Employee Attrition Analysis: A comprehensive Analysis and Strategy Development

A data driven solution to uncover the roots of employee attrition

Shaily Sahay
New Jersey Institute of Technology

Table of Contents

Background

Attrition, also known as turnover, refers to the process of employees leaving their jobs, either voluntarily or involuntarily. It is a common issue faced by many organizations, and it is challenging to identify the reasons why employees leave because, more often than not, it is a mix of several factors. It is important to identify the main factors influencing attrition, and also quantify their importance so that targeted solutions could be devised to mitigate the turnover rate.

High levels of attrition can negatively impact an organization by reducing productivity, increasing costs, and lowering morale among remaining employees. As a result, a detailed analysis of employee data is warranted to gain insights into this issue and develop strategies for improving retention.

The analysis and visualization of the data will be done using **Python** programming and **Tableau**.

Scope of Study

The analysis of the project at hand is based on employee data maintained by the HR department of a large company. This is a public dataset and has been obtained from Kaggle, the link of which is given below:

Source of data: https://www.kaggle.com/datasets/jacksonchou/hr-data-for-analytics

Data Description

Following points should be noted about the dataset:

- Each row of the dataset represents an employee, identified by the 'EmployeeID' attribute.
- There is no time-frame defined for the dataset explicitly. But since we are finding the yearly attrition, we will assume the data is for one financial year.
- The dataset consists of 12000 observations and 11 attributes.
    - 5 numerical attributes →
    - 4 ordinal attributes
    - 2 nominal attribute

- The main attribute of interest in the set is 'left', with the domain:
    - 0 = People who are working in the company
    - 1 = People who have left the company

  The effect of each independent feature on ‘left’ attribute will be observed to identify factors responsible for attrition

- No skewed distributions were observed in the dataset

Methodology

Data Preprocessing

The preprocessing involved the following steps:

- The ‘satisfaction_range’ attribute was added to represent the levels of satisfaction as groups. For the purpose of this project, following ranges have been chosen:
    - 'Low' (0 to 0.3)
    - 'Medium' (0.3 to 0.6)
    - 'High' (0.6 to 1.0)

```
ALTER TABLE EmployeeAttrition
Add Satisfaction_Range Varchar(10)
/

UPDATE EmployeeAttrition
SET Satisfaction_Range = (
                            WITH SatisfactionRangeTable as
                            (
                                SELECT EMPLOYEEID,
                                    CASE
                                        WHEN SATISFACTION_LEVEL >= 0.0 AND SATISFACTION_LEVEL <= 0.3 THEN 'low'
                                        WHEN SATISFACTION_LEVEL > 0.3 AND SATISFACTION_LEVEL <= 0.6 THEN 'medium'
                                        WHEN SATISFACTION_LEVEL > 0.6 AND SATISFACTION_LEVEL <= 1.0 THEN 'high'
                                        ELSE 'NA'
                                    END AS Satisfaction_Range
                                FROM EmployeeAttrition
                            )

                            SELECT Satisfaction_Range
                            FROM SatisfactionRangeTable
                            WHERE EmployeeAttrition.EMPLOYEEID = SatisfactionRangeTable.EMPLOYEEID
                        );
/
```

- For better readability, name of the following attributes were changed:

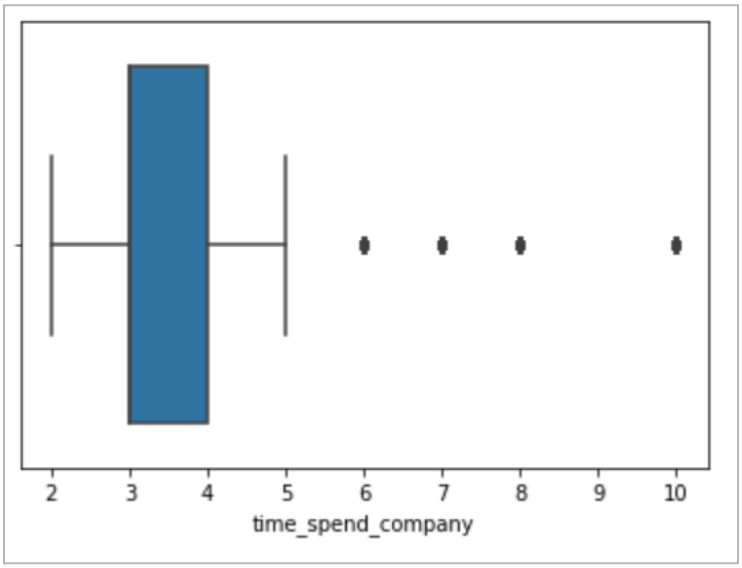
- 'left' was changed to 'turnover'. From this point onward, this attribute will be referred to as ‘turnover’.
- 'sales' was changed to 'dept'

