IBM HR Analytics Employee Attrition & Performance

Analysis Report

Attrition Analytics – EDA & Predictive Modelling

Objective: The objective of the present report is to study factors like salary, satisfactory level, growth opportunities, facilities, policies and procedures, recognition, appreciation, suggestions of the employee's by which it helps to know the Attrition level in the organizations and factors relating to retain them. This study also helps to find out where the organizations are lagging in retaining.

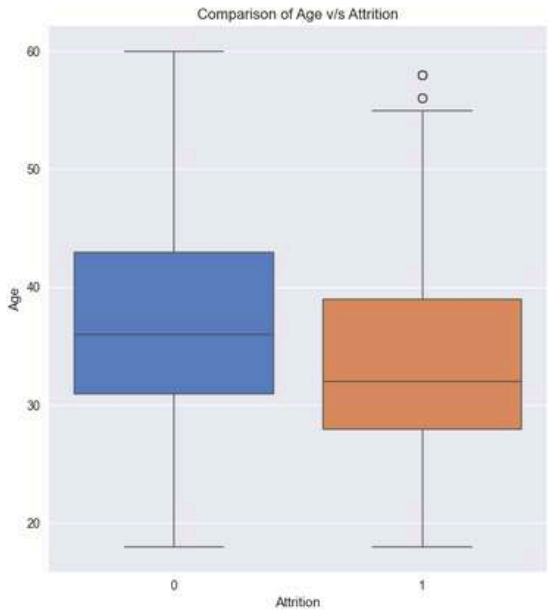
Exploratory Data Analysis

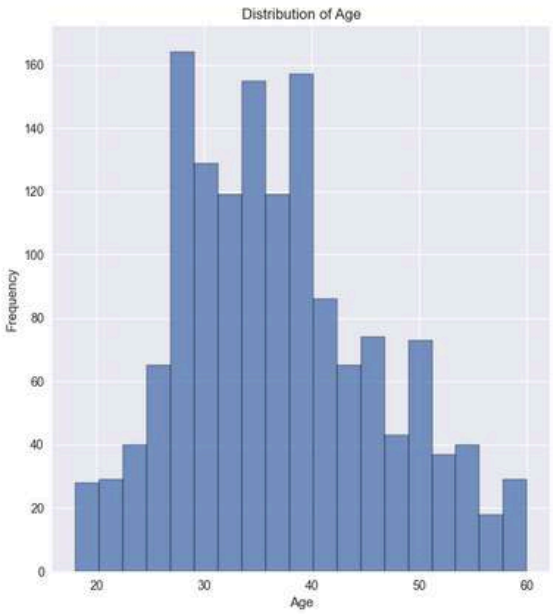
Features in the dataset:

- Age
- Attrition
- BusinessTravel
- DailyRate
- Department
- DistanceFromHome
- Education
- EducationField
- EmployeeCount
- EmployeeNumber
- EnvironmentSatisfaction
- Gender
- HourlyRate
- JobInvolvement
- JobLevel
- JobRole
- JobSatisfaction
- MaritalStatus
- MonthlyIncome
- MonthlyRate
- NumCompaniesWorked
- Over18
- OverTime
- PercentSalaryHike
- PerformanceRating
- RelationshipSatisfaction
- StandardHours
- StockOptionLevel
- TotalWorkingYears
- TrainingTimesLastYear
- WorkLifeBalance
- YearsAtCompany
- YearsInCurrentRole
- YearsSinceLastPromotion
- YearsWithCurrManager

- **Over18**, **EmployeeCount** and **StandardHours** columns have only 1 unique value and **EmployeeNumber** has 1470 unique values which suggests that they are not very useful for us. So. we are going to drop them.
- For removing other, not useful columns, we need to visualize the importance of features on Attrition.

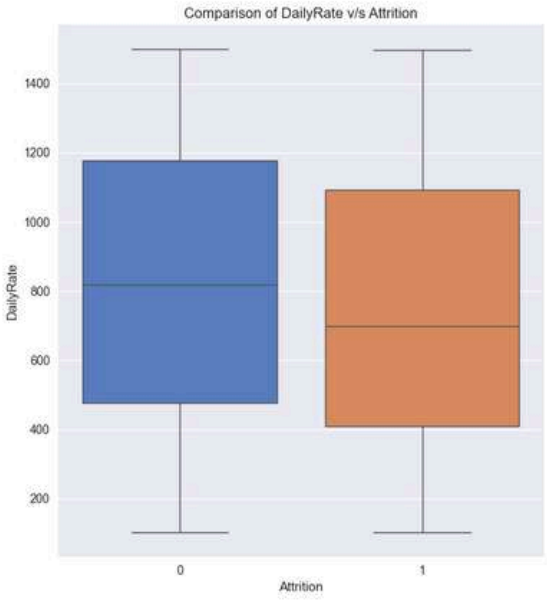
Attrition vs Age:

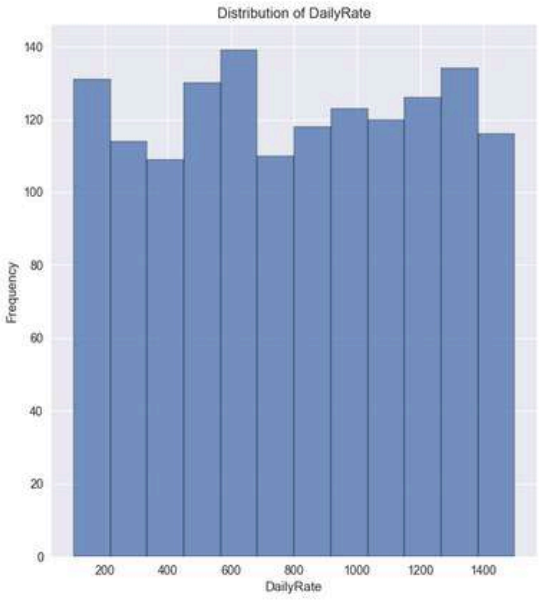




- From boxplot, it's clear that the median age of employees in the company falls within the range of 30 to 40 years.
- The youngest employee is 18 years old, while the oldest is 60 years old.
- From Age Comparison boxplot, it's evident that a significant proportion of those who left the company are under 40 years old.
- Conversely, among those who didn't leave the company, the majority are between the ages of 32 and 40 years old.

