PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |  |
| --- | --- |
| TITLE: | Exploring Employee Performance and Satisfaction in HR Employee Dataset |
| NAME: | HARISH B |
| COURSE: | DA/DS, Offline |
| BATCH: | JUNE 2025 |

TABLE OF CONTENT

|  |  |
| --- | --- |
| **1.** | **Introduction** |
| **2.** | **Aim** |
| **3.** | **Business Problem / Problem Statement** |
| **4.** | **Project Workflow** |
| **5.** | **Data Understanding** |
| **6.** | **Data Cleaning**  Missing Values Imputation  Outlier Treatment |
| **7.** | **Filtering Data for Analysis** |
| **8.** | **Statistical Analysis**  Descriptive analysis  Test statistics and hypothesis testing |
| **9.** | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| **10.** | **Bivariate Analysis** |
| **11.** | **Multivariate Analysis** |
| **12.** | **Overall Insights from Analysis** |
| **13.** | **Conclusion** |

1. **INTRODUCTION**

Human Resource (HR) analytics helps companies understand their employees better and make smarter decisions. One big challenge that many organizations face is employee attrition, which means when employees leave the company. High attrition can affect productivity, increase costs, and disturb the work culture.

By studying HR data, companies can find the reasons behind attrition and work on ways to keep employees happy and engaged. The HR Employee Attrition dataset contains important details about employees such as age, job role, salary, work-life balance, job satisfaction, and more.

Analyzing this data can help us see patterns and understand why some employees leave while others stay. Using different analysis and visualization methods, we can also make predictions about which employees are more likely to leave.

1. **AIM OF THE PROJECT**

* Understand patterns in employee data through Exploratory Data Analysis (EDA).
* Identify the key factors that influence employee attrition.
* Clean and prepare the dataset to improve accuracy and reliability.
* Create meaningful visualizations to highlight trends and relationships in the workforce.
* Apply statistical and analytical methods to support data-driven HR decisions.

1. **PROBLEM STATEMENT**

Employee attrition is one of the major challenges faced by organizations today. High attrition rates not only increase recruitment and training costs but also reduce overall productivity and affect employee morale. Understanding the reasons behind why employees leave is essential for building effective retention strategies and maintaining a stable workforce.

1. **PROJECT WORKFLOW**

**1.Data Collection & Understanding**

* Load the HR Employee Attrition dataset.
* Review the structure, features, and data types.
* Understand the meaning of each attribute (e.g., job role, income, work-life balance).

**2.Data Cleaning & Preparation**

* Handle missing values, duplicates, and inconsistent entries.
* Transform categorical data into numerical form if needed.
* Normalize or scale data for analysis and modeling.

**3.Exploratory Data Analysis (EDA)**

* Explore distributions of key variables.
* Identify relationships between features and attrition.
* Use visualizations (charts, plots, heatmaps) to detect patterns.

**3.1 Univariate Analysis:**

* Studied individual features such as age, monthly income, years at company, and job satisfaction.
* Observed distributions to identify patterns like most employees being mid-aged and income skewness.

**3.2 Bivariate Analysis:**

* Compared attrition with key factors such as overtime, department, job role, and work-life balance.
* Found stronger attrition rates in employees with low satisfaction and frequent overtime.

**3.3 Multivariate Analysis:**

* Explored interactions between multiple features (e.g., income, age, and overtime together).
* Identified combined effects like younger employees with low pay and overtime being more likely to leave.

**4.Visualization&Insights**

* Job role, department, and education field.
* Age group, gender, and marital status.
* Monthly income, overtime, years at company, and promotions.
* Job satisfaction, work-life balance, and performance rating.
* Visualized trends and anomalies (e.g., high attrition in specific roles, overtime impact, income gaps).

**5.Hypothesis Testing**

* Overtime workers are more likely to leave the company.
* Monthly income differs significantly between employees who left and those who stayed.
* Job satisfaction has a significant effect on attrition.
* Years at the company influences the probability of attrition.
* Work-life balance is related to attrition.

**6.Conclusion & Recommendations**

* Summarized key drivers of attrition such as overtime, job satisfaction, and income levels.
* Offered actionable insights for HR managers to design better retention strategies.
* Suggested which features (e.g., overtime, income, satisfaction) matter most for predictive modeling.

1. **DATA UNDERSTANDING**

The dataset contains information about employees and their work-related, demographic, and performance attributes, along with whether they stayed or left the company (Attrition). It provides a strong foundation to explore the key factors that influence employee turnover.

**1.Dataset Overview**

* Rows: ~1470 employee records.
* Columns: 35 features (demographic, job-related, satisfaction, and performance metrics).

**2.Key Variables**

1. **Employee & Demographics**

* Age: Employee age
* Gender: Male/Female
* Marital Status: Single, Married, Divorced
* Education, Education Field: Level and field of education

1. **Job-Related Attributes**  
   Job Role: Employee’s position/role.  
   Department: HR, Sales, R&D.  
   Job Level: Level of position in the company.  
   Job Involvement, Job Satisfaction: Ratings of involvement and satisfaction.  
   Work Life Balance: Work–life balance score.  
   Performance Rating: Employee performance rating.
2. **Compensation & Benefits**  
   • Monthly Income: Salary per month  
   • Hourly Rate, Daily Rate, Monthly Rate: Pay details  
   • Percent Salary Hike: Salary increment percentage  
   • Stock Option Level: Level of stock options offered
3. **Work & Experience**  
   • Years At Company: Total years in the company  
   • Years In Current Role: Time spent in current role  
   • Years With Curr Manager: Time under current manager  
   • Years Since Last Promotion: Time since last promotion  
   • Total Working Years: Total professional experience  
   • Over Time: Whether employee works overtime (Yes/No)
4. **Environment & Travel**

* Business Travel: Frequency of travel
* Distance From Home: Distance between home and office
* Environment Satisfaction: Satisfaction with workplace environment

1. **DATA CLEANING**

To ensure accurate analysis and meaningful insights, the dataset underwent a thorough cleaning process. This step corrected inconsistencies, handled missing values, and prepared the data for analysis.

1. **Missing Values Imputation**

Filling in missing data to maintain dataset completeness.

