Employee Attrition and Analysis Prediction

This project aims to uncover the factors that influence employee attrition and predict which employees are most likely to leave the company.

Problem Statement:

Acme Corporation, a leading tech company, is facing a growing challenge with employee turnover. The HR department is increasingly concerned about the rising rate of attrition, which negatively affects team dynamics, disrupts project continuity, and undermines overall company morale. To address this issue, Acme Corporation seeks to utilize data analytics and machine learning to better understand the drivers of employee turnover and to predict which employees are most at risk of leaving in the near future.

Difference Between Attrition and Layoff?

While attrition happens voluntarily when employees choose to leave the company, layoff refers to the involuntary termination of employees by the employer, often due to financial constraints, restructuring, or other business-related reasons. In attrition, the position may remain unfilled, whereas in a layoff, the employer actively decides to downsize the workforce.

About the Dataset:

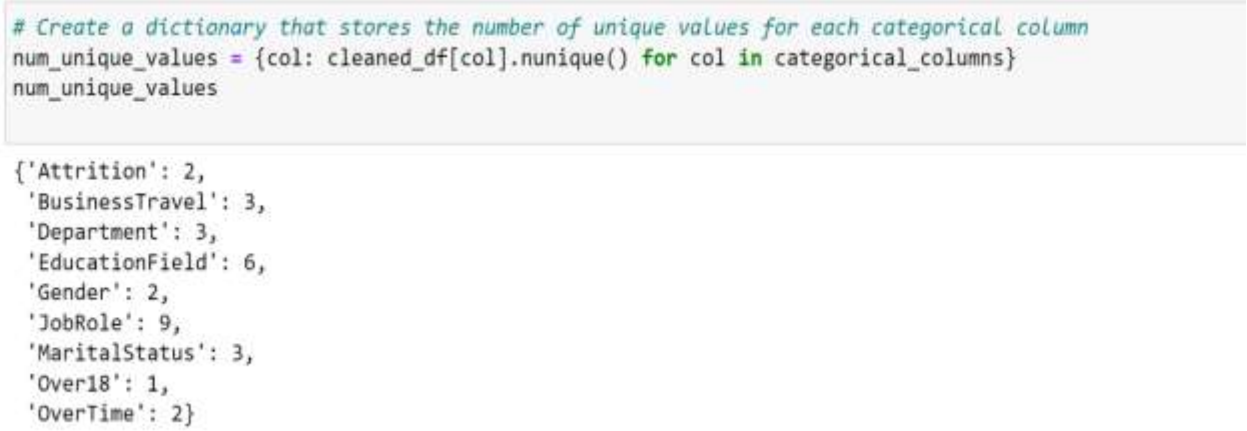
Acme Corporation has provided historical data encompassing employee demographics, job satisfaction, work environment, performance metrics, and turnover status. This dataset spans the past five years and includes information on employees who have left the company as well as those who are still employed. The dataset comprises various features that shed light on employee characteristics, job satisfaction, and performance.

The dataset contains 1,470 rows and 35 columns.

Data Cleaning:

The dataset was found to have no missing values and no duplicate entries. It primarily consists of numerical data (integer and float types), making it well-suited for analytical tasks. Additionally, a few columns are categorical (object data type), and there are no inconsistencies in data types. The categorical columns in the dataset are: Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, and OverTime.

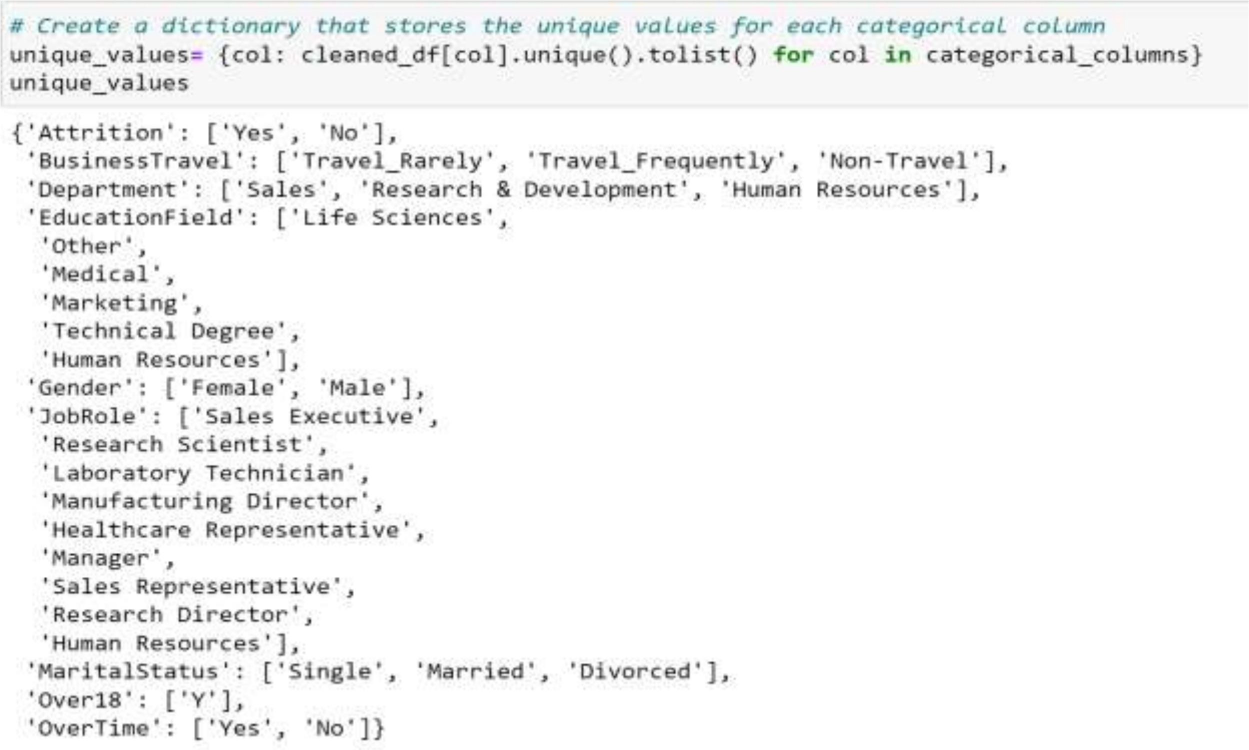
The following code was used to determine the number of values in the categorical columns:



```
# Create a dictionary that stores the number of unique values for each categorical column
num_unique_values = {col: cleaned_df[col].nunique() for col in categorical_columns}
num_unique_values
```

```
{'Attrition': 2,
 'BusinessTravel': 3,
 'Department': 3,
 'EducationField': 6,
 'Gender': 2,
 'JobRole': 9,
 'MaritalStatus': 3,
 'Over18': 1,
 'OverTime': 2}
```

