Predicting Employee Attrition

A Data-Driven Approach to Employee Retention

Enkel Mezini

Gallup Workplace -
Shane McFeely and Ben Wigert

"52% of voluntarily exiting employees say their manager or organization could have done something to prevent them from leaving their job."

https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx

Which of the following issues best describes your primary reason for leaving your previous job?

Top four themes for leaving a job in 2024 are:

Engagement and Culture	37%
Wellbeing and Work-Life Balance	31%
Pay and Benefits	16%
Managers and Leaders	9%

Gallup - https://www.gallup.com/467702/indicator-employee-retention-attraction.aspx

Employee Description

____

-Age
-Gender
-Education
-Education Field
-Number Comp. Worked
-Marital Status
-Distance From Home

Objective Employee Data

____

-Business Travel
-Daily Rate
-HourlyRate
-Department
-Years At Company
-Years In Current Role
-Years Since Promotion
-Years With Cur. Manager
-Monthly Income
-Monthly Rate
-Over Time
-Job Level
-Job Role
-Percent Salary Hike
-Stock Option Level
-Total Working Years
-Training Times Last Year

Subjective Employee Data

____

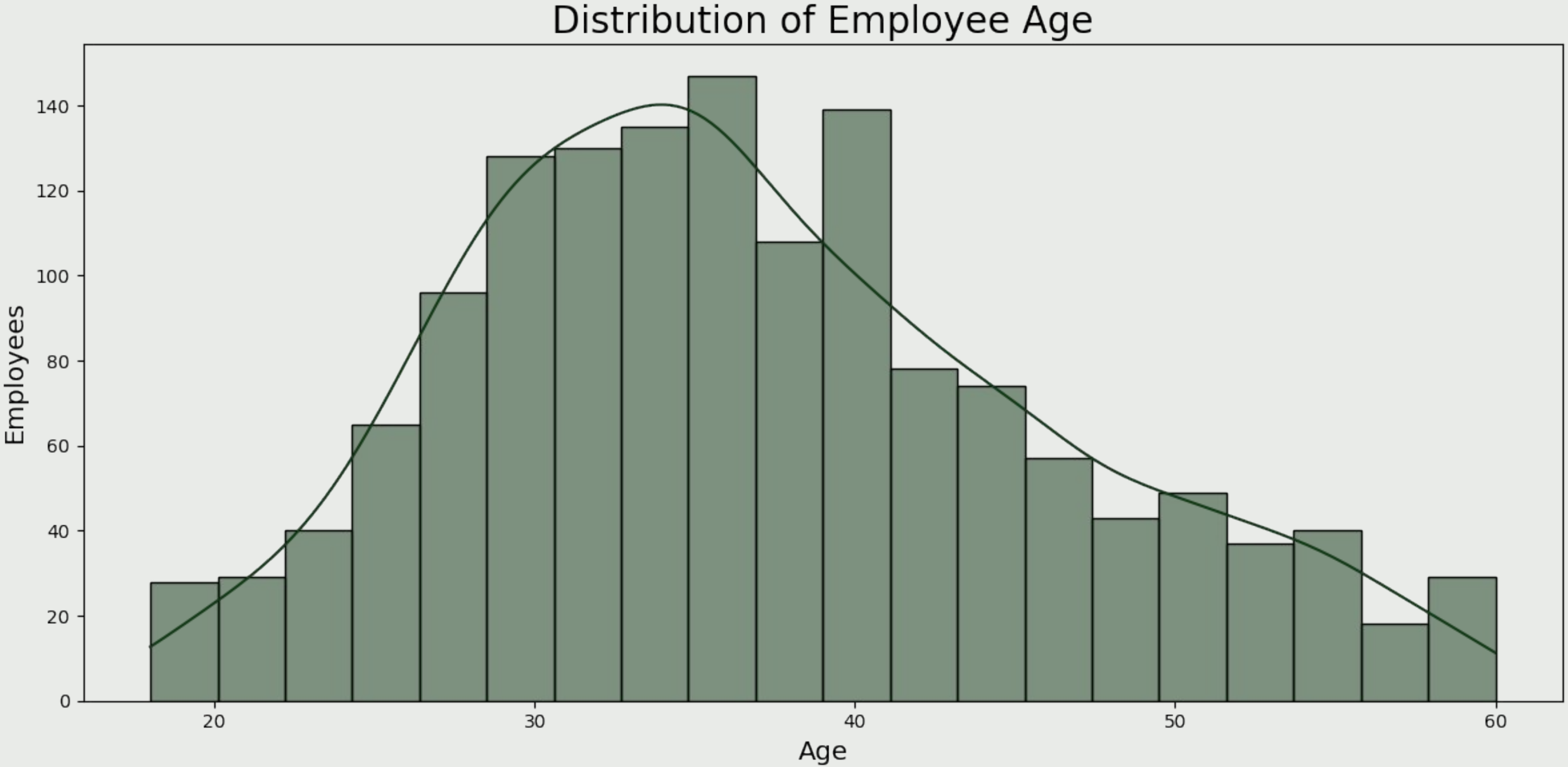
-Environment Satisfaction
-Job Involvement
-Job Satisfaction
-Performance Rating
-Relationship Satisfaction
-Work Life Balance