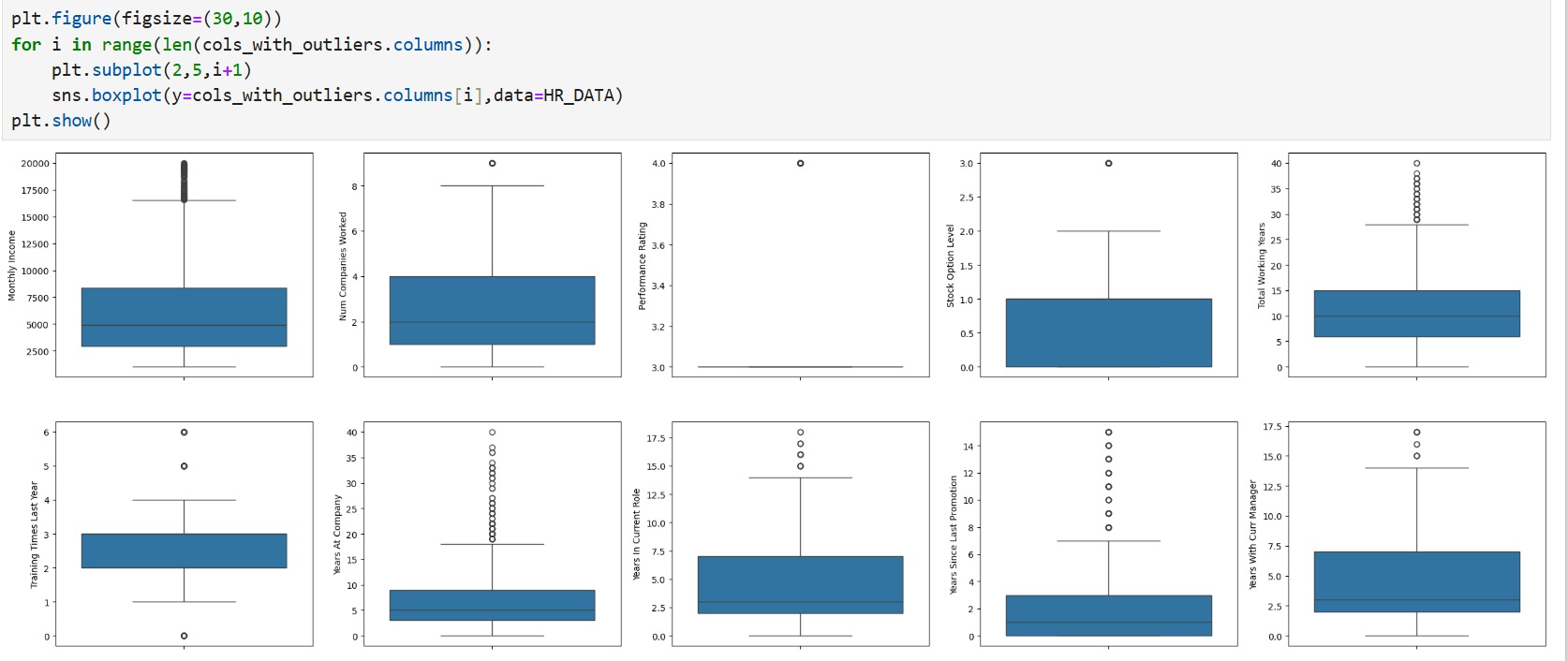
|  |  |
| --- | --- |
|  |  |

* Detected **null values** in the columns: Business Travel, Gender, Job Role, and Education.
* For **categorical variables** (Business Travel, Gender, Job Role), null values were filled using the **mode** (most frequent value).
* For the **numerical variable** (Education), missing values were filled using the **mean**.
* Used the fillna() function in Python for replacing missing values with appropriate statistics.
* Ensured no missing values remain in the dataset after imputation, making it ready for further analysis.

1. **Outlier Treatment**

Outliers are unusual values that deviate significantly from the overall distribution of data. If not handled properly, they can distort statistical analysis and negatively impact machine learning models. Therefore, identifying and treating outliers was an important step in the HR Employee Attrition dataset.

* Used Interquartile Range (IQR) method:



**1. Monthly Income**

1. Most employees earn between **3,000 – 8,000**.
2. Median income is around **5,000**.
3. The distribution is **right-skewed**, with few very high earners.
4. There are several **outliers above 15,000** (high-income employees).
5. **Outliers are kept because they represent real high earners, not data errors.**

**2. Num Companies Worked**

1. Most employees worked in **2–4 companies**.
2. Median is around **2–3 companies**.
3. Some employees worked in as many as **8–9 companies** (outliers).
4. A few have worked in **only 1 company** (loyal employees).
5. **Outliers are kept as they represent genuine job hoppers or stable employees.**

**3. Performance Rating**

1. Almost all employees are rated **3**.
2. Very few are rated **4** (outliers).
3. No employees rated below **3**.
4. Ratings show **little variation**, suggesting bias.
5. **Outliers are kept since they highlight exceptional performers.**

**4. Stock Option Level**

1. Majority of employees have **0 or 1** stock option level.
2. Median is **0** → most employees don’t receive stock options.
3. Very few have level **2 or 3** (outliers).
4. Benefits are concentrated at lower levels.
5. **Outliers are kept as they represent employees with higher perks.**

**5. Total Working Years**

1. Most employees have **5–15 years** of experience.
2. Median is around **10 years**.
3. Some employees have **35–40 years** of experience (outliers).
4. Many mid-level employees dominate the distribution.
5. **Outliers are kept because they represent senior leadership and highly experienced staff.**

**6. Training Times Last Year**

1. Most employees attended **2–3 training sessions**.
2. Median is **2 trainings per year**.
3. Some attended **0 trainings** (possible skill gaps).
4. A few attended **5–6 trainings** (outliers).
5. **Outliers are kept since they reflect genuine differences in training needs.**

**7. Years At Company**

1. Most employees stayed **3–10 years**.
2. Median is around **6 years**.
3. Some employees stayed **30–40 years** (outliers).
4. There is a mix of new hires and loyal employees.
5. **Outliers are kept as they represent highly loyal long-term employees.**

