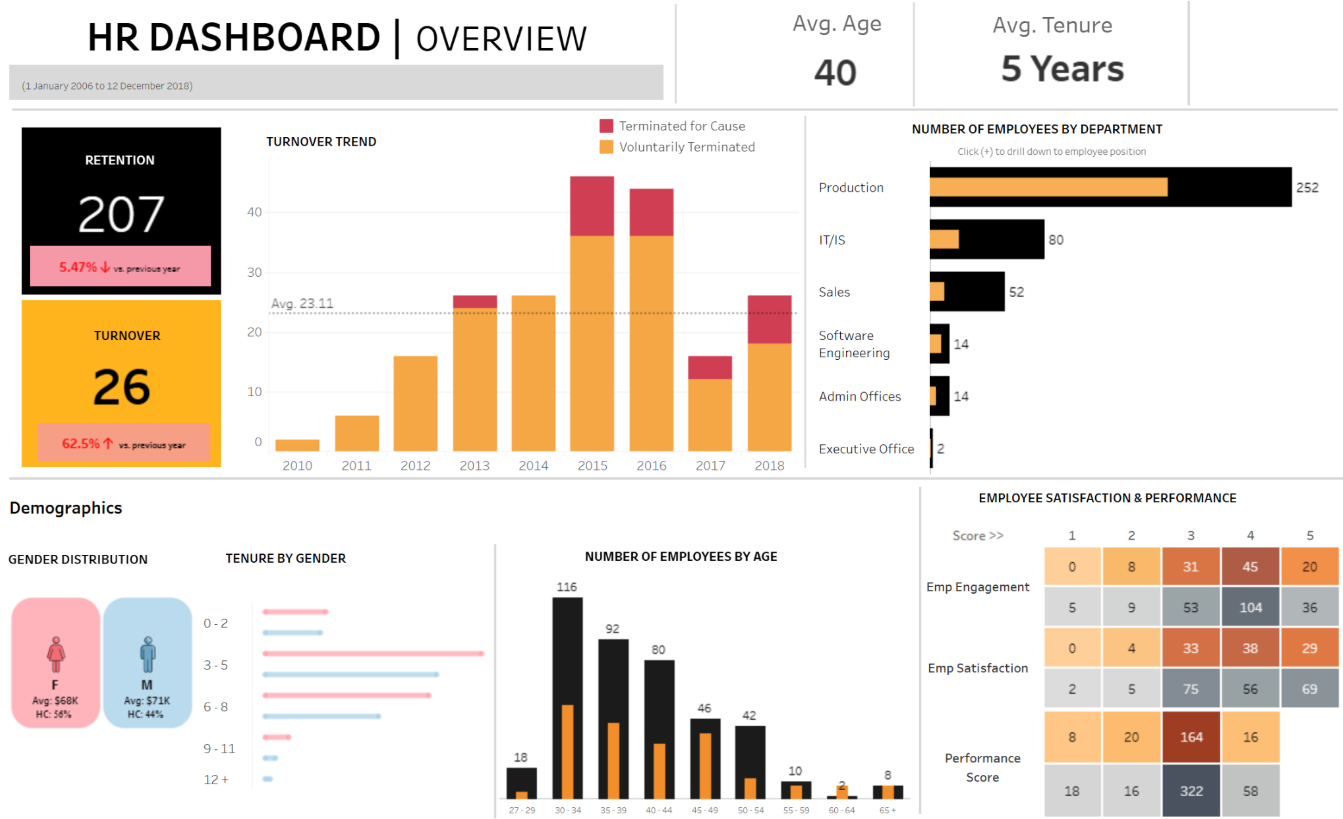


Figure 5: Tableau Interactive Dashboard ([Link](https://public.tableau.com/views/HR_dashboard_17441545885240/HROverview?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link))

### **Recommendations**

Based on the findings, several recommendations are proposed to improve organizational health and retention:

First, HR should proactively engage employees identified as high risk through personalized interventions, such as career coaching and stay interviews. Second, satisfaction initiatives should be prioritized, as engagement is directly linked to performance. Third, compensation policies should continue to be monitored to ensure ongoing pay equity. Fourth, HR should regularly review departmental trends and manager performance to identify patterns associated with higher turnover. Finally, recruitment strategies should be optimized by focusing on sources with better retention outcomes. Further analyses, visualizations, and suggestions have been published in my [github](https://github.com/HyItsAngela/HR_Analytics), particularly in the analysis notebook.

### **Conclusion**

This case study illustrates the power of People Analytics in transforming HR practices. By integrating statistical analysis, predictive modeling, and dashboard development, the project delivers actionable insights that support data-driven workforce management. The experience reinforces the value of an end-to-end data science workflow, from data cleaning and feature engineering to model deployment and strategic communication. Through this project, I have developed a deeper understanding of how data science can be used to drive organizational impact and enhance HR practices.