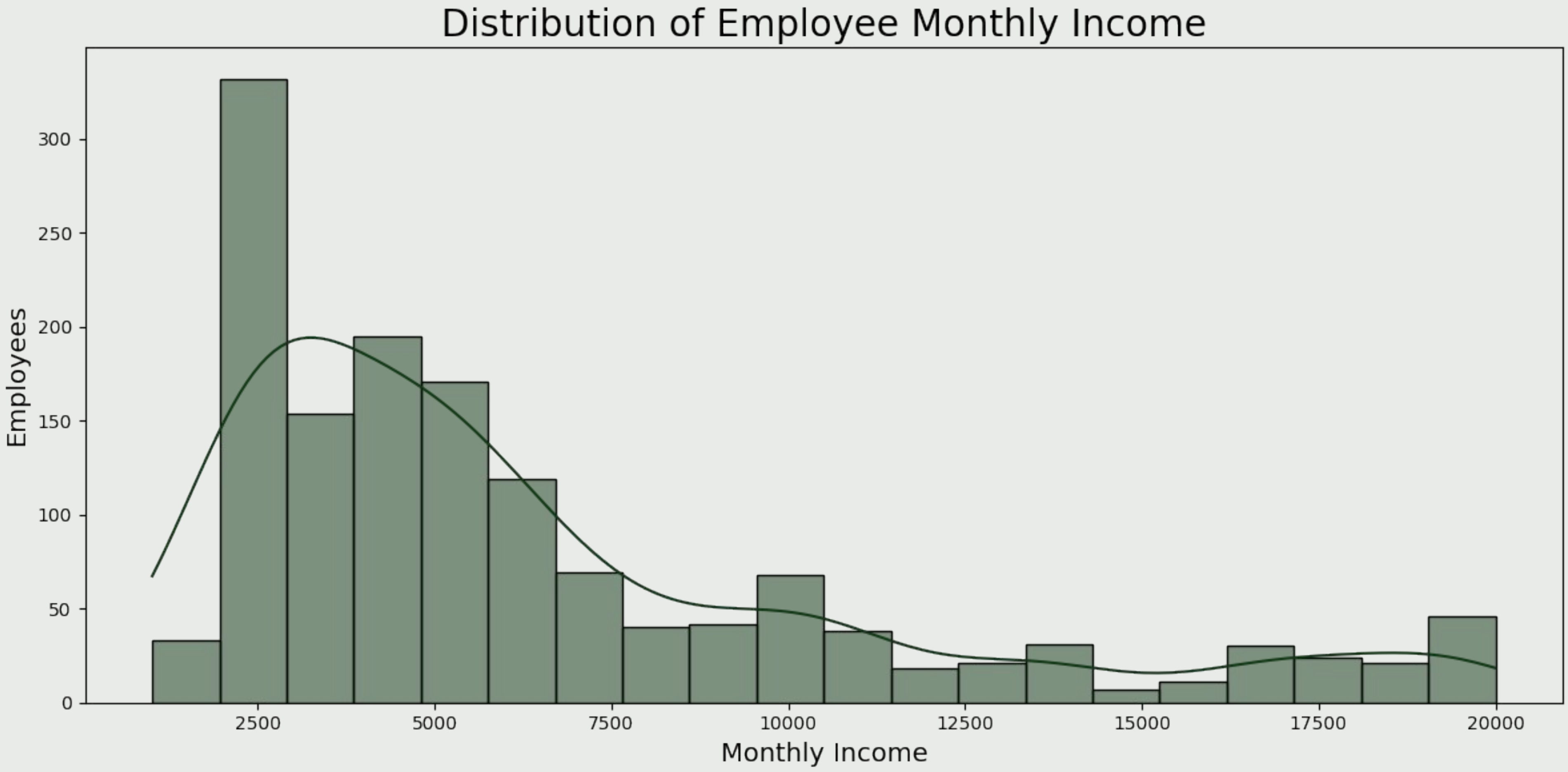
Target Variable

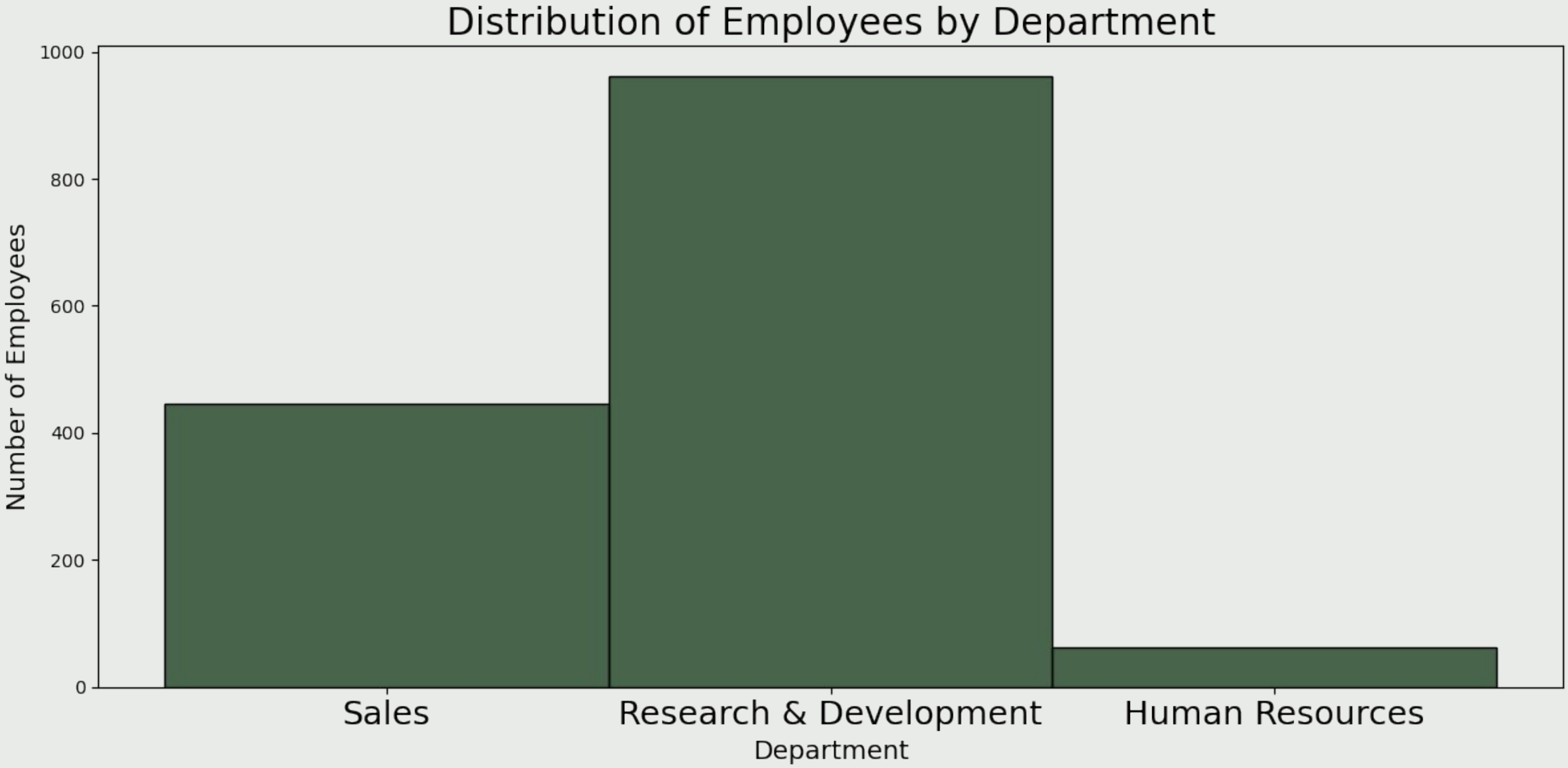
____

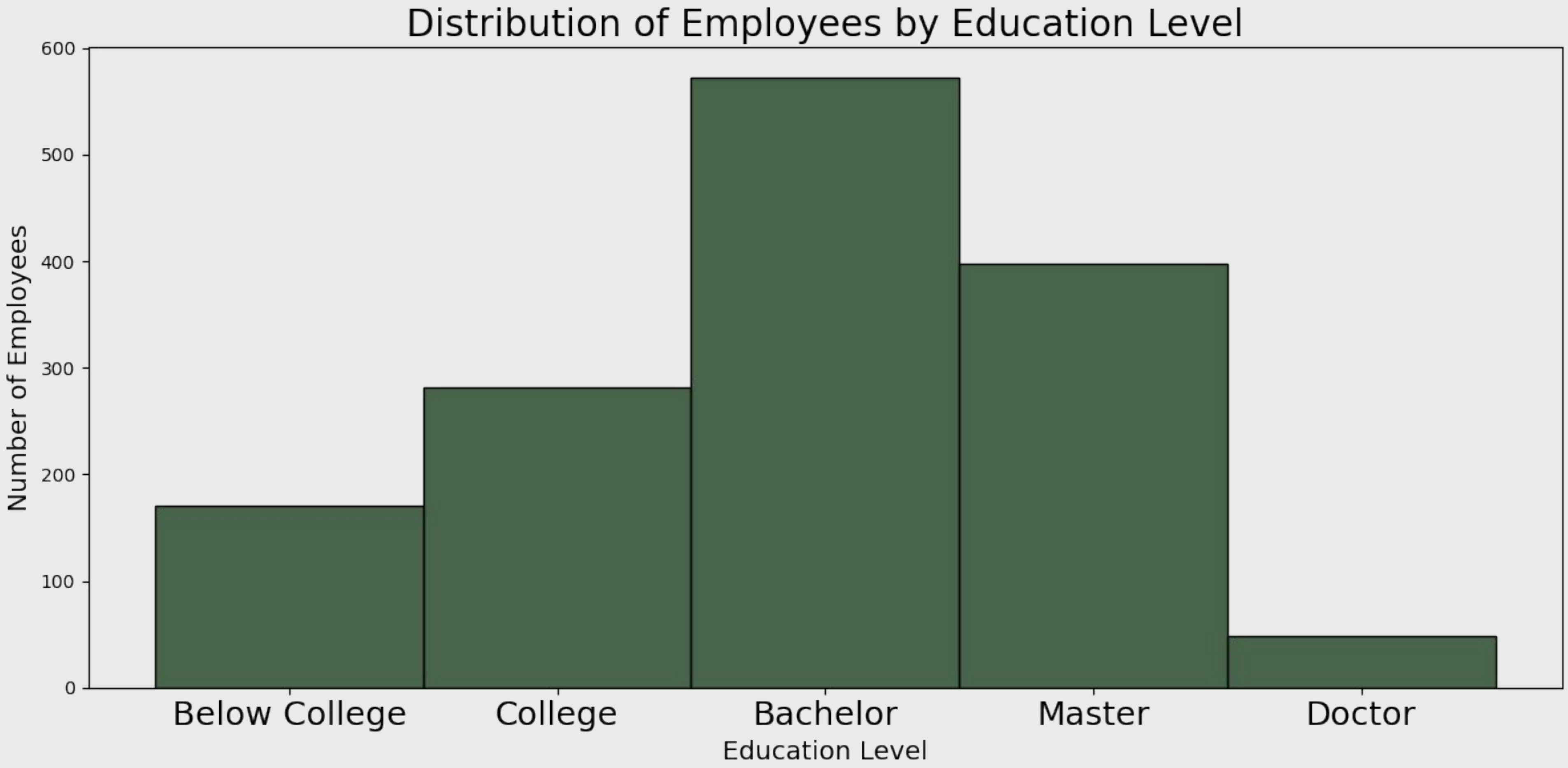
-Attrition

Employee Demographics

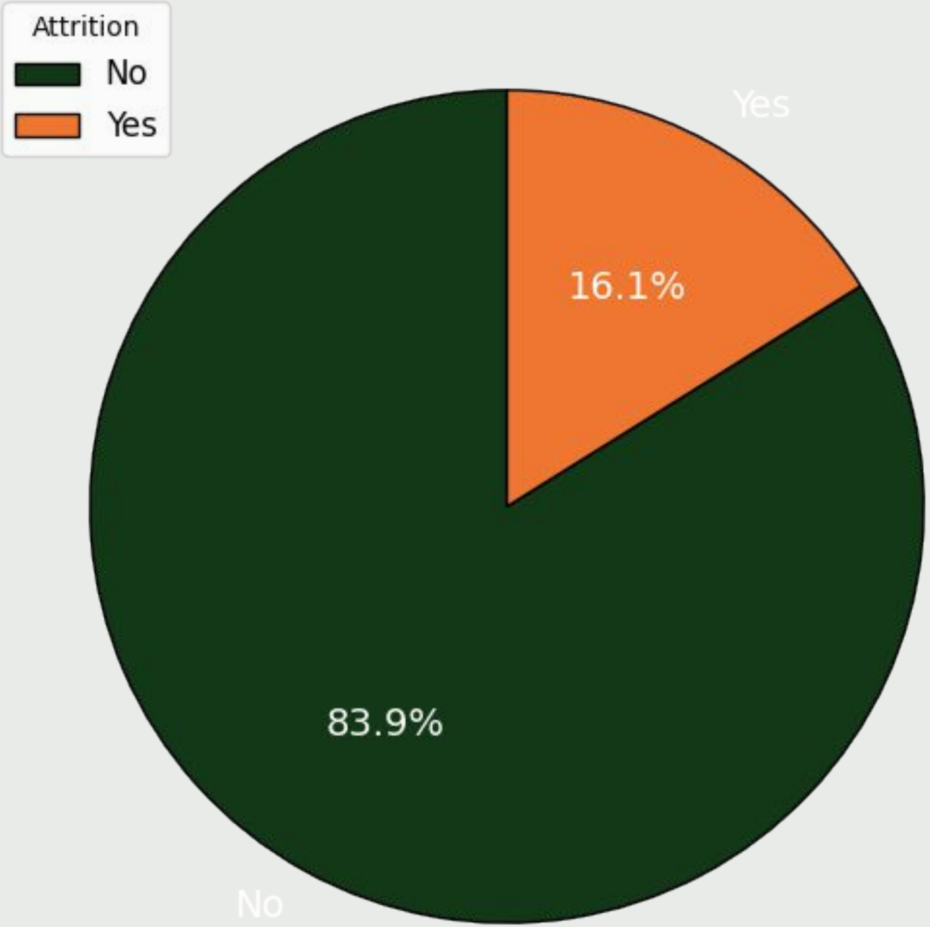

Distribution of Employee Age
Employees
0
20
40
60
80
100
120
140
20
30
40
50
60
Age


