# CP1: EMPLOYEE TURNOVER DATA ANALYSIS

Final Report

# About dataset

- Simulated employee data of a large company facing problem of high turnover
- 14999 samples of employees classified into two categories: Left and Stayed
- Other columns
  - Salary
  - Satisfaction level
  - Last Evaluation
  - Number of projects
  - Average monthly hours
  - Time spent in the company
  - Work accidents
  - Promoted in the last five years
  - Department
- Data can be downloaded at: https://www.kaggle.com/ludobenistant/hr-analytics

# Limitations

- Since this is a simulated dataset, it is not possible improve the model as new data becomes available

- The dataset is imbalanced, there are more samples available for employees that stayed vs. left (Based on 24% Turnover, 24% of records correspond to employees that left and the remaining correspond to the ones that stayed)

# Cleaning and Wrangling

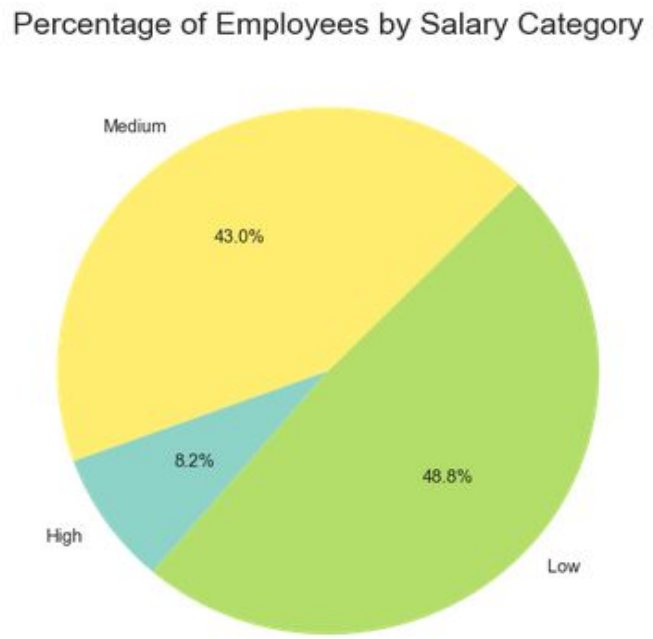
- Checked data for null values, Nan, unexpected formats (using regex)
- Renamed columns with easy to understand names
- Checked if the data is tidy. Tidy data follows these rules
  - Each variable must have its own column.
  - Each observation must have its own row.
  - Each value must have its own cell. Classified columns into Qualitative vs. Quantitative

# EXPLORATORY DATA ANALYSIS

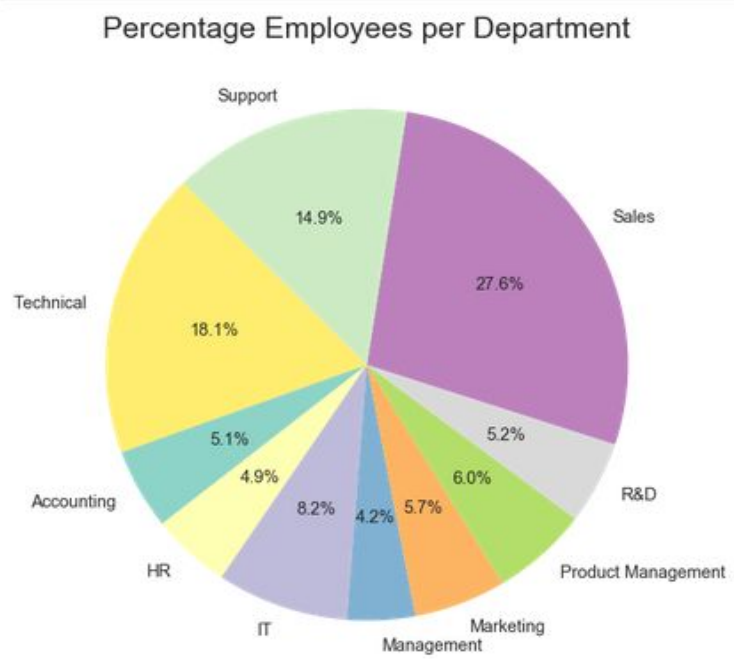
*“The greatest value  of a picture is when it forces us to see what we never expected to see”- John Tukey*