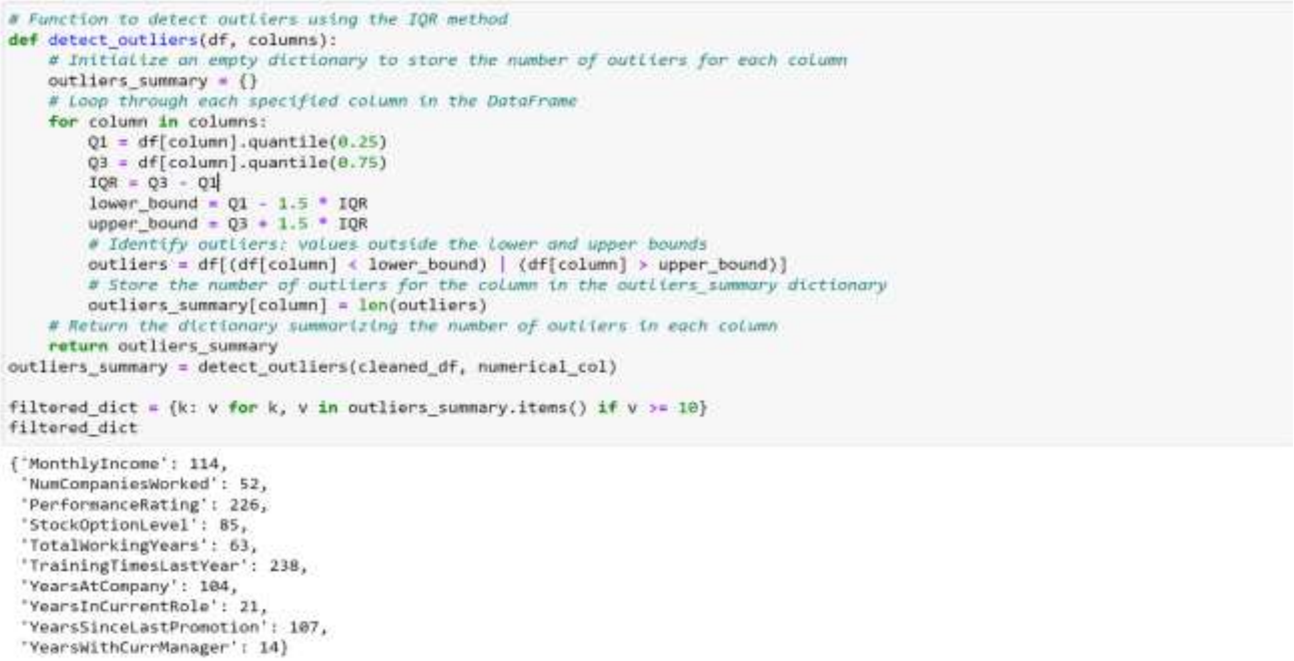
The distinct values within the categorical columns are listed below:



```
# Create a dictionary that stores the unique values for each categorical column
unique_values= {col: cleaned_df[col].unique().tolist() for col in categorical_columns}
unique_values
```

```
{'Attrition': ['Yes', 'No'],
 'BusinessTravel': ['Travel_Rarely', 'Travel_Frequently', 'Non-Travel'],
 'Department': ['Sales', 'Research & Development', 'Human Resources'],
 'EducationField': ['Life Sciences',
  'Other',
  'Medical',
  'Marketing',
  'Technical Degree',
  'Human Resources'],
 'Gender': ['Female', 'Male'],
 'JobRole': ['Sales Executive',
  'Research Scientist',
  'Laboratory Technician',
  'Manufacturing Director',
  'Healthcare Representative',
  'Manager',
  'Sales Representative',
  'Research Director',
  'Human Resources'],
 'MaritalStatus': ['Single', 'Married', 'Divorced'],
 'Over18': ['Y'],
 'OverTime': ['Yes', 'No']}
```

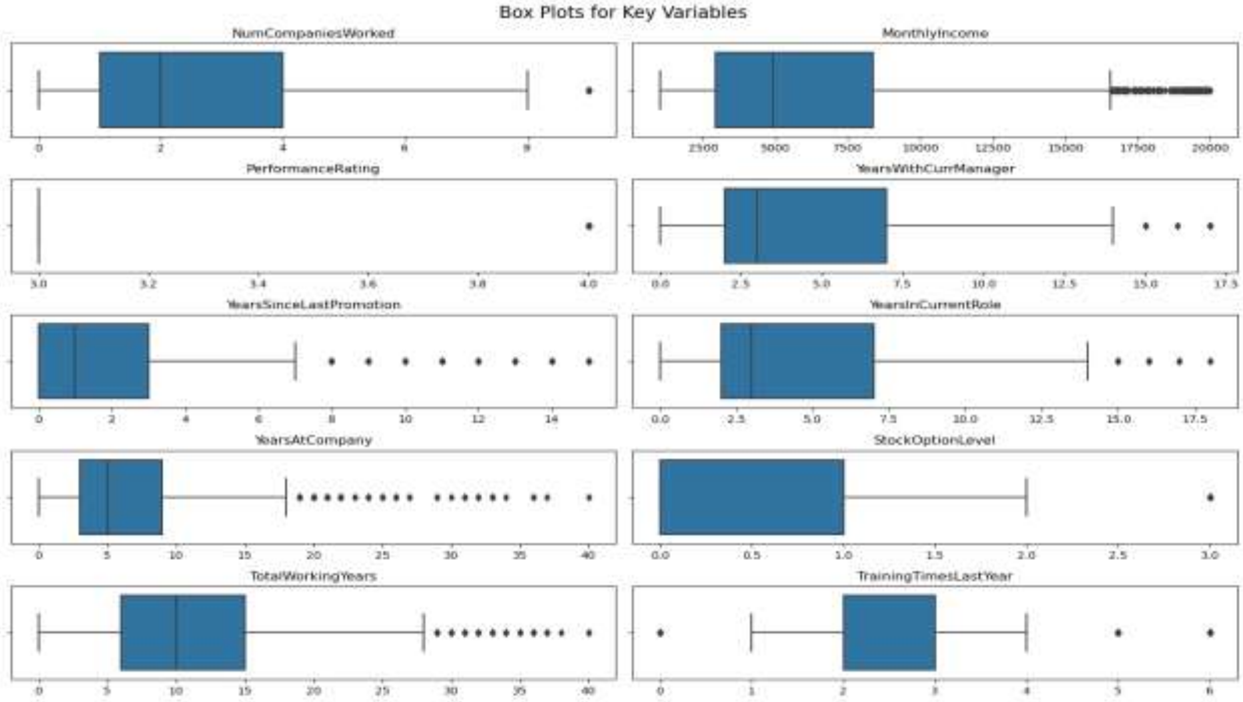
Ten columns in the dataset were identified to contain outliers. These outliers were detected using the Interquartile Range (IQR) method. The code used for identifying these outliers and the affected columns is provided below:



```
# Function to detect outliers using the IQR method
def detect_outliers(df, columns):
    # Initialize an empty dictionary to store the number of outliers for each column
    outliers_summary = {}
    # Loop through each specified column in the DataFrame
    for column in columns:
        Q1 = df[column].quantile(0.25)
        Q3 = df[column].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        # Identify outliers: values outside the lower and upper bounds
        outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)]
        # Store the number of outliers for the column in the outliers_summary dictionary
        outliers_summary[column] = len(outliers)
    # Return the dictionary summarizing the number of outliers in each column
    return outliers_summary
outliers_summary = detect_outliers(cleaned_df, numerical_col)

filtered_dict = {k: v for k, v in outliers_summary.items() if v >= 10}
filtered_dict
```

```
{'MonthlyIncome': 114,
 'NumCompaniesWorked': 52,
 'PerformanceRating': 226,
 'StockOptionLevel': 85,
 'TotalWorkingYears': 63,
 'TrainingTimesLastYear': 238,
 'YearsAtCompany': 104,
 'YearsInCurrentRole': 21,
 'YearsSinceLastPromotion': 107,
 'YearsWithCurrManager': 14}
```

After identifying the outliers, the values in these columns were visualized to assess whether they were related or if they indicated potential data entry errors.



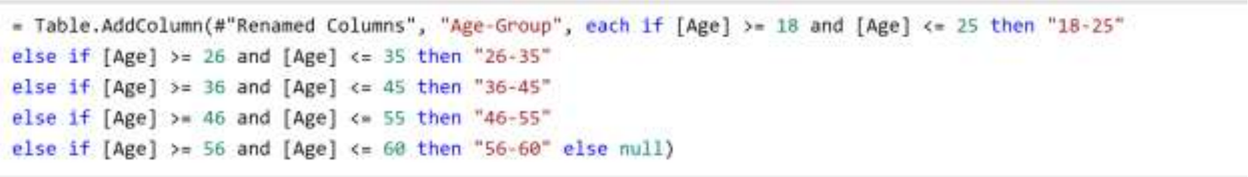
It was observed that the outliers in the dataset are valid data points, representing significant observations such as long-tenured employees or high earners. These outliers play an essential role in influencing key metrics, and therefore, it is advisable to retain them for further analysis and dashboard creation. These outliers will be particularly valuable when constructing comprehensive dashboards in Power BI, enabling detailed insights into crucial employee segments.

