Analysis of Employee Attrition and Demographics

**Introduction**

Employee attrition is a critical issue for organizations, affecting productivity, morale, and financial performance. This report aims to explore the reasons behind employee attrition using data from an HR dataset. We will examine key factors such as job satisfaction, work-life balance, and relationship satisfaction, alongside demographic information such as age, monthly income, and total working years. The analysis will be conducted using Power BI for data exploration and visualization and Python for data encoding and predictive analysis.

**Methodology**

* **Part 1: Python**
* Data Import and Initial Exploration

Loaded the dataset and inspected its structure, data types, and missing values. Generated summary statistics to understand the distributions of numerical features.

* Data Cleaning

Handled missing values by removing rows with null entries. Removed duplicate entries to ensure data integrity.

* Data Exploration and Outlier Detection

Used histograms and box plots to visualize data distributions. Identified outliers using the Interquartile Range (IQR) and Z-score methods.

* Data Encoding

Encoded binary categorical variables using Label Encoding. Encoded multi-class categorical variables using One-hot Encoding.

* Data Labelling

Labelled the target variable (Attrition) for classification by using binary encoding.

* Data Splitting

Split the dataset into training and testing sets using a 70-30 split.

* Model Fitting

Trained a logistic regression model to predict employee attrition. Evaluated the model using a classification report, confusion matrix, and accuracy score.

* Save Pre-processed Data

Saved the cleaned and pre-processed dataset to a CSV file for future use.

* **Part 2: Power BI**
* Data Import and Initial Exploration

The dataset was imported into Power BI Desktop using the Text/CSV option. The file was loaded to begin the analysis. The dataset was inspected for its structure, data types, and initial missing values using the Data View in Power BI.

* Data Cleaning

In the Power Query Editor, rows with missing values were identified and handled. Depending on the context, rows with missing critical information were removed, or missing values were imputed using mean or median values. Duplicate rows were identified and removed to ensure data integrity.

* Data Exploration and Outlier Detection

Histograms for numerical features such as Age, Monthly Income, and Total Working Years were created to understand the distribution of data.

* Creating Key Metrics

A card visual was created to display the overall turnover rate. A card visual was used to show the average job satisfaction score. Another card visual was created to display the average performance rating. Slicers were added for Department, Business Travel, and Education Field to allow users to filter the data dynamically.

* Attrition Analysis Visuals

Bar charts and pie charts were created to visualize reasons for attrition based on Job Satisfaction, Work Life Balance, and Relationship Satisfaction.

* Demographic Analysis Visuals

Tables and cards were used to display turnover rates by Department, Gender, and Marital Status.

* Predictive Analytics with BI

The trained logistic regression model was integrated into Power BI using Python scripts to provide predictive insights. Line charts and stacked bar charts were used to visualize predicted turnover trends over time.

* User Interaction and Dashboard Layout

Tooltips and Drillthroughs: Detailed tooltips were added to visuals, and drillthrough actions were enabled to allow users to explore data points in more detail.

**Results**

* **Data Quality**

The dataset was clean with no missing values, ensuring the completeness of the analysis. Also there were no duplicate entries, maintaining data integrity.