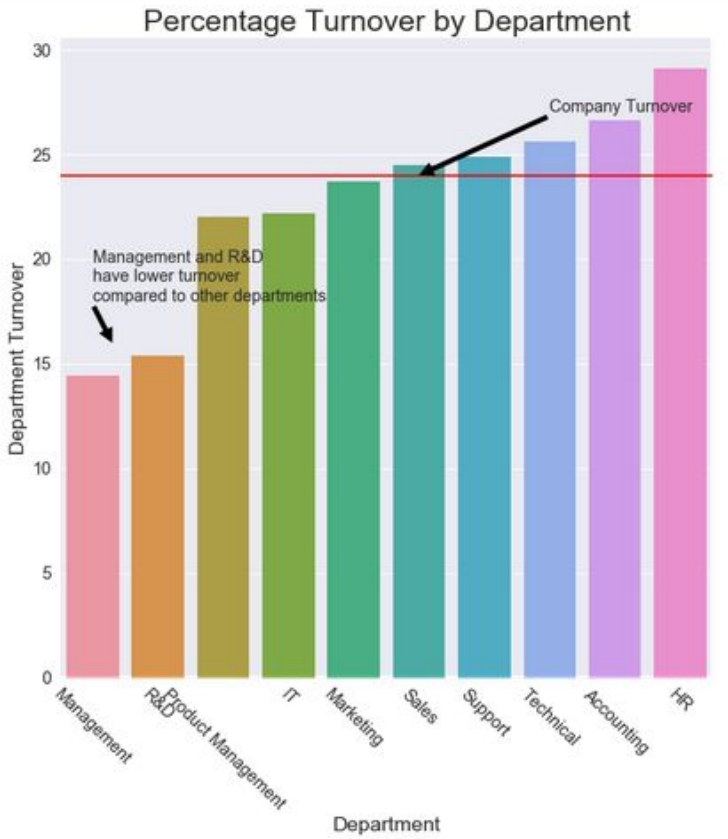
# Salary

- Majority of employees have Low/Medium salaries
- Low/Medium salaried employees have significantly higher turnover




