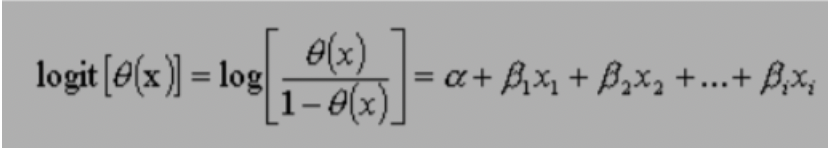
# **EMPLOYEE RISK ATTRITION ASSESMENT**

# **USING LOGISTIC REGRESSION ANALYSIS** **SACRED HEART UNIVERSITY** **AI FOR BUSINESS BUAN-617-FO**

**PROF: EASIN HASAN**  
  
**PROJECT BY**  
**MUKESH KORRI – 0935730**

**ABSTRACT**  
This paper will cover the usage of logistic regression as a modeling algorithm to predict employee attrition risk within a company based on employee data. In this project, I will be covering my analysis and approach through different process flows in the data science pipeline. The main goal is to understand the reasonings behind employee turnover and to come up with a model to classify an employee’s risk of attrition. A recommendation for a retention plan was created, which incorporates some best practices for employee retention at different risk levels of attrition.  
  
**INTRODUCTION**  
“You don’t build a business. You build people, and people build the business.” - Zig Ziglar   
Long-term success, a healthy work environment, and high employee retention are all signs of a successful company. In a sense, it’s the employees who make the company. It’s the employees who do the work. It is the employees who shape the company’s culture. But when a company experiences a high rate of employee turnover, then something goes wrong. This can lead the company to huge monetary losses by these innovative and valuable employees.  
  
When companies experience a high rate of turnover, their investments in their employees are drained away, which includes a loss of salary, benefits, bonuses, training, and other expensive resources. Companies that maintain a healthy organization and culture are always a good sign of future prosperity. Recognizing and understanding what factors that were associated with employee turnover will allow companies and individuals to limit this from happening and may even increase employee productivity and growth. These predictive insights give managers the opportunity to take corrective steps to build and preserve their successful business.  
  
**OBJECTIVE**  
The company wants to understand what factors contributed most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies for targeted employees. Overall, the implementation of this model will allow management to create better decision-making actions  
  
**OVERVIEW OF LOGISTIC REGRESSION**

Logistic Regression commonly deals with the issue of how likely an observation is to belong to each group. This model is commonly used to predict the likelihood of an event occurring. In contrast to linear regression, the output of logistic regression is transformed with a logit function. This makes the output either 0 or 1. This is a useful model to take advantage of for this problem because we are interested in predicting whether an employee will leave (0) or stay (1).   
  
Another reason for why logistic regression is the preferred model of choice is because of its interpretability. Logistic regression predicts the outcome of the response variable (turnover) through a set of other explanatory variables, also called predictors. In context of this domain, the value of our response variable is categorized into two forms: 0 (zero) or 1 (one). The value of 0 (zero) represents the probability of an employee not leaving the company and the value of 1 (one) represents the probability of an employee leaving the company.  
  
Logistic Regression models the probability of ‘success’ is as:

  
  
The equation above shows the relationship between the dependent variable (success), denoted as (θ) and independent variables or predictor of event, denoted as xi. Where α is the constant of the equation and, β is the coefficient of the predictor variables

## **O.S.E.M.N. PIPELINE**

I will be following a typical data science pipeline, which is called “OSEMN” (pronounced awesome):

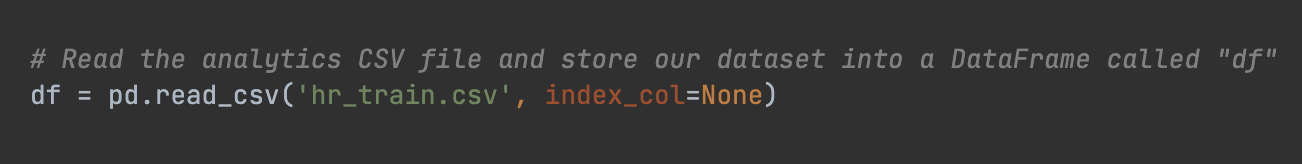
1. Obtaining the data is the first approach in solving the problem.
2. Scrubbing or cleaning the data is the next step. This includes data imputation of missing or invalid data and fixing column names.
3. Exploring the data will follow right after and allow further insight of what our dataset contains. Looking for any outliers or weird data.
4. Modeling the data will give us our predictive power on whether an employee will leave.
5. Interpreting the data is last. With all the results and analysis of the data, what conclusion is made? What factors contributed most to employee turnover? What relationship of variables was found?

## **OBTAINING THE DATA**

The data was found from the “Human Resources Analytics” dataset provided by Kaggle’s website:

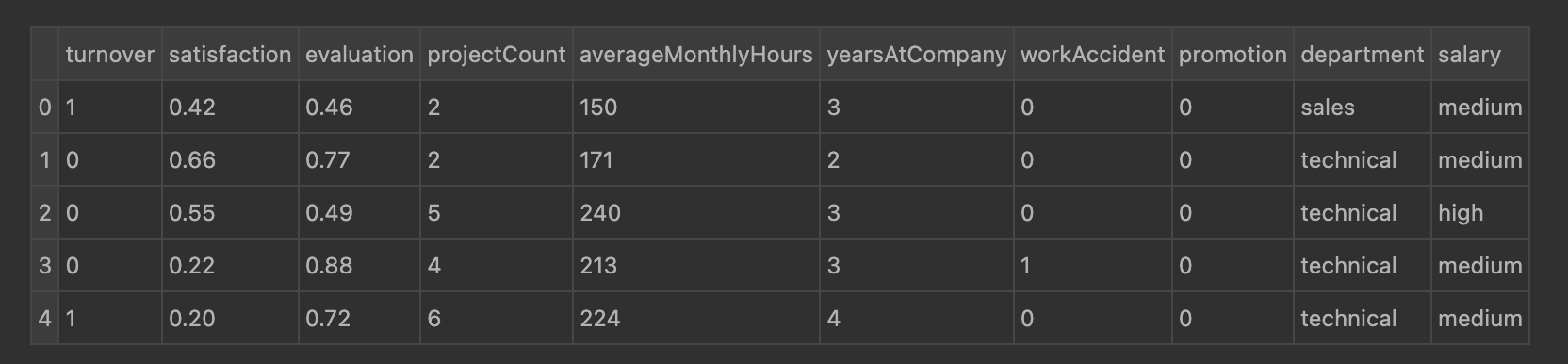
<https://www.kaggle.com/datasets/kavitajambhale/hr-employer-leaving-the-company-prediction>This dataset is a simulation of a hypothetical company, which means that the features and observations used are all made up to mimic a real-world scenario. The number of observations given from the dataset contains 10,499 employee information.

I will be using Python as the programming language for the analysis:

  
  
**DATA PREPERATION/CLEANING**

Typically, data preparation/cleaning requires a lot of work and can be a very tedious procedure. This dataset from Kaggle is clean and contains no missing values. But still, I will have to examine the dataset to make sure that everything else is readable and that the observation values match the feature names appropriately.  
  