- No missing values were found in the dataset

```
In [6]: ## Check for null values:

        df.isna().sum()

Out[6]: EmployeeID              0
        satisfaction_level      0
        last_evaluation         0
        number_project          0
        average_montly_hours    0
        time_spend_company      0
        Work_accident           0
        left                    0
        promotion_last_5years   0
        sales                   0
        salary                  0
        dtype: int64
```

*** There are no missing values ***

- Although we have some extreme values in 'time_spend_company', they simply reflect that few employees worked for long years (6 to 10 years) in the company. Since this information is valuable, we will not delete these rows. Also, imputation does not make logical sense for this attribute, so the outlying rows will remain unchanged.



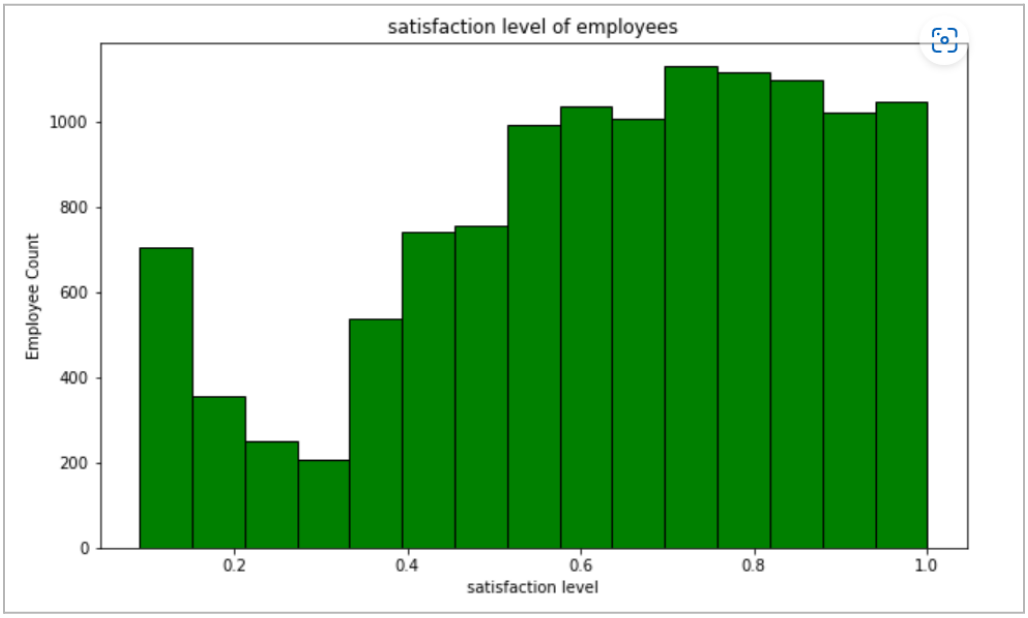
Data Analysis

The process of analysis was broken down into following sections:

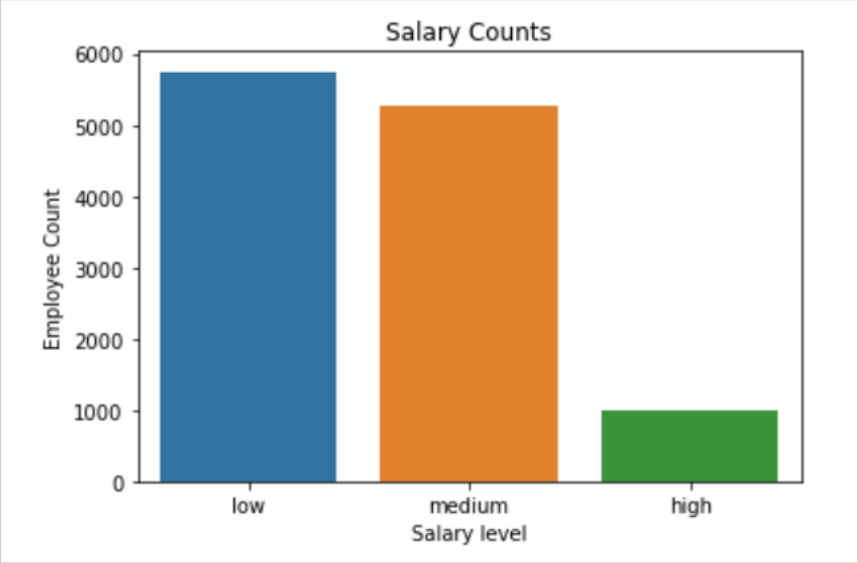
- Univariate analysis → To summarize the main characteristics of each attribute and describe its distribution, central tendency, dispersion, and shape.
- Bivariate analysis → To determine if there is a relationship or association between each independent variable, and the dependent variable ‘turnover’
- Multivariate analysis → To understand the relationships and dependencies among important features and the dependent variable ‘turnover’