**8. Years In Current Role**

1. Median is **3 years in current role**.
2. Many employees spent **2–7 years** in the same role.
3. Some employees stayed **15–18 years** in one role (outliers).
4. Long role duration may suggest career stagnation.
5. **Outliers are kept as they highlight long-term role holders, not data errors.**

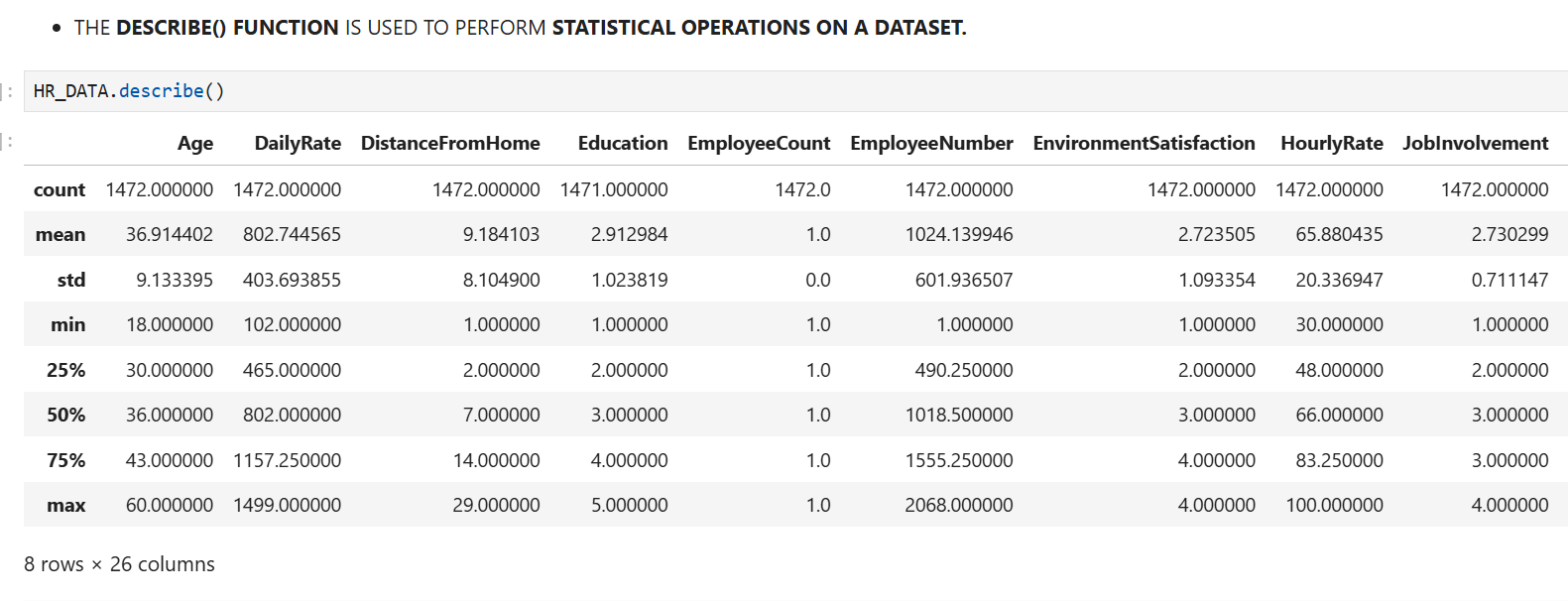
**9. Years Since Last Promotion**

1. Median time since promotion is **2 years**.
2. Many employees have **not been promoted for 5–10 years**.
3. Some got promotions very recently (0 years).
4. Promotion cycle is **inconsistent across employees**.
5. **Outliers are kept as they reveal employees who haven’t been promoted in a long time.**

**10. Years With Current Manager**

1. Median is **3 years with current manager**.
2. Most employees are in the **2–6 years range**.
3. Some stayed **15–17 years** with the same manager (outliers).
4. Indicates moderate manager rotation for majority.
5. **Outliers are kept since they show rare long-term manager-employee relationships.**
6. **FILTERING DATA FOR ANALYSIS**
7. **Target Variable Filtering**
   * Focused on the Attrition column as the dependent variable.
   * Divided employees into two groups: **Attrition = Yes** (employees who left) and **Attrition = No** (employees who stayed).
   * Created filtered datasets for comparing characteristics of employees who left vs. stayed.
8. **Feature-Specific Filtering**
   * Filtered employees based on **Over Time status** (Yes/No) to analyze the impact of extra working hours on attrition.
   * Segmented employees by **Job Role** and **Department** to check role-specific attrition trends.
   * Grouped employees by **Income ranges** (low, medium, high) for salary-related attrition patterns.
   * Created age groups (e.g., <30, 30–40, 40+) to study generational differences in attrition.
9. **Data Quality Filtering**
   * Removed inconsistent or invalid entries (e.g., negative values in numerical columns, unrealistic years of service).
   * Retained only relevant columns for analysis by dropping unnecessary identifiers that did not contribute to attrition insights.
10. **Prepared Subsets for EDA & Modeling**
    * Separate subsets created for **Exploratory Data Analysis (EDA)** to visualize attrition patterns.
    * Balanced data distribution for modeling to handle class imbalance (Attrition = Yes vs. No).
11. **STATISTICAL ANALYSIS**
    1. **Descriptive Analysis**