ETL in Power BI:

After cleaning the dataset in Jupyter Notebook, it was downloaded as an Excel file and subsequently imported into Power BI, where the extraction process was completed. Following the import, the data transformation phase commenced. During the transformation phase, the first row of the dataset was promoted to serve as headers. Some columns were renamed. The columns EmployeeCount, Over18, and StandardHours were removed because they contained only a single, constant value, making them irrelevant for analysis. Additionally, six custom columns were created to group the data based on specific criteria, facilitating more effective analysis.

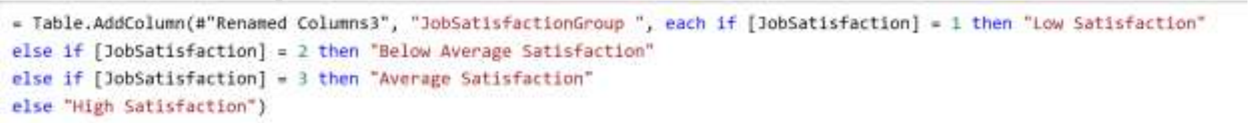
The custom columns along Power Query M Language, are as follows:

1. **Age Group**



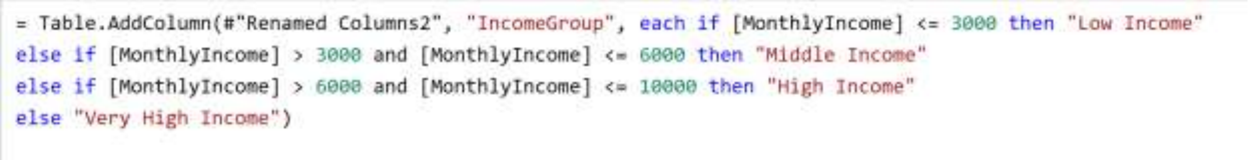
```
= Table.AddColumn(#"Renamed Columns", "Age-Group", each if [Age] >= 18 and [Age] <= 25 then "18-25"
else if [Age] >= 26 and [Age] <= 35 then "26-35"
else if [Age] >= 36 and [Age] <= 45 then "36-45"
else if [Age] >= 46 and [Age] <= 55 then "46-55"
else if [Age] >= 56 and [Age] <= 60 then "56-60" else null)
```

2. **Job Satisfaction Group**



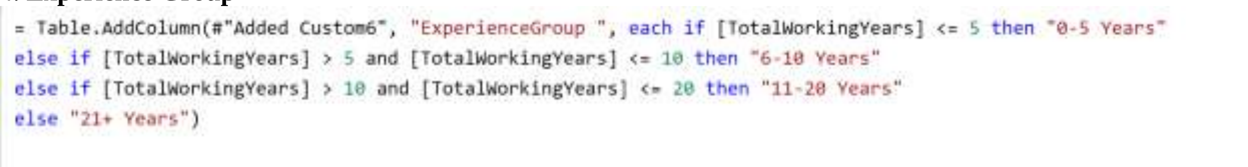
```
= Table.AddColumn(#"Renamed Columns3", "JobSatisfactionGroup ", each if [JobSatisfaction] = 1 then "Low Satisfaction"
else if [JobSatisfaction] = 2 then "Below Average Satisfaction"
else if [JobSatisfaction] = 3 then "Average Satisfaction"
else "High Satisfaction")
```

3. **Income Group**



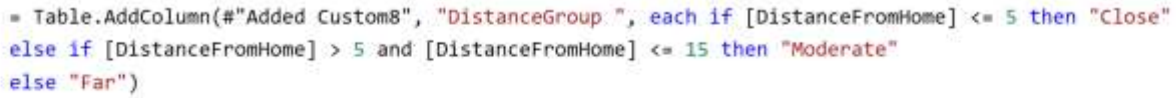
```
= Table.AddColumn(#"Renamed Columns2", "IncomeGroup", each if [MonthlyIncome] <= 3000 then "Low Income"
else if [MonthlyIncome] > 3000 and [MonthlyIncome] <= 6000 then "Middle Income"
else if [MonthlyIncome] > 6000 and [MonthlyIncome] <= 10000 then "High Income"
else "Very High Income")
```

4. **Experience Group**



```
= Table.AddColumn(#"Added Custom6", "ExperienceGroup ", each if [TotalWorkingYears] <= 5 then "0-5 Years"
else if [TotalWorkingYears] > 5 and [TotalWorkingYears] <= 10 then "6-10 Years"
else if [TotalWorkingYears] > 10 and [TotalWorkingYears] <= 20 then "11-20 Years"
else "21+ Years")
```

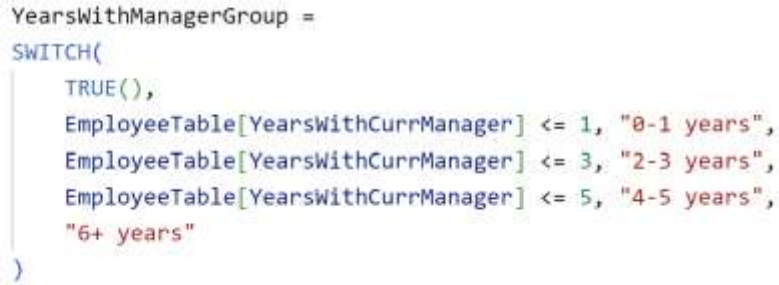
5. **Distance From Home Group**



```
= Table.AddColumn(#"Added Custom8", "DistanceGroup ", each if [DistanceFromHome] <= 5 then "Close"
else if [DistanceFromHome] > 5 and [DistanceFromHome] <= 15 then "Moderate"
else "Far")
```

The final column was created using a DAX query. The DAX query is provided below:

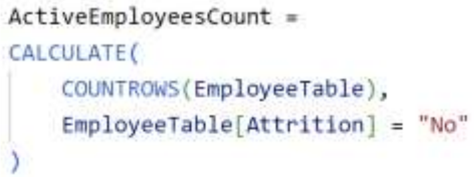
6. **Years With Manager Group**



```
YearsWithManagerGroup =
SWITCH(
    TRUE(),
    EmployeeTable[YearsWithCurrManager] <= 1, "0-1 years",
    EmployeeTable[YearsWithCurrManager] <= 3, "2-3 years",
    EmployeeTable[YearsWithCurrManager] <= 5, "4-5 years",
    "6+ years"
)
```

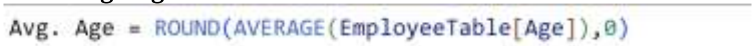
After completing the data transformation, the data was loaded into Power BI, where the data modeling process began. Since there was only a single table, no relationships needed to be established. With the data model in place, the next step was to create DAX measures to enhance the analysis. A total of five DAX measures were created, and the DAX measures, along with their corresponding code, are provided below:

1. **Active Employees**



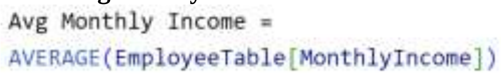
```
ActiveEmployeesCount =
CALCULATE(
    COUNTROWS(EmployeeTable),
    EmployeeTable[Attrition] = "No"
)
```

2. **Average Age**



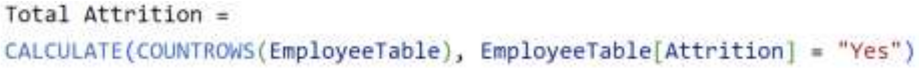
```
Avg. Age = ROUND(AVERAGE(EmployeeTable[Age]),0)
```

3. **Average Salary**



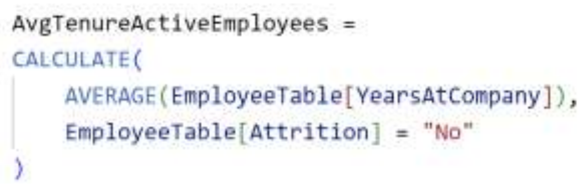
```
Avg Monthly Income =
AVERAGE(EmployeeTable[MonthlyIncome])
```

4. **Total Attrition**



```
Total Attrition =
CALCULATE(COUNTROWS(EmployeeTable), EmployeeTable[Attrition] = "Yes")
```

5. **Average Tenure of Active Employees**



```
AvgTenureActiveEmployees =
CALCULATE(
    AVERAGE(EmployeeTable[YearsAtCompany]),
    EmployeeTable[Attrition] = "No"
)
```