Univariate analysis

- There are 12000 employees in 'XYZ Corp', and 2000 have left the company in the last year (almost 17% attrition rate). The company has 10 departments, with 'sales', 'technical', and 'support' departments comprising 61% of the total employees.
- The average satisfaction level is around 62%, with 58% of the employees highly satisfied, and 12% having low satisfaction. This tells us that people are generally satisfied working in this company, but there is still room for improvement as 42% of the employees feel some level of dissatisfaction.



- The average years that people stay in the company is 3.36 years, with the majority of employees spending 2-3 years. However there is a high variation in this value, with minimum value of 2 years and maximum of 10 years.
- The company has high performers, with 53% of the employees getting a high evaluation (> 0.7) last year. The performance average for the company is around 71%. This is an area that needs attention - are high performers leaving the company? Are they satisfied with the company?
- The salary range is highly skewed, with only 8% of the employees falling under the high salary range. This could be due to the fact that generally high compensation is offered to select few positions - the upper management and senior technical managers. However, with 53% of employees receiving very high evaluations, this number seems to be small. Could this be a reason for dissatisfaction among employees? Should the salary offering of the company be revised to offer better compensation to high performers.

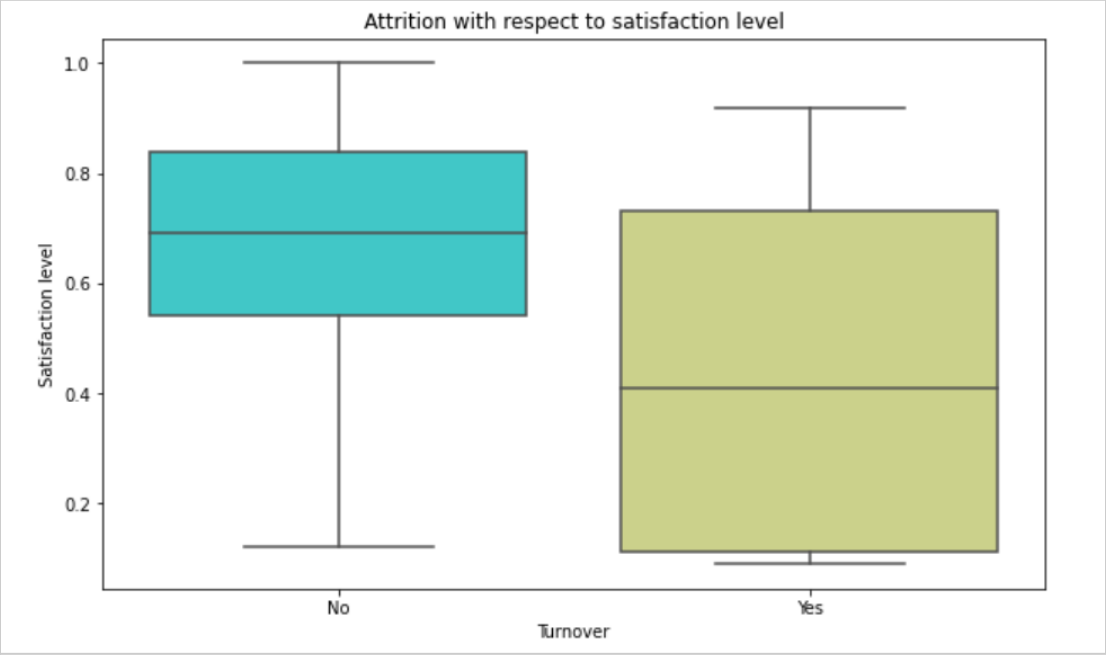


- 'Average monthly hours' for this company is quite high, where 71% of the employees work for more than the standard working hours (160 hours monthly). It follows a bi-modal distribution, peaking at around 160 hours and 260 hours.