**This involved:**

* Conversion of data into categorical data, such as the “department” and “salary” features
* Renaming of features for better readability
* Feature selection with a decision tree classifier
* Check to see if there are any missing values in the dataset



**The following independent variables were used in the model:**

• Satisfaction: An employee’s level of satisfaction in percentage

• Evaluation: An employee’s evaluation score in percentage

• Project Count: The number of projects the employee has done

• Average Monthly Hours: The total monthly hours an employee worked

• Years At Company: The number of years an employee was at the company

• Work Accident: Whether an employee had an accident or not. Where 0 (zero) means no and 1 (one) means yes

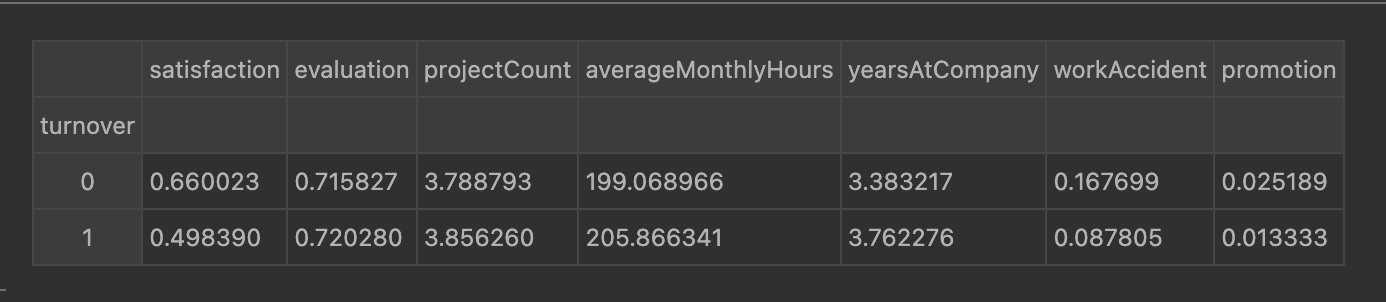
• Promotion: Whether an employee has had a promotion within the last five years. Where 0 (zero) means no and 1 (one) means yes

• Department: The type of department an employee worked under. Which includes sales, accounting, hr., technical, support, management, IT, product management, and marketing.

• Salary:The type of salary an employee gets, which ranges from low, medium, or high.  
  
**EXPLORATORY DATA ANALYSIS**

1. **Statistical Overview:**   
   Here are some important numbers to keep in mind of the dataset:

* There is 10,499 employees and 10 independent variables
* Turnover rate: 24%
* Mean satisfaction: 0.61

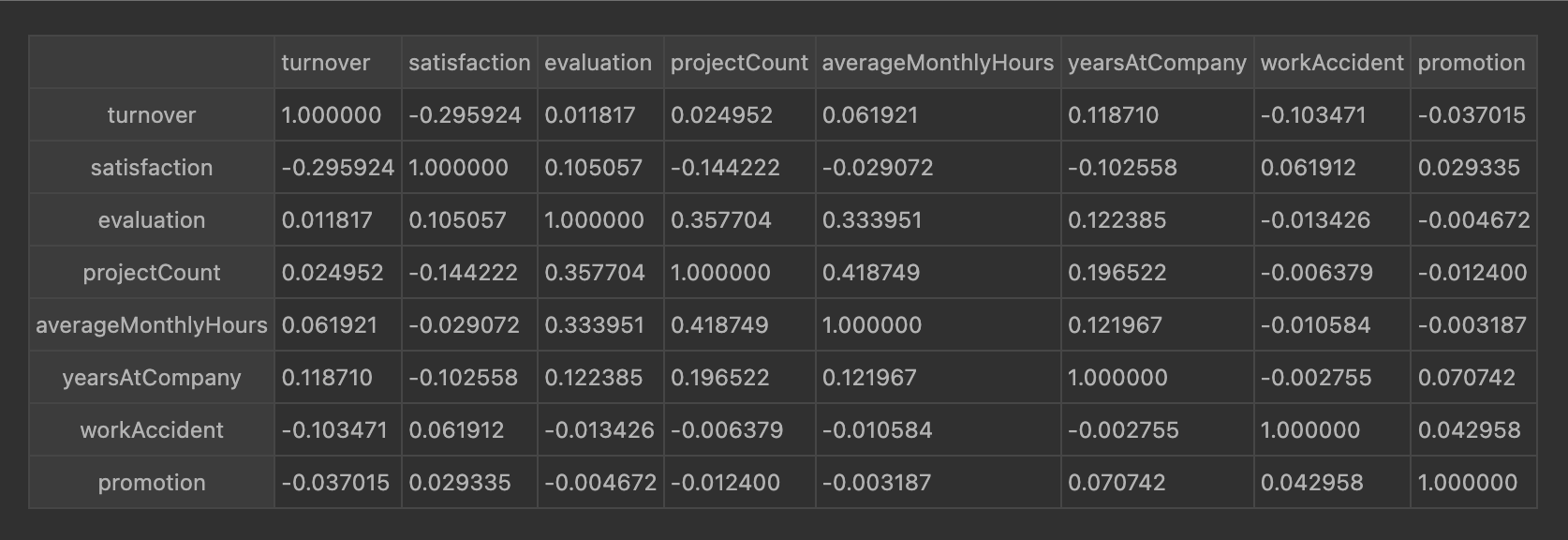


**b. Correlation Matrix & Heatmap**

**Summary:**

From the heatmap, there is a positive (+) correlation between the variables: projectCount, averageMonthlyHours, and evaluation. Which means that the employees who worked more hours and did more projects had higher evaluations.

For the negative (-) relationships, the most important feature that correlated with our target variable (turnover) is satisfaction. This should support our initial intuition that employees who tend to quit would normally have a lower satisfaction level.