Dashboard Creation and Analysis:

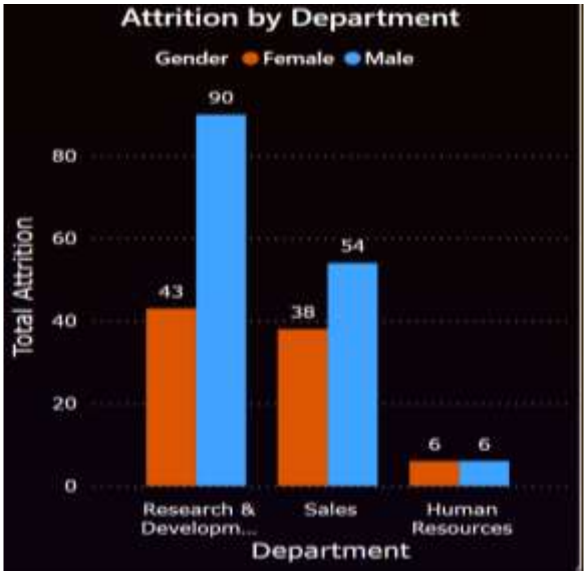
After completing the measures, the focus shifted to building the dashboard. To guide the dashboard creation process, several problem statements were formulated.

These **problem statements** are as follows:

1. How does employee attrition vary across different departments within the company, and what role does gender play in this variation?
2. What is the relationship between employee attrition and their years of job experience?
3. How does job satisfaction combined with income level impact employee attrition?
4. Does marital status affect employee attrition, and if so, how?
5. How do age and frequency of business travel influence employee attrition?
6. How does working overtime correlate with employee attrition?
7. How does the length of time an employee has been with their current manager affect attrition?
8. Which job roles are most susceptible to employee attrition?
9. How does the distance an employee lives from work impact attrition?
10. Which educational backgrounds are associated with higher attrition rates?
11. How does the perception of work-life balance impact employee attrition?
12. How do years at the company and the time since the last promotion influence employee attrition?

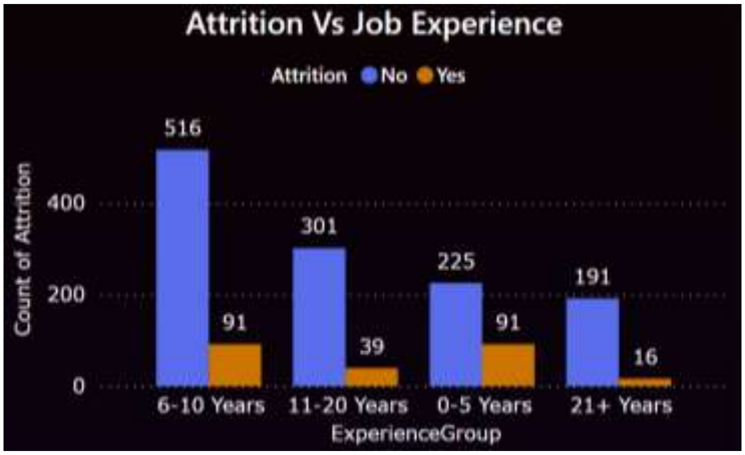
Insights:

1.



- The Research & Development department has the highest attrition overall, with a notable difference between male (90) and female (43) employees.
- The Sales department also experiences significant attrition, but the gap between male (54) and female (38) employees is smaller.
- Human Resources has the least attrition, with equal numbers of male and female employees (6 each) leaving.

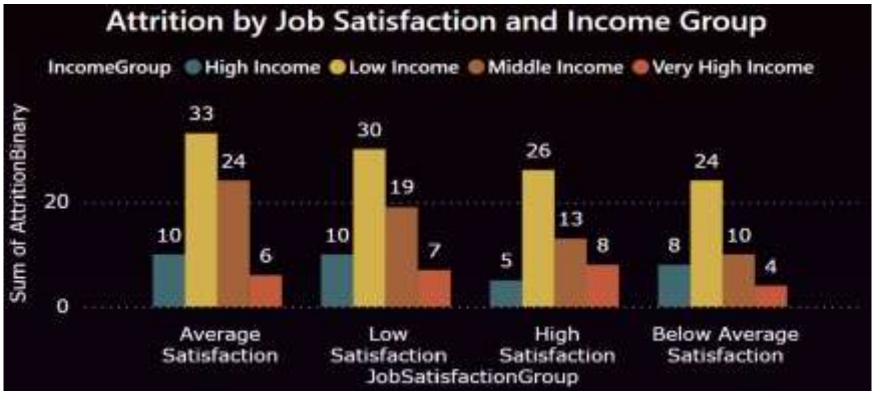
2.



- Employees with 6-10 years of experience have the highest attrition rate (91), indicating a potential issue with retaining mid-level employees.
- Employees with 0-5 years of experience also show a high attrition count (91), which could suggest challenges in retaining newer employees.

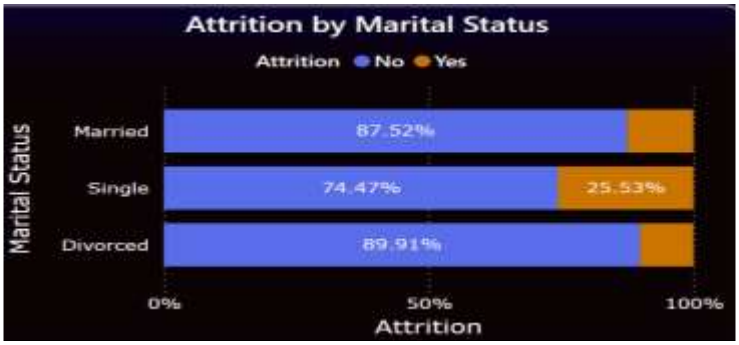
- Attrition significantly decreases as employees gain more than 20 years of experience, with only 16 leaving.

3.



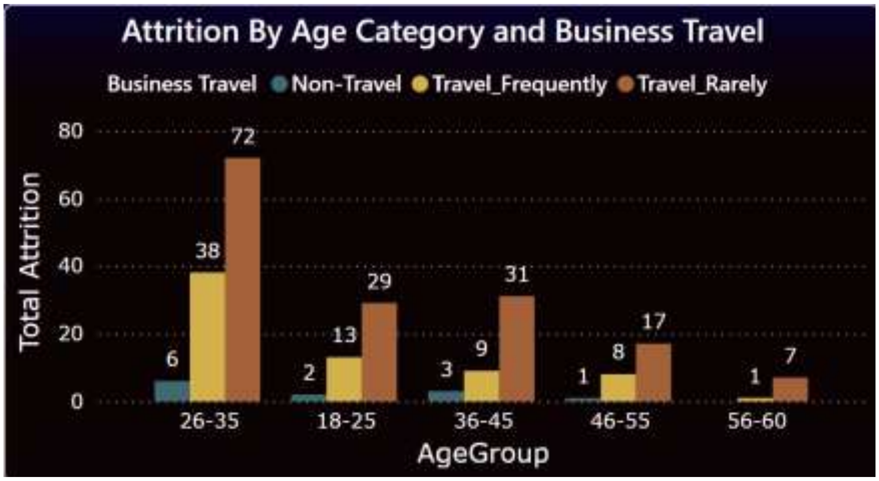
- High attrition rates are most pronounced in the Low Income group across all job satisfaction levels, with the highest rate (33%) occurring in the Average Satisfaction category.
- The Very High Income group consistently shows the lowest attrition rates across all job satisfaction levels, suggesting higher pay may contribute to employee retention regardless of job satisfaction.
- Surprisingly, the High Satisfaction category does not always correspond to the lowest attrition rates, particularly for the Middle Income group where it shows higher attrition (26%) compared to their Below Average Satisfaction rate (24%).

4.



- Single employees have the highest attrition rate (25.53%), which may indicate that single employees feel less tied to their current job or location.
- Married employees have a lower attrition rate (87.52% staying), possibly due to greater stability or commitment to the job.
- Divorced employees have the lowest attrition rate (89.91% staying), which may suggest that they prioritize job stability during life transitions.