- Promotion is another factor that seems worrying as less than 3% of the employees received promotion in the last 5 years. The promotion policy demands a revision.

Bivariate analysis

- As expected, there is a greater turnover percentage among employees with lesser satisfaction (less than 0.5).

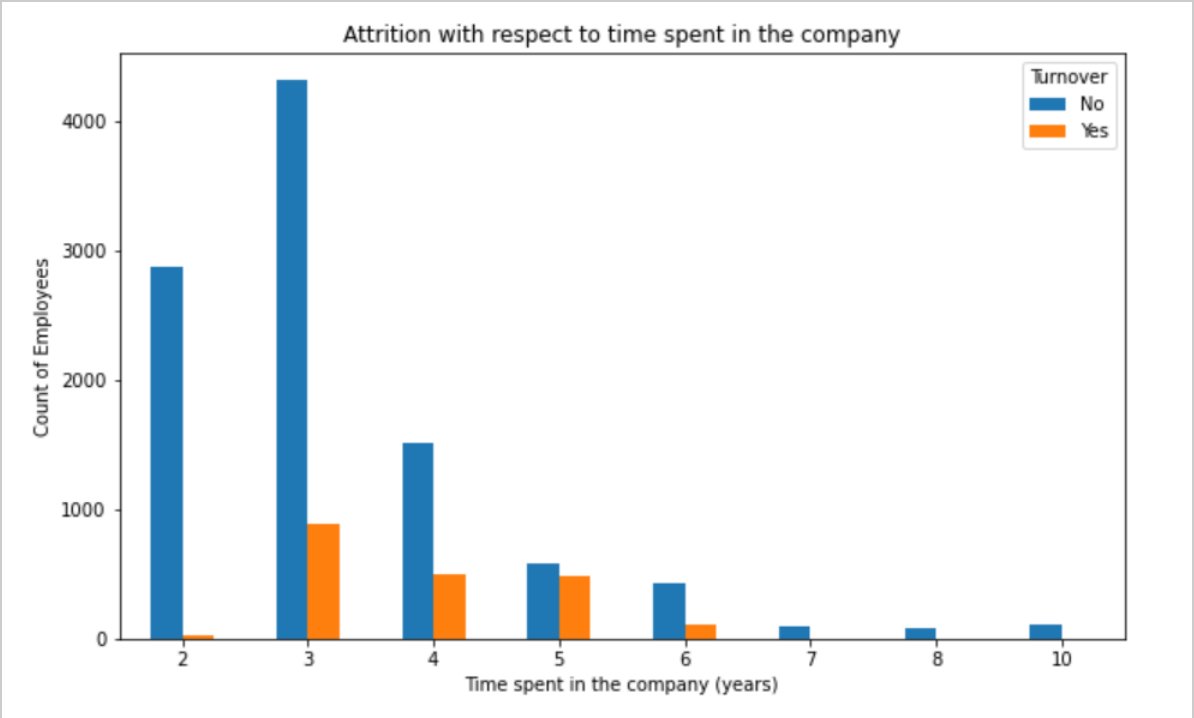


However, we do have 550 employees who were highly satisfied (more than 0.7) who left the company. This is alarming. Of all the employees who left, more than 25% were highly satisfied. These employees' exit interview feedback should be observed closely to know the reason behind their attrition.