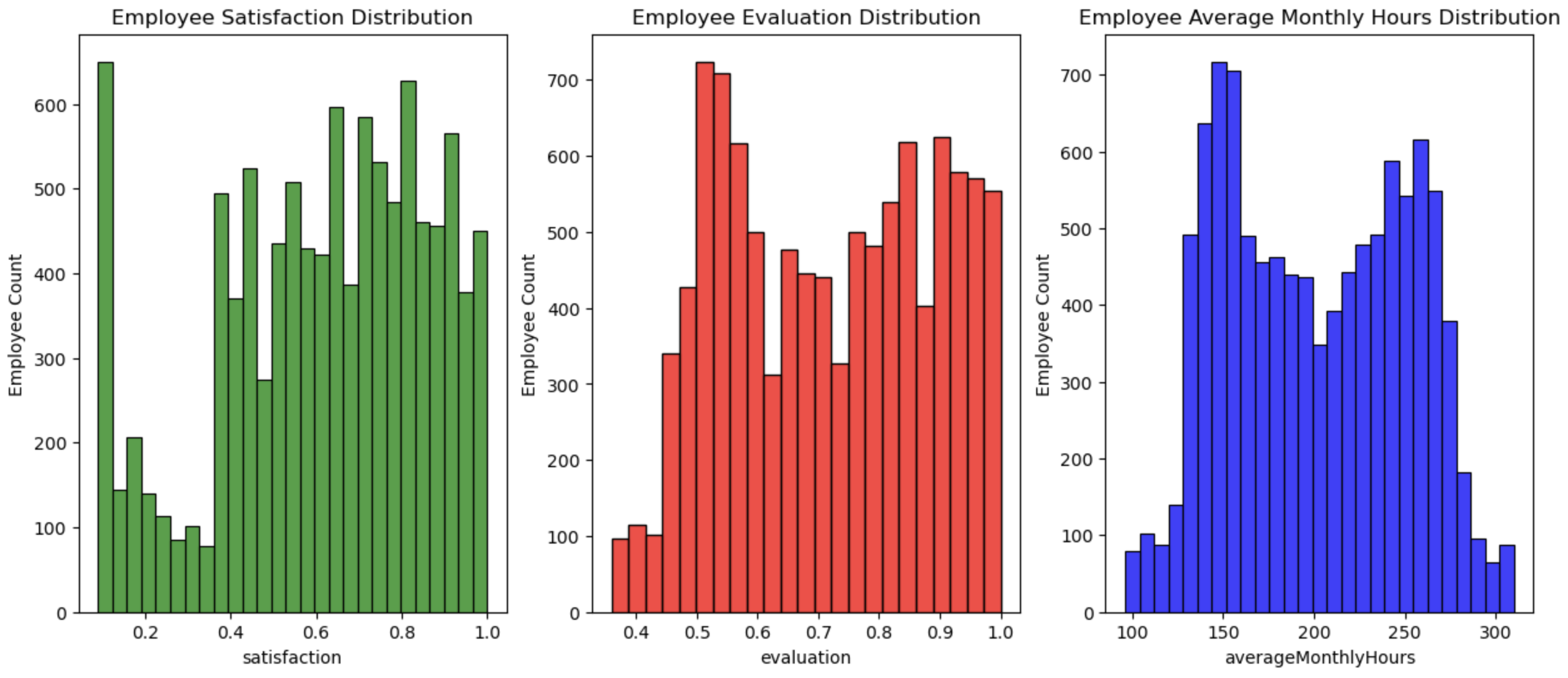


**Histogram (Satisfaction / Evaluation / Average Monthly Hours)**

**Summary:**

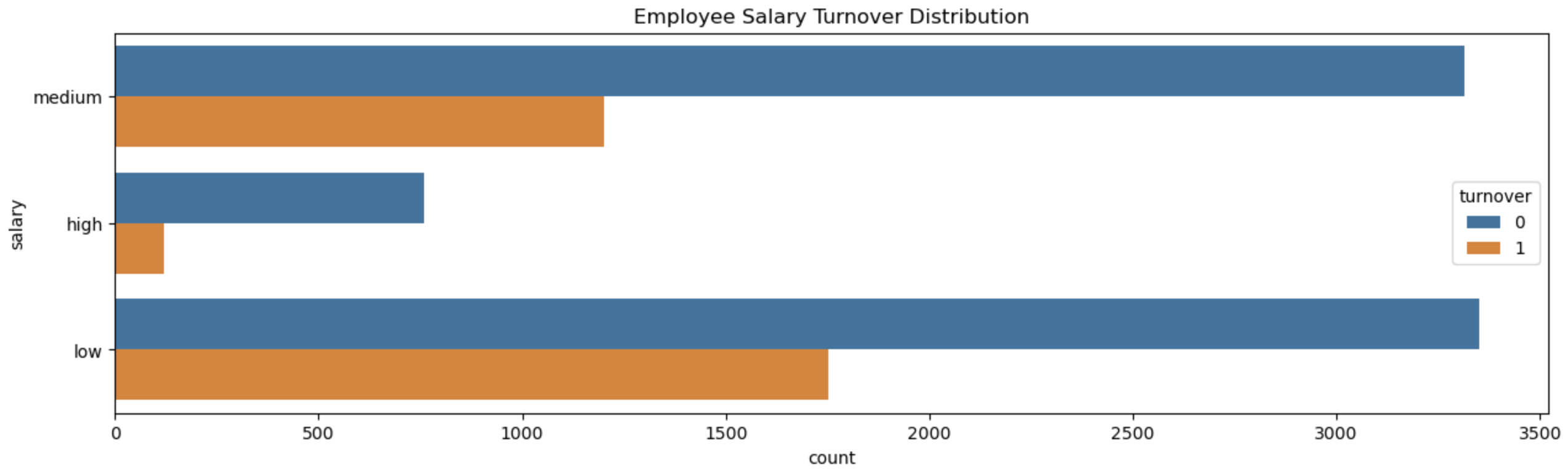
Let us examine the distribution on some of the employee’s features:

* **Satisfaction**: There are three distributions for employee satisfaction in the dataset. One group falls within satisfaction level of (0–0.3), another within satisfaction level of (0.3-0.5), and one from (0.5-1).
* **Evaluation:** There are three distributions for employee evaluation. One group falls within the lower spectrum of (0-0.55), the middle spectrum of (0.55-.7), and the higher spectrum of (.7-1).
* **Average Monthly Hours:** There are three distributions for employees, average monthly hours. Those that work 100-150 hours, another group that works 150-250 hours, and a third group who works 250-300 hours.
* The evaluation and average monthly hours feature both share a similar distribution, which indicates high collinearity of the features.

