Employee Attrition Analysis in IT Companies

Project by: Karthi Raju

Institution: Anudip Foundation

Project Description

Understanding employee turnover patterns based on job roles, experience, and salary.

1. Abstract:

Employee attrition is a significant issue in IT companies, affecting workforce stability and increasing hiring costs. This project analyzes key factors contributing to employee turnover, such as salary, job roles, and experience levels.

Data preprocessing includes handling missing values, filtering inconsistent records, and standardizing attributes. Exploratory Data Analysis (EDA) helps identify trends, while visualization techniques (bar charts, scatter plots, histograms) make insights more comprehensible.

Findings indicate that employees with lower salaries and mid-career experience levels are more likely to leave. By applying this analysis, IT companies can design better retention strategies and reduce attrition rates.

2. Introduction:

Employee attrition negatively impacts IT companies by increasing recruitment costs and reducing productivity. A high turnover rate also affects business growth and long-term planning.

This project aims to analyze attrition patterns using data analytics, focusing on salary, experience, and job roles. By using Python and libraries like Pandas, Seaborn, and Matplotlib, this study extracts valuable insights that can help HR professionals develop data-driven retention strategies.

3. Problem Statement:

High employee attrition in the IT industry affects workforce planning and talent management. Understanding why employees leave is crucial for implementing effective HR strategies.

This study seeks to identify the major contributors to employee turnover using real-world data analysis.

4. Objectives:

The primary objectives of this project are:

- To analyze attrition trends in IT companies.
- To determine the impact of salary, job role, and experience on employee turnover.
- To visualize attrition trends using data analytics.
- To provide insights for HR professionals to improve retention strategies.

5. Methodology:

5.1 Data Collection:

The dataset includes employee records containing attributes such as job roles, salary, experience, and attrition status. The data is sourced from HR databases and publicly available workforce datasets.

5.2 Data Preprocessing:

- Handling missing values using dropna() and fillna(0).
- Filtering out inconsistent or incorrect records.
- Standardizing numerical attributes (salary, experience) for better analysis.

5.3 Data Analysis:

- **Statistical Analysis:** Identifying patterns in salary and attrition rates.
- **Correlation Analysis:** Examining the relationship between experience levels and attrition.
- **Categorical Data Analysis:** Analyzing job roles and department-specific attrition trends.

5.4 Data Visualization:

Data visualization is essential for interpreting trends effectively. The following techniques are used:

- **Bar Charts:** Showing attrition rates based on salary groups.
- **Scatter Plots:** Displaying the relationship between experience and attrition.
- **Histograms:** Illustrating salary distributions for employees who left vs. stayed.

5.5 Implementation Code:

This section provides the Python implementation of the Employee Attrition Analysis. The analysis was performed using NumPy, Pandas, Matplotlib, and Seaborn for data processing and visualization.

5.5.1 Importing Required Libraries:

The following libraries were used for data handling, visualization, and statistical analysis:

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns
```

5.5.2 Loading and Displaying the Dataset

The dataset is loaded into a Pandas DataFrame, and the first few rows are displayed:

```
data = pd.read_csv('/content/employees.csv')

data.head()

print(data)  # Display dataset
```

```
    First Name  Gender  Start Date Last Login Time  Salary  Bonus %  \
0      Douglas    Male    8/6/1993        12:42 PM   97308    6.945
1       Thomas    Male   3/31/1996         6:53 AM   61933    4.170
2        Maria  Female   4/23/1993        11:17 AM  130590   11.858
3        Jerry    Male    3/4/2005         1:00 PM  138705    9.340
4        Larry    Male   1/24/1998         4:47 PM  101004    1.389
..         ...     ...         ...             ...     ...      ...
995      Henry     NaN  11/23/2014         6:09 AM  132483   16.655
996    Phillip    Male   1/31/1984         6:30 AM   42392   19.675
997    Russell    Male   5/20/2013        12:39 PM   96914    1.421
998      Larry    Male   4/20/2013         4:45 PM   60500   11.985
999     Albert    Male   5/15/2012         6:24 PM  129949   10.169

    Senior Management                  Team
0                True             Marketing
1                True                   NaN
2               False               Finance
3                True               Finance
4                True       Client Services
..                ...                   ...
995             False          Distribution
996             False               Finance
997             False               Product
998             False  Business Development
999              True                 Sales