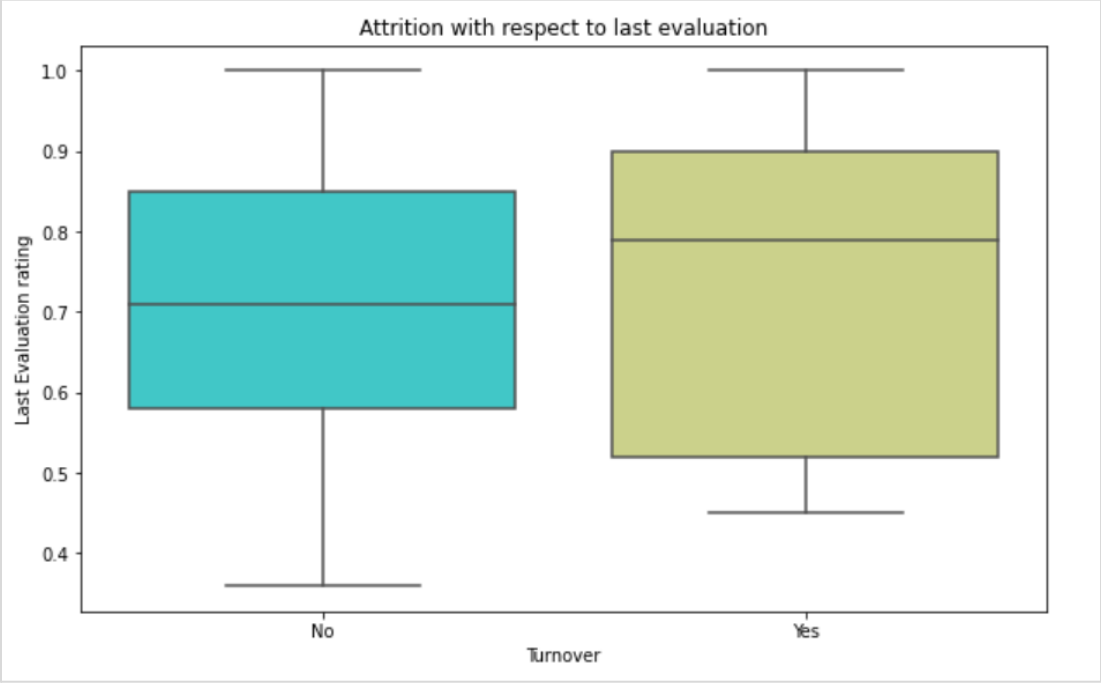
```
# Breakdown of Attribution data with respect to satisfaction_range

# Creating crosstab
print(pd.crosstab(df['turnover'], df['satisfaction_range']))

# Creating barplot
ax = pd.crosstab(df['satisfaction_range'],df['turnover']).plot(kind='bar', rot=0)
ax.legend(["No", "Yes"], title='Turnover')

plt.title('Attrition with respect to satisfaction range')
plt.xlabel('satisfaction Range')
plt.ylabel('Count of Employees')
plt.show()
```

```
satisfaction_range  Low  Medium  High
turnover
0                   867    2742  6391
1                   537     913   550
```

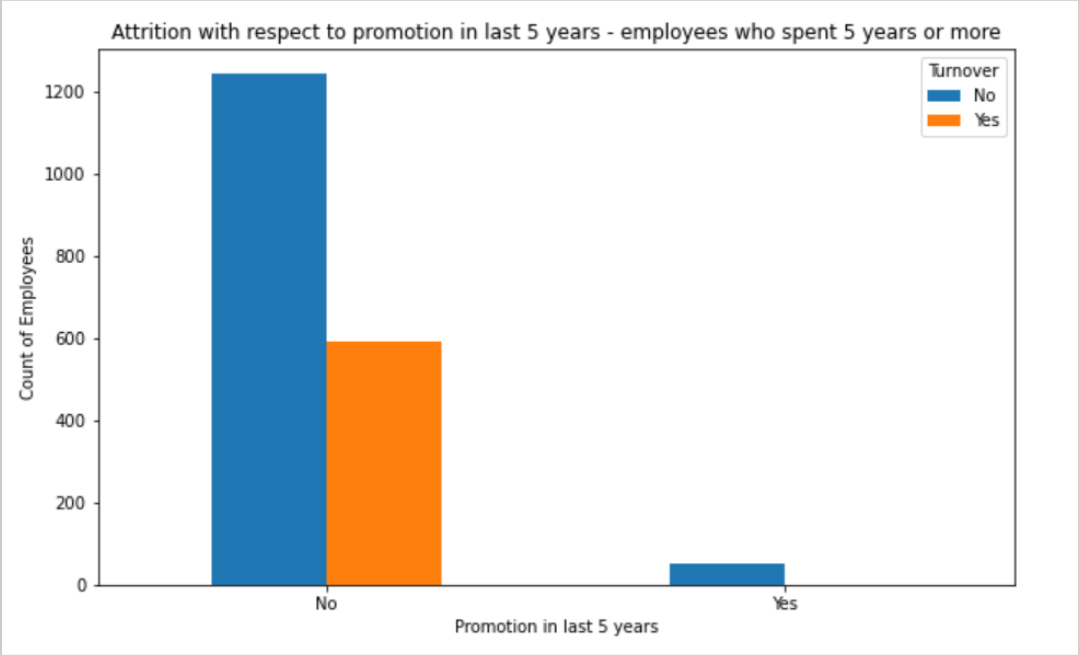
- Maximum attrition is seen among people who have been in the company for 3-5 years. This could be due to factors like:
    - Low salary growth
    - Not finding the job challenging



- The mean evaluation of turnover employees is higher than that of current employees. This means that employees who left were on average better performers than those who stayed.