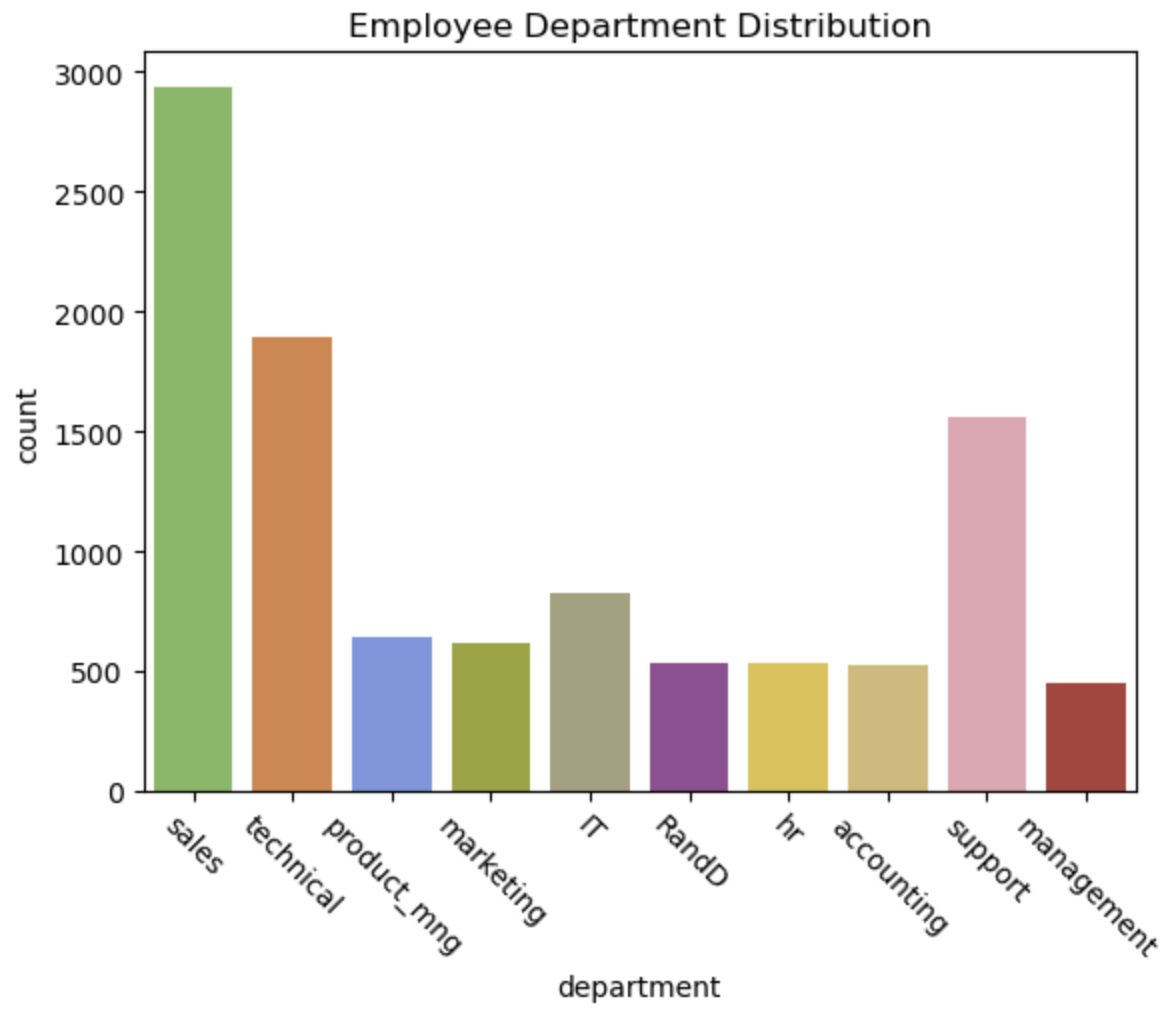
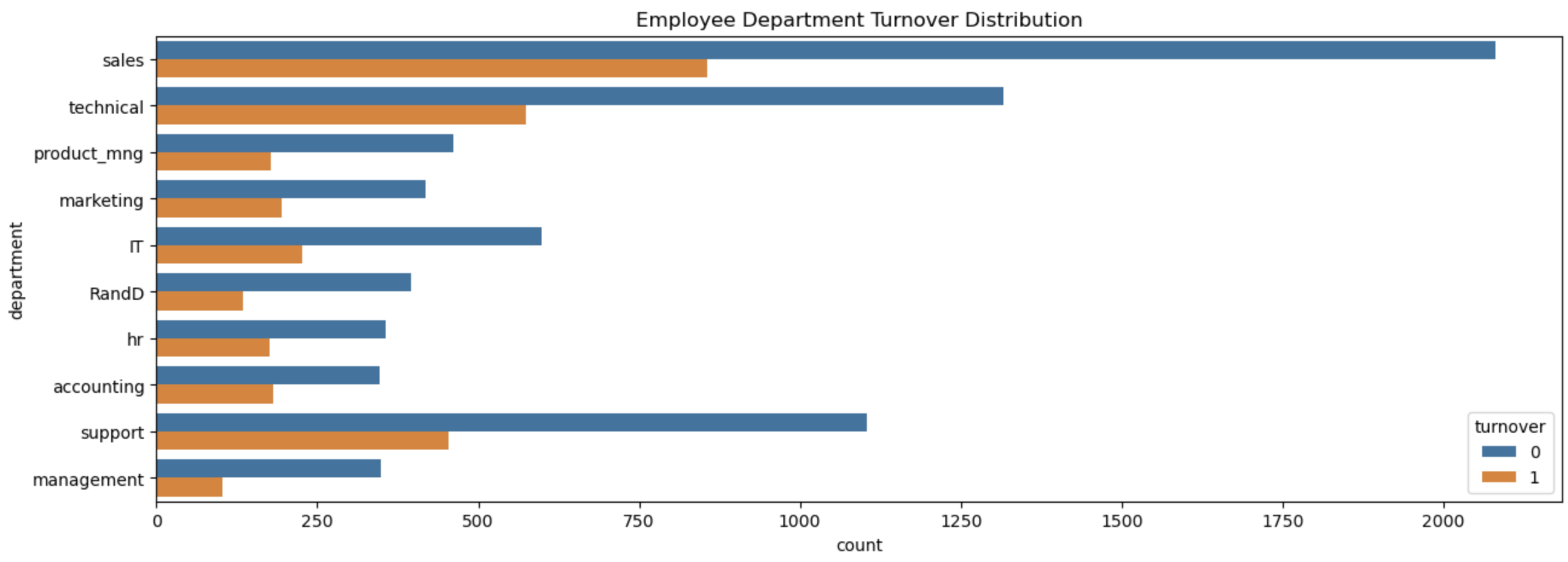
  
**Salary V.S. Turnover**

**Summary:**

* Majority of employees who left either had low or medium salary
* Only a few employees left with high salary

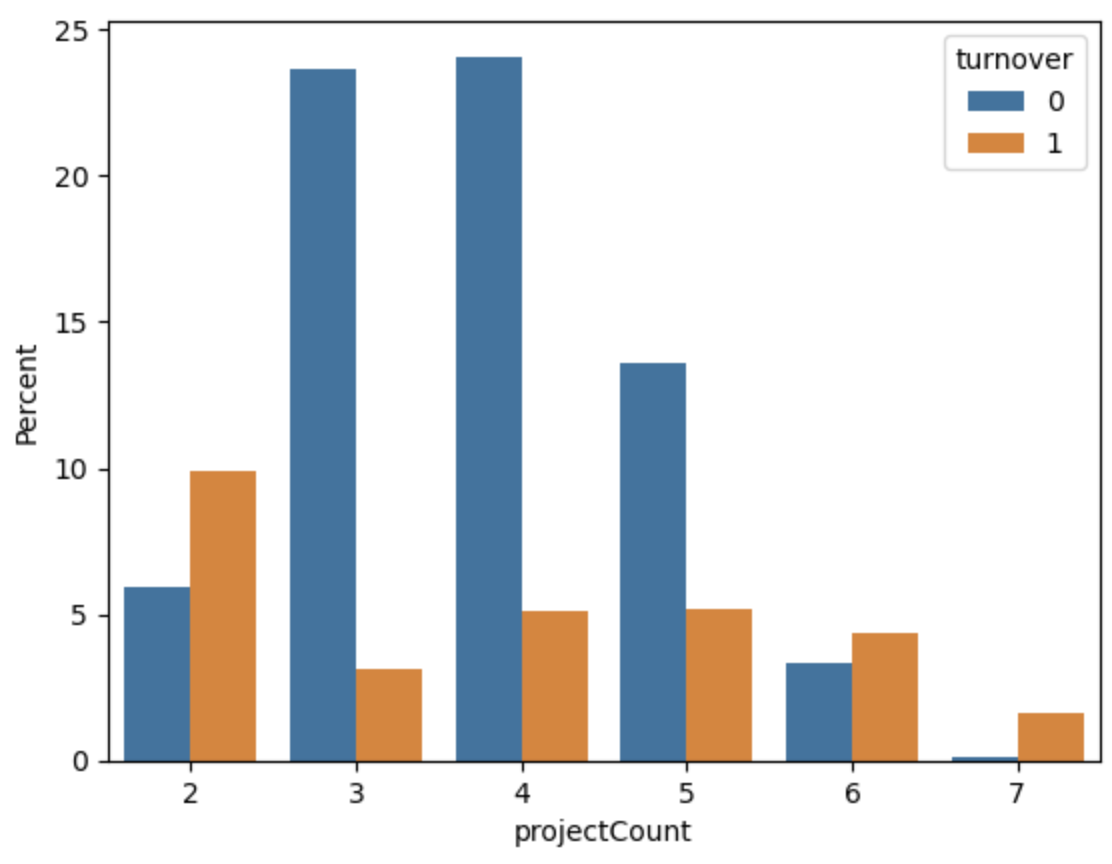
  
**Department V.S. Turnover**

**Summary:**

The top three departments with the most employees are sales, technical, and support.  
  