[1000 rows x 8 columns]
```

5.5.3 Data Cleaning

Handling missing values using dropna() and fillna() to ensure a clean dataset:

cleaned_data = data.dropna() # Drops rows with missing values

print(cleaned_data)

```
    First Name  Gender Start Date Last Login Time  Salary  Bonus %  \
0      Douglas    Male   8/6/1993        12:42 PM   97308    6.945
2        Maria  Female  4/23/1993        11:17 AM  130590   11.858
3        Jerry    Male   3/4/2005         1:00 PM  138705    9.340
4        Larry    Male  1/24/1998         4:47 PM  101004    1.389
5       Dennis    Male  4/18/1987         1:35 AM  115163   10.125
..         ...     ...        ...             ...     ...      ...
994     George    Male  6/21/2013         5:47 PM   98874    4.479
996    Phillip    Male  1/31/1984         6:30 AM   42392   19.675
997    Russell    Male  5/20/2013        12:39 PM   96914    1.421
998      Larry    Male  4/20/2013         4:45 PM   60500   11.985
999     Albert    Male  5/15/2012         6:24 PM  129949   10.169

    Senior Management                  Team
0                True             Marketing
2               False               Finance
3                True               Finance
4                True       Client Services
5               False                 Legal
..                ...                   ...
994              True             Marketing
996             False               Finance
997             False               Product
998             False  Business Development
999              True                 Sales

[764 rows x 8 columns]
```

filled_data = cleaned_data.fillna(0) # Fills remaining missing values with 0

filled_data = filled_data[(filled_data != 0).all(axis=1)] # Removes any rows with 0 values

print(filled_data)

```
    First Name  Gender  Start Date Last Login Time  Salary  Bonus %  \
0      Douglas    Male    8/6/1993        12:42 PM   97308    6.945
3        Jerry    Male    3/4/2005         1:00 PM  138705    9.340
4        Larry    Male   1/24/1998         4:47 PM  101004    1.389
6         Ruby  Female   8/17/1987         4:20 PM   65476   10.012
8       Angela  Female  11/22/2005         6:29 AM   95570   18.523
..         ...     ...         ...             ...     ...      ...
991       Rose  Female   8/25/2002         5:12 AM  134505   11.051
992    Anthony    Male  10/16/2011         8:35 AM  112769   11.625
993       Tina  Female   5/15/1997         3:53 PM   56450   19.040
994     George    Male   6/21/2013         5:47 PM   98874    4.479
999     Albert    Male   5/15/2012         6:24 PM  129949   10.169

     Senior Management             Team
0                 True        Marketing
3                 True          Finance
4                 True  Client Services
6                 True          Product
8                 True      Engineering
..                 ...              ...
991               True        Marketing
992               True          Finance
993               True      Engineering
994               True        Marketing
999               True            Sales