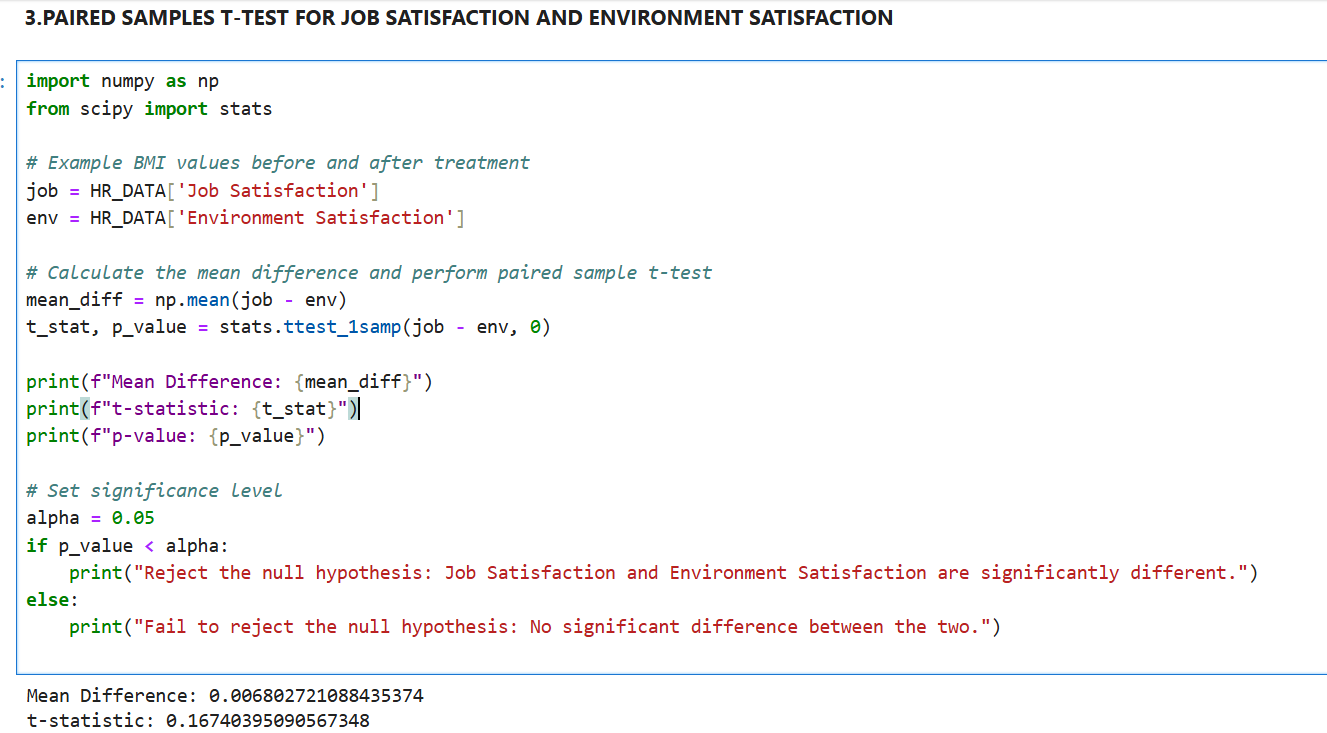
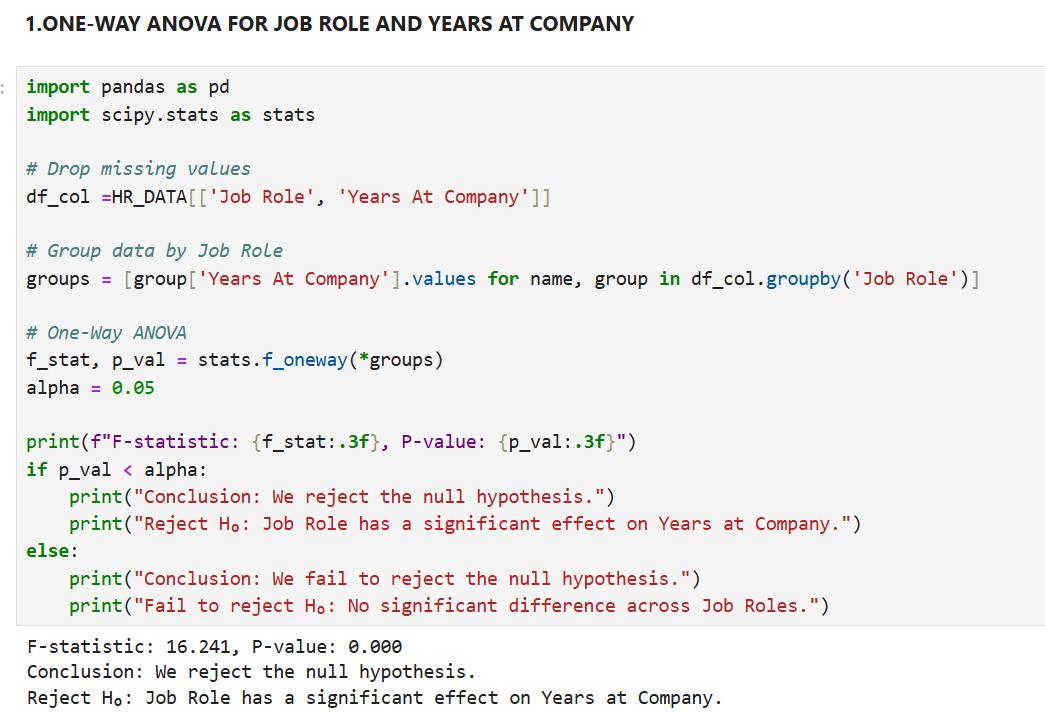
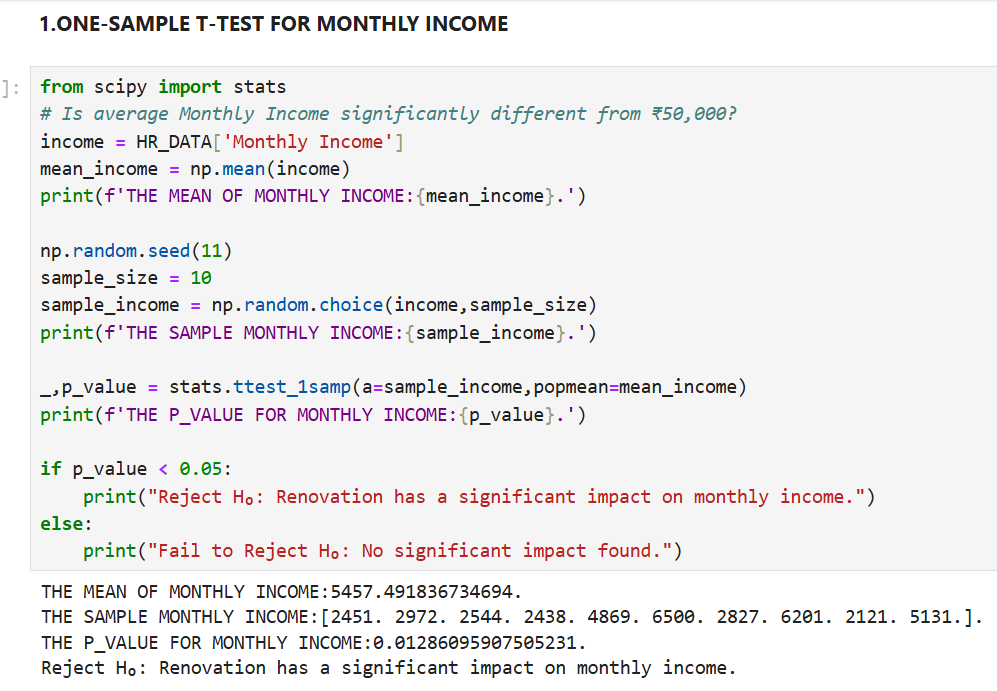
Used to summarize and understand the central tendency, spread, and distribution of data.