# Department

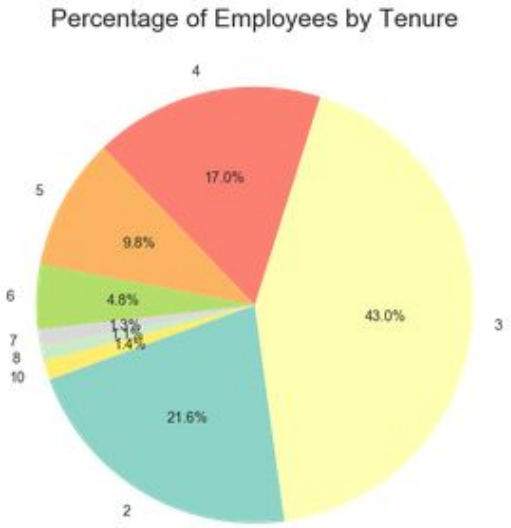


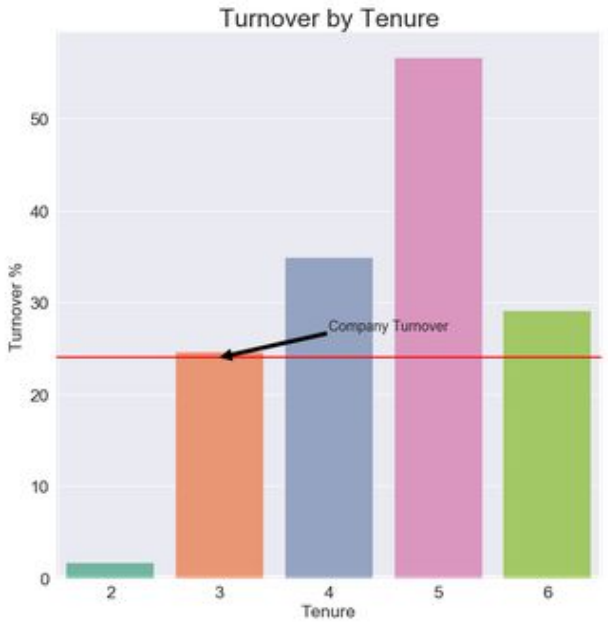



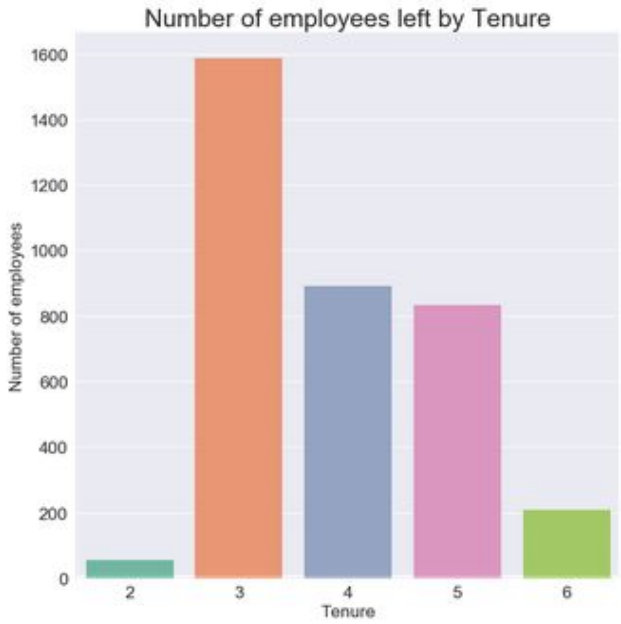

- Majority of employees are in Sales, Technical and Support departments.
- R&D and Management have a lower turnover compared to other departments.

# Tenure

- Most employees have a tenure between two to five years
- Most employees leave during their third to sixth year of service

