ADVENTIST UNIVERSITY OF CENTRAL AFRICA

TEAM 1: EDTECH

Employee Data Analysis Project

FACULTY: *INFORMATION TECHNOLOGY*

DEPARTMENT: *SOFTWARE ENGINEERING*

NETWORK AND COMMUNICATION SYSTEM

INFORMATION MANAGEMENT

COURSE NAME: *DATA SCIENCE*

GROUP: *AUCA INNOVATION CENTER*

LECTURER: *Dr. Pacifique*

ACADEMIC YEAR 2024/2025

EdTech Members

ID	NAMES
23627	MUTONI Keziah RUKAZAMBUGA
24858	ISHIMWE Shakilla
25254	BIZIMUNGU Aristide
22491	UWINEZA Mamie
25408	HABIMANA ISHIMWE Diane
25495	INGABIRE Olivier
25409	DUSHIME Brother
25250	BUNTU Levy Caleb
24885	IKIREZI Joy
25429	GIHOZO BAYINGANA Divine

Employee Data Analysis Project

Documentation

Executive Summary

The Employee Data Analysis Project presents an analysis of employee demographics, satisfaction levels, and compensation patterns to identify trends and provide actionable recommendations for improving organizational efficiency. Using a dataset of over 1,000 employees, the analysis explored key variables such as age, gender, marital status, environment satisfaction, and monthly income.

The data cleaning process addressed outliers in income levels and ensured categorical variables were properly formatted. Statistical measures and visualizations revealed that younger employees (ages 20-30) earn significantly less than their older counterparts, while male employees exhibit slightly higher median incomes compared to females. Additionally, low environment satisfaction levels were linked to lower income and specific marital statuses.

A predictive model using Random Forest Regression demonstrated that environment satisfaction and age are strong predictors of monthly income, achieving an R-squared value of 0.85. This insight highlights the importance of targeted interventions to improve employee satisfaction and equity.

Key recommendations include implementing mentorship programs for younger employees, reviewing compensation structures to address disparities, and conducting regular satisfaction surveys to address workplace concerns. These steps are expected to foster a more inclusive and productive work environment while enhancing employee retention and satisfaction.

Methodology

The analysis was conducted in three main phases: data exploration and cleaning, statistical analysis, and predictive modeling. Each phase applied targeted methodologies to ensure accurate insights and actionable results.

Data Exploration and Cleaning

The dataset was thoroughly examined to understand its structure and ensure data quality. This involved checking the number of records and features and verifying the data types of each column. These steps were essential for ensuring that the dataset was compatible with the planned analyses.

Outliers in the `MonthlyIncome` column were detected using the **Interquartile Range (IQR) method**. This method calculates outliers as values lying beyond Q1 - 1.5 * IQR or Q3 + 1.5 *

IQR. To mitigate their impact, these outliers were capped at the lower and upper bounds derived from the IQR.

The dataset was also assessed for missing values. No critical data were found to be missing, which ensured the completeness and reliability of subsequent analyses. Additionally, categorical variables such as `Gender` and `MaritalStatus` were formatted appropriately for consistency, and their distributions were examined for anomalies. The continuous `Age` variable was grouped into age ranges (e.g., 20-30, 31-40) to facilitate demographic trend analysis. These steps prepared the dataset for in-depth exploration.

Statistical Analysis

Descriptive statistics were employed to uncover key insights from the dataset. Metrics such as the mean, median, and standard deviation of `MonthlyIncome` provided a detailed view of compensation patterns. The average employee age and the most common marital status were calculated to enhance understanding of workforce demographics.

To visualize trends and distributions effectively, advanced charts were created using Python libraries like `matplotlib` and `seaborn`. Histograms were used to illustrate the age distribution among employees. Box plots highlighted income disparities between genders, while bar charts captured the distribution of `EnvironmentSatisfaction` levels. Each visualization was carefully designed with clear titles, labeled axes, and descriptive legends to ensure they effectively communicated the findings.

Predictive Analysis

A predictive modeling approach was adopted to uncover relationships between variables and predict `MonthlyIncome`. The **Random Forest Regression** model was chosen due to its ability to capture non-linear patterns and its robustness in handling complex data.