Distribution of Employee Monthly Income
Employees
Monthly Income
0
50
100
150
200
250
300
2500
5000
7500
10000
12500
15000
17500
20000

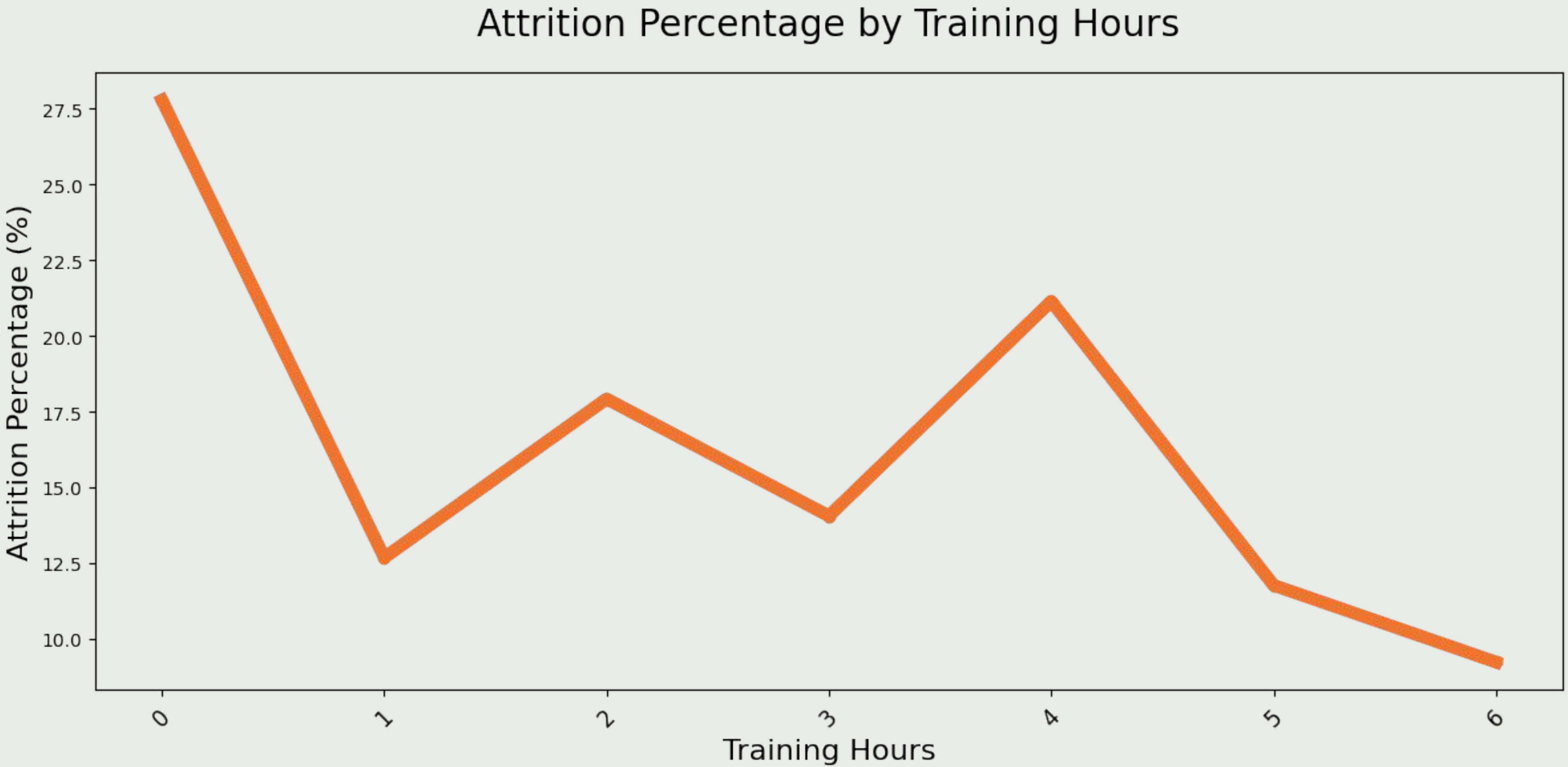

Distribution of Employees by Department
Number of Employees
1000
800
600
400
200
0
Sales
Research & Development
Human Resources
Department