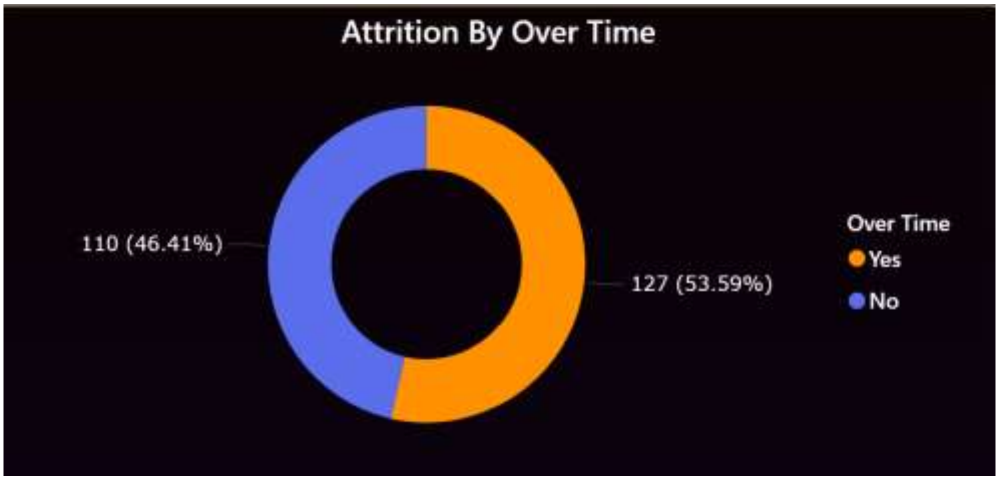
5.



- Employees aged 26-35 who travel rarely have the highest attrition rate (72), suggesting that infrequent travel might contribute to dissatisfaction among younger professionals.
- The 36-45 age group also shows significant attrition, particularly among those who travel rarely (31), which could indicate burnout or dissatisfaction despite the infrequent travel.

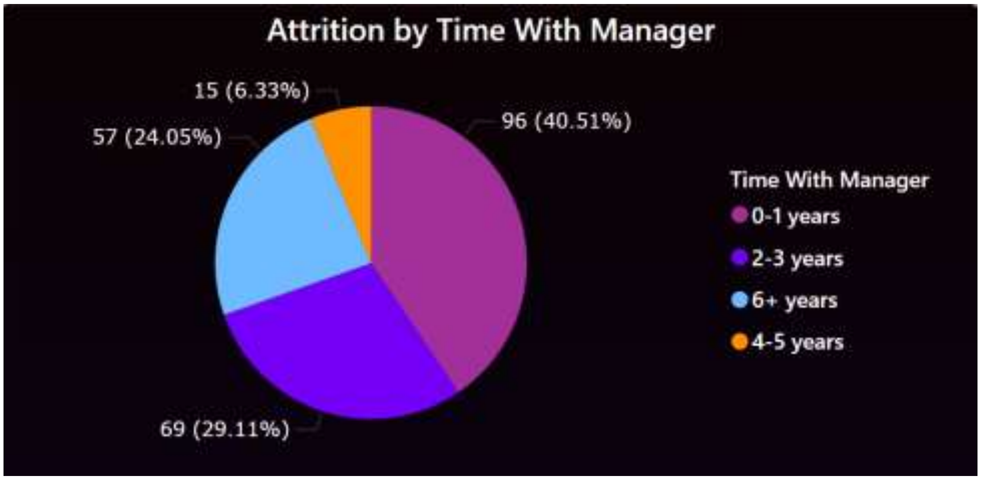
- Attrition is much lower among employees over 46, regardless of travel frequency, suggesting that older employees may have settled into their roles.

6.



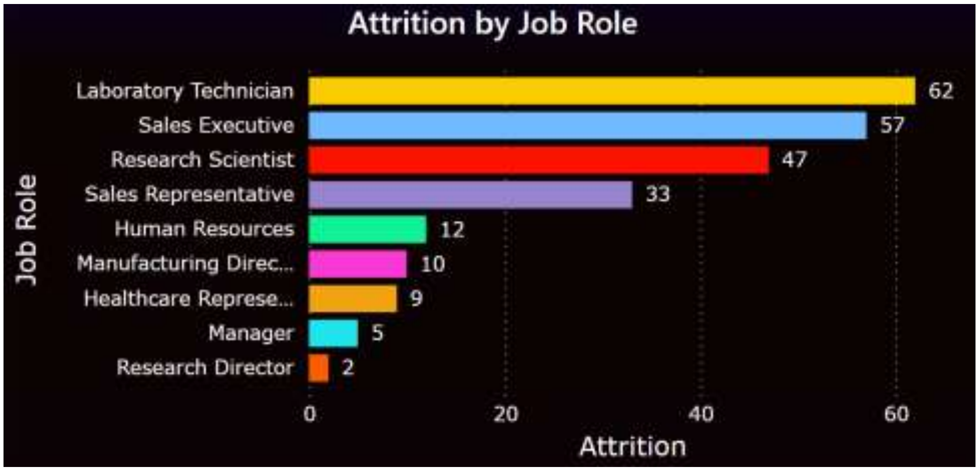
- A significant portion of employees who have to work overtime, (127) indicating that overtime might be a contributing factor to attrition.
- However, a slightly higher number of employees who do not work overtime also leave (110), suggesting that factors other than overtime, such as job satisfaction or work-life balance, might also play significant roles in attrition.

7.



- Employees who have been with their current manager for 0-1 years have the highest attrition rate (96), suggesting that newer relationships with managers might be a critical period for retention efforts.
- Attrition decreases as the length of time with the manager increases, with those who have been with their manager for more than 4-5 years showing the lowest attrition.

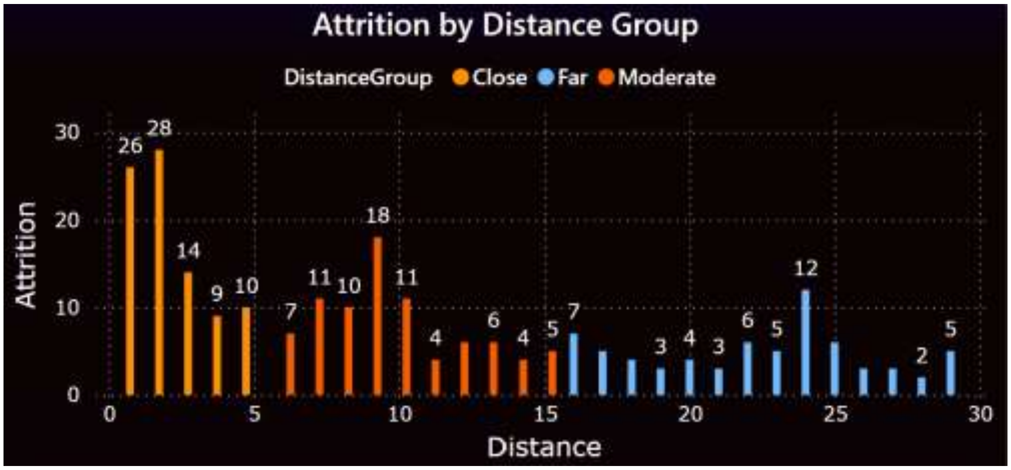
8.



- Laboratory Technicians have the highest attrition (62), followed closely by Sales Executives (57) and Research Scientists (47), suggesting that these roles might have inherent challenges that lead to higher turnover.

