**Introduction**

A large company called Company X is facing the problem of high turnover. The Sample dataset consists of 14,999 rows of employee data classified into the ones that left and one that have not left. Along with the classification, following employee details are available (features). This data is simulated.

1. Salary
2. Satisfaction level
3. Last Evaluation
4. Number of projects
5. Average monthly hours
6. Time spent in the company
7. Work accidents
8. Promoted in the last five years
9. Department

The goal of project is to understand the factors causing employees to leave and also predict the employees that are highly likely to leave.

**Limitations of dataset**

* 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. the ones that left

**Cleaning and Wrangling**

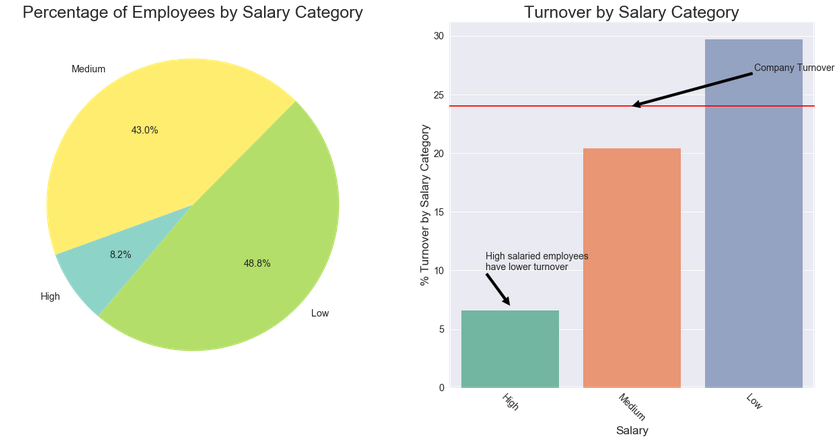
1. Checked data for null values, Nan, unexpected formats (using regex)
2. Renamed columns with easy to understand names
3. Checked if the data is tidy. Tidy data follows these rules
   1. Each variable must have its own column.
   2. Each observation must have its own row.
   3. Each value must have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

**Exploratory data analysis**

I took the approach of visualizing left vs. stayed employees by each feature to understand if there were any stark differences between the two groups. Following inferences can be drawn:

**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.