- Barring just 1 employee, every other employee who left did not receive promotion in the last 5 years. This is compelling evidence that lack of promotion, among other factors, is a big reason that employees leave.



Multivariate Analysis

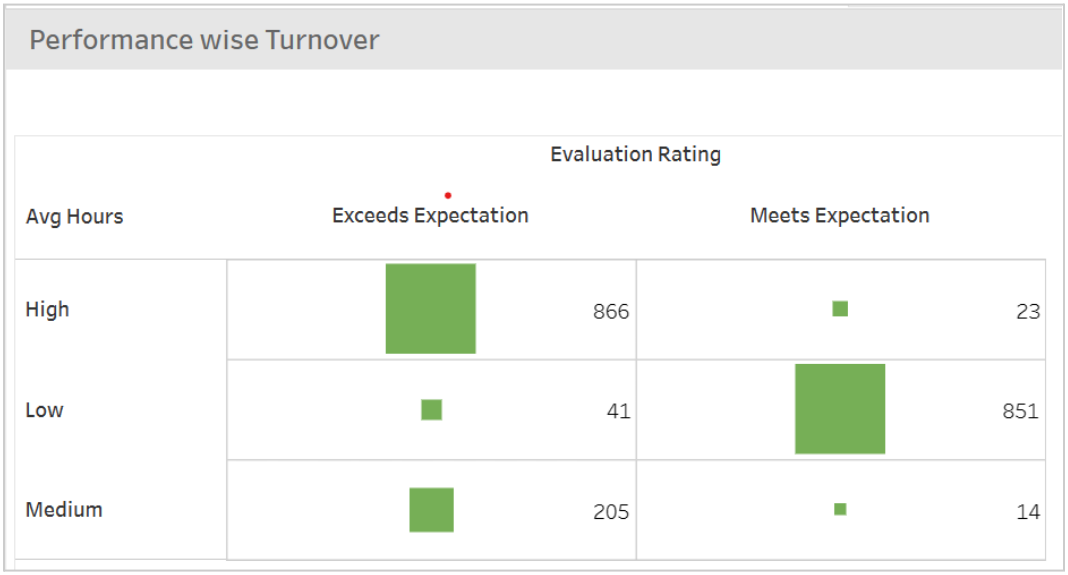
From the above analysis, there are two important issues that stand out and need further analysis:

1. **The mean evaluation rating of turnover employees is higher than that of retained employees. Why did good performers leave?**
2. **25% of the turnover employees were highly satisfied. Why did they leave?**

Performance wise turnover

The performance wise turnover paints an interesting picture. Only those employees left the company who received a ‘Meets Expectation’ or a ‘Exceeds Expectation’ rating in the last evaluation.

Also, among employees who received a very high rating (Exceeds Expectation), the majority of them were working for long hours.



This is a serious issue that needs immediate redressal. **Good performers are leaving the company due to overwork**, whereas low performers continue to work in the company. In fact, out of 2000 turnovers, 1042 were top performers, and overworked:

Satisfaction wise turnover

To understand why highly satisfied people leave the company, it was important to identify the top factors affecting attrition. Following methods were used to achieve this:

- **Visual plots**
- **ANOVA test**: To identify important ‘numerical’ attributes affecting attrition

```
# Using ANOVA for NUMERICAL features

# configure to select all features
fs_an = SelectKBest(score_func=f_classif, k='all')

# learn relationship from training data
fit_an = fs_an.fit(x_train, y_train)
dfscores_an = pd.DataFrame(fit_an.scores_)

# concat two dataframes for better visualization
featureScores_an = pd.concat([dfcolumns,dfscores_an],axis=1)
featureScores_an.columns = ['Specs','Score']  #naming the dataframe columns

print(featureScores_an.nlargest(cols,'Score')) # print all columns in descending order of score
```

```
                   Specs        Score
0             EmployeeID  6997.326549
1      satisfaction_level  1357.776516
5      time_spend_company   293.220228
6           Work_accident   146.991314
8                  salary   145.848530
4    average_montly_hours    51.232356
7   promotion_last_5years    19.222374
3          number_project    10.328191
9                  RandD     9.333941
12            management     3.734636
11                    hr     3.243634
2        last_evaluation     2.196806
15                 sales     1.445641
17             technical     0.518912
13             marketing     0.100529
14           product mng     0.077085
```

Since ANOVA is used for selecting NUMERICAL features, we'll ignore the CATEGORICAL features from the above list. Following NUMERICAL features have highest correlations with the dependent variable 'turnover'.