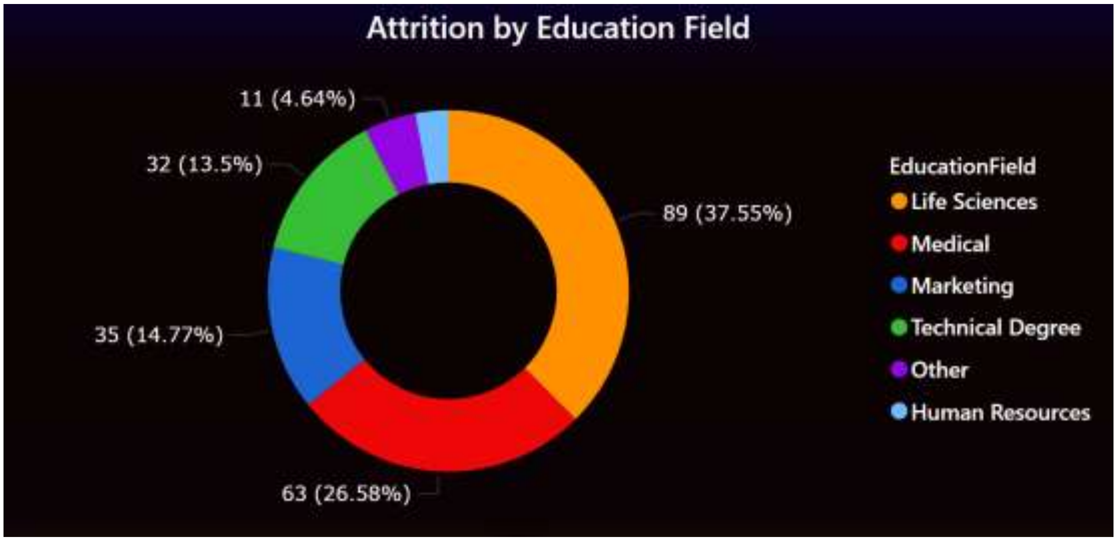
- Higher-level positions like Managers and Research Directors have the lowest attrition rates, which could indicate higher job satisfaction or better compensation at these levels.

9.



- Employees living close to work (0-5 distance units) have the highest attrition, particularly those very close (0 units) with 26-28 leaving, suggesting that proximity does not necessarily correlate with retention.
- Attrition decreases as the distance increases, but there is a slight uptick for those living 15-20 units away, which might indicate a balance where commute becomes inconvenient.
- Employees living far away (25-30 units) have the lowest attrition, which could suggest that those willing to commute long distances are more committed to their roles.

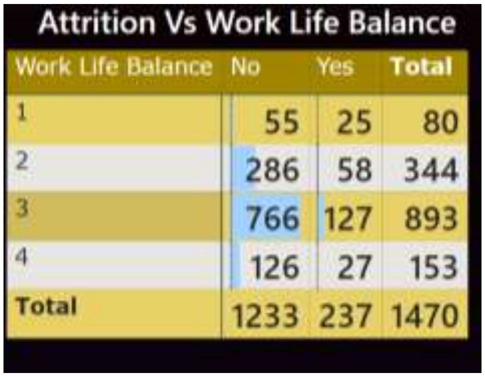
10.



- Employees with a background in Life Sciences have the highest attrition rate (89), suggesting that this field might be particularly challenging or that employees might have better opportunities elsewhere.

- Those in Medical (63) and Technical Degree (35) fields also show higher attrition, while employees in Marketing and Human Resources have lower attrition rates, possibly indicating better job satisfaction or stability in these fields.

11.



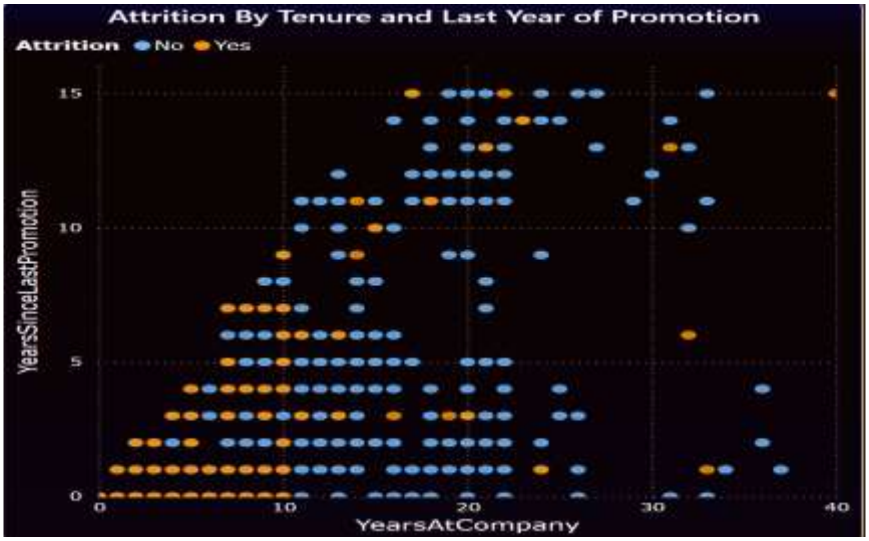
Attrition Vs Work Life Balance

Work Life Balance	No	Yes	Total
1	55	25	80
2	286	58	344
3	766	127	893
4	126	27	153
Total	1233	237	1470

- Employees who rated their work-life balance as "3" (on a scale where 1 is the lowest and 4 is the highest) have the highest overall count (893), with a significant portion of them (127) experiencing attrition.
- The lowest attrition is observed among those who rated their work-life balance as "4" (27), suggesting that a higher work-life balance rating correlates with better employee retention.

- Interestingly, those with a "1" rating (the lowest) have a smaller attrition count (25), possibly because this group is smaller overall or because those with very low work-life balance may leave quickly, leaving a self-selected group of those who tolerate it.

12.



- The plot shows employees with tenure (Years At Company) ranging from 0 to about 40 years, with a higher concentration of employees in the 0-20 year range.

- Most recent promotions (Years Since Last Promotion) appear to cluster in the 0-5 year range, regardless of tenure. This suggests that promotions are more frequent early in an employee's career or shortly after joining the company.

- There seems to be a higher concentration of attrition (orange dots) among employees with shorter tenure (0-10 years) and those who haven't been promoted recently (0-5 years since last promotion).

- Employees with longer tenure (20+ years) generally show less attrition, as indicated by more blue dots in this area. This suggests that long-term employees are more likely to stay with the company.

Conclusion:

Based on the analysis, we can draw several important conclusions about employee attrition patterns in the company:
Tenure and Attrition: There's a clear correlation between employee tenure and attrition rates. Employees with shorter tenure (0-10 years) are more likely to leave the company, while those with longer tenure (20+ years) show significantly lower attrition rates. This suggests that the company should focus on retention strategies for employees in their early to mid-career stages.

Departmental and Gender Differences: Attrition rates vary significantly across departments, with Research & Development experiencing the highest attrition, followed by Sales. There's also a notable gender disparity, particularly in R&D, where male employees leave at a higher rate than females. This indicates a need for targeted retention strategies for specific departments and addressing potential gender-related issues.

Income and Job Satisfaction: There's a strong inverse relationship between income and attrition. Employees in the Low Income group show the highest attrition rates across all job satisfaction levels, while the Very High Income group consistently has the lowest attrition. This suggests that competitive compensation is a crucial factor in retention. Interestingly, high job satisfaction doesn't always correlate with lower attrition, especially in the Middle Income group. This implies that factors beyond job satisfaction, such as career growth opportunities, may influence retention.

Work-Life Balance: Employees who rate their work-life balance as good (3 out of 4) form the largest group but also show significant attrition. Those with the highest work-life balance rating (4) have the lowest attrition, emphasizing the importance of maintaining a healthy work-life balance for retention.

Promotion Patterns: Most promotions occur within the first 5 years of employment, regardless of overall tenure. This pattern, combined with higher attrition in the 6-10 year experience range, suggests that the company might be losing valuable mid-level talent. Implementing clearer career progression paths and opportunities for long-term growth could help address this issue.