[381 rows x 8 columns]
```

5.5.4 Data Visualization

5.5.4.1 Employee Count by Gender

A bar chart representing the number of employees categorized by gender:

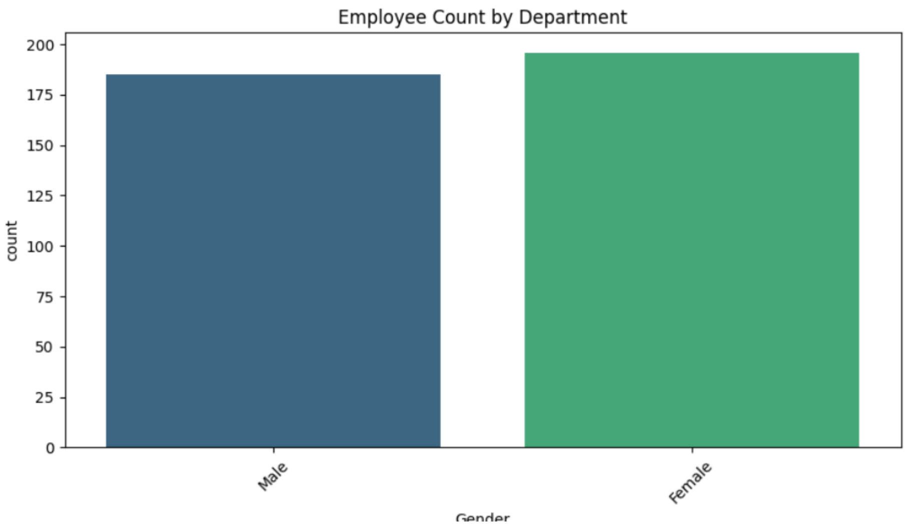
plt.figure(figsize=(10,5))

sns.countplot(x='Gender', data=filled_data, palette='viridis')

plt.xticks(rotation=45)

plt.title('Employee Count by Gender')

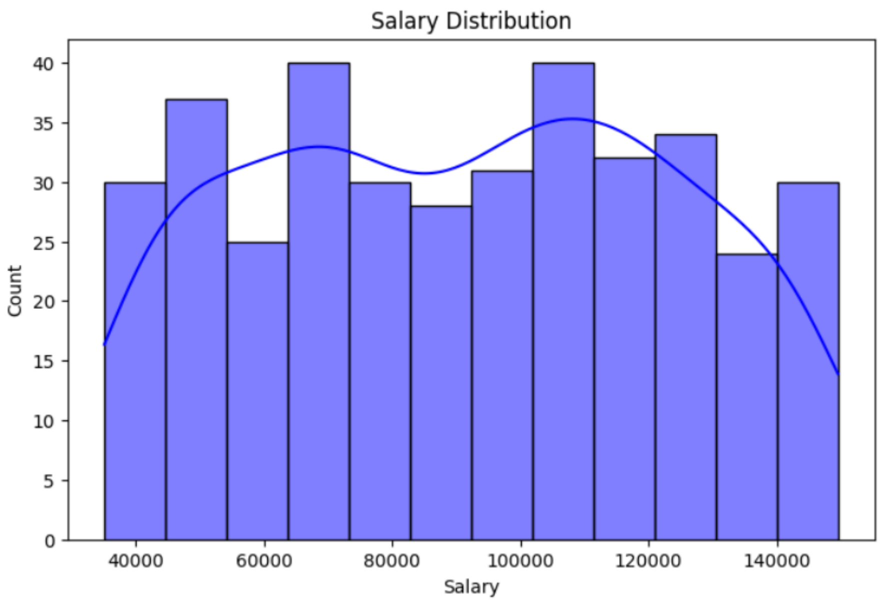
plt.show()



5.5.4.2 Salary Distribution

A histogram illustrating salary distribution:

```
plt.figure(figsize=(8,5))
sns.histplot(filled_data['Salary'], bins=12, kde=True, color='blue')
plt.title('Salary Distribution')
plt.xlabel('Salary')
plt.ylabel('Count')
plt.show()
```

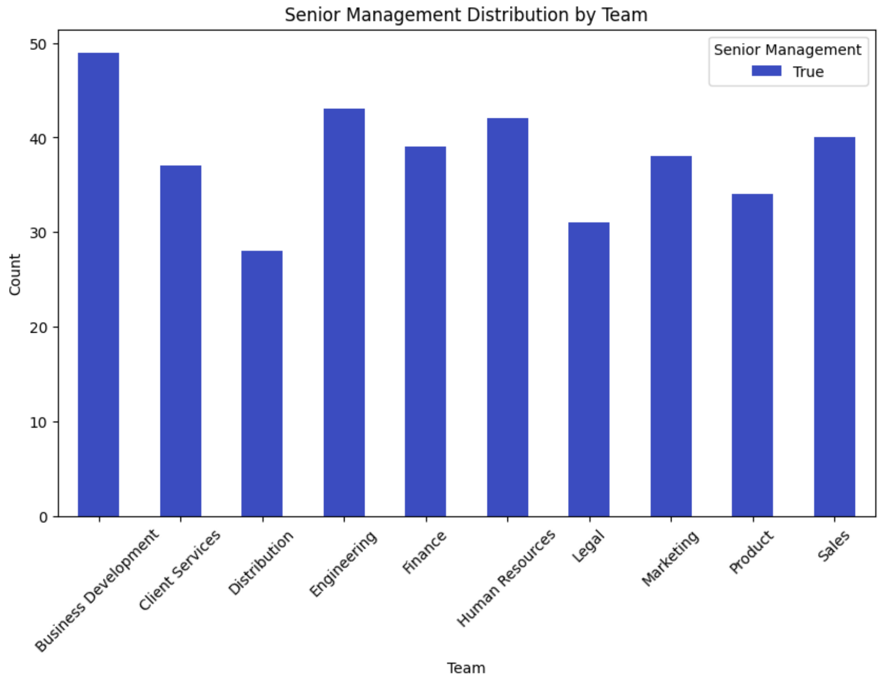


5.5.4.3 Senior Management Count by Team

A stacked bar chart showing the distribution of senior management across different teams:

```
plt.figure(figsize=(10,6))

senior_mgmt_counts = filled_data.groupby(['Team', 'Senior Management']).size().unstack()


# Stacked bar chart

senior_mgmt_counts.plot(kind='bar', stacked=True, colormap='coolwarm', figsize=(10,6))

plt.xlabel("Team")

plt.ylabel("Count")

plt.title("Senior Management Count by Team")

plt.legend(title="Senior Management")

plt.show()
```

<Figure size 1000x600 with 0 Axes>



5.5.5 Additional Analysis and Visualizations

5.5.5.1 Hiring Trends Over Time

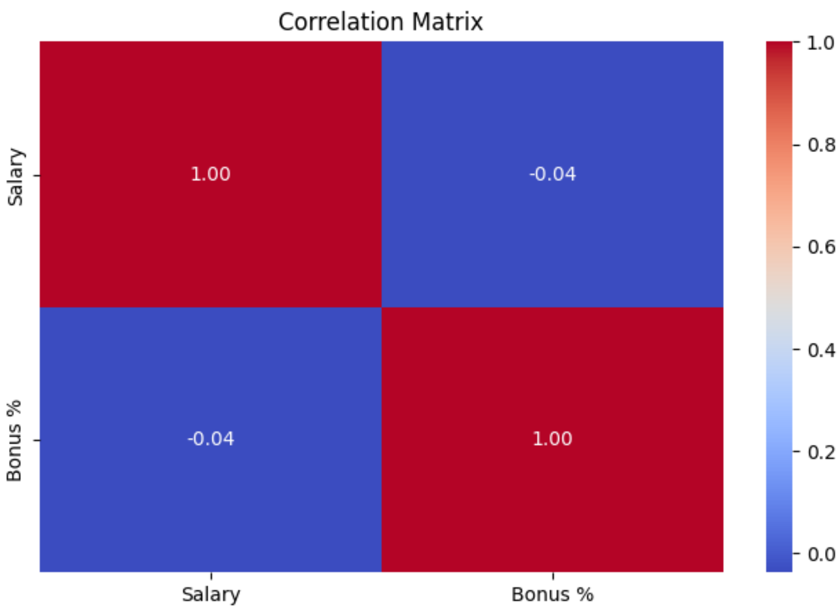
This visualization shows the number of hires per year to identify hiring patterns over time.

```
# Convert 'Start Date' to datetime format
filled_data['Start Date'] = pd.to_datetime(filled_data['Start Date'])


# Count hires per year
filled_data['Year'] = filled_data['Start Date'].dt.year
hiring_trends = filled_data['Year'].value_counts().sort_index()


# Plot hiring trends
plt.figure(figsize=(10,5))
plt.plot(hiring_trends.index, hiring_trends.values, marker='o', linestyle='-',
color='blue')
plt.xlabel("Year")
plt.ylabel("Number of Hires")
plt.title("Hiring Trends Over Time")
plt.grid(True)
plt.show()
```

5.5.5.2 Correlation Matrix

A heatmap representing the correlation between numerical variables in the dataset:

```
# Select only numeric columns
numeric_data = filled_data.select_dtypes(include=['number'])


# Plot correlation heatmap
plt.figure(figsize=(8,5))
sns.heatmap(numeric_data.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