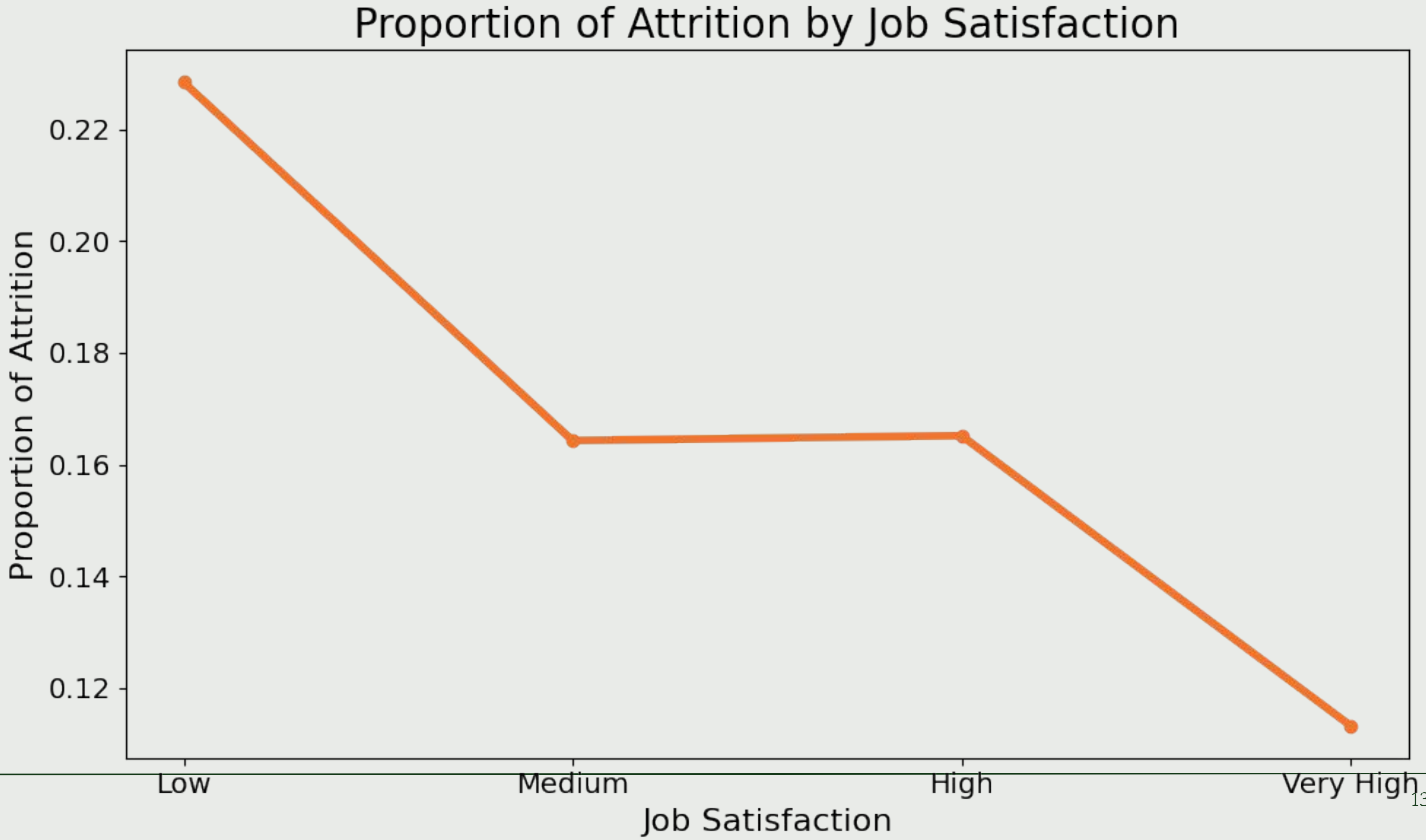

Distribution of Employees by Education Level
Number of Employees
0
100
200
300
400
500
600
Below College
College
Bachelor
Master
Doctor
Education Level

Total Employee Attrition Rate

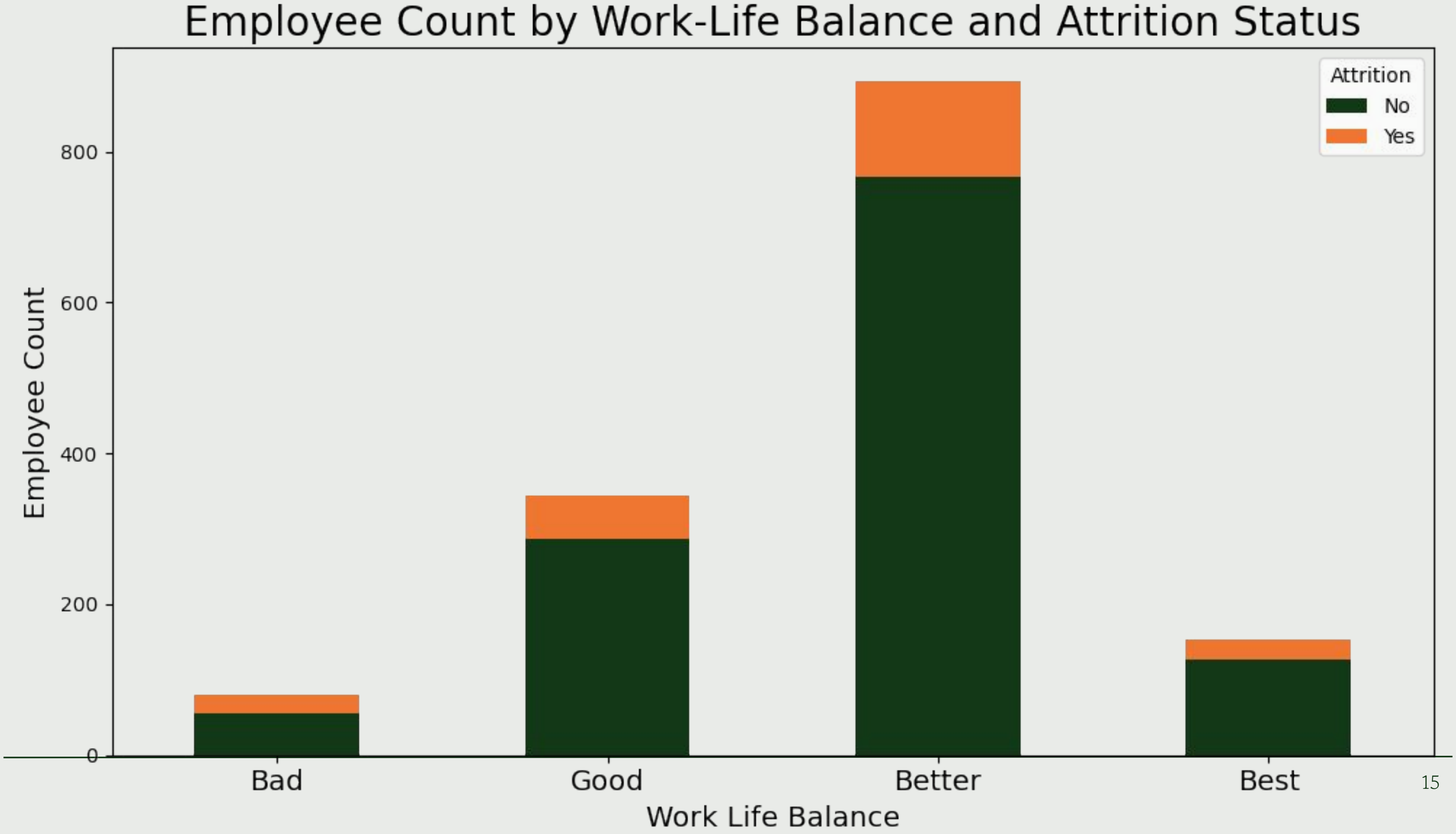


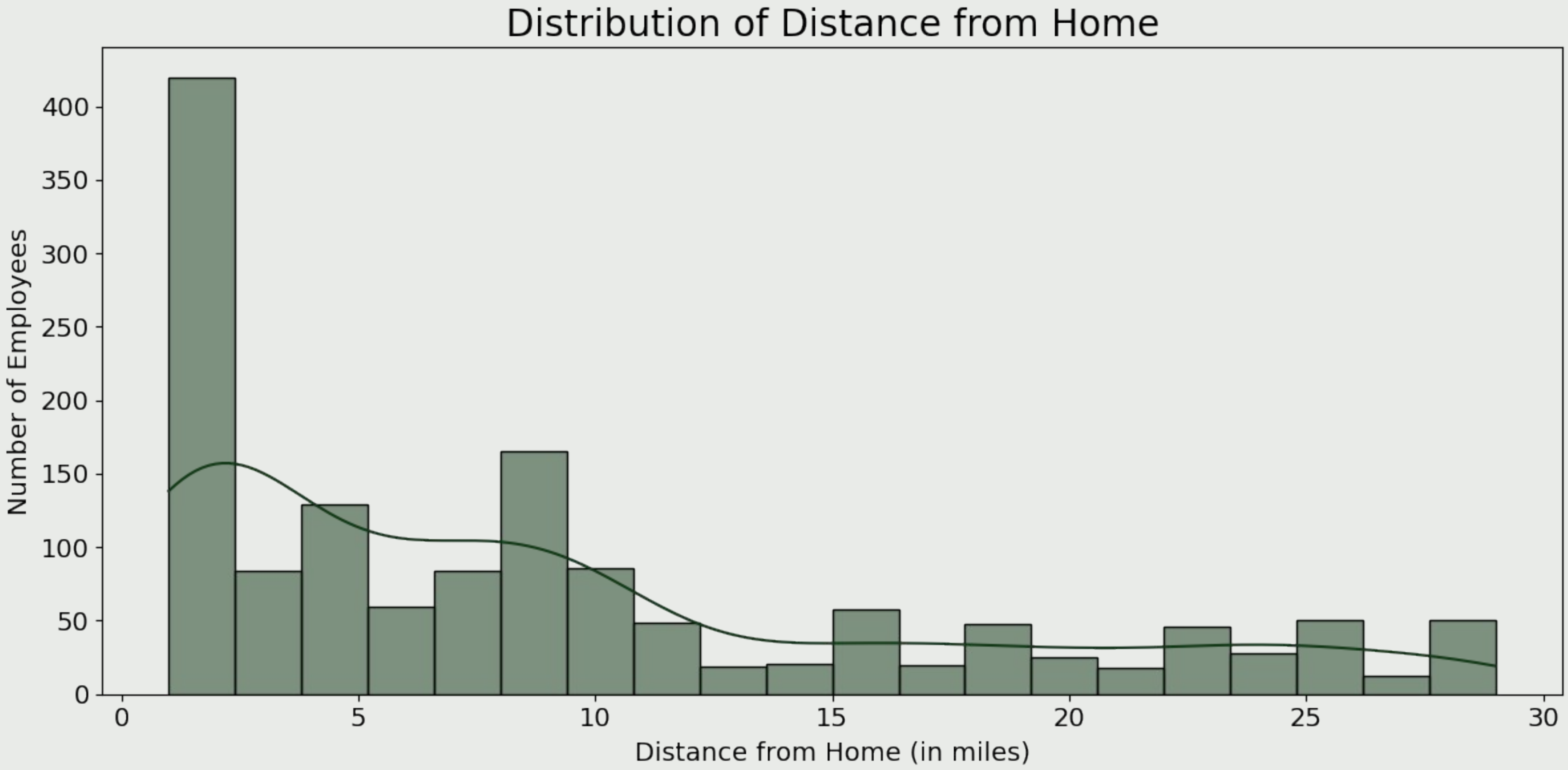
Engagement & Culture


Attrition Percentage by Training Hours
Attrition Percentage (%)
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
0
1
2
3
4
5
6
Training Hours