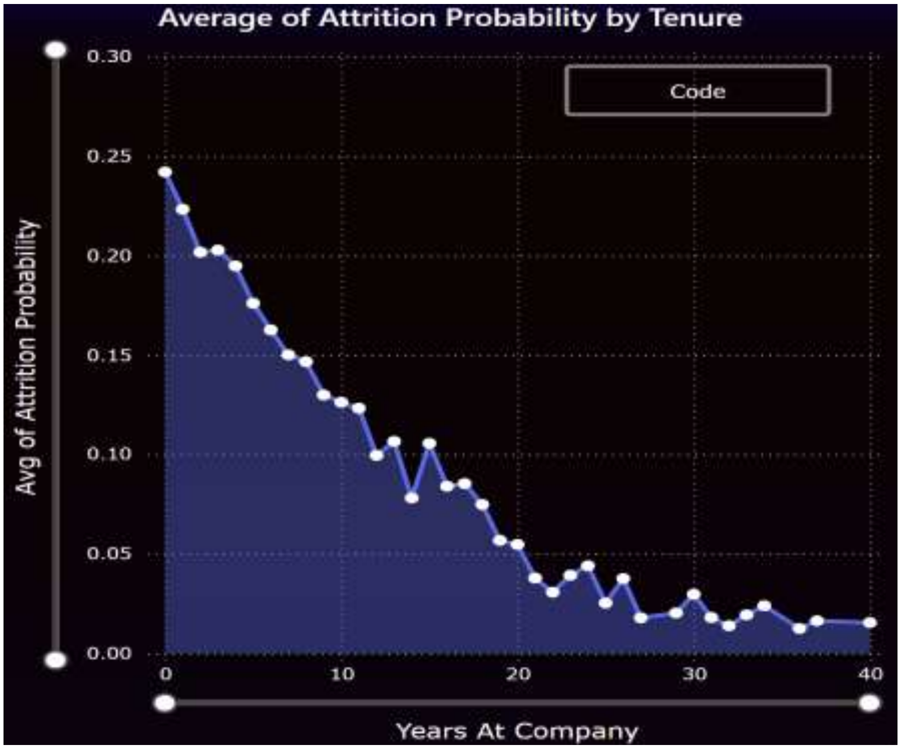
Educational Background and Marital Status: Employees with backgrounds in Life Sciences, Medical, and Technical fields show higher attrition rates. This could indicate industry-specific challenges or more competitive external opportunities for these groups. Single employees have the highest attrition rate, possibly due to greater flexibility or fewer personal commitments tying them to their current job. In conclusion, the company faces multifaceted attrition challenges that require a comprehensive approach to employee retention.

Key areas for focus include:

- Developing targeted retention strategies for employees in their first 10 years, especially in R&D and Sales departments.
- Addressing potential gender disparities in certain departments.
- Ensuring competitive compensation, particularly for lower and middle-income groups. Enhancing work-life balance initiatives.
- Creating more robust career development and promotion opportunities beyond the initial years of employment.
- Tailoring retention strategies to address the specific needs of employees based on their educational background and personal circumstances.

By addressing these areas, the company can work towards reducing attrition rates, retaining valuable talent, and creating a more stable and satisfied workforce.

Predictive Analysis:
1.



Description

This analysis employs a logistic regression model to predict employee attrition using key factors such as Years At Company, Job Satisfaction, Work-Life Balance, and Environment Satisfaction. The model was trained on a dataset where the target variable 'Attrition' was mapped to a binary outcome ('Yes' = 1, 'No' = 0). The prediction results were added to the dataset as a new column representing the probability of each employee leaving the company.

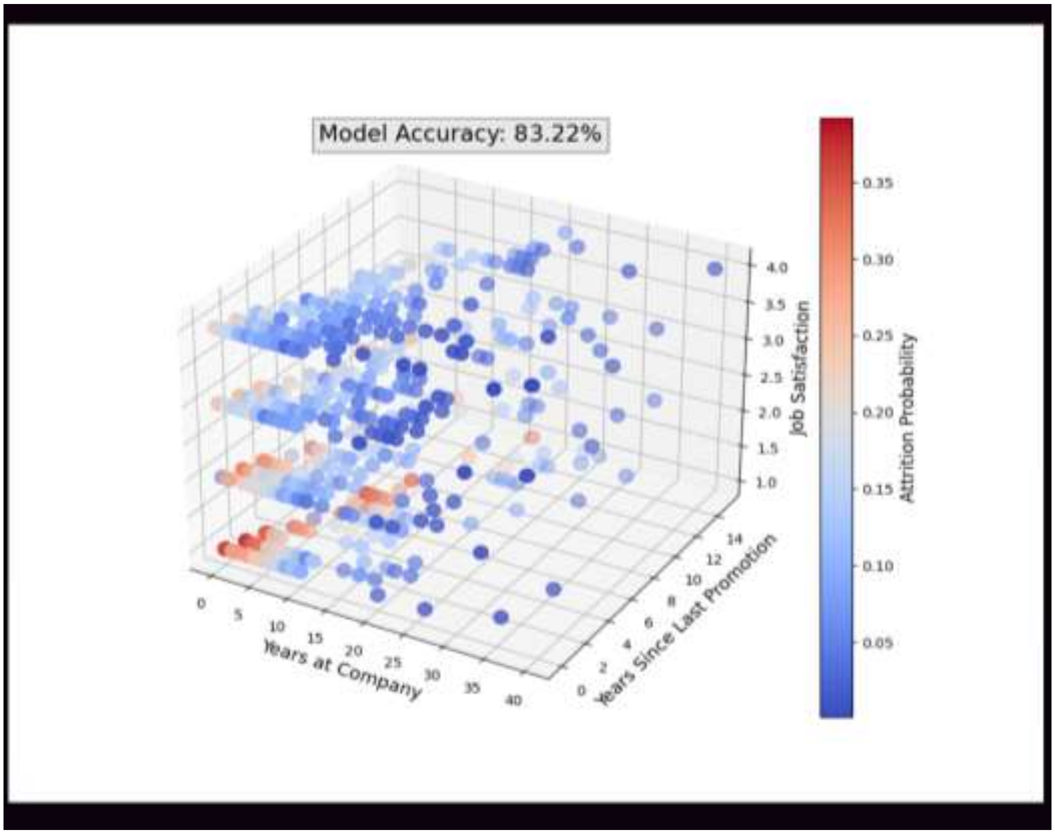
Upon analyzing the dataset, a significant class imbalance was observed in the 'Attrition' column, where 83.88% of the data represented employees who did not leave the company, and only 16.12% represented those who did. This imbalance suggests that the model may be biased towards predicting 'No' for attrition. Such imbalances can impact the accuracy and reliability of predictions, particularly in cases where detecting the minority class ('Yes' for attrition) is critical.

Insights

- The logistic regression model revealed that the probability of attrition is inversely related to the tenure of the employee. Employees with shorter tenures have significantly higher probabilities of leaving, with the probability decreasing as tenure increases. This trend suggests that newer employees are at a higher risk of attrition.

- The inclusion of factors such as Job Satisfaction, Work-Life Balance, and Environment Satisfaction in the model highlights their expected impact on attrition. While these variables are not directly visualized, it is inferred that lower satisfaction in these areas correlates with higher attrition probabilities, especially among employees with shorter tenures.
- The significant class imbalance in the dataset indicates that attrition is relatively rare compared to employee retention. This imbalance necessitates careful consideration during model training, as it could lead to a model that under-predicts attrition cases. Techniques such as resampling or adjusting class weights could be employed to address this issue.
- Long-term employees exhibit much lower probabilities of attrition, indicating that they are more likely to remain with the company. This could be due to increased job security, satisfaction, or loyalty to the company over time.

2.



Description:

This analysis predicts employee attrition probability using a logistic regression model based on features like years at the company, job satisfaction, years since last promotion, and monthly income. The dataset shows a significant class imbalance, with 83.88% of employees marked as "No" for attrition and 16.12% as "Yes." This imbalance can affect the model's accuracy, potentially leading to a bias toward predicting non-attrition.

The model achieved an accuracy of 83.22%**,** indicating reasonable effectiveness in predicting attrition. The inclusion of MonthlyIncome as a feature adds depth to the model, helping to capture more factors that may influence an employee's decision to leave.

Similarities with Previous Code:

- **Logistic Regression Model**: Both analyses use logistic regression to predict attrition probabilities.
- **Dataset Structure**: The structure remains consistent, with attrition mapped as a binary target and similar features used for training.