1. satisfaction_level
2. time_spend_company
3. average_montly_hours

- **Chi-square test**: To identify important ‘categorical’ attributes affecting attrition

```
# apply SelectKBest class to get feature importance values

# Using 'Chi-square test' for CATEGORICAL features

fs_chi = SelectKBest(score_func=chi2, k='all')
fit = fs_chi.fit(x_train,y_train)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(x_train.columns)

#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns

cols = len(x_train.columns)
print(featureScores.nlargest(cols,'Score')) # print all columns in descending order of score
```

```
                   Specs         Score
0             EmployeeID  8.098970e+06
4    average_montly_hours  6.034258e+02
5      time_spend_company  1.507195e+02
6           Work_accident  1.225708e+02
1      satisfaction_level  1.097588e+02
8                  salary  9.599003e+01
7   promotion_last_5years  1.885816e+01
9                   RandD  8.785666e+00
3          number_project  3.700692e+00
12             management  3.600161e+00
11                     hr  3.076661e+00
15                  sales  1.058090e+00
17              technical  4.216269e-01
13              marketing  9.487204e-02
```

Since Chi-square is used for selecting CATEGORICAL features, we'll ignore the NUMERICAL features from the above list. Following CATEGORICAL features have good correlation with the dependent variable 'turnover'.

1. Work_accident
2. Salary

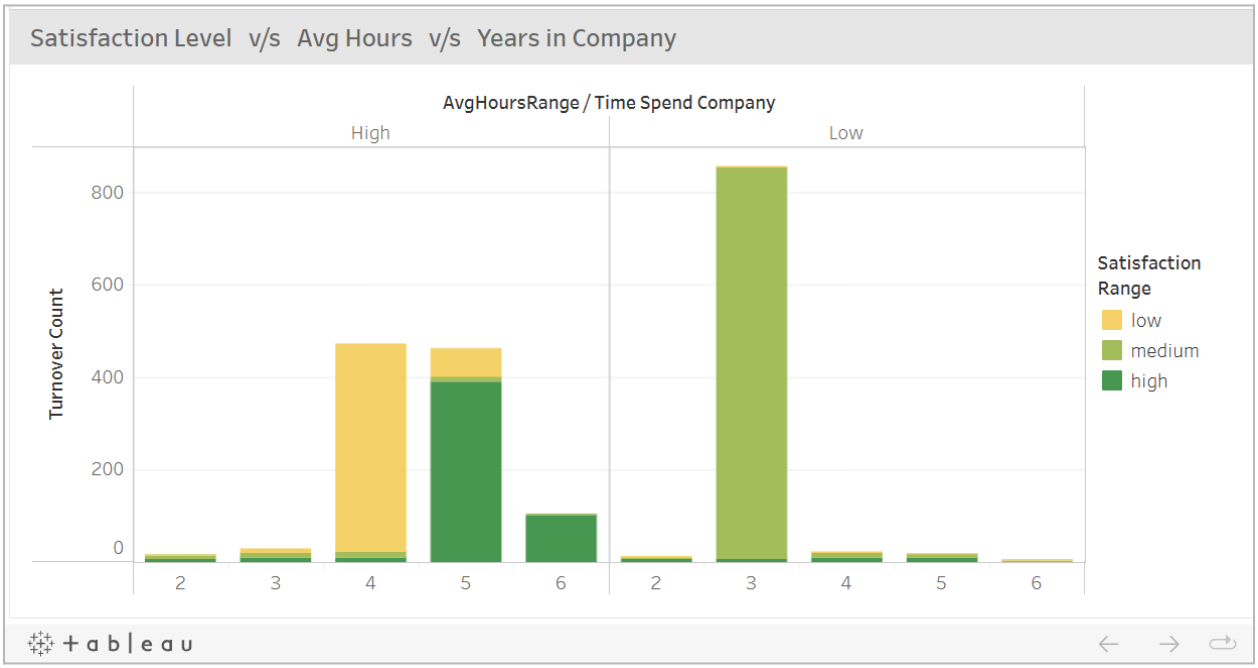
- **features_importances_** : This method gives weighted importance of all the features that the Machine Learning model used to make predictions for the given dataset.