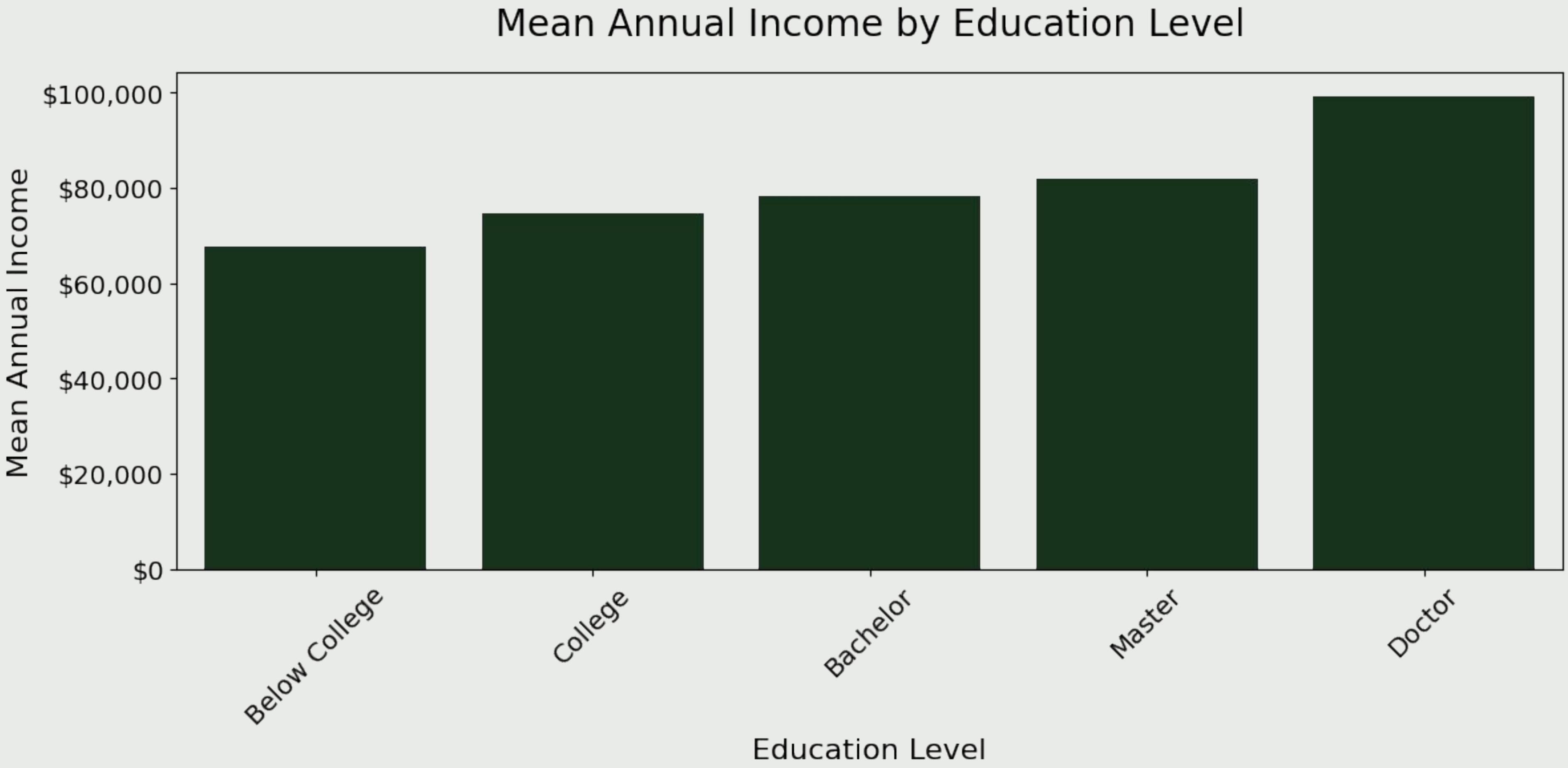

Proportion of Attrition by Job Satisfaction
Proportion of Attrition
0.22
0.20
0.18
0.16
0.14
0.12
Low
Medium
High
Very High
Job Satisfaction