* **Data Encoding**

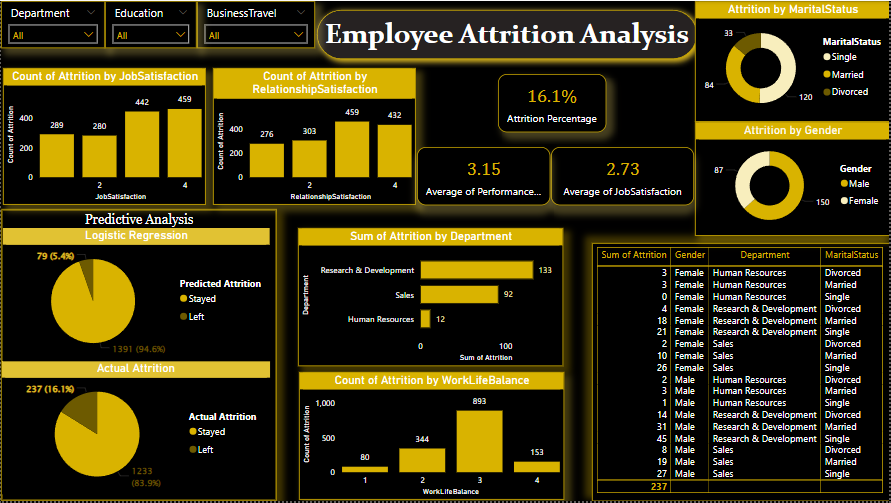
Columns such as Gender, Attrition, etc. were encoded using label encoding to convert them into numerical format and columns such as Business Travel, Department, Education Field, etc. were encoded using one-hot encoding to handle multi-class categories effectively**.**

* **Logistic Regression Model**

The logistic regression model was trained to predict employee attrition. The model achieved an accuracy of 85.71%. The precision for predicting non-attrition (Class 0) was high at 0.88, while the precision for predicting attrition (Class 1) was lower at 0.45. The recall rates were 0.97 for non-attrition and 0.15 for attrition, indicating that the model was more effective at identifying employees who stayed rather than those who left. The overall performance highlights the challenge of predicting attrition, given the imbalance in the dataset.

The confusion matrix reveals the model's performance in predicting employee attrition. It shows that the model correctly identified 369 employees who did not leave the company (true positives for non-attrition) and misclassified 11 employees who left the company as staying (false positives for non-attrition). For employees who actually left the company, the model correctly identified only 9 of them (true positives for attrition), while it failed to recognize 52 employees who left, mistakenly predicting that they would stay (false negatives for attrition). This indicates the model is more accurate in predicting employees who will stay rather than those who will leave, highlighting a challenge in detecting actual attrition cases.

* **Visualizations**



The overall attrition rate is 16.1%, indicating that a significant portion of the workforce has left the company. Job satisfaction levels play a crucial role in employee retention, with those rating their job satisfaction as 1 experiencing higher attrition compared to those with ratings of 3 or 4. Similarly, low relationship satisfaction correlates with higher attrition rates, emphasizing the importance of positive interpersonal relationships at work.

Key metrics reveal an average performance rating of 3.15, suggesting generally high performance among employees, while the average job satisfaction score of 2.73 indicates there is room for improvement in this area.

Demographic insights show that single employees (84) have the highest attrition rates, followed by married (120) and divorced employees (33). Gender analysis indicates that female employees (150) leave at higher rates than their male counterparts (87), suggesting potential gender-specific challenges within the organization.

Departmental insights reveal that the Research & Development (133) and Sales (100) departments have the highest attrition rates, while Human Resources has the lowest (12), indicating better retention strategies in place.

Work-life balance is a critical factor, with employees rating their balance as 1 showing the highest attrition count (893). Those with better work-life balance (ratings 3 and 4) experience lower attrition rates, highlighting the importance of maintaining a healthy work-life balance for retention.

The logistic regression model predicts an attrition rate of 5.49%, which significantly underestimates the actual attrition rate of 16.19%. The confusion matrix reveals that the model accurately predicts most non-attrition cases but struggles with predicting actual attrition, reflected in the lower precision and recall for attrition prediction. This suggests that while the model is useful, it requires further refinement for better predictive accuracy.

**Conclusion**

In conclusion, the analysis identifies several key factors influencing employee attrition: job satisfaction, relationship satisfaction, demographics, departmental affiliation, and work-life balance. Addressing these areas can help the company develop targeted strategies to improve retention. Enhancing job and relationship satisfaction, supporting work-life balance, and focusing on high-attrition departments can significantly reduce turnover. Additionally, refining the predictive model can provide more accurate forecasts, enabling proactive management of potential attrition risks.