  
  
**Project Count V.S. Turnover**

**Summary:**

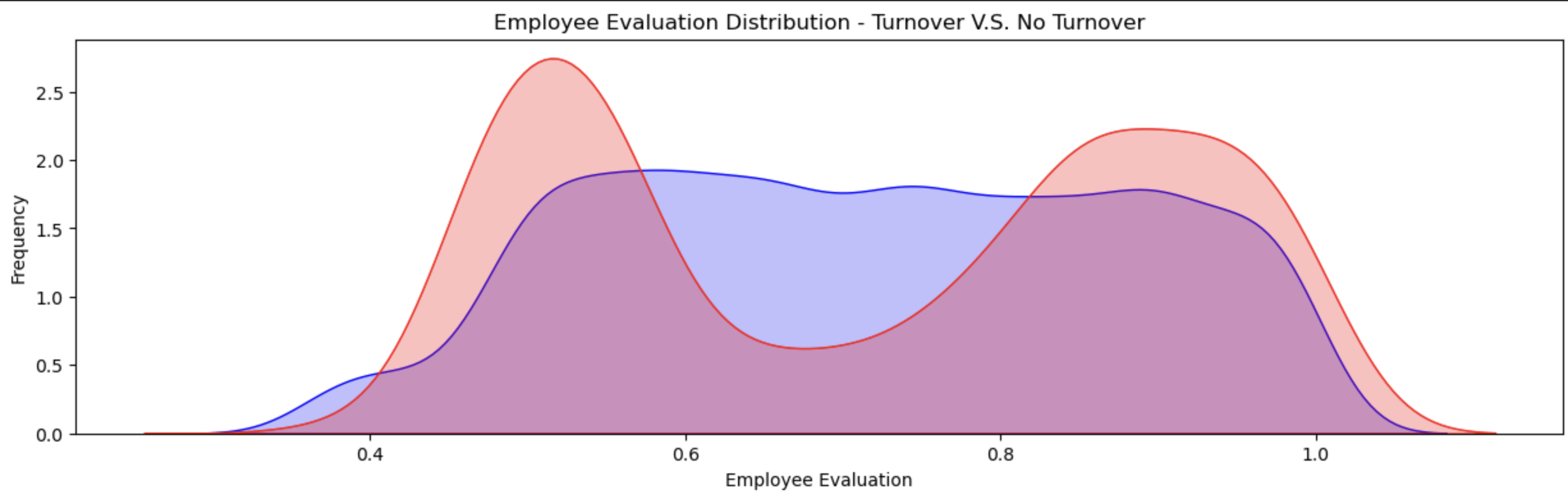
* More than half of employees with 2,6, and 7 projects left the company
* Majority of the employees who did not leave had 3,4, and 5 projects
* All employees with 7 projects left the company
* There is an increase in turnover as project count increases



**Evaluation V.S. Turnover**

**Summary:**

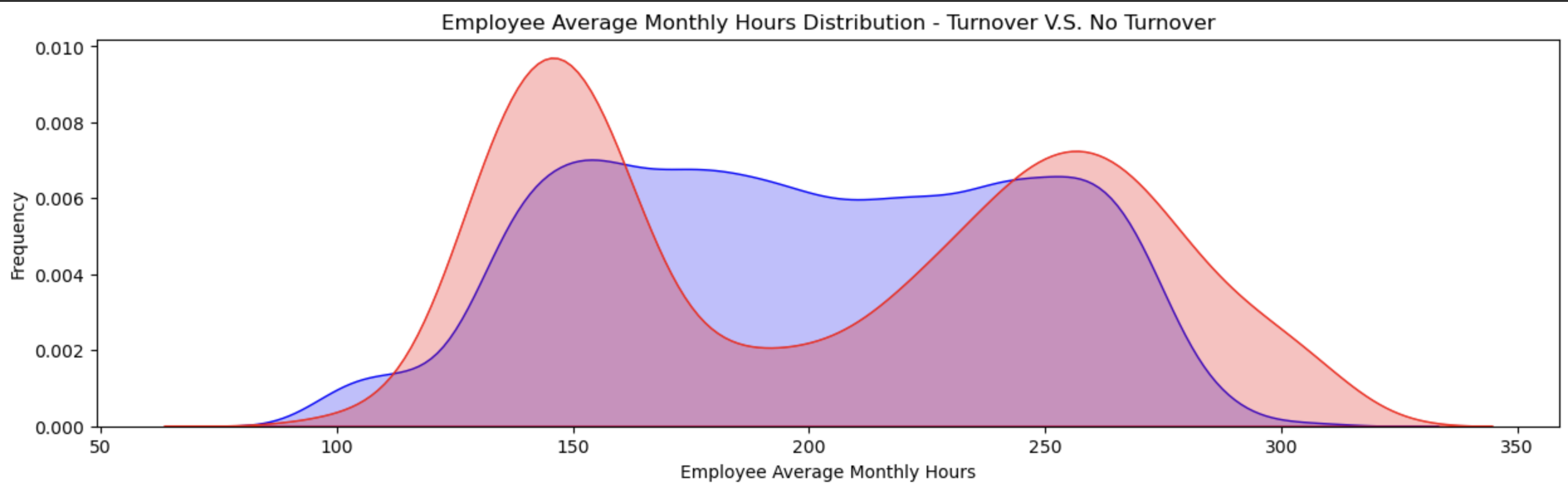
* There is a bimodal distribution for employees that left the company
* Employees with low evaluation levels (0.2-0.6) and high evaluation levels (0.8-1) were the bulk of employee turnover
* Employees with evaluation levels (0.6-0.8) had the smallest turnover rate