Using the features_importances_ report of the Random Forest model, among all the factors, **satisfaction_level', 'time_spend_company'** and **'average_monthly_hours'** were identified to be the biggest factors leading to employee turnover.

```
# Get most important features

# feature importance

feat_imp_rf = pd.Series(rf_model.feature_importances_,index=x_train_imp.columns)
print(feat_imp_rf)
```

```
satisfaction_level       0.383302
time_spend_company       0.200326
average_montly_hours     0.203321
number_project           0.190245
Work_accident            0.008520
salary                   0.012720
promotion_last_5years    0.001566
dtype: float64
```

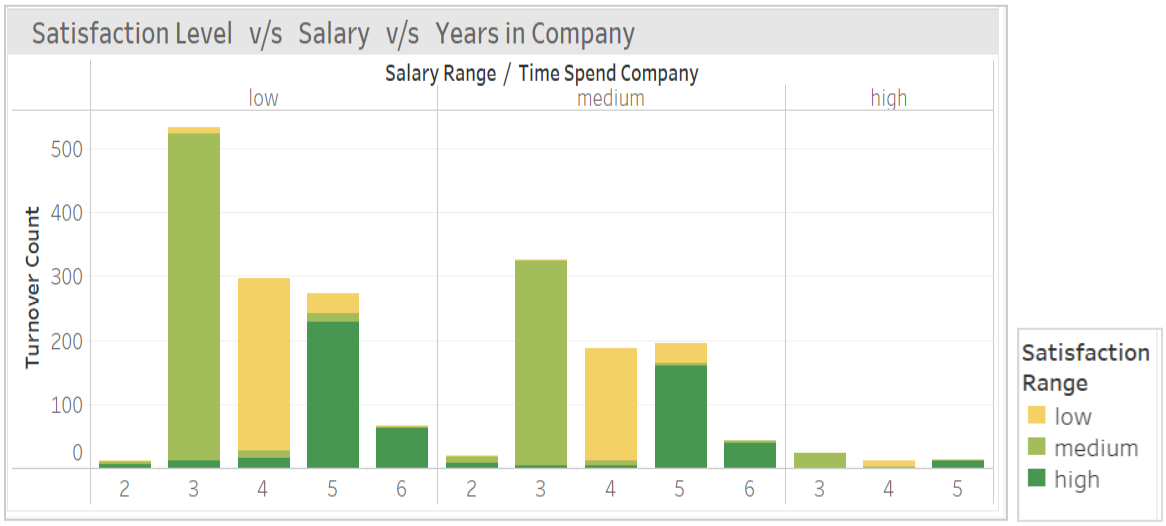
Analyzing these 3 factors together:



The above plot clearly shows that **highly satisfied employees who left were working long hours, and had been in the company for 5-6 years**.

Also, there seems to be a **level of dissatisfaction among employees who were in the company for 3 to 4 years.**

Analyzing salary range for this group:



Employees who were in the company for 3-4 years had a low-medium salary range.

Key takeaways and Potential solutions:

Looking at the data, following are the key points:

- It seems that many people are leaving because of low levels of satisfaction, not getting promoted and over-work.
- Insufficient compensation is a big reason for dissatisfaction among employees.
- A common factor among most turnover employees is lack of promotion, due to which employees do not find their jobs rewarding, leading to their attrition.
- Highest attrition is in the HR department, followed by Accounting and Technical teams.

Based on the above identified factors, following solutions should be put in place:

- **Restrict work hours** → The HR policy needs a change to implement stricter working hours, not exceeding the expected average of 160 hours per month.
- **Better compensation** → A common factor among dissatisfied employees was low to medium salary ranges. The company can follow industry standards as criteria to drive its cost to company and salary guidelines.
- **Promotion policy review** → With more than 99% of turnover employees not getting promoted in 5 years, the promotion policy needs an immediate review to address the extremely low rate of promotion in the company.
- **Department oriented** → Efforts should be more focussed on departments seeing higher attrition rates - HR, Accounting and Technical.

Project file links

GitHub repository

https://github.com/shailysahay/EmployeeAttrition

Tableau Dashboards

Department KPIs and Performance wise turnover:
https://public.tableau.com/views/EmployeeAttrition_Dash1/Dash1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Tops Factors Analysis:
https://public.tableau.com/views/EmployeeAttrition_16742326659470/DASH_2?:language=en-US&:display_count=n&:origin=viz_share_link