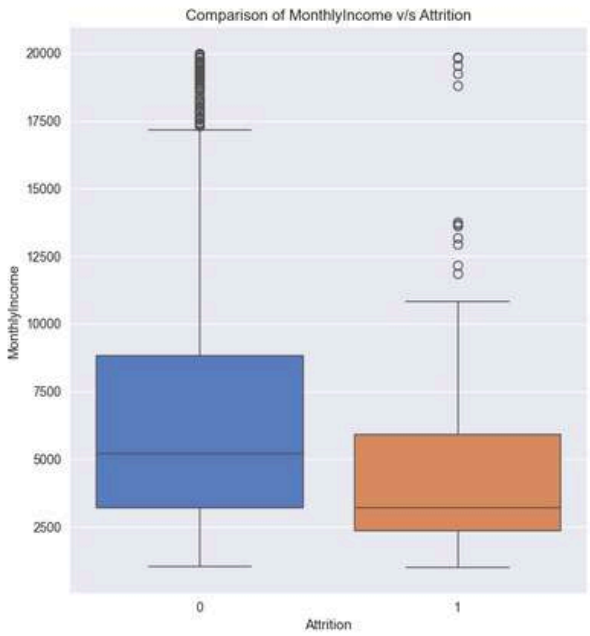
Attrition vs Daily Rate:

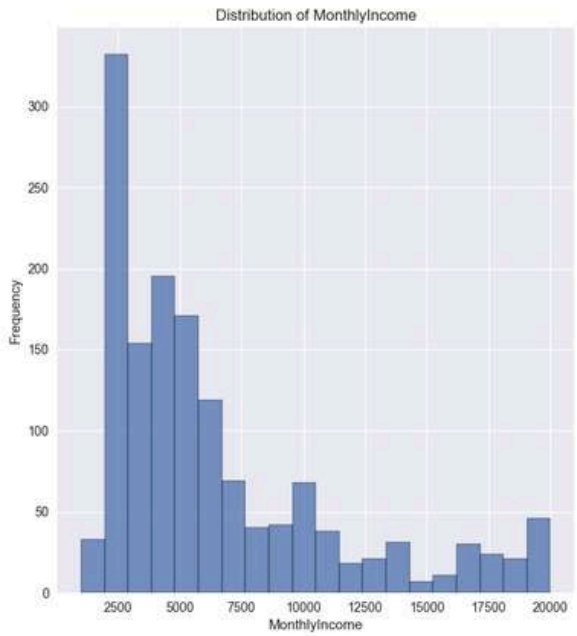




- As it can be observed from the plot, that employees with lower 'DailyRate' are more prone to the attrition.

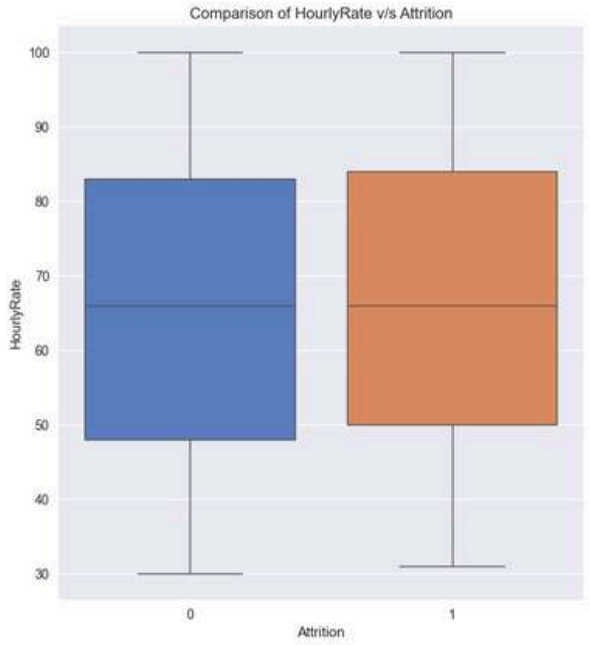
Attrition vs Monthly Income:

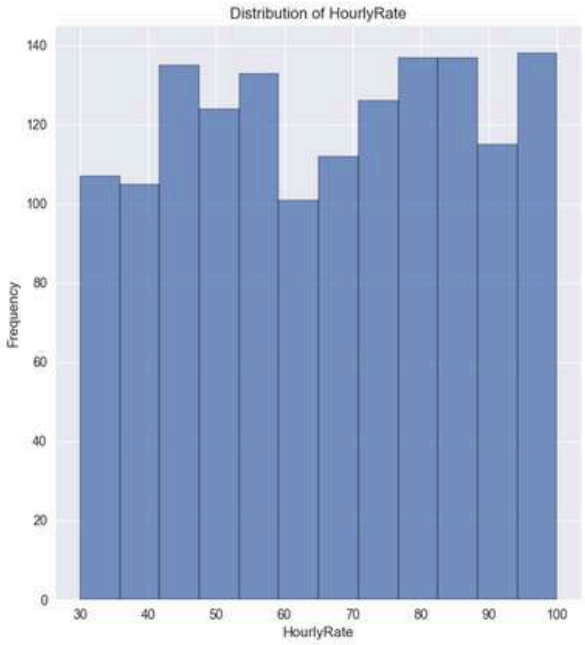




- Same trend goes with Monthly Income also, the employees with lower 'MonthlyIncome' are more likely to leave the company.