**Average Monthly Hours V.S. Turnover**

**Summary:**

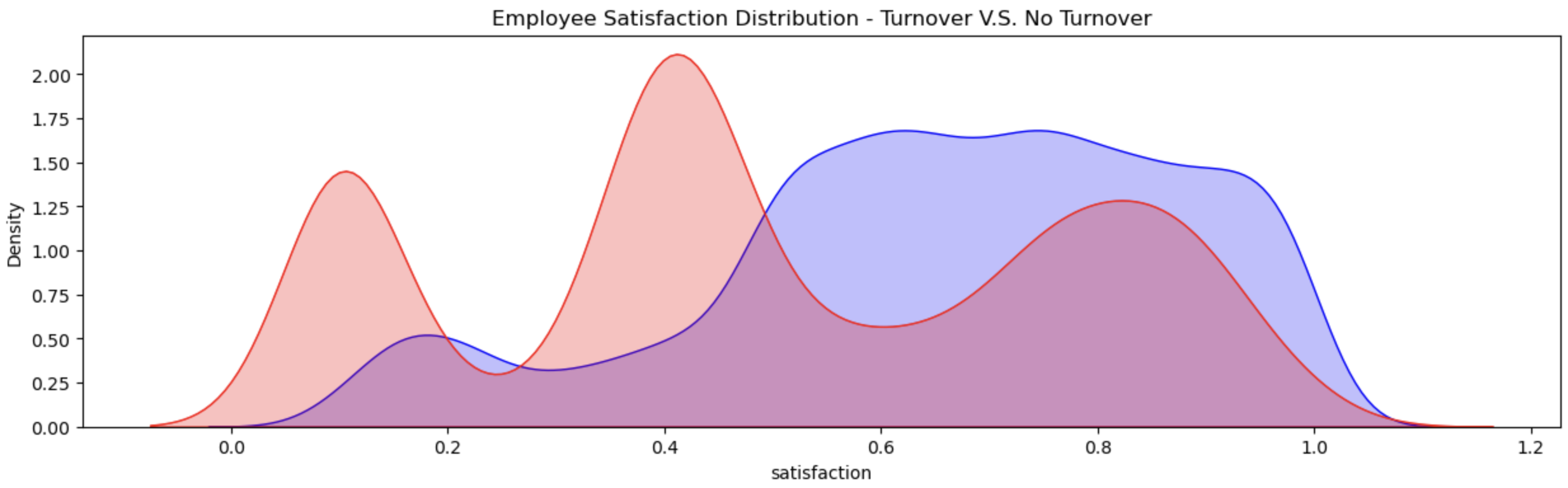
* There is another bimodal distribution for employees that left the company
* Employees who had less hours of work (~150 hours or less) left the company more
* Employees who had more hours of work (~250 hours or more) left the company more
* Employees who left were underworked or overworked



**Satisfaction V.S. Turnover**

**Summary:**

* There is a tri-modal distribution for employees that left the company
* Employees left with low satisfaction levels of (0-0.2)
* Employees left with low satisfaction levels of (0.3-0.5)
* Employees left with high satisfaction levels of (0.7-1)

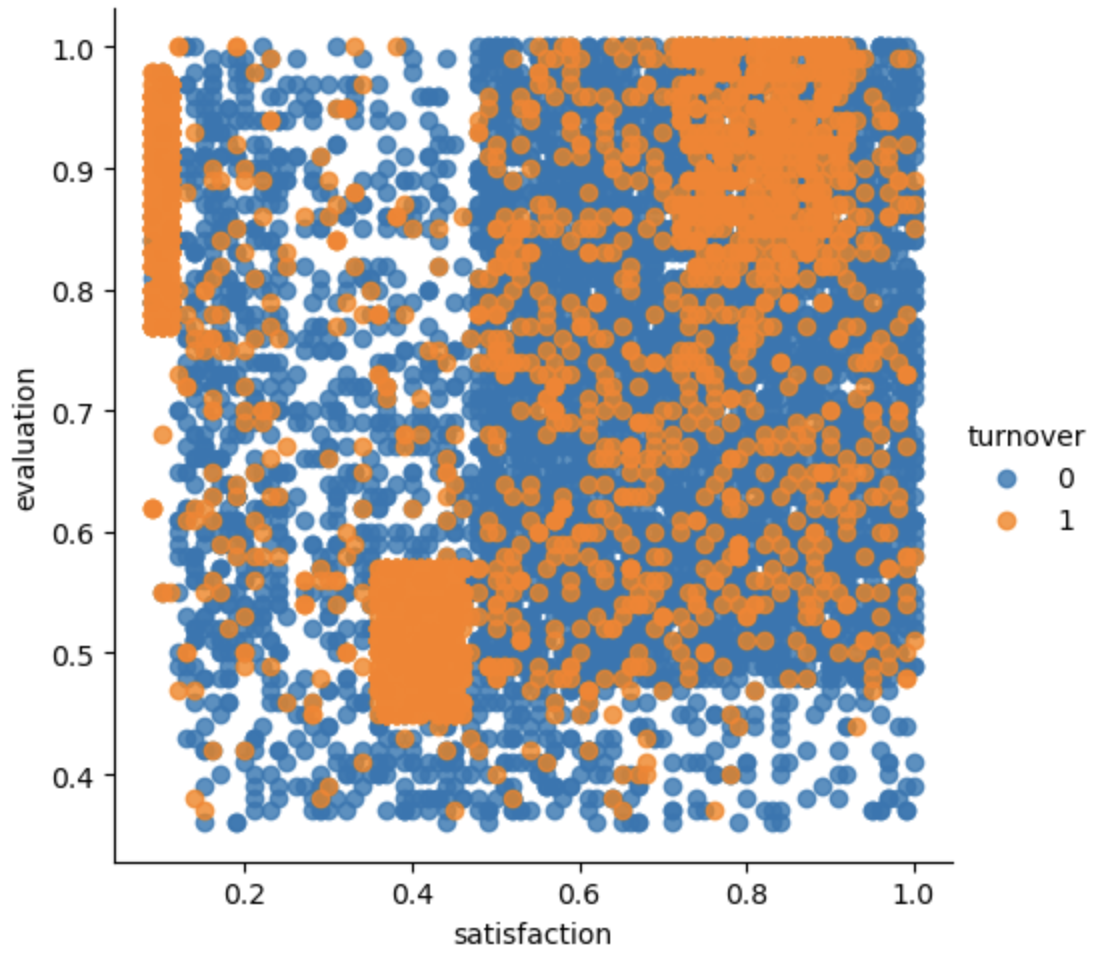


**Evaluation V.S. Satisfaction**

**Summary:**

There are three distinct clusters for employees who left the company

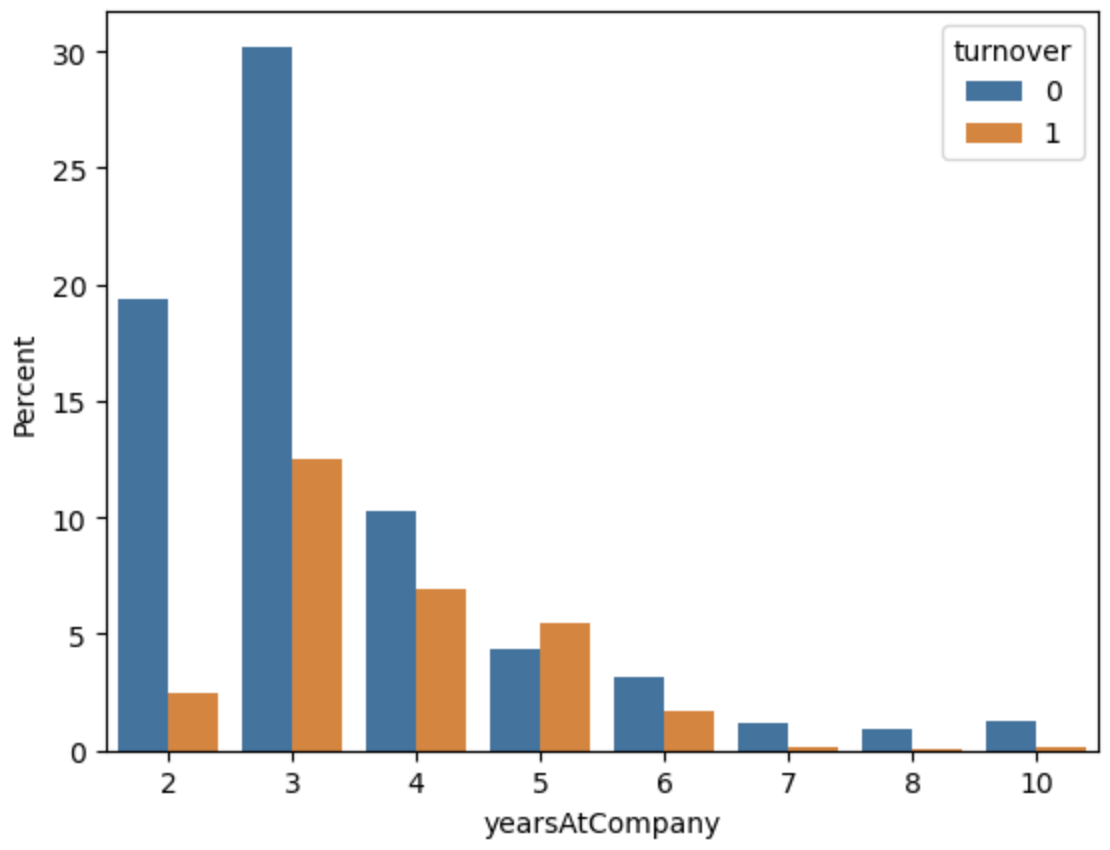
* Cluster 1 (“Overworked” Employees): These employees had satisfaction scores below 0.2 and evaluation scores above 0.75. Employees here were evaluated highly and felt bad at work.
* Cluster 2 (“Under Performing Employees”): These employees had satisfaction scores between (0.35-0.5) and evaluation scores below 0.6. Employees here were evaluated poorly and felt bad at work. This is a typical reason employees leave.
* Cluster 3 (“Ideal Worker”): These employees had satisfaction scores between (0.7-1) and evaluation scores of (0.8-1). Employees here were evaluated highly and felt satisfied at work.