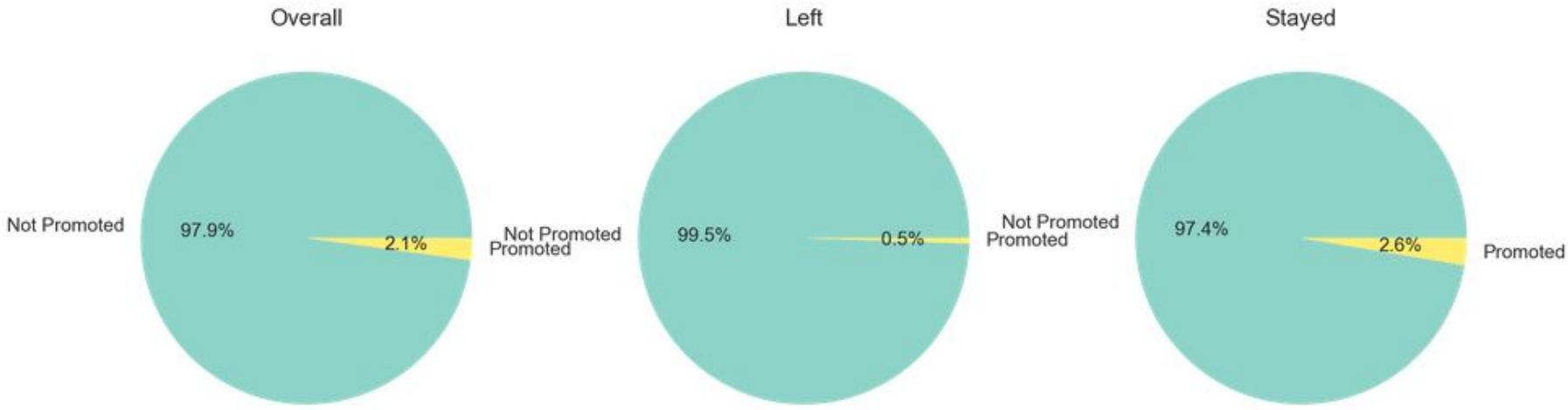


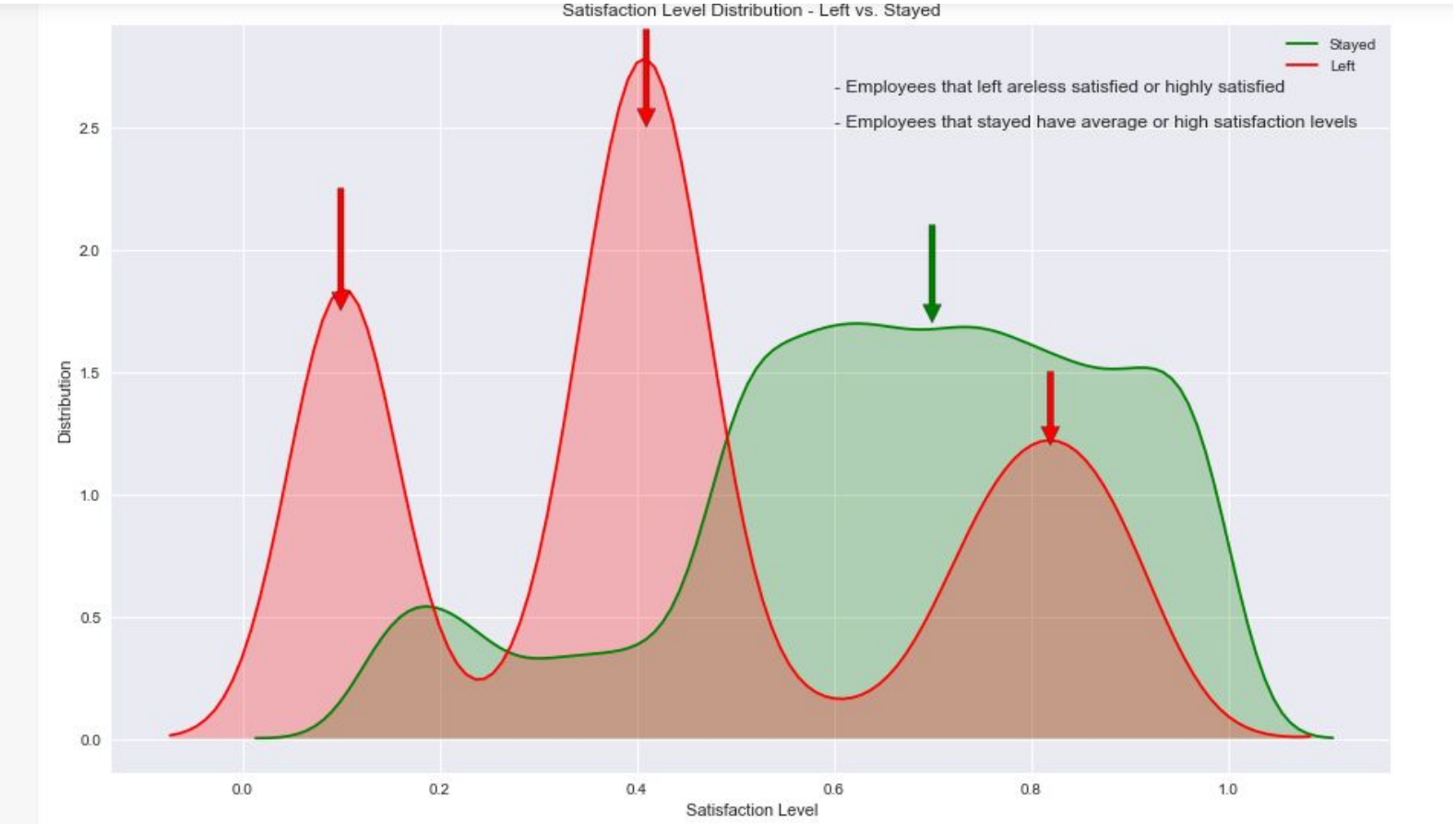
# Promotion in last 5 years

- Very few employees have been promoted in the last 5 years overall
- % of promotions is better among the employees that stayed