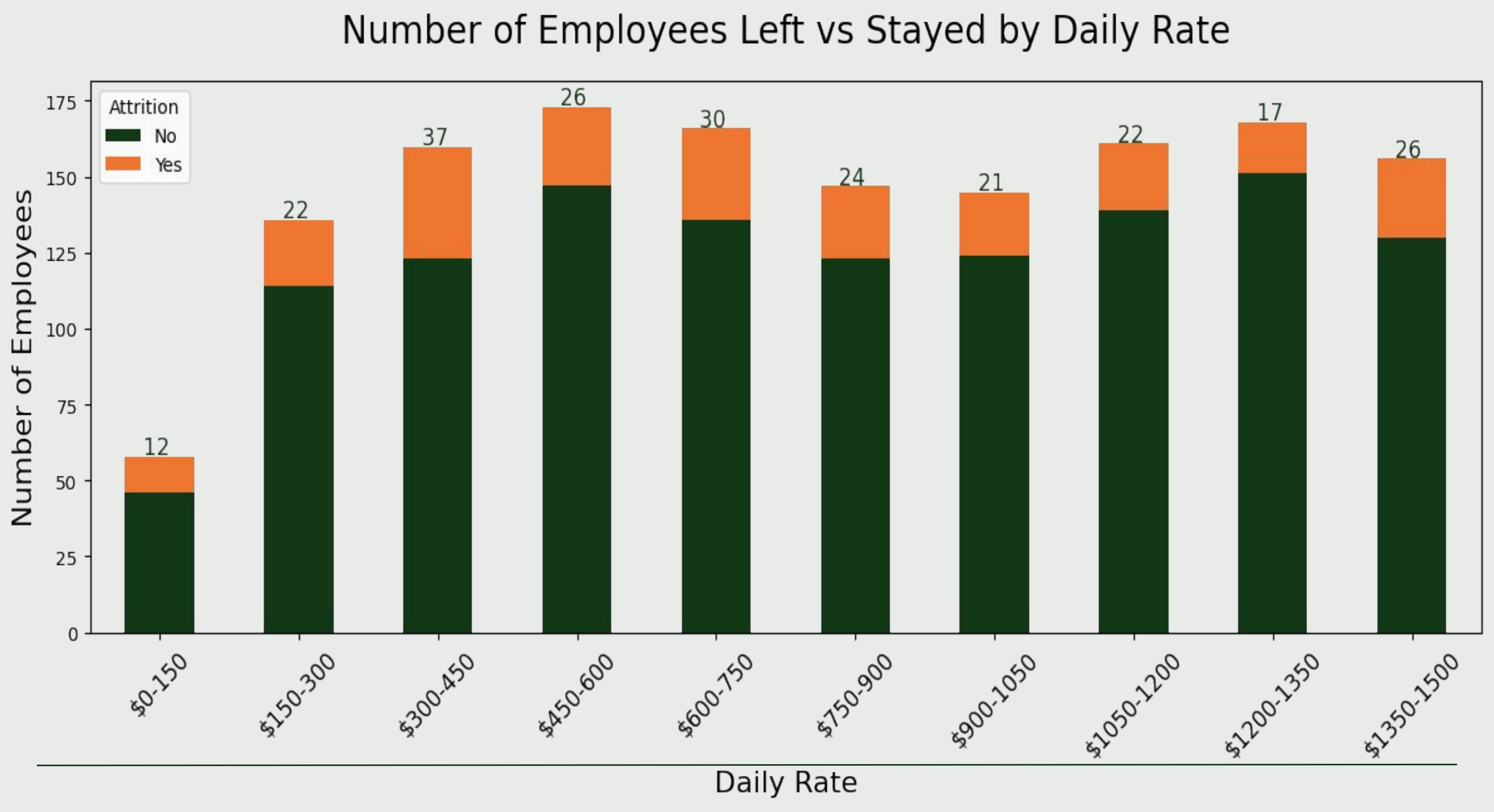
Wellbeing & Work-Life Balance

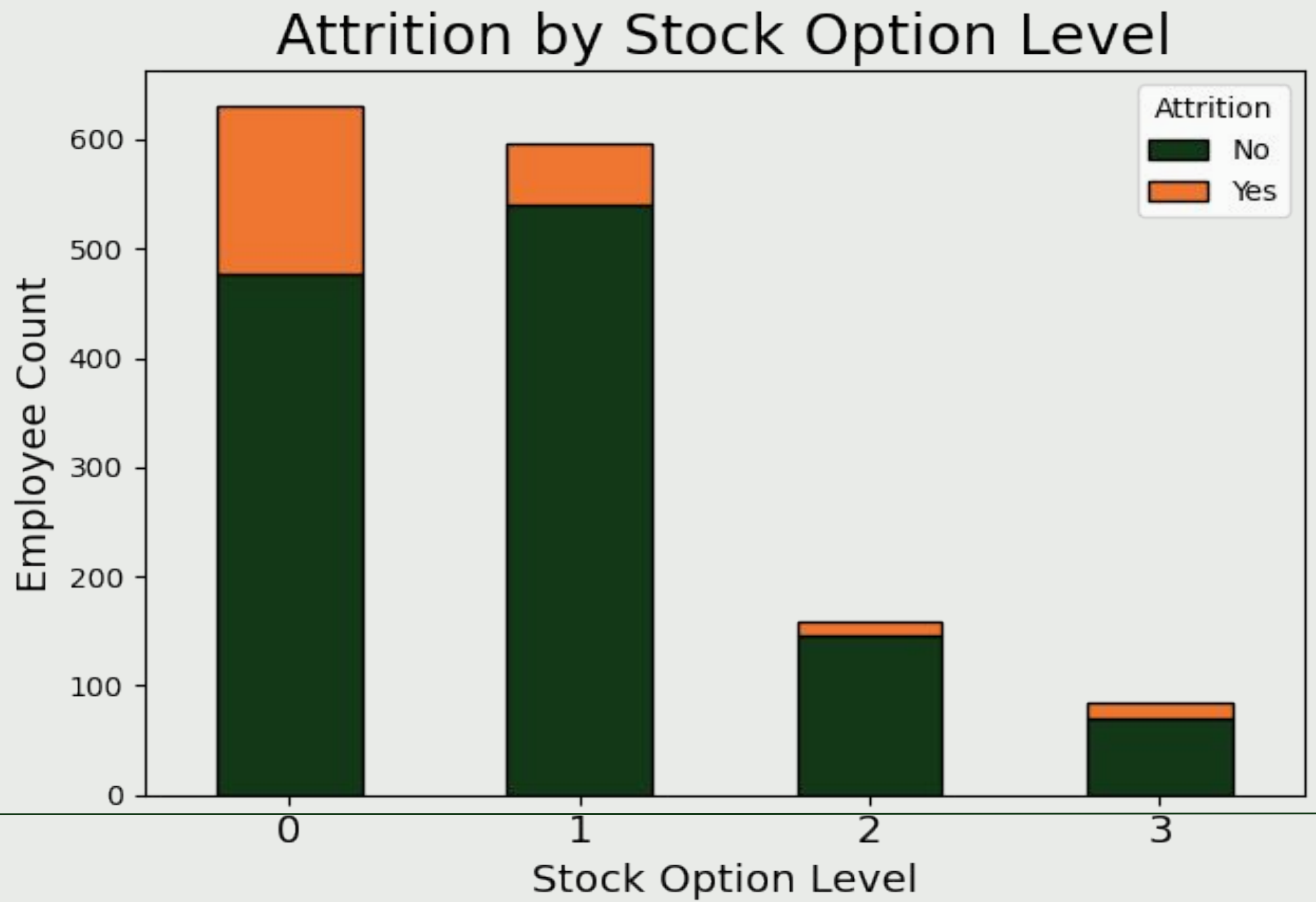

Employee Count by Work-Life Balance and Attrition Status
Attrition
No
Yes
800
600
400
200
0
Employee Count
Bad
Good
Better
Best
Work Life Balance

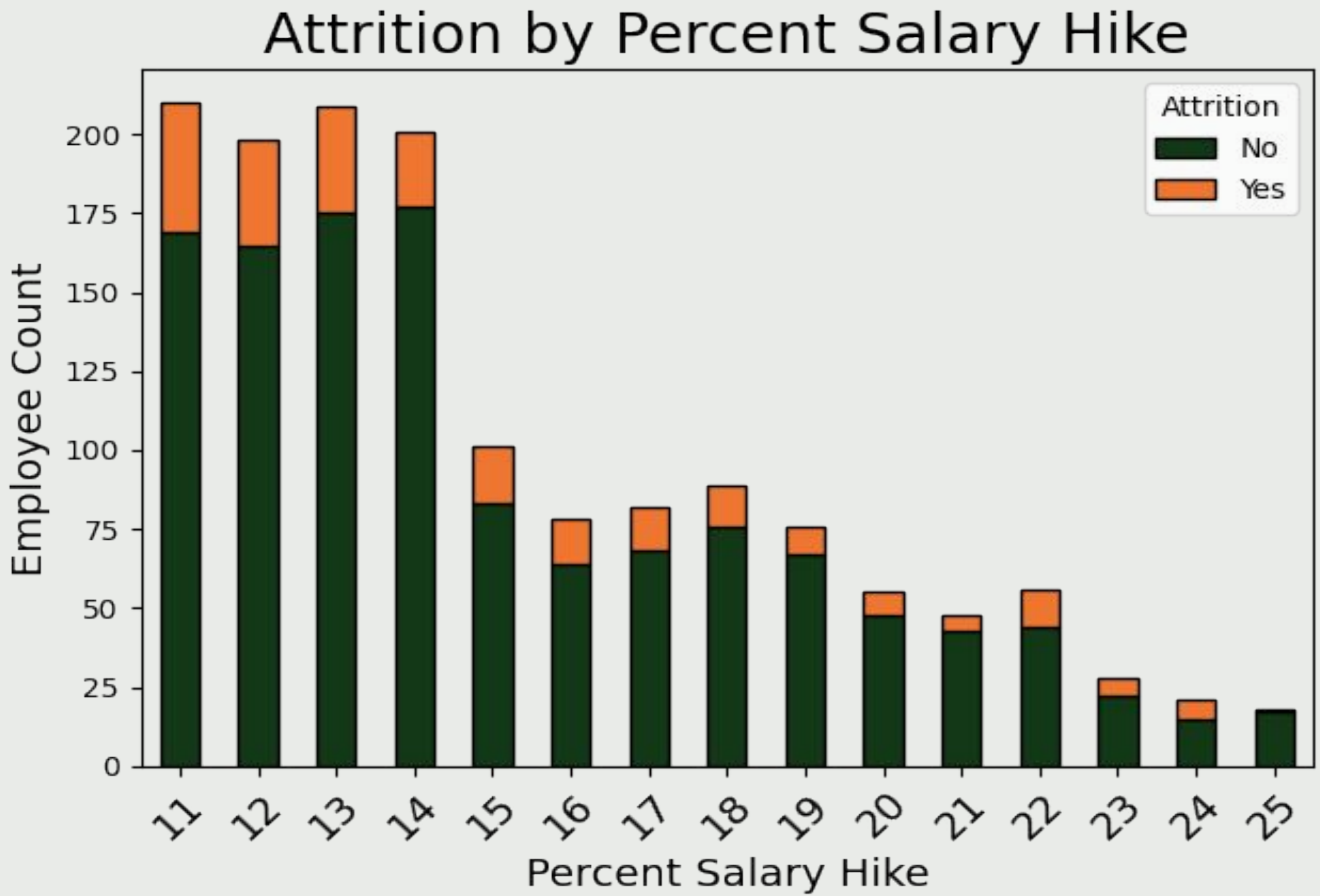

Distribution of Distance from Home
Number of Employees
0
50
100
150
200
250
300
350
400
0
5
10
15
20
25
30
Distance from Home (in miles)

Pay and Benefits


Mean Annual Income by Education Level
Mean Annual Income
$100,000
$80,000
$60,000
$40,000
$20,000
$0
Below College
College
Bachelor
Master
Doctor
Education Level