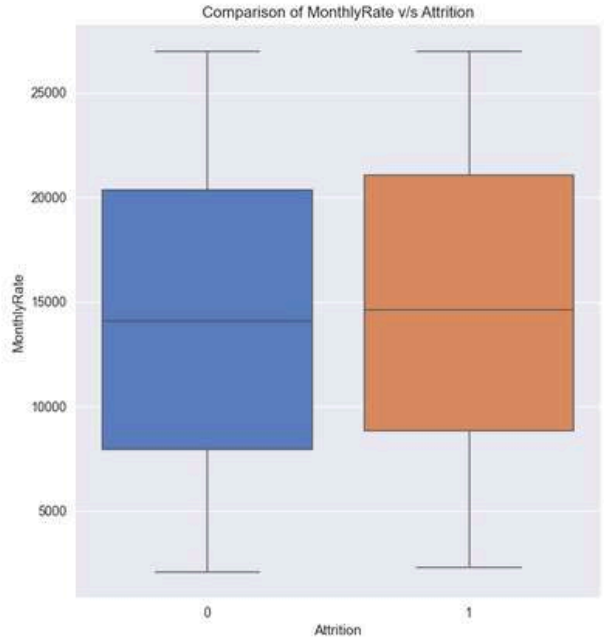
Attrition vs Hourly Rate:

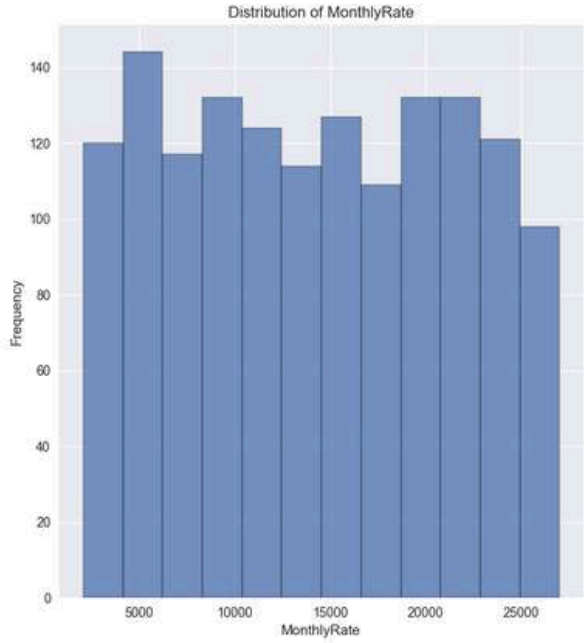




- Same trend goes with Monthly Income also, the employees with lower 'MonthlyIncome' are more likely to leave the company.

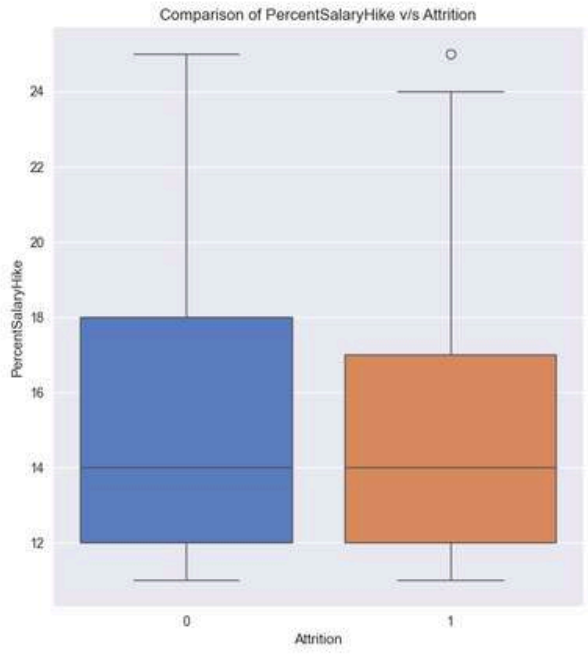
Attrition vs Monthly Rate:

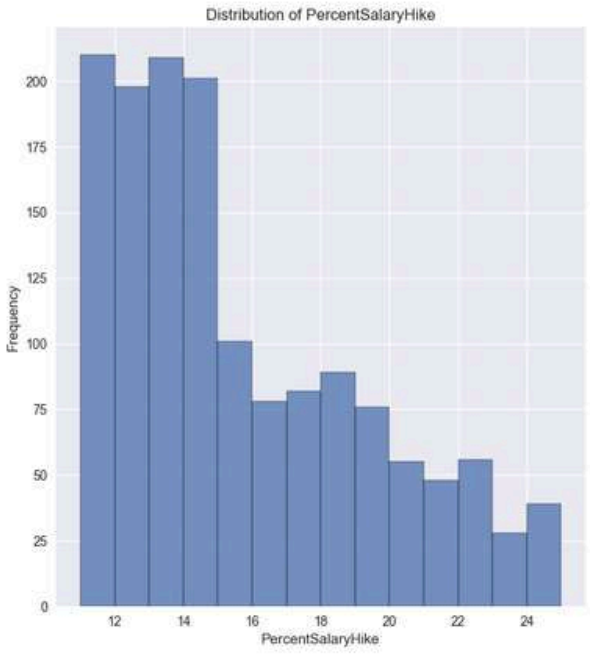


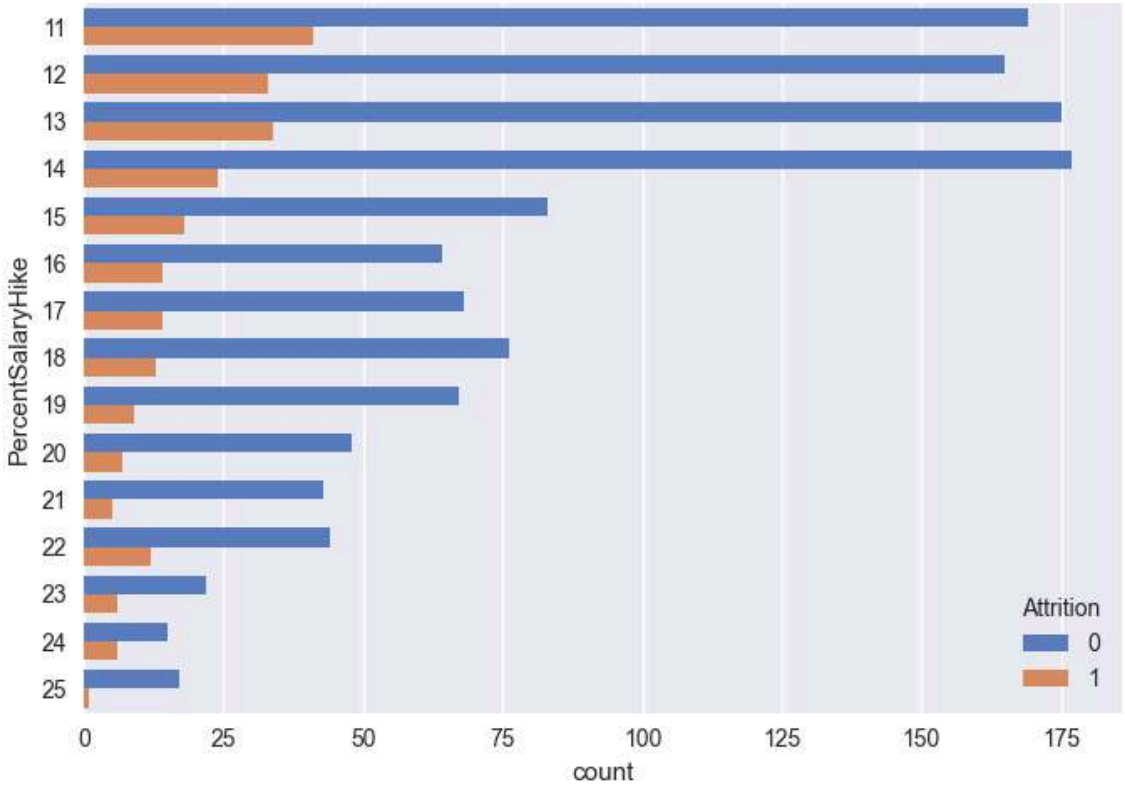


- For 'HourlyRate' and 'MonthlyRate', we can see that there is no significant relation between them and Attrition.
- Employees with high 'HourlyRate' and high 'MonthlyRate' are also leaving the company, which says there might be some other reason for leaving the company, so both are considered not significant to the Attrition.

Attrition vs Percent Salary Hike:

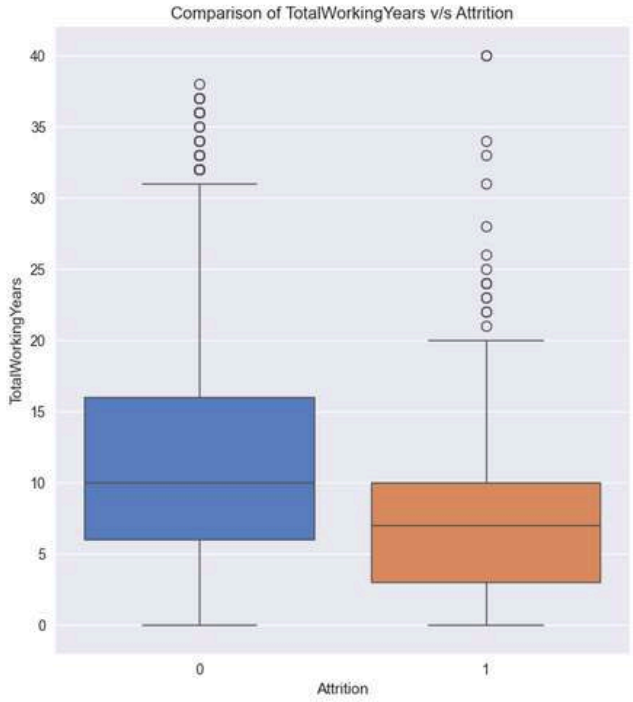


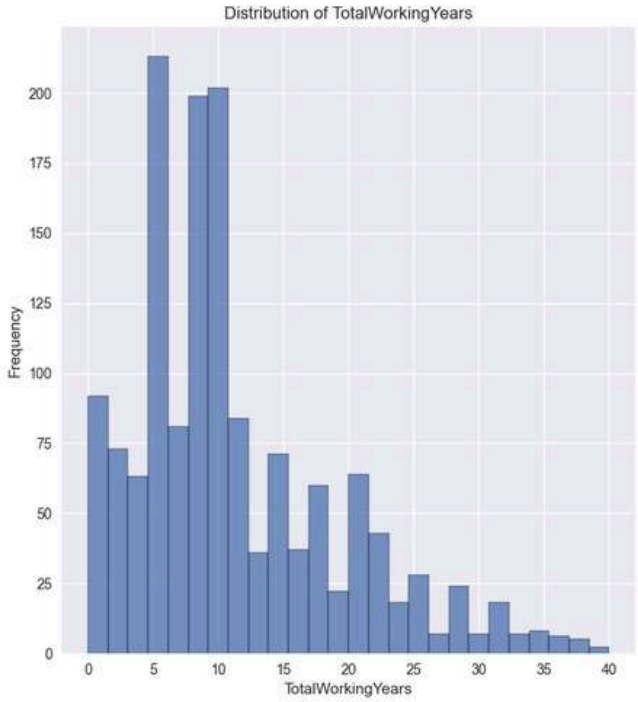


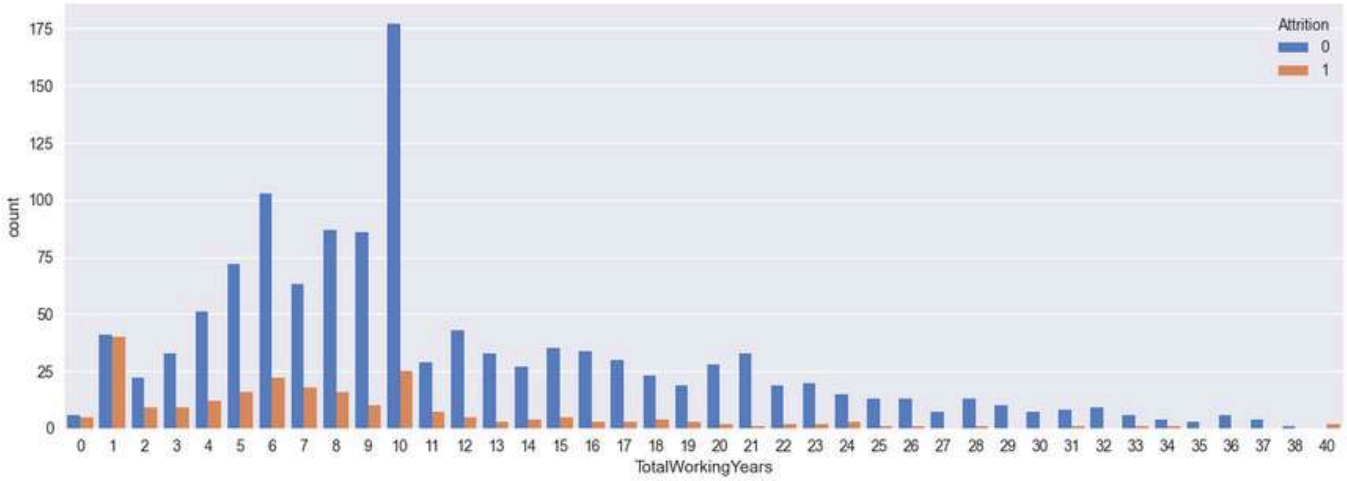


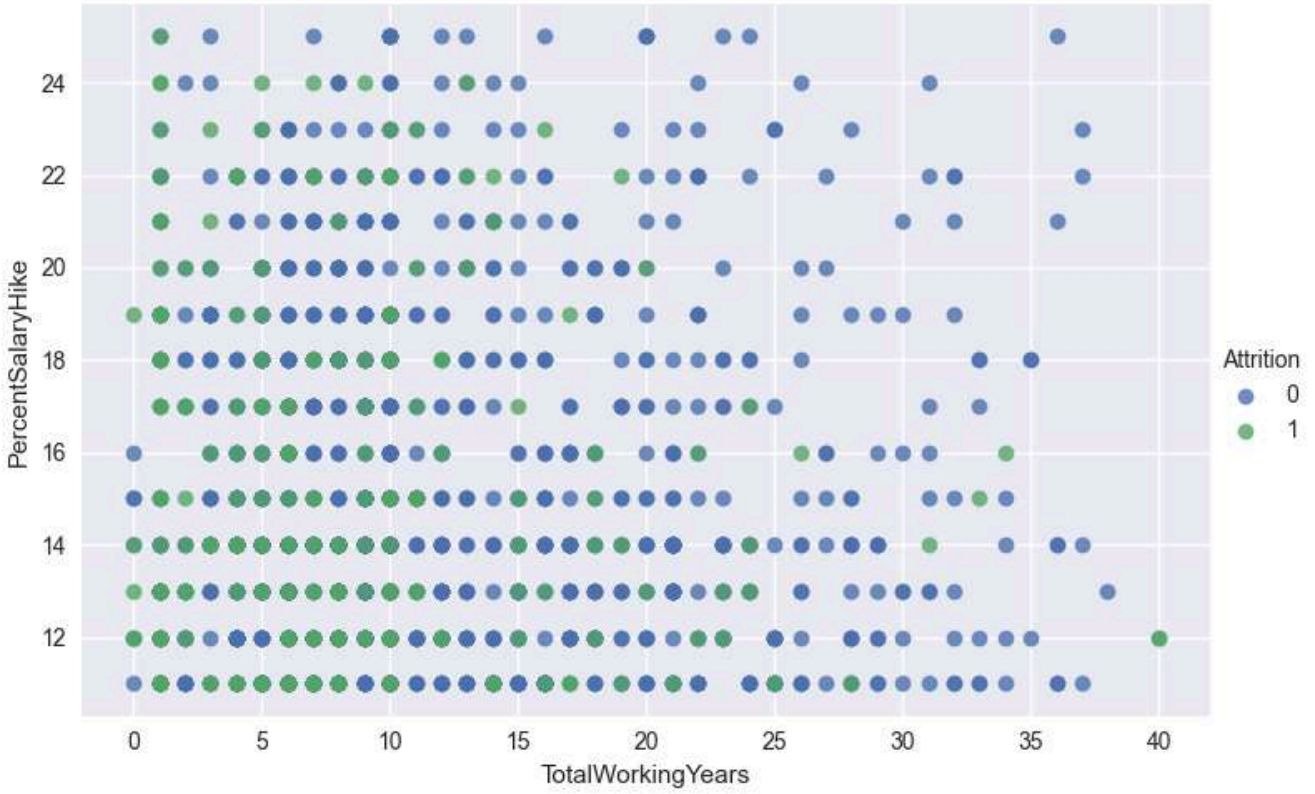
- Employees with high 'PercentSalaryHike' also left the company, but the number is very low.
- As expected, the employees with low PercentSalaryHike are more prone to Attrition.