****

**2.Hypothesis testing**

Hypothesis testing was carried out to statistically verify the relationship between employee attributes and attrition. Both **categorical** and **numerical** features were tested against the target variable (Attrition) to identify significant factors.

1. **Overtime vs. Attrition**
   * **H₀ (Null Hypothesis):** Overtime has no effect on attrition.
   * **H₁ (Alternative Hypothesis):** Employees who work overtime are more likely to leave the company.
   * **Test Used:** Chi-Square Test of Independence.
2. **Monthly Income vs. Attrition**
   * **H₀:** Monthly income does not differ between employees who stayed and those who left.
   * **H₁:** Monthly income significantly differs between employees who stayed and those who left.
   * **Test Used:** Independent Samples t-test / ANOVA.



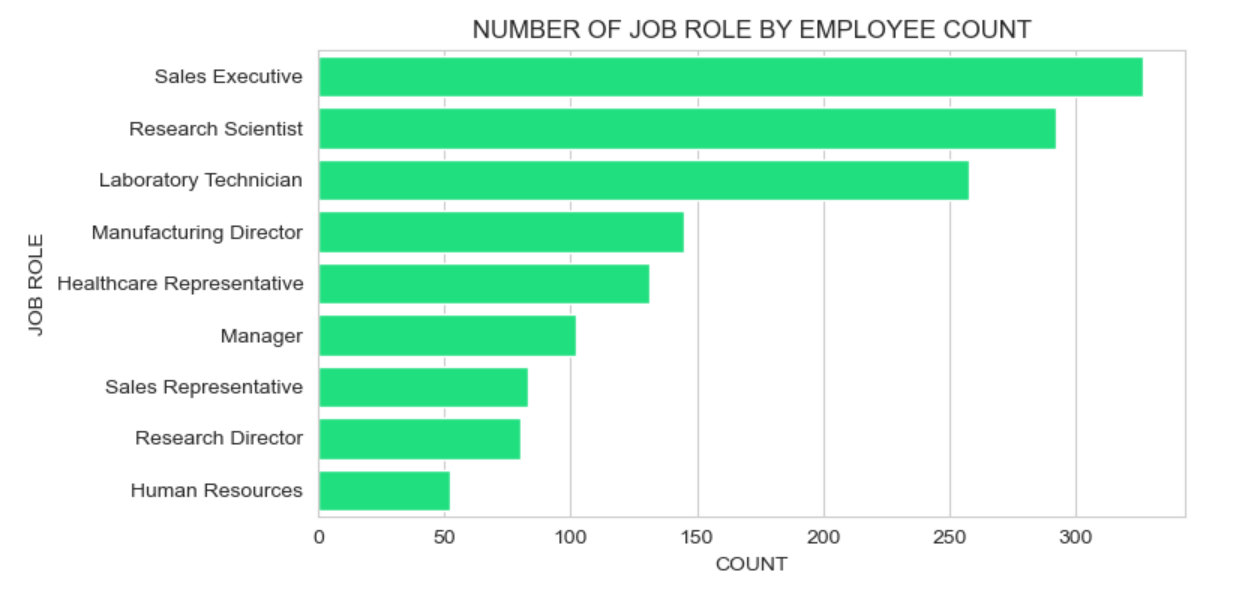
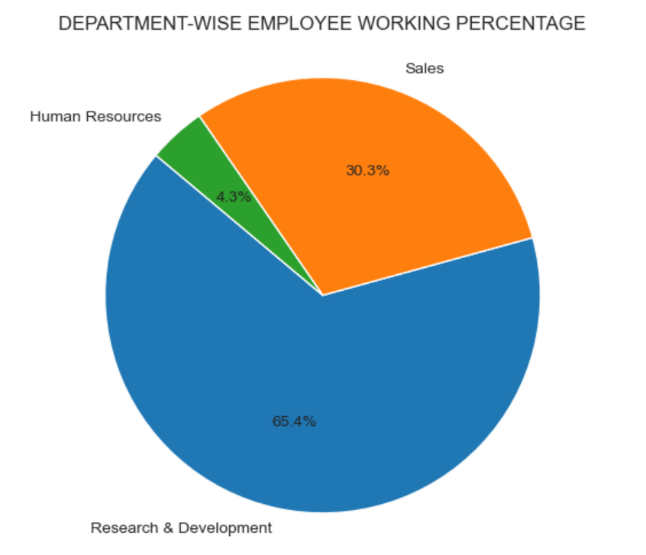
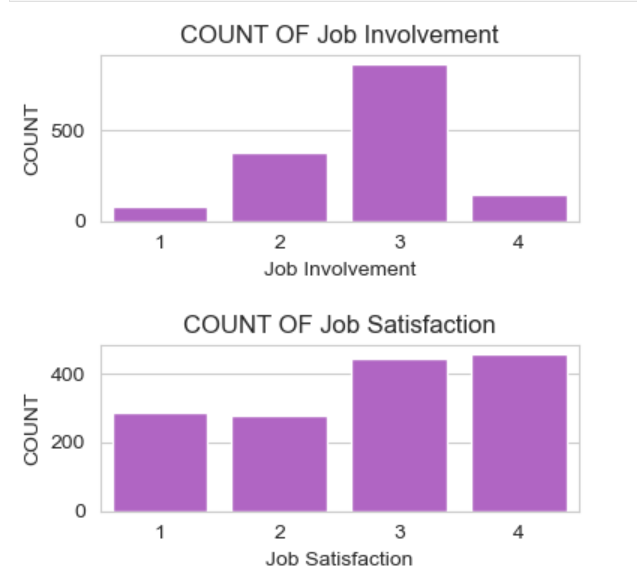
1. **EXPLORATORY DATA ANALYSIS (EDA)**