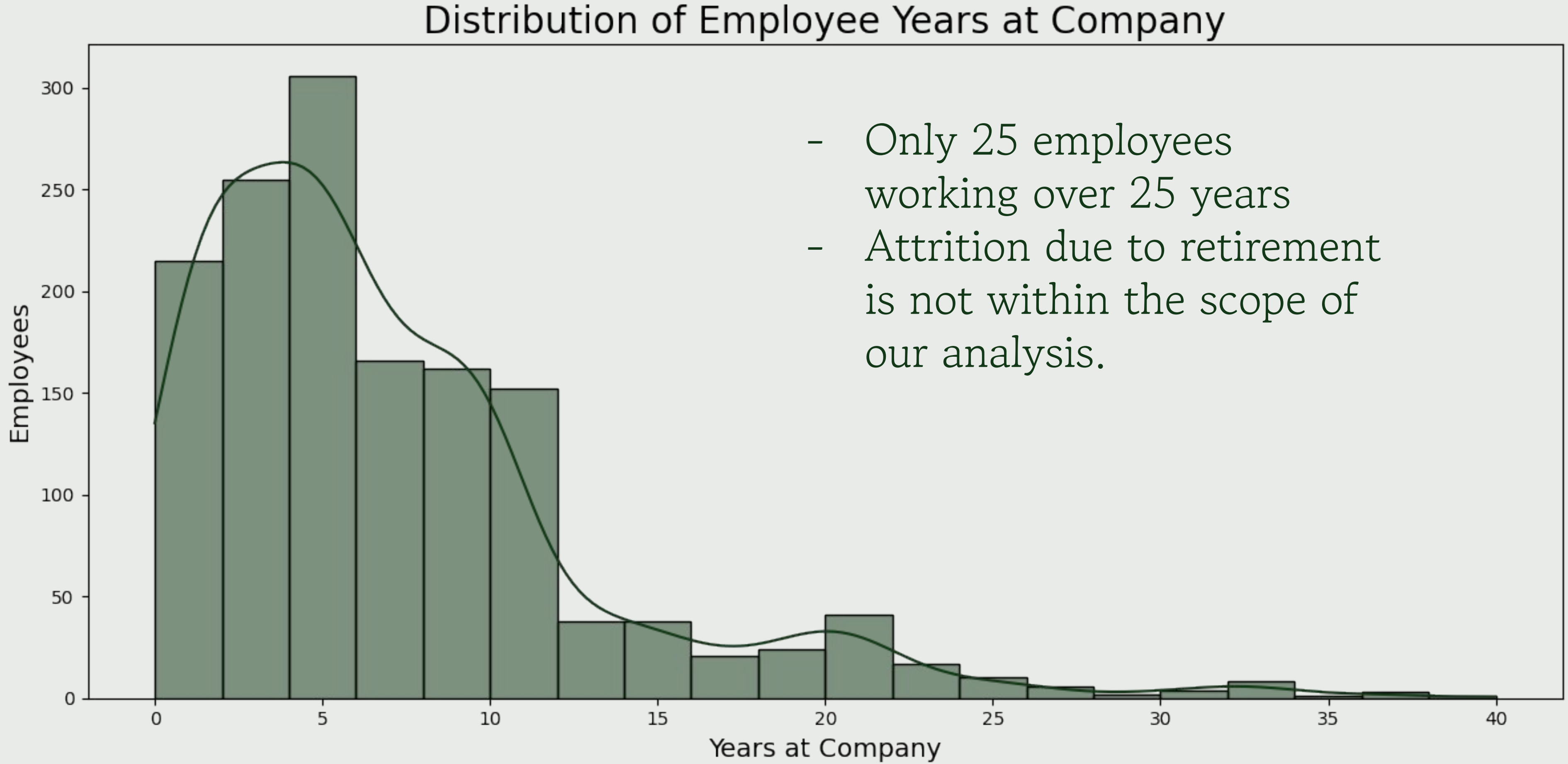

Number of Employees Left vs Stayed by Daily Rate
Attrition
No
Yes
Number of Employees
Daily Rate
0
25
50
75
100
125
150
175
12
22
37
26
30
24
21
22
17
26
$0-150
$150-300
$300-450
$450-600
$600-750
$750-900
$900-1050
$1050-1200
$1200-1350
$1350-1500


Attrition by Stock Option Level
Attrition
No
Yes
Employee Count
0
100
200
300
400
500
600
0
1
2
3
Stock Option Level


Attrition by Percent Salary Hike
Attrition
No
Yes
Employee Count
0
25
50
75
100
125
150
175
200
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Percent Salary Hike

The Retirement Problem

- Attrition rises quickly after 25 years of employment likely do to retirement



- Only 25 employees working over 25 years
- Attrition due to retirement is not within the scope of our analysis.

Objective:

Explore how predictive models **built on our themes** provides actionable insights to decrease employee turnover/attrition and associated costs.

Hypothesis:

Use machine learning to combine theme-specific models for stronger predictions

Our Themes

01	02	03
Engagement and Culture	Wellbeing and Work-Life Balance	Pay and Benefits

Gallup - https://www.gallup.com/467702/indicator-employee-retention-attraction.aspx

Simplified and Enhance

Engagement and Culture	Wellbeing and Work-Life Balance	Pay and Benefits
• Advancement, development or career opportunities • Workplace Culture • Insufficient Training • Unrealistic job expectations • Not Treated with respect	• Relocation • Work-Life Balance • Physical Working Conditions • Personal Reasons (Family, Medical, Etc.)	• Pay • Benefits

Gallup - https://www.gallup.com/467702/indicator-employee-retention-attraction.aspx

Logistic Regression

Nature of the Problem

Binary classification problem where the outcome is either

- ❖ "Yes" (employee leaves)
- ❖ "No" (employee stays)

Probabilistic Output

Probability Scores for predictions which can be useful for making decisions based on the likelihood of attrition.

Efficiency

Quickly **train** and **deploy** the model making it suitable for real-time HR analytics

Hypotheses for Predictive Model

- Job Satisfaction: Employees with lower job satisfaction are more likely to leave the company.
- Monthly Income: **Employees with higher monthly income are less likely to leave the company.**
- Overtime: Employees who work overtime are more likely to leave the company.

Steps to take

- **Load the Dataset**: We'll load the dataset
- **Select Relevant Columns**: We'll use **Monthly_Income** as the feature and **Attrition** as the target variable.
- **Split Data**: We'll split the data into training and testing sets.
- **Train Logistic Regression Model**: We'll initialize and train the **Logistic Regression** model.
- **Evaluate the Model**: We'll evaluate the model using **confusion matrix** (heatmap) and **classification report** (distribution plot).

Support Vector Machine

Support Vector Machine (SVM)

- A supervised training model ideal for classification tasks that can be easily interpreted for our dataset.
- It trains high dimensional data that can give us better information by learning complex relationships.
- It helps us analyze various factors by finding the optimal hyperplane separating the two classes.

PROS:

- Effective for high dimensional data with multiple variables.
- It can potentially find a strong boundary within the different classes from the hyperplane.

CONS:

- It is not probabilistic by nature compared to other generic models like Logistic Regression.
- It can have limited scalability due to the size of the dataset.

Support Vector Machine (SVM)

These are the variables that are likely to relied upon training SVM:

OverTime – Employees who are experiencing burnouts or poor working conditions are highly likely to leave.

MaritalStatus – Employees who are not married can take risks and relocate for better prospects.

DistanceFromHome - Employees who commute for a longer period of time to work could be less satisfied with their job.

JobRole - Employees who feel their job is monotonous and mundane will lose motivation and look for other places.

Department - Employees who consider their job role is not acknowledged within the department are more likely to leave their job.

Decision Tree

Decision Tree

Decision Trees are great for **visual explanations** and spotting threshold effects. Decision Trees are powerful and flexible models — they can handle both categorical and numerical variables without needing much preprocessing.

Pros

- Interpretability & Transparency
- Handles Mixed Data Types
- Captures Nonlinear Relationships
- Fast to Train and Easy to Visualize

Cons

- Overfitting Risk
- Instability
- Not Always the Most Accurate
- Doesn't Handle Imbalanced Data Well

Decision Tree

Best Subset of Variables for Decision Tree (Employee Attrition)

Age – Younger employees may be more mobile.

DistanceFromHome – Long commutes can lead to attrition.

NumCompaniesWorked – Can indicate job-hopping behavior.

TotalWorkingYears – Shows overall career experience.

YearsAtCompany – Key indicator of tenure.

YearsSinceLastPromotion – Career stagnation can lead to turnover.