**Years at Company V.S. Turnover**

**Summary:**

More than half of the employees with 4 and 5 years left the company



## **CLASS IMBALANCE (PRECISION/RECALL)**

This dataset is an example of a class imbalance problem because of the skewed distribution of employees who did and did not leave. More skewed the class → accuracy breaks down. In this case, evaluating our model’s algorithm based on accuracy is the wrong thing to measure. We would have to know the different errors that we care about and correct decisions. Accuracy does not measure an important concept that must be considered in this type of evaluation: False Positive and False Negative errors.

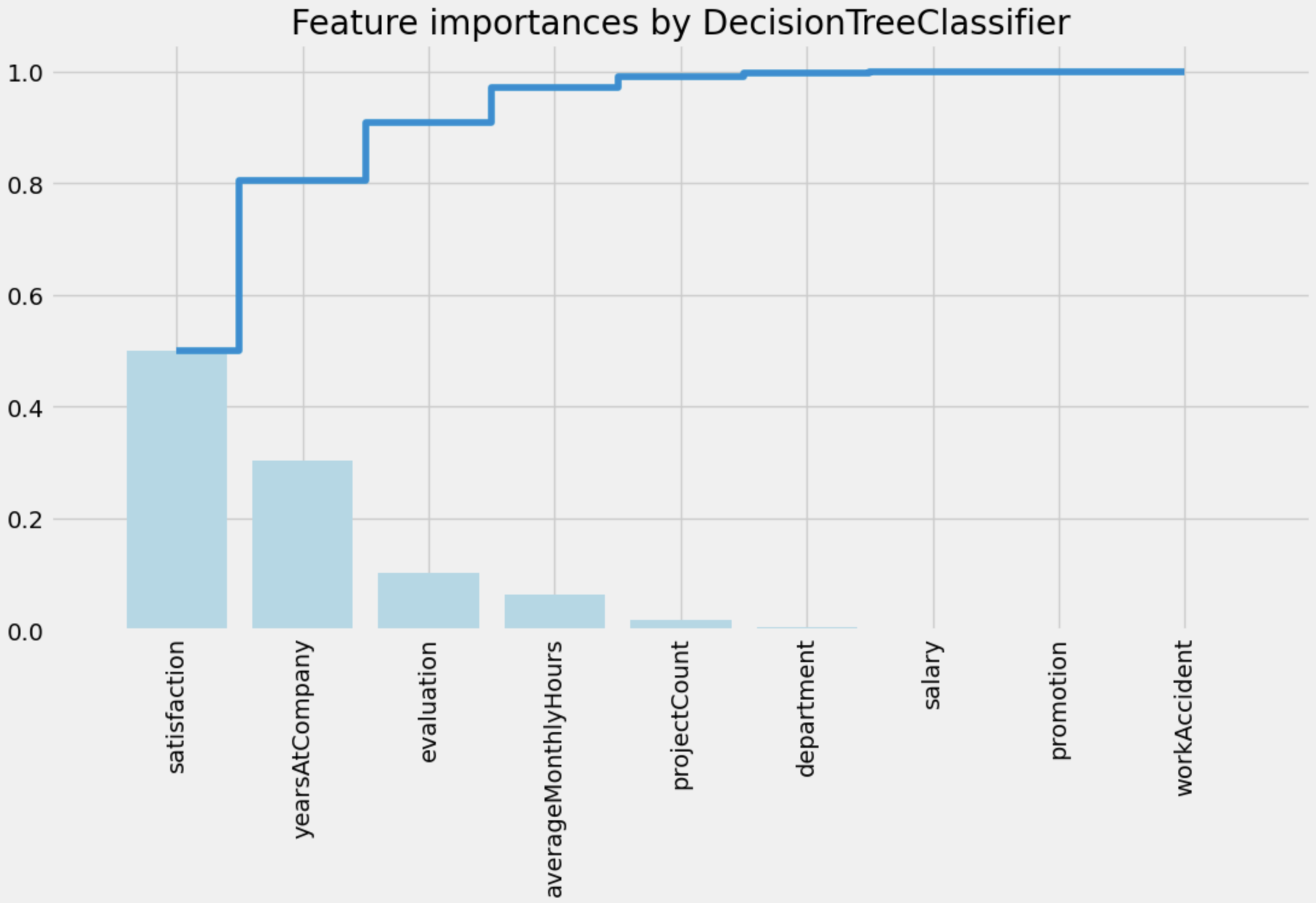
In this problem what type of errors do we care about more? False Positives or False Negatives?

False Positives (Type I Error): You predict that the employee will leave, but do not. Consider the employee turnover domain where Human Resources treats them because they think they will leave the company within a month, but the employee does not. This is a false positive. This mistake could be expensive, inconvenient, and time-consuming for both Human Resources and employees, but can be seen as a worthwhile investment for relational growth.   
  
False Negatives (Type II Error): You predict that the employee will stay but leave. Compare this with the opposite error, where Human Resources does not give treatment/incentives to the employees, and they do leave. This is a false negative. This type of error is more detrimental because the company lost an employee, which could lead to great setbacks and more money to rehire  
  
Depending on these errors, different costs are weighed based on the type of employee being treated. For example, if it is a high-salary employee then would we need a costlier form of treatment? What if it is a low-salary employee? The cost for each error is different and should be weighed accordingly.

In our employee retention problem, rather than simply predicting whether an employee will leave the company within a certain period, we would much rather have an estimate of the probability that he/she will leave the company. We would rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances.