Attrition vs Total Working Years:



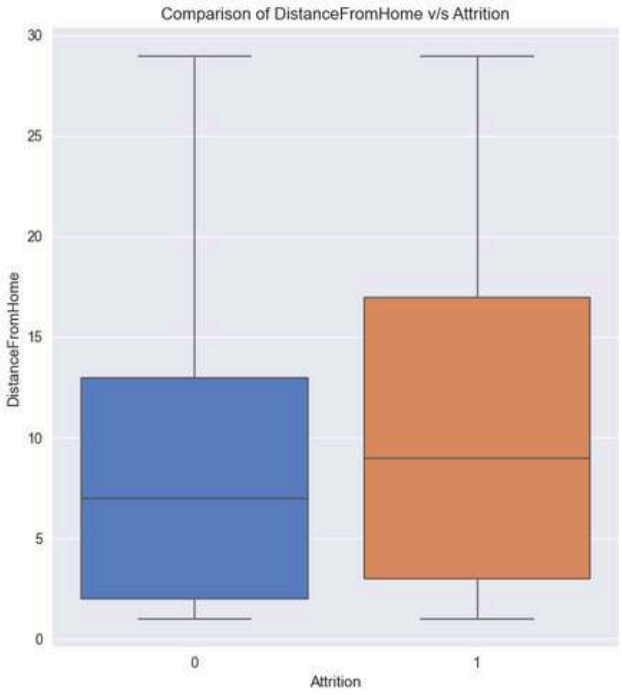


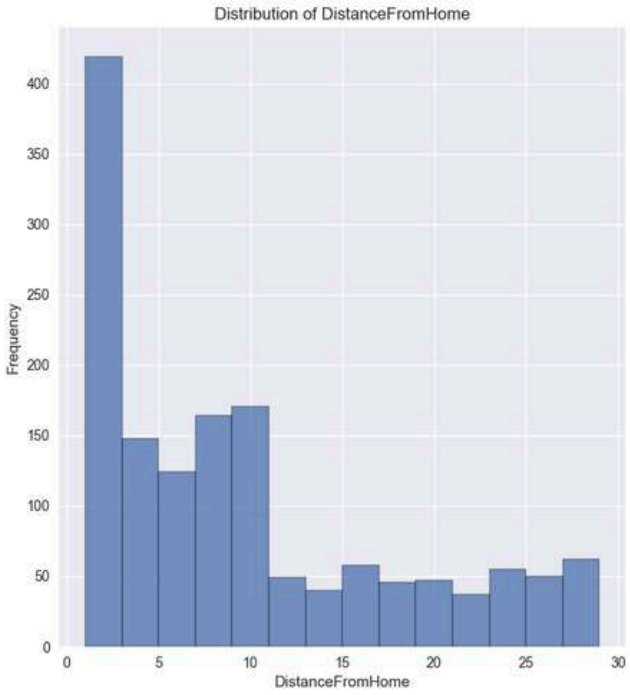


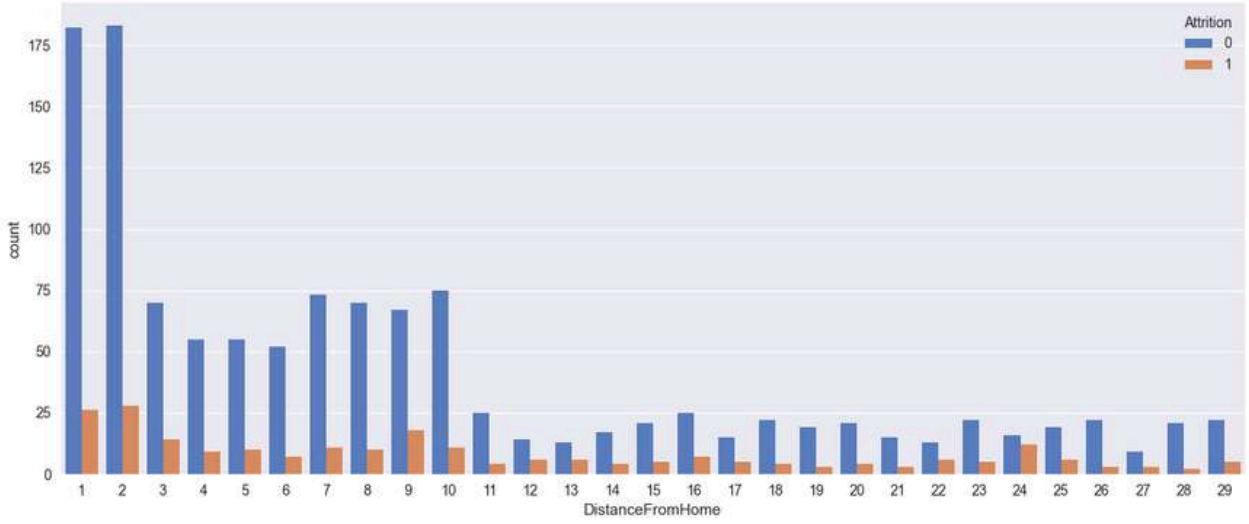


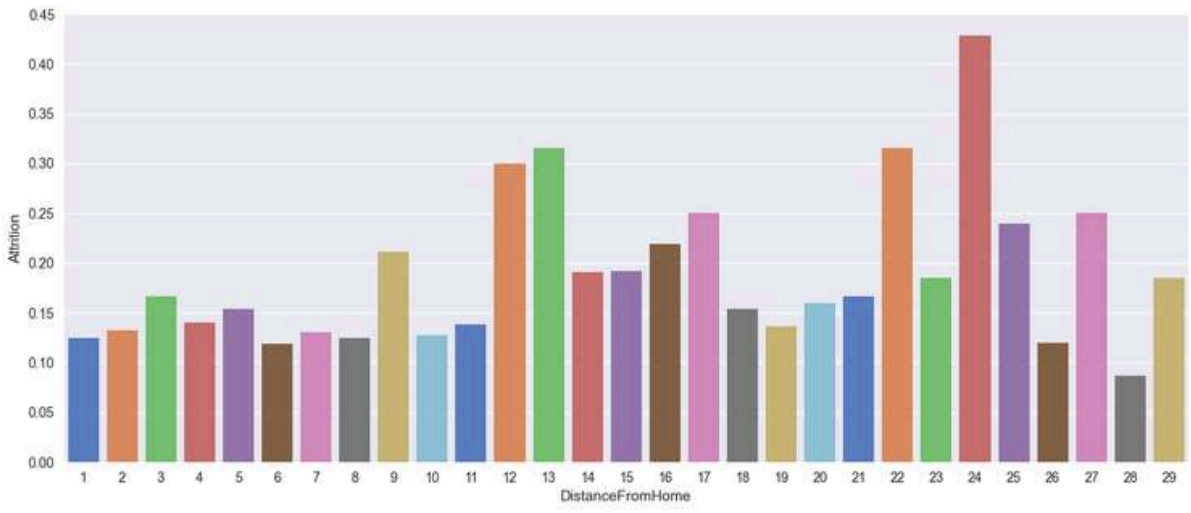
- It can be seen from the lmplot of 'PercentSalaryHike' vs 'TotalWorkingYears' that employees with fewer years of experience are more prone to attrition irrespective of the percent of salary hike. There is no fixed relationship between the percent salary hike and total working years.
- Attrition is not seen among the employees having more than 20 years of experience if their salary hike is more than 18%, even if the salary hike is below 18% attrition among the employees is very low.