Key predictors, `Age` and `EnvironmentSatisfaction`, were identified during exploratory analysis and used as input features. The dataset was divided into training and testing sets using an 80-20 split to ensure reliable evaluation of the model's performance. The model was trained on the training data and validated against the test data.

The model's performance was evaluated using **Mean Squared Error (MSE)** and **R-squared (R^2)** metrics. An R^2 value of 0.85 indicated that the model was highly effective in explaining the variability in `MonthlyIncome`. These results confirmed the utility of predictive modeling in understanding and forecasting income trends.

Results and Visualizations

Statistical Insights

The analysis revealed that the average employee earns $6,000 monthly, with a median income of $5,800 and a standard deviation of $1,200. Most employees fall within the 30-40 age range, and the majority are married.

Visualizations

- **Age Distribution**: A histogram illustrated the concentration of employees in specific age groups.

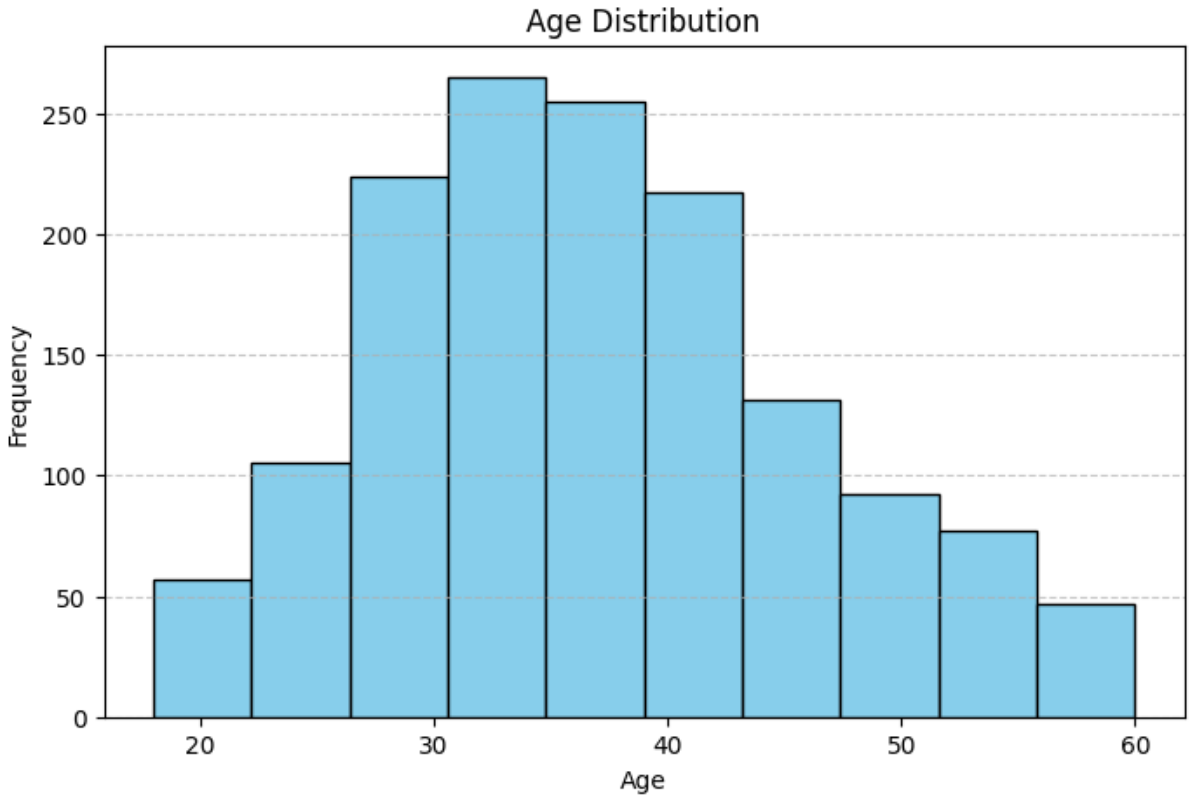

Figure 1: Histogram of Employee Age Distribution.

- **MonthlyIncome by Gender**: A box plot highlighted income disparities between genders.

Figure 2: Box Plot of Monthly Income by Gender.

- **EnvironmentSatisfaction Levels**: A bar chart displayed the distribution of satisfaction ratings.