## **FEATURE IMPORTANCE** **Summary:**

By using a decision tree classifier, it could rank the features used for the prediction. The top three features were employee satisfaction, yearsAtCompany, and evaluation. This is helpful in creating our model for logistic regression because it will be more interpretable to understand what goes into our model when we utilize fewer features.



**Important Features:**

* Satisfaction
* Years at Company
* Evaluation

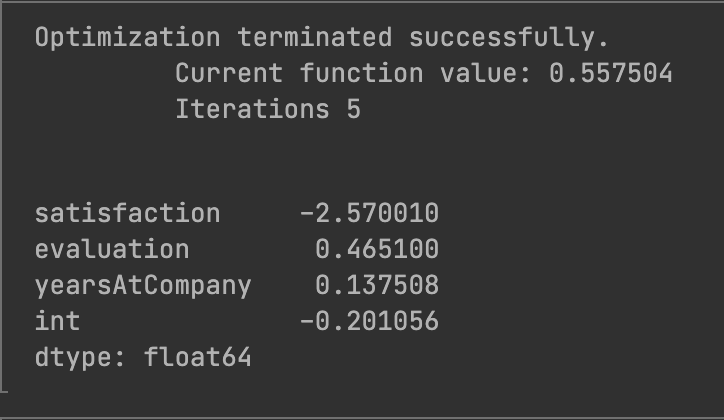
**MODELING**  
With the elimination of the other variables, I will be using the three most key features to create our model: Satisfaction, Evaluation, and YearsAtCompany.

## **MODELING** With the elimination of the other variables, I will be using the three most key features to create our model: Satisfaction, Evaluation, and YearsAtCompany.

Following overall equation was developed:

Employee Turnover Score = Satisfaction\*(-2.570010) +

Evaluation\*(0.465100) + YearsAtCompany\*(0.137508) + (-0.201056)



The values above are the coefficient assigned to each independent variable. The constant -0.201056 represents the effect of all uncontrollable variables

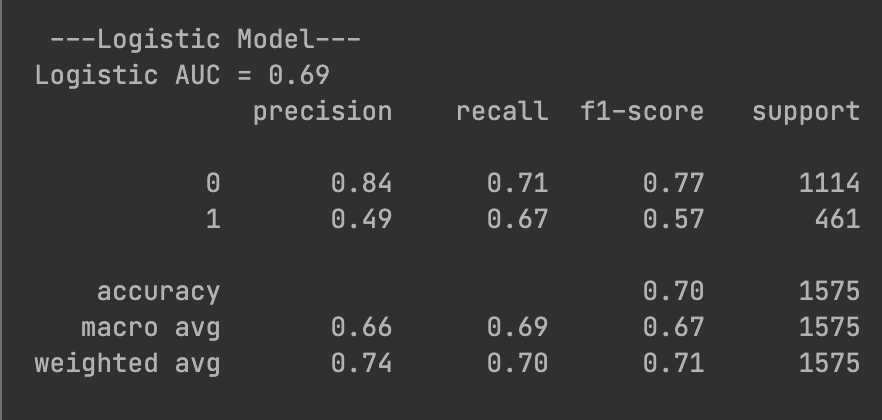
**For Example:**

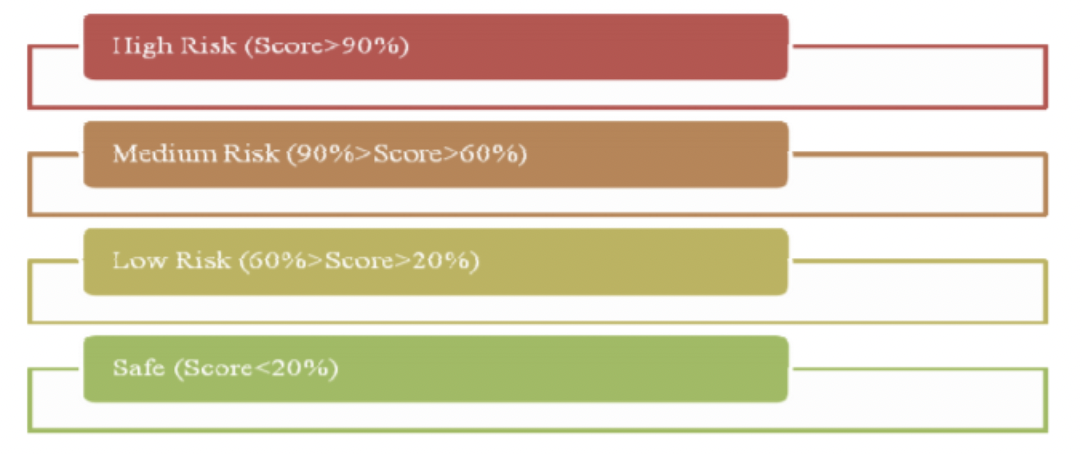
If you were in these employee values into the equation:

* Satisfaction: 0.7
* Evaluation: 0.8
* YearsAtCompany: 3

Employee Turnover Score = (0.7) \*(-2.570010) + (0.8) \*(0.465100) + (3) \*(0.137508) + (-0.201056) = 1.215459 = 12%

**Result:**  
This employee would have a 12% chance of leaving the company. This information can then be used to form our retention plan.  
**TEST EVALUATION**



**RETENTION PLAN**  
  


With the logistic regression model, we can now use our scores and evaluate the employees through different scoring metrics. Each zone is explained here:

* Safe Zone (Green) – Employees within this zone are considered safe.
* Low Risk Zone (Yellow) – Employees within this zone are to be taken into consideration for potential turnover. This is more of a long-term track.
* Medium Risk Zone (Orange) – Employees within this zone are at risk of turnover. Action should be taken and monitored accordingly.
* High Risk Zone (Red) – Employees in this zone have the highest chance of turnover. Action should be taken immediately.

## **CONCLUSION** This paper outlines the different analysis done on the employee dataset and the usage of a Logistic regression model to make predictive insights on the probability of an employee to turnover. This model can be applied throughout the company's various departments and be used to help make better decisions about employee retention. The model should be updated periodically and include additional features for it to make more accurate predictions.

**Summary of our analysis:**

* Employees left when they are underworked (less than 150hr/month or 6hr/day)
* Employees left when they are overworked (more than 250hr/month or 10hr/day)
* Employees with either high or low evaluations should be taken into consideration for high turnover rate
* Employees with low to medium salaries are the bulk of employee turnover
* Employees that had 2,6, or 7 project count was at risk of leaving the company
* Employee satisfaction is the highest indicator of employee turnover.
* Employee that had 4 and 5 yearsAtCompany should be taken into consideration for high turnover rate

Employee satisfaction, yearsAtCompany, and evaluation were the three biggest factors in determining turnover

## **FUTURE WORK** This problem is about people's decisions. When modeling the data, we should not be using this predictive metric as a solution decider. But we can use this to arm people with much better relevant information for better decision making. We would have to conduct more experiments or collect more data about the employees in order to come up with a more accurate finding. I would recommend gathering more variables from the database that could have more impact on determining employee turnover and satisfaction such as their distance from home, gender, age, etc.

## **REFERENCES**

<https://datascienceplus.com/building-a-logistic-regression-in-python-step-by-step/>

<https://journalistsresource.org/tip-sheets/research/regression-analysis-primer-for-journalists><https://analytics4all.org/2016/05/11/python-logistic-regression/>