We hypothesize that decision tree models can effectively identify hierarchical patterns in employee data — such as tenure, overtime, job satisfaction, and promotion — that distinguish employees who are likely to leave from those who are likely to stay.

HYPOTHESIS

Use machine learning to combine theme-specific models for stronger predictions

Variable Breakdown

Engagement and Culture

- BusinessTravel
- Department
- EducationField
- EnvironmentSatisfaction
- JobInvolvement
- JobRole
- JobSatisfaction
- PerformanceRating
- RelationshipSatisfaction
- YearsInCurrentRole
- YearsWithCurrManager

Wellbeing & Work-Life Balance

- Age
- DistanceFromHome
- Gender
- MaritalStatus
- OverTime
- TotalWorkingYears
- TrainingTimesLastYear
- WorkLifeBalance
- YearsAtCompany
- YearsSinceLastPromotion
- NumCompaniesWorked

Pay And Benefits

- DailyRate
- HourlyRate
- MonthlyIncome
- MonthlyRate
- PercentSalaryHike
- StockOptionLevel
- Education
- JobLevel

Theme Specific Model Results

Engagement and Culture

Best results achieved using **Decision Tree**

No Attrition	**210**	**37**
Attrition	**25**	**22**
	No Attrition	Attrition

Accuracy
78.91%

Wellbeing & Work-Life Balance

Best results achieved using **SVM**

No Attrition	**194**	**53**
Attrition	**21**	**26**
	No Attrition	Attrition

Accuracy
74.83%

Pay and Benefits

Best results achieved using **Decision Tree**

No Attrition	**177**	**70**
Attrition	**30**	**17**
	No Attrition	Attrition

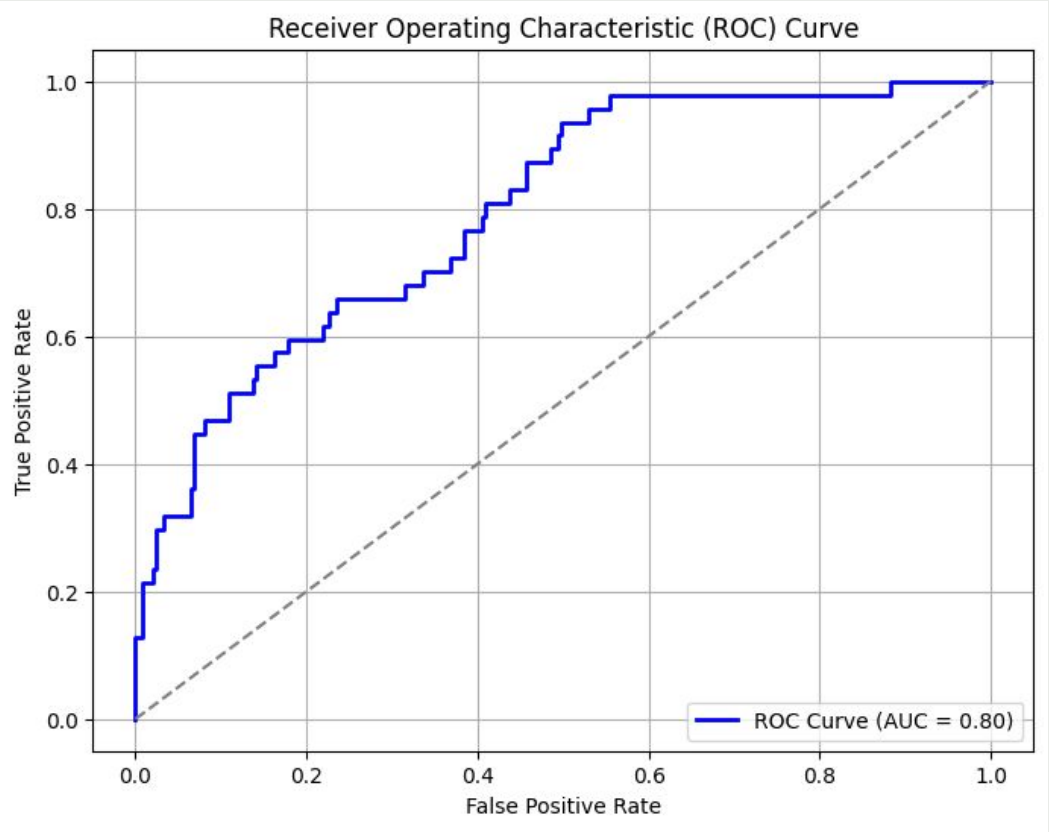
Accuracy
65.99%

Final Model Results

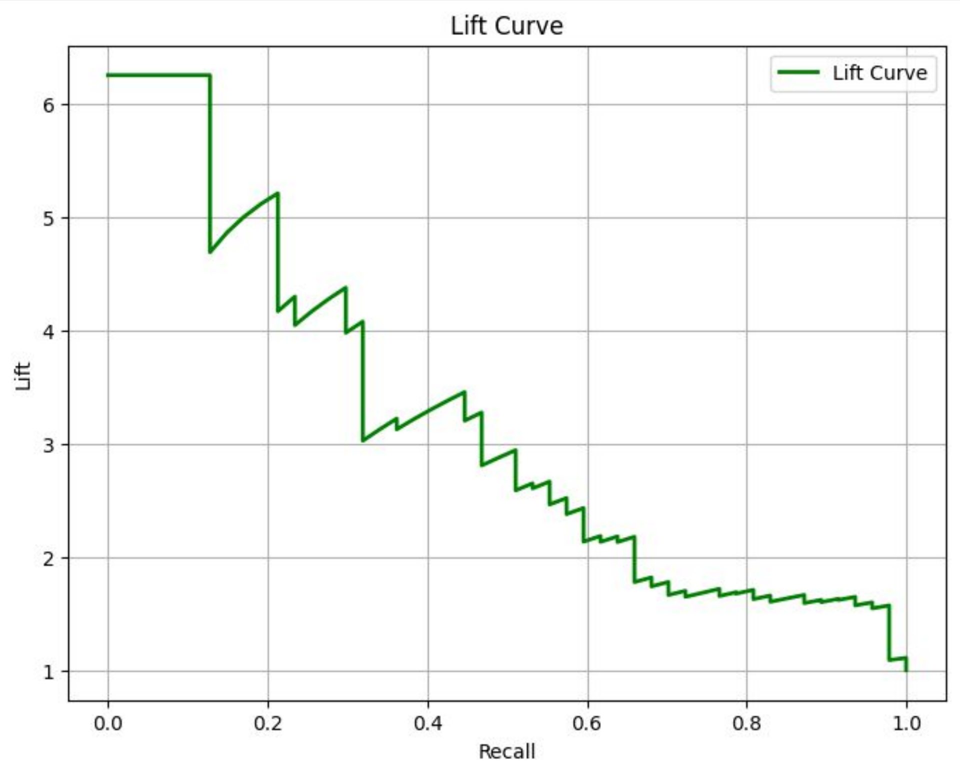
Stacking Model Performance

No Attrition	**230**	**17**
Attrition	**27**	**20**
	No Attrition	Attrition

Our stacking model, which combines logistic regression, decision tree, and SVM, has achieved an accuracy rate of **85%**. This means that out of every 100 predictions made by our model, 85 are correct.



- The AUC of **0.80** indicates that our model is quite good at distinguishing between the two groups we're trying to predict.
- If we randomly picked one employee who will leave and one who won't, there's an 80% chance our model will correctly identify which is which.
- This accuracy allows us to make more informed decisions about resource allocation for retention strategies, ultimately saving costs and improving employee retention



- This lift curve shows how much more effective our model is at identifying potential leaver.
- We are over **six (6)** times more likely to find someone who will actually leave.
- The lift gradually decreases, as expected. This suggests that the model's precision in identifying likely leavers diminishes as we try to capture a larger proportion of them.

Business Impact

- **Loss of Intellectual Capital, Client Risk and Delivery Delays**
  High attrition or layoffs in key roles (e.g., R&D, AI, cloud) leads to deep knowledge drain and longer onboarding cycles.

- **Reduced Innovation Velocity**
  Disengaged or departing talent slows product development, especially in competitive spaces like quantum computing and AI.

- **Increased Rehiring & Training Costs**
  Estimated at ~$100K per employee, with higher costs for specialized or senior positions.

Conclusion

Recommendations

1. **Predictive Retention Dashboards:**
    - Use your model in HR systems to flag high-risk employees monthly & Automate alerts for HRBPs (HR Business Partners).
2. **Tailored Interventions**:
    - If risk is due to engagement: manager coaching or rotation.
    - If pay-related: off-cycle review or retention package.
3. **Pilot & Iterate**:
    - Start with one region or department.
    - Measure attrition rate reduction and engagement uplift