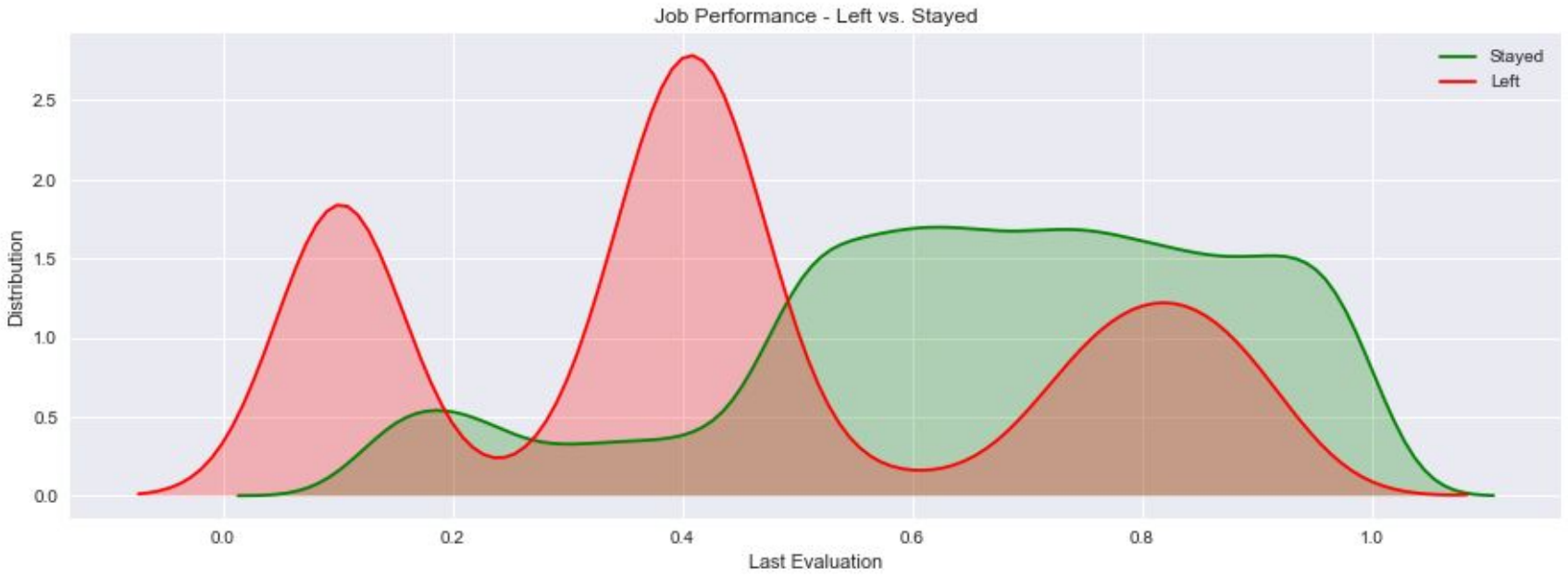
# Satisfaction Level

- Employees that left are either less satisfied or highly satisfied
- There are very few employees among the ones that stayed that are less satisfied