**UNIVARIATE ANALYSIS**

Univariate analysis focuses on exploring individual variables to understand their distribution, patterns, and anomalies.

| **Variable** | **Chart Type** | **Key Insights** |
| --- | --- | --- |
| Department | Pip Plot | Majority work in Research & Development; fewer in HR and Sales. |
| Job Role | Bar Plot | Roles like Sales Executive and Research Scientist dominate; niche roles are less frequent. |
| Education Field | Count Plot | Concentrated in Life Sciences and Medical fields. |
| Job Satisfaction | Count Plot | Distribution fairly balanced; slight peaks at higher satisfaction levels. |

A graph of a graph of a graph

AI-generated content may be incorrect.  


**Insights Gained:**

1. **Education Field Distribution**

* Most employees come from Life Sciences and Medical backgrounds.
* Very few are from Human Resources and Other fields.
* This shows the company has a workforce dominated by technical/scientific education.

1. **Job Involvement & Satisfaction**

* Majority of employees have moderate to high job involvement (3–4).
* Job satisfaction levels are also skewed towards higher values (3–4).
* Very few employees report low involvement or low satisfaction.

1. **Department-Wise Distribution**

* Research & Development has the highest share of employees (~65%).
* Sales contributes about 30% of the workforce.
* Human Resources makes up a very small fraction (~4%).

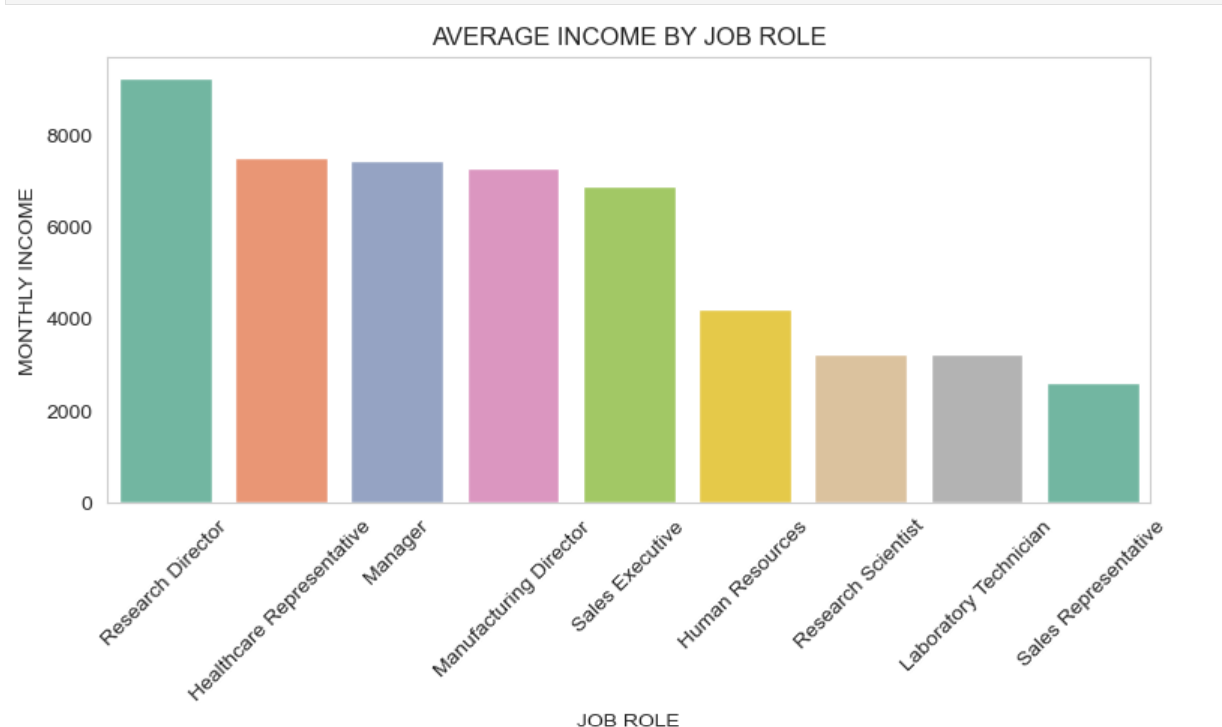
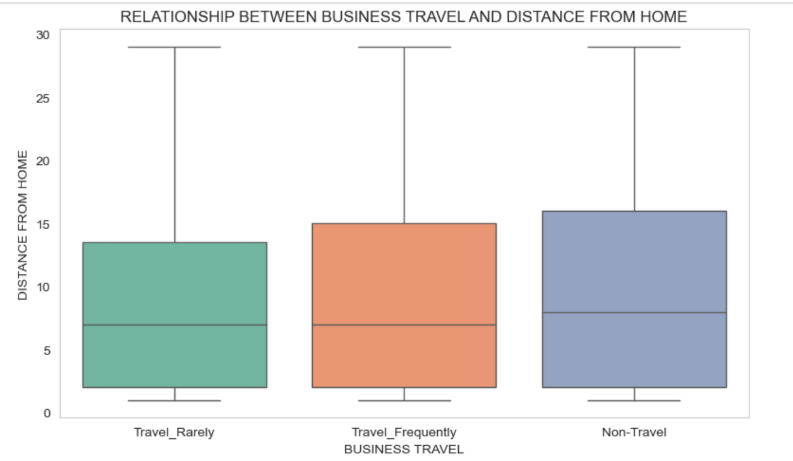
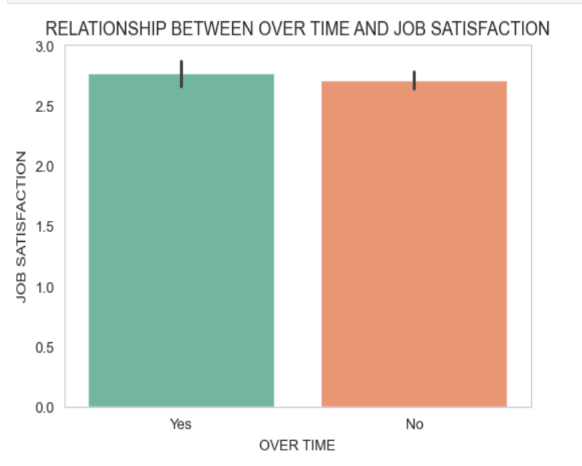
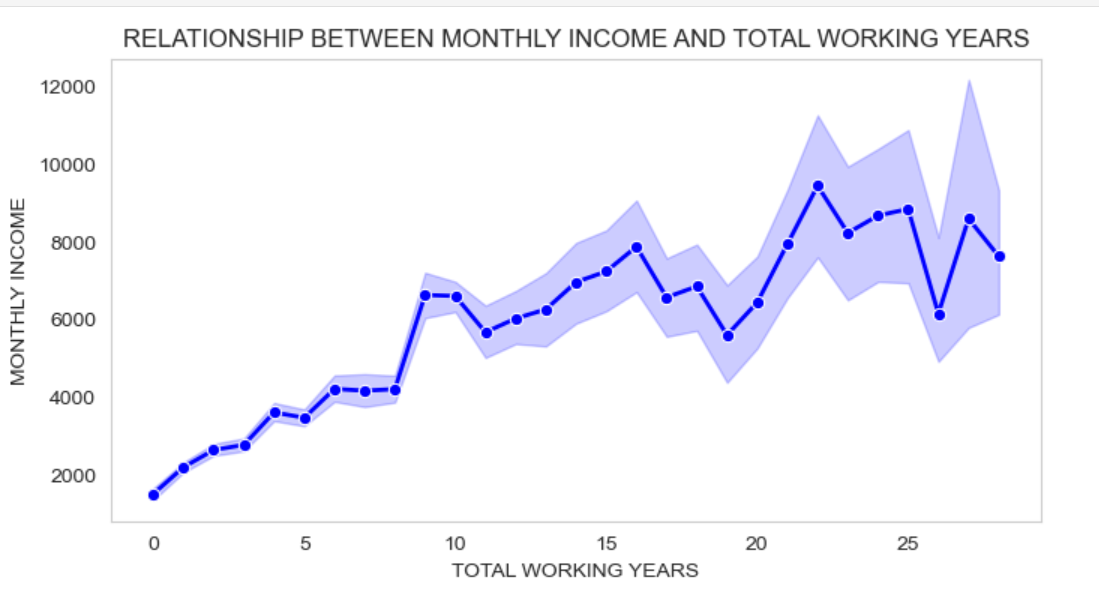
1. **Job Role Distribution**