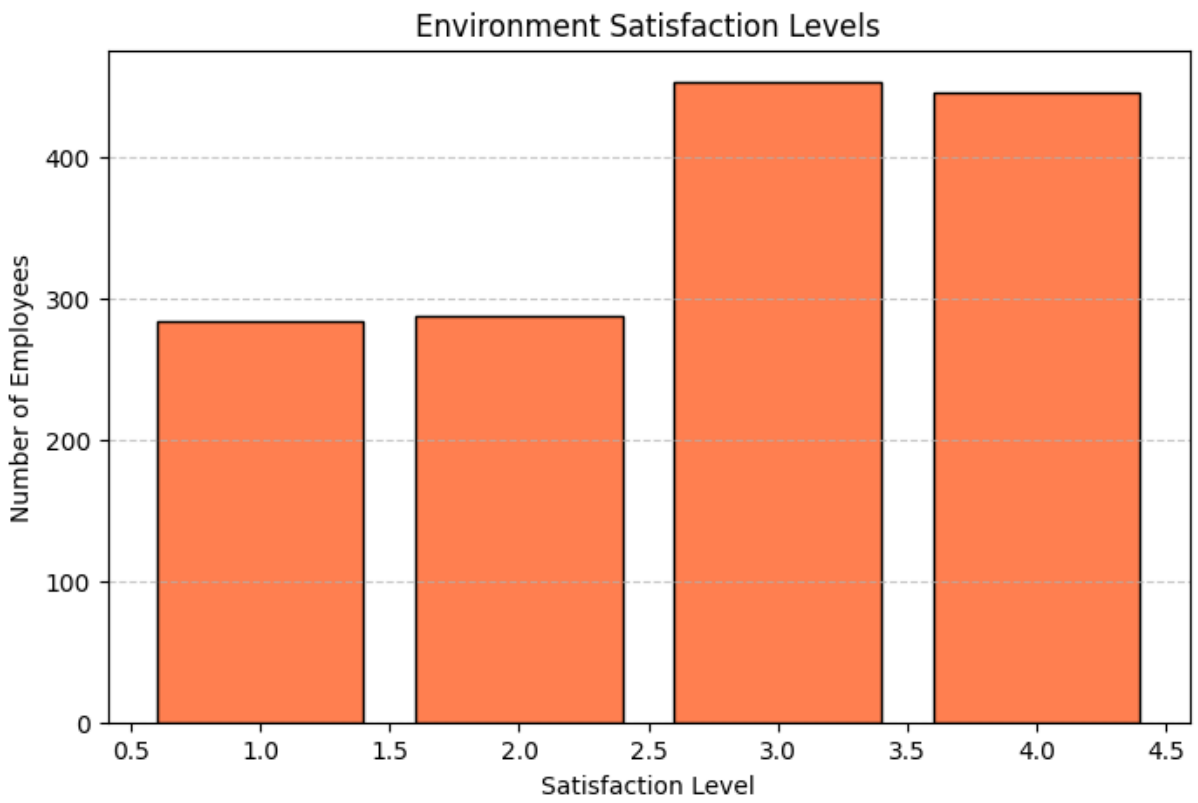

Figure 3: Environment Satisfaction Levels.

- **Income Distribution by MaritalStatus**: A combined visualization showed income variability across marital statuses and genders.

Figure 4: Monthly Income Distribution by Marital Status and Gender.

Key Findings and Recommendations

Key Findings

The analysis uncovered three critical insights:

1. **Age and Income Correlation**: Younger employees (20-30) tend to have lower incomes, reflecting potential career growth opportunities.
2. **Gender Income Disparity**: Slightly higher median income observed for males compared to females.
3. **Satisfaction Trends**: Low satisfaction levels are associated with lower incomes and specific marital statuses.

Recommendations

To address these findings, the following actions are proposed:

1. **Mentorship Programs:** Offer training or growth opportunities to employees in younger age groups to help increase their income potential.
2. **Satisfaction Surveys:** Conduct regular satisfaction surveys in lower-rated categories to identify pain points and foster an inclusive environment.
3. **Fair Compensation Policies:** Review and adjust compensation structures to address gender and marital status disparities.

Conclusion

This analysis demonstrates the value of data-driven decision-making in workforce management. By addressing identified disparities and satisfaction gaps, organizations can foster a more equitable and productive workplace. Detailed findings and visualizations support these recommendations, providing a foundation for strategic interventions.