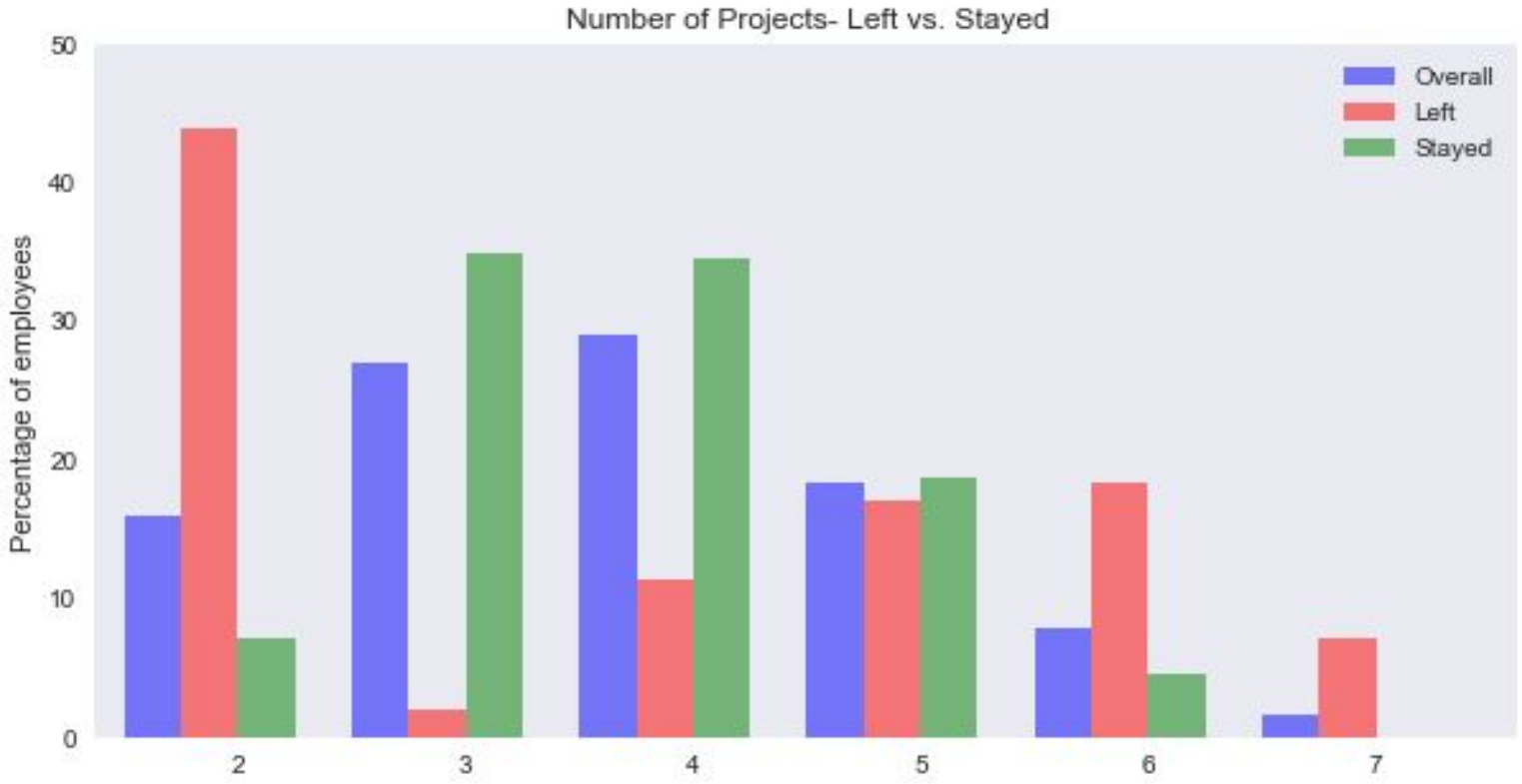
# Job Performance

- Employees that left worked too few or too many hours
- Employees that stayed are evenly distributed across a wide range of average hours worked