Differences:

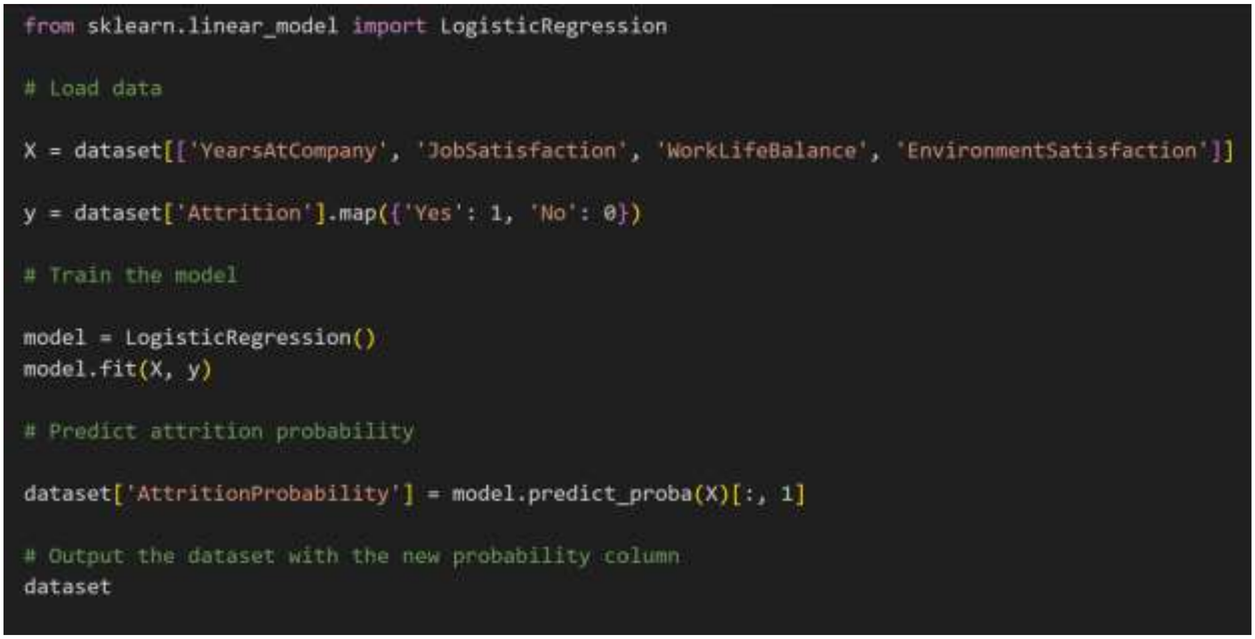
- **Feature Selection**: The current model introduces MonthlyIncome, providing additional context for attrition predictions.

Insights:

- **Longer Tenure, Lower Attrition**: Employees with longer tenures, higher job satisfaction, or recent promotions are less likely to leave, indicating more stability and satisfaction in their roles.
- **Higher Attrition for Newer Employees**: Employees with shorter tenures or those lacking recent promotions are more likely to leave, highlighting the need for early career development and regular promotions to improve retention.
- **Impact of Job Satisfaction**: Higher job satisfaction is strongly correlated with lower attrition probability, emphasizing the importance of improving employee satisfaction to reduce turnover.

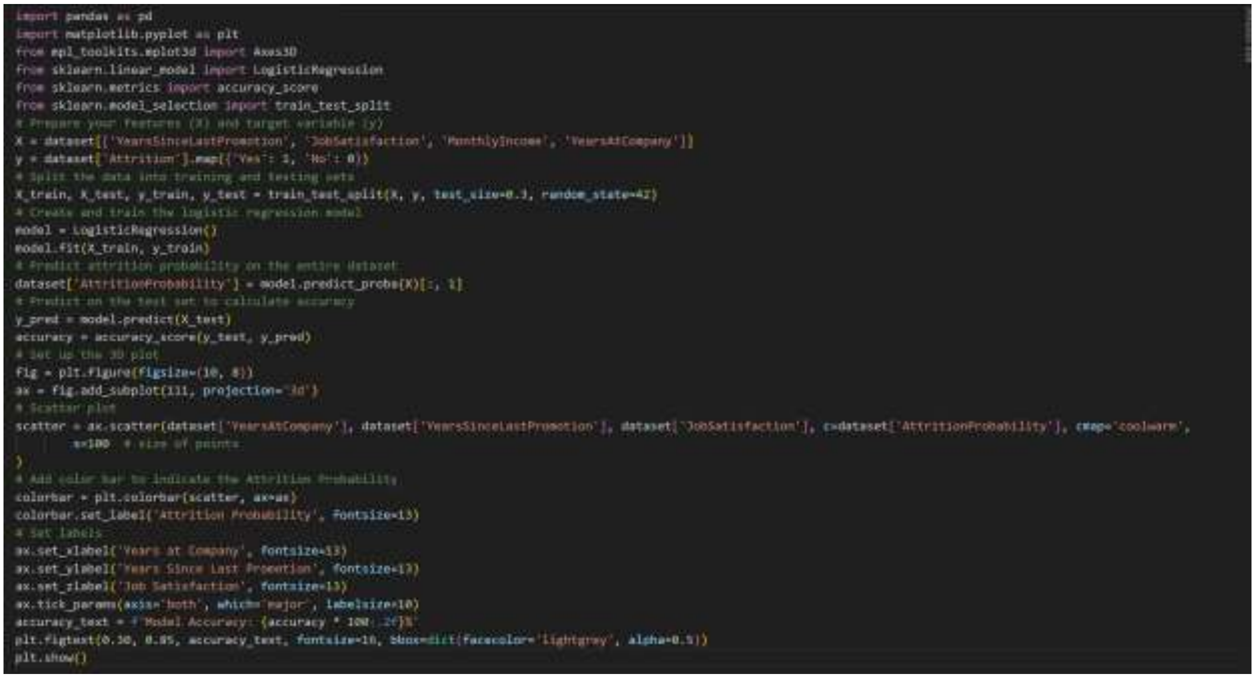
Code:

1.



```
from sklearn.linear_model import LogisticRegression

# Load data

X = dataset[['YearsAtCompany', 'JobSatisfaction', 'WorkLifeBalance', 'EnvironmentSatisfaction']]

y = dataset['Attrition'].map({'Yes': 1, 'No': 0})

# Train the model

model = LogisticRegression()
model.fit(X, y)

# Predict attrition probability

dataset['AttritionProbability'] = model.predict_proba(X)[:, 1]

# Output the dataset with the new probability column
dataset
```

2.



```
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
# Prepare your features (X) and target variable (y)
X = dataset[['YearsSinceLastPromotion', 'JobSatisfaction', 'MonthlyIncome', 'YearsAtCompany']]
y = dataset['Attrition'].map({'Yes': 1, 'No': 0})
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
# Create and train the logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)
# Predict attrition probability on the entire dataset
dataset['AttritionProbability'] = model.predict_proba(X)[:, 1]
# Predict on the test set to calculate accuracy
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
# Set up the 3D plot
fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')
# Scatter plot
scatter = ax.scatter(dataset['YearsAtCompany'], dataset['YearsSinceLastPromotion'], dataset['JobSatisfaction'], c=dataset['AttritionProbability'], cmap='coolwarm',
        s=100  # size of points
)
# Add color bar to indicate the Attrition Probability
colorbar = plt.colorbar(scatter, ax=ax)
colorbar.set_label('Attrition Probability', fontsize=13)
# Set labels
ax.set_xlabel('Years at Company', fontsize=13)
ax.set_ylabel('Years Since Last Promotion', fontsize=13)
ax.set_zlabel('Job Satisfaction', fontsize=13)
ax.tick_params(axis='both', which='major', labelsize=10)
accuracy_text = f'Model Accuracy: {accuracy * 100:.2f}%'
plt.figtext(0.30, 0.85, accuracy_text, fontsize=16, bbox=dict(facecolor='lightgrey', alpha=0.5))
plt.show()
```

Link to the POWERBI Dashboard:

https://app.powerbi.com/links/KzJWXw3lE8?ctid=6ef33993-a22d-4072-84ca-afc10ca48665&pbi_source=linkShare&bookmarkGuid=2dd69e1e-3ec9-4455-933f-3829af7a7644