Attrition vs Distance From Home:



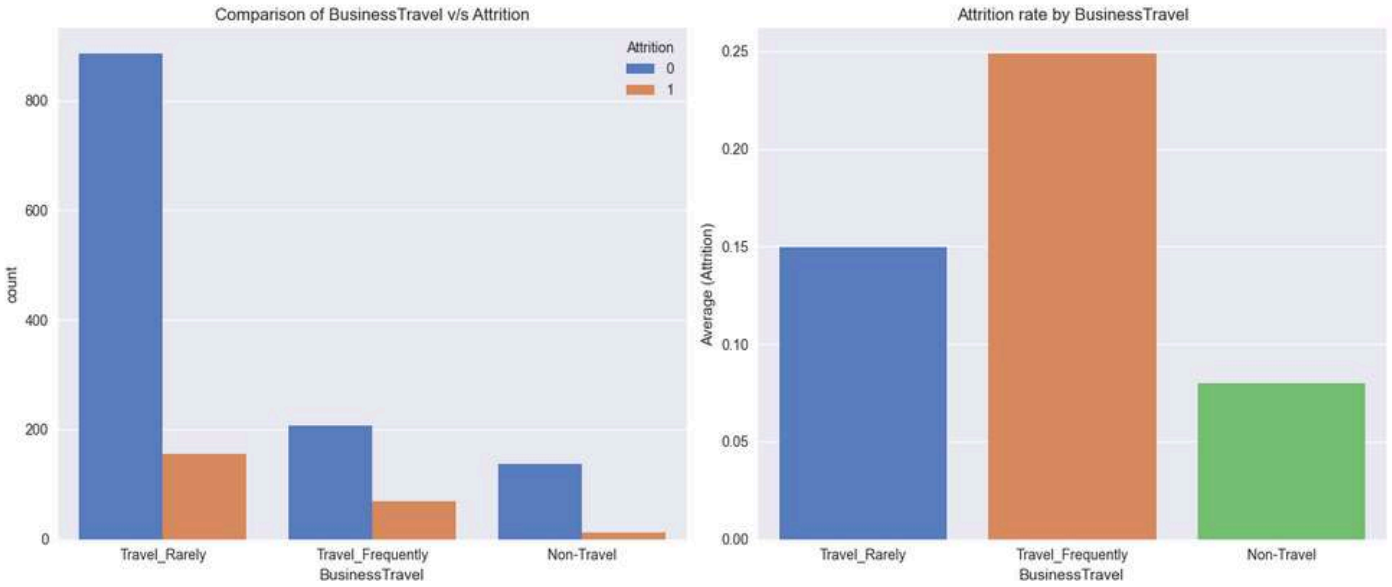






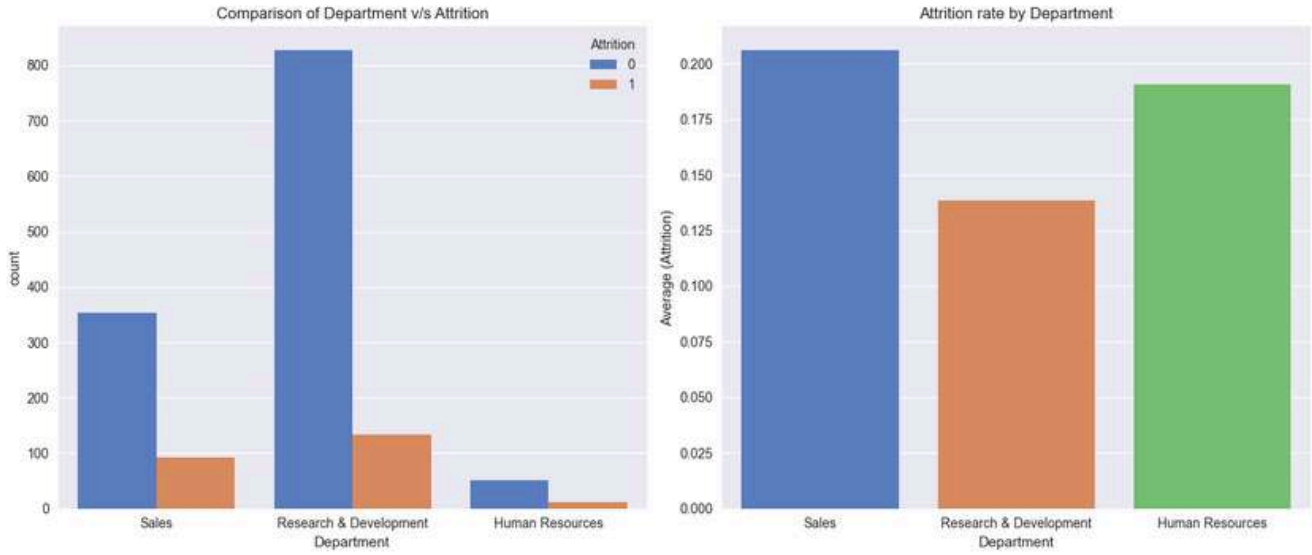
- There is a higher number of people who reside near to company and hence the attrition levels are lower for distances less than 10.
- With the increase in distance from home, the attrition rate also increases.
- One more interesting observation is that more than 85% of employees who live within nearly 10 km distance from the company, didn't leave the company.