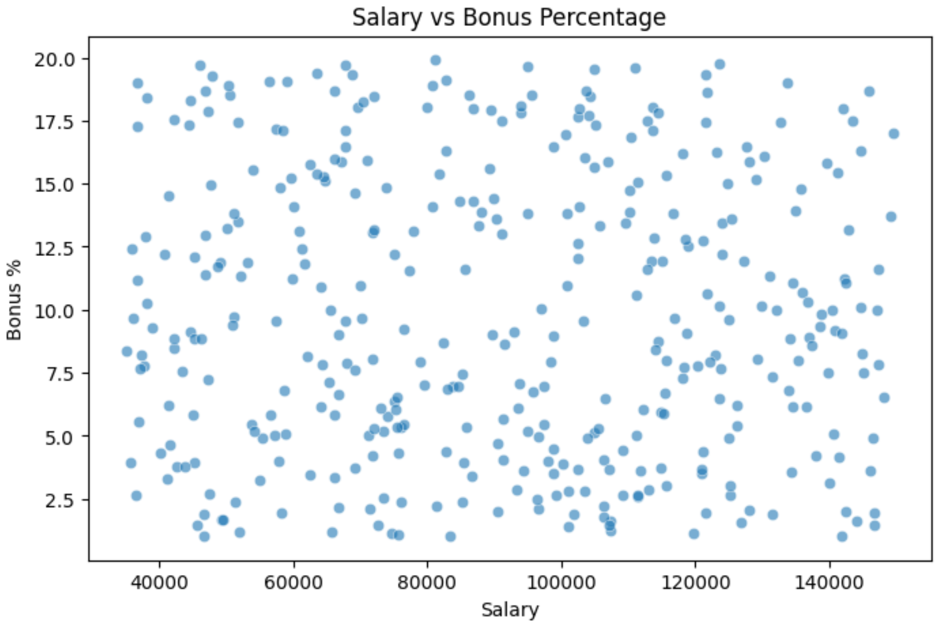


5.5.5.3 Salary vs Bonus Percentage

A scatter plot to analyze the relationship between salary and bonus percentage:

```
plt.figure(figsize=(8,5))
sns.scatterplot(x=filled_data['Salary'], y=filled_data['Bonus %'], alpha=0.6)
```

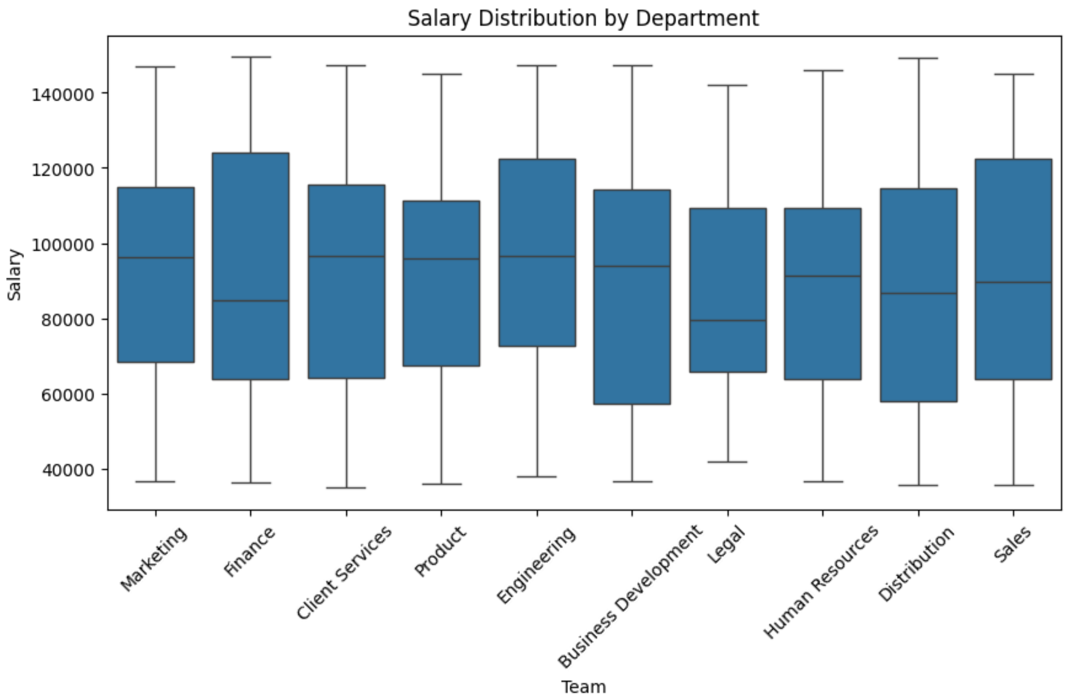
```
plt.title('Salary vs Bonus Percentage')
plt.xlabel('Salary')
plt.ylabel('Bonus %')
plt.show()
```



5.5.5.4 Salary Distribution by Department

A boxplot showing how salaries are distributed across different departments:

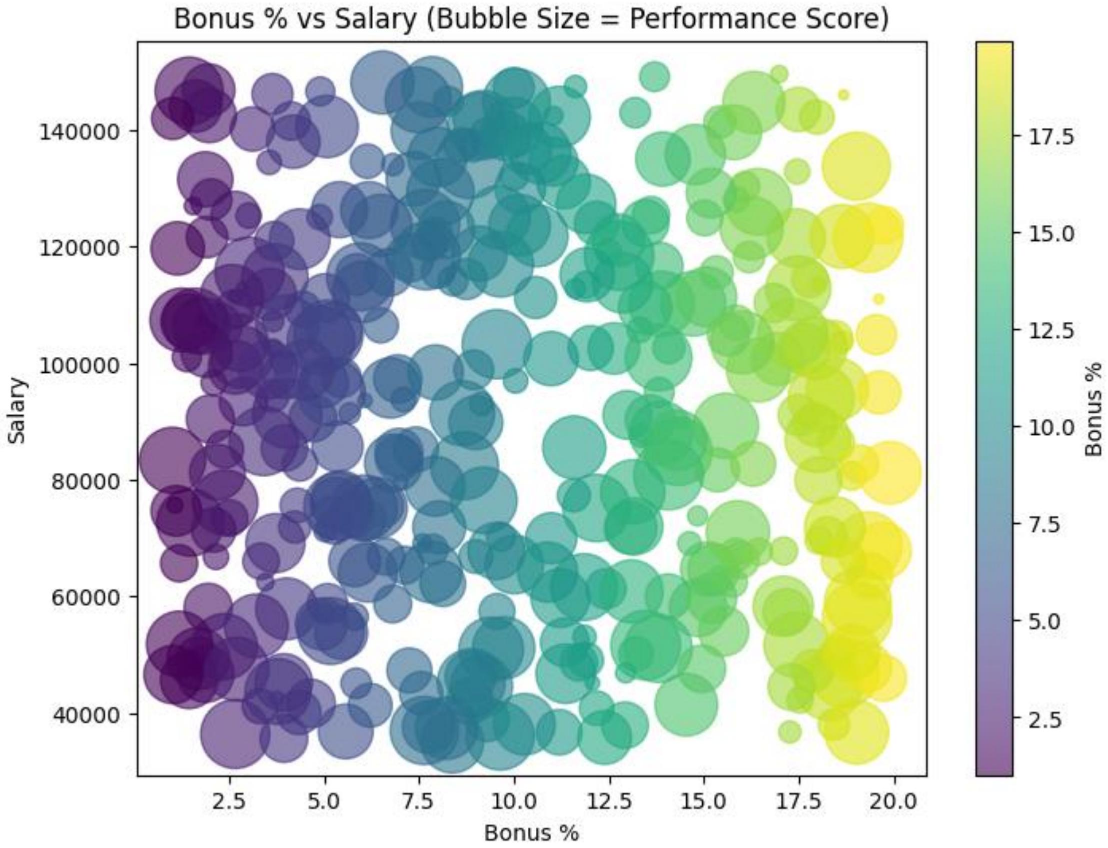
```
plt.figure(figsize=(10,5))
sns.boxplot(x='Team', y='Salary', data=filled_data)
plt.xticks(rotation=45)
plt.title('Salary Distribution by Department')
plt.show()
```



5.5.5.5 Bonus % vs Salary (Bubble Chart)

A **bubble chart** to analyze the relationship between **bonus percentage and salary**, where bubble size represents performance scores.

```
x = filled_data['Bonus %']

y = filled_data['Salary']

colors = filled_data['Bonus %']

sizes = 10 * np.random.randint(100, size=len(filled_data))


plt.figure(figsize=(8,6))

plt.scatter(x, y, c=colors, s=sizes, alpha=0.6, cmap='viridis')

plt.colorbar(label="Bonus %")

plt.xlabel("Bonus %")

plt.ylabel("Salary")

plt.title("Bonus % vs Salary (Bubble Size = Performance Score)")

plt.show()
```



7. Findings & Discussion:

Key Insights from Data Analysis:

- **Salary Impact:** Employees in lower salary brackets have a higher attrition rate.
- **Experience Level Influence:** Mid-career professionals (3-7 years of experience) tend to leave the most.
- **Job Role-Specific Trends:** Employees in software development and testing roles show higher turnover than managerial positions.
- **Benefits & Incentives:** Employees receiving additional bonuses or benefits have lower attrition rates.

These findings suggest that organizations must focus on competitive salaries, career growth opportunities, and incentives to improve employee retention.

8. Conclusion & Future Scope:

Conclusion:

This project successfully identifies critical factors influencing employee attrition in IT companies. The analysis highlights the importance of salary, experience, and job roles in employee turnover.

By leveraging data-driven insights, IT firms can reduce attrition rates, enhance employee satisfaction, and optimize HR strategies for long-term workforce stability.

Future Scope:

- Expanding the dataset to include more industries for broader analysis.
- Implementing **Machine Learning models** for predictive attrition analysis.
- Exploring additional factors such as work-life balance and company culture.