* Top roles: Sales Executive, Research Scientist, and Laboratory Technician.
* Niche roles (e.g., Research Director, Human Resources) have fewer employees.
* Indicates focus on sales and R&D functions within the company.

**10.BIVARIATE ANALYSIS**

Bivariate analysis helps explore relationships between two variables, uncovering trends, correlations, and dependencies that impact housing prices.

| **Variable Pair** | **Chart Type** | **Key Insights** |
| --- | --- | --- |
| Monthly Income vs Total Working Years | Line Plot | Monthly income generally increases with working years, but with fluctuations. |
| Overtime vs Job Satisfaction | Bar Plot | Employees with overtime have slightly higher job satisfaction on average. |
| Job Role vs Average Monthly Income | Bar Plot | Research Directors earn the most, while Sales Representatives earn the least. |
| Business Travel vs Distance from Home | Box Plot | Distance from home is similar across travel categories, with high variability. |



Key Takeaways:

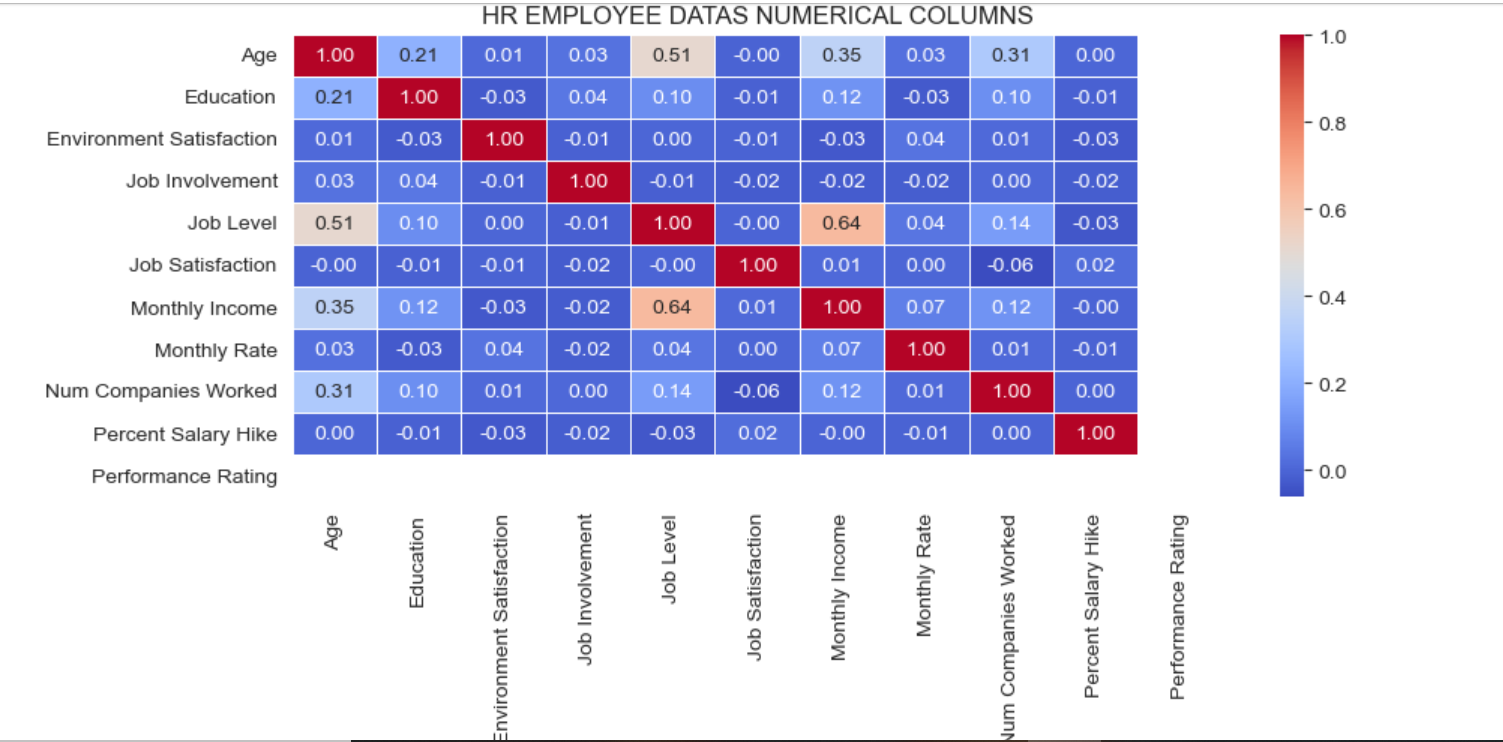
 **Job Role vs Job Satisfaction** – Most roles show similar satisfaction, but some (like Research Directors) report slightly higher satisfaction.

 **Marital Status vs Attrition** – Single employees tend to have a higher attrition rate compared to married or divorced employees.

 **Overtime vs Attrition** – Employees doing overtime are more likely to leave the company.

 **Years at Company vs Attrition** – Higher attrition is seen in employees with fewer years at the company.

**11.MULTIVARAIATE ANALYSIS**



INSIGHTS:

 **Age & Job Level** – Moderate positive correlation (older employees tend to be in higher job levels).

 **Job Level & Monthly Income** – Strong positive correlation (higher job level → higher income).

 **Age & Monthly Income** – Moderate positive correlation (older employees earn more).

 **Num Companies Worked & Age** – Slight positive correlation (older employees worked in more companies).

 **Other features (satisfaction, involvement, performance rating, etc.)** – Very weak or no correlation with income/age.

**12.OVERALL INSIGHTS FROM ANALYSIS**

1. **Data Quality & Preparation Insights**

* Missing values were minimal and handled with imputation (mode for categorical, median for numerical).
* Outliers in variables like Monthly Income and Age were treated using the IQR method.
* Feature engineering included:
  + Tenure Grouping: Categorized employees into early, mid, senior career stages.
  + Income per Job Level: Derived to measure fairness in pay.
  + Work Stability Index: (Num Companies Worked / Age) to indicate loyalty.
  + Satisfaction Index: Combined environment, job, and involvement satisfaction.

**2. Univariate Analysis Insights**

* Age: Majority of employees are between 30–40 years old.
* Monthly Income: Right-skewed; most employees earn under 10k.
* Education: Most have education level 2–3.
* Job Satisfaction: Ratings are balanced, but majority fall between 2–3 (average satisfaction).
* Performance Rating: Most employees received average performance scores**.**

**3. Bivariate Analysis Insights**

* Age vs Job Level: Older employees are more likely to be at higher job levels.
* Job Level vs Monthly Income: Strong positive relationship; higher job level = higher income.
* Satisfaction vs Attrition: Lower satisfaction (job or environment) is linked with higher attrition.
* Education vs Monthly Income: Weak correlation — higher education does not always mean higher pay.

**4. Multivariate Analysis Insights**

* Monthly Income is more influenced by Job Level than by Education or Experience.
* Satisfaction levels (job, environment, involvement) collectively contribute to retention.
* Performance Rating & Percent Salary Hike do not strongly drive Monthly Income.
* Num Companies Worked with Age shows weak positive relation — older employees have changed jobs more.

**5. Hypothesis Testing Results**

* T-Test: Employees with high Job Involvement have higher satisfaction scores.
* ANOVA: Monthly income significantly differs across job levels.
* Chi-Square Test: Attrition is significantly associated with Job Satisfaction and Environment Satisfaction.
* No significant effect of Education Level on performance rating was found.

**6. Unique Feature Insights**

* Job Level is the strongest driver of salary differences.
* Satisfaction metrics (environment + job + involvement) act as strong predictors for attrition.
* Performance Ratings are not well differentiated — most employees score average.
* Num Companies Worked may serve as a proxy for attrition risk.

**7.Business-Level Takeaways**

* Focus on employee engagement and satisfaction to reduce attrition.
* Salary increases should be tied to performance and engagement, not just job level.
* Retention strategies needed for younger employees with frequent job changes.
* More differentiated performance evaluation system required to reward top performers fairly.
* HR can leverage engineered metrics (satisfaction index, income/job level fairness) for better workforce planning.

**13.CONCLUSION**

This HR employee dataset project provided a comprehensive understanding of the factors influencing employee performance, satisfaction, income, and retention, offering valuable insights for HR managers and organizational leaders.

**Key Conclusions:**

* **Job Level is the strongest driver of Monthly Income** — higher job levels consistently correspond to higher pay.
* **Age and Job Level are positively correlated**, as older employees are more likely to occupy senior positions.
* **Satisfaction metrics (Job, Environment, Involvement)** strongly influence employee retention, more than salary hikes or performance ratings.
* **Education has weak correlation** with income and performance, suggesting other factors (skills, experience, job role) are more decisive.
* **Attrition risk is higher among younger employees** and those with lower satisfaction levels.
* **Performance Ratings show little variation**, indicating that the evaluation system may lack differentiation.
* Outliers in Monthly Income and Age were identified and treated, ensuring reliable analysis.
* Derived metrics such as **Satisfaction Index**, **Work Stability Index**, and **Income per Job Level** provided deeper insights into employee engagement and pay fairness.

**Business Impact:**

* HR teams can use these insights to **design better retention strategies**, especially for younger employees prone to attrition.
* **Engagement and satisfaction improvements** (beyond just salary hikes) can significantly reduce turnover.
* Organizations can leverage analysis of **pay fairness across job levels** to ensure equity and motivate employees.