Attrition vs Business Travel:



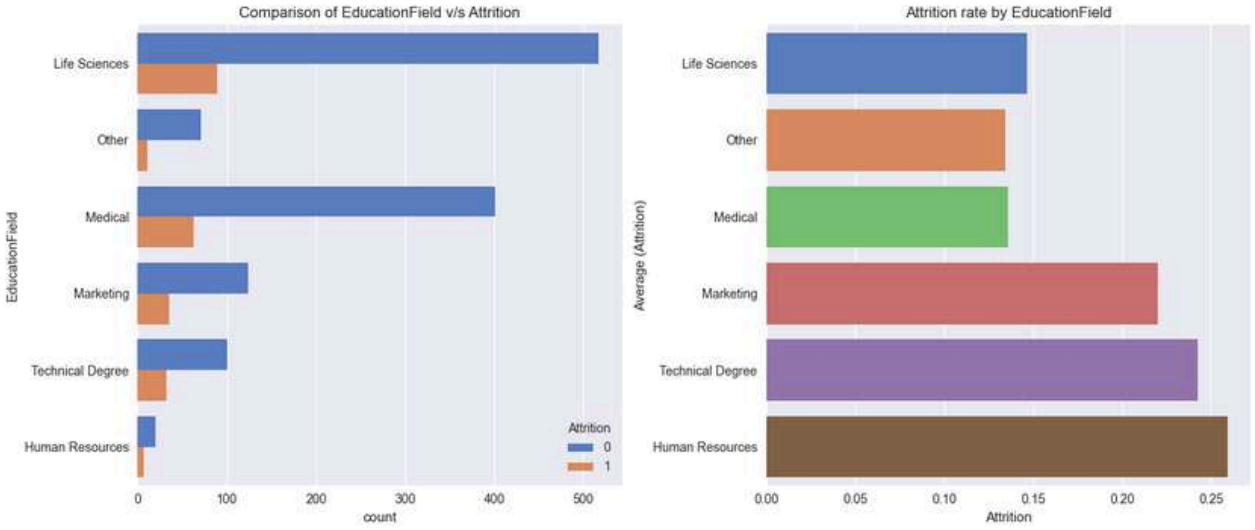
- There are more number of employees who travel rarely compared to that of employees who travel frequently.
- In the case of employees who travel frequently around 25% of them have left the company, in the case of employees who travel rarely around 15% of them have left the company and in Non-Travel cases, less than 10% of employees have left the company, but in last two cases attrition rate doesn't vary significantly on travel.

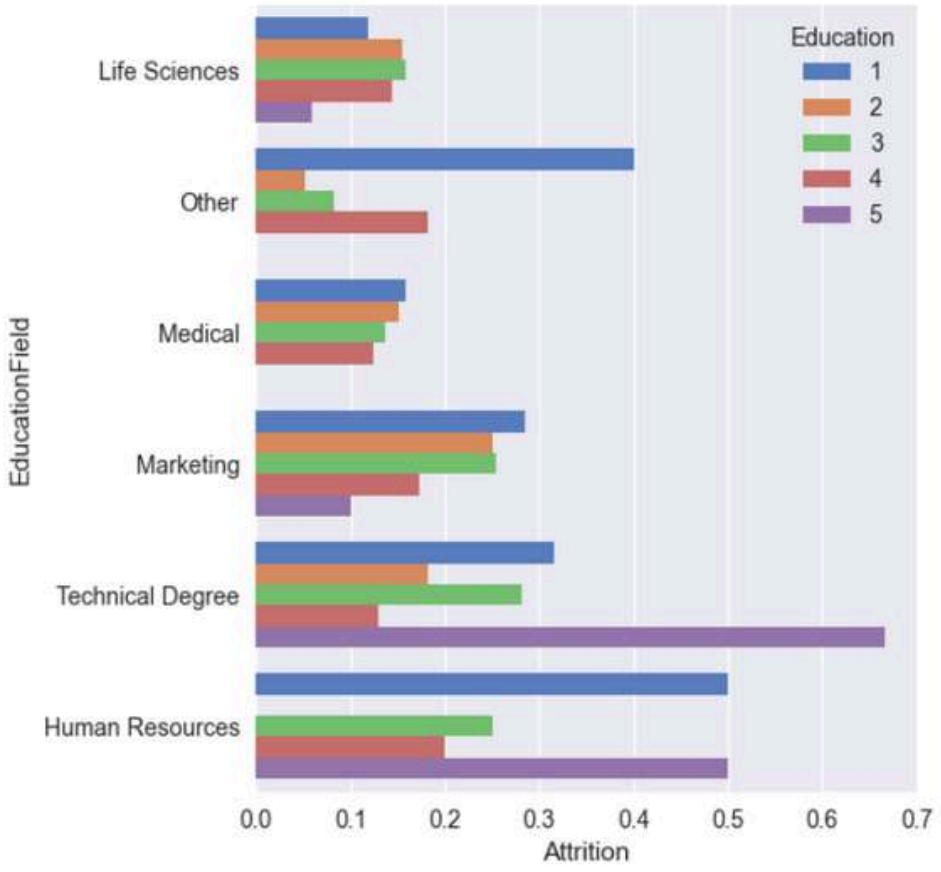
Attrition vs Department:



- R & D Department has highest number of employees, followed by Sales Department, and least being HR Department.
- Sales Department has seen highest attrition level (around 20%) among all, followed by HR Department (around 18%).
- It's really bad condition for HR Department, because the number in total is already less, on top of that Attrition rate is also high.
- It can be said that HR Department and variables related to it has a big impact on the Attrition.