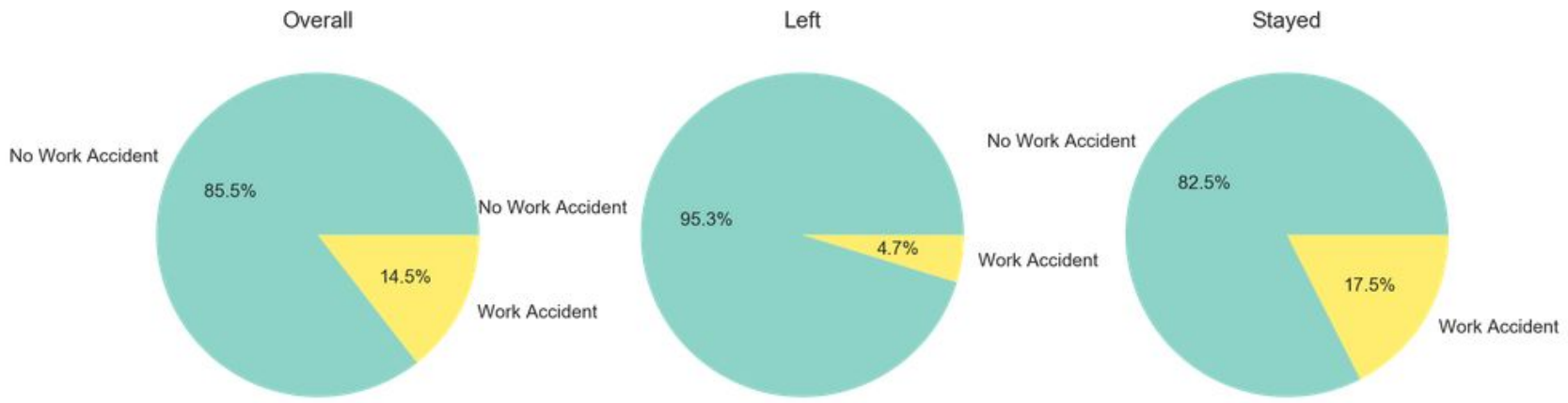
# Number of projects

- Employees that left worked on too less or too many projects



# Work Accidents

- Employees that left had fewer accidents compared to the ones that stayed



# Inferential Statistics

- Used Chi-squared test to determine if the differences in groups were statistically significant
- Consequently ran pair-wise chi squared tests to understand if there were any groups where the differences were significantly higher
- All tests confirmed the inferences drawn from visual EDA

# Machine Learning

# Machine Learning Methods

- Machine Learning Algorithms used for classification
  - Logistic Regression
  - Random Forest
  - XGBoost
- Methods used to handling imbalanced data
  - SMOTE
  - ADASYN
  - Random Under Sampling

# Results

| Method | Test Score | Precision(0) | Precision(1) | Recall(0) | Recall(1) |
|---|---|---|---|---|---|
| Baseline using Logistic Regression | 80.48 | 0.83 | 0.64 | 0.93 | 0.4 |
| Regularized Logistic Regression model (Ridge) | 80.51 | 0.83 | 0.64 | 0.93 | 0.4 |
| Regularized Logistic Regression model (Lasso) | 80.59 | 0.83 | 0.65 | 0.93 | 0.41 |
| Random Forest Classifier | 98.96 | 0.99 | 0.98 | 0.99 | 0.97 |
| XGBoost Classifier | 97.2 | 0.98 | 0.95 | 0.98 | 0.93 |
| SMOTE+ Logistic Regression | 75.68 | 0.92 | 0.49 | 0.74 | 0.8 |
| SMOTE+ Random Forest Classifier | 98.83 | 0.99 | 0.98 | 0.99 | 0.97 |
| SMOTE+ XGBoost Classifier | 97.2 | 0.98 | 0.95 | 0.98 | 0.93 |
| ADASYN+ Logistic Regression | 75.23 | 0.95 | 0.49 | 0.71 | 0.89 |
| ADASYN+ Random Forest Classifier | 98.53 | 0.99 | 0.97 | 0.99 | 0.97 |
| ADASYN+ XGBoost Classifier | 97.2 | 0.98 | 0.95 | 0.98 | 0.93 |
| Random Under Sampler+ Logistic Regression | 75.23 | 0.95 | 0.71 | 0.49 | 0.89 |
| Random Under Sampler+ Random Forest Classifier | 98.53 | 0.99 | 0.97 | 0.99 | 0.97 |
| Random Under Sampler+ XGBoost Classifier | 97.2 | 0.98 | 0.95 | 0.98 | 0.93 |

# Findings

- Most employees do not stay beyond 3 to 5 years of service.
- Employees that left typically fell into these categories:
  - Worked too many hours or worked too few hours
  - Less satisfied or highly satisfied
  - High performers or low performers
- Turnover higher among low to medium salaried employees
- Among departments, Sales, Technical and Support are have the highest Turnover.
- With given data Random Forest and XGBoost have the best performance

# Recommendations

- Place emphasis on employees that are in their first few years of employment i.e. 1 to 5 years of service
- Ensure employees are balanced in their workload in projects, number of hours worked etc.
- High performers and low performers are at the highest risk for turnover. Low performers need to be put on a performance improvement plan. High performers need to be incentivized with promotions, better pay and challenging projects.
- Evaluate pay structures especially for low and medium salaried employees. Ensure salary and benefits are not the reason for high turnover.
- Create more incentives and opportunities in the departments with high turnover: Sales, Technical and Support
- Choose RandomForest or XGBoost classifier to predict employee turnover. The code also contains models that tackle class imbalances and hyper parameter tuning to prevent overfitting. As new data becomes available, evaluate the models and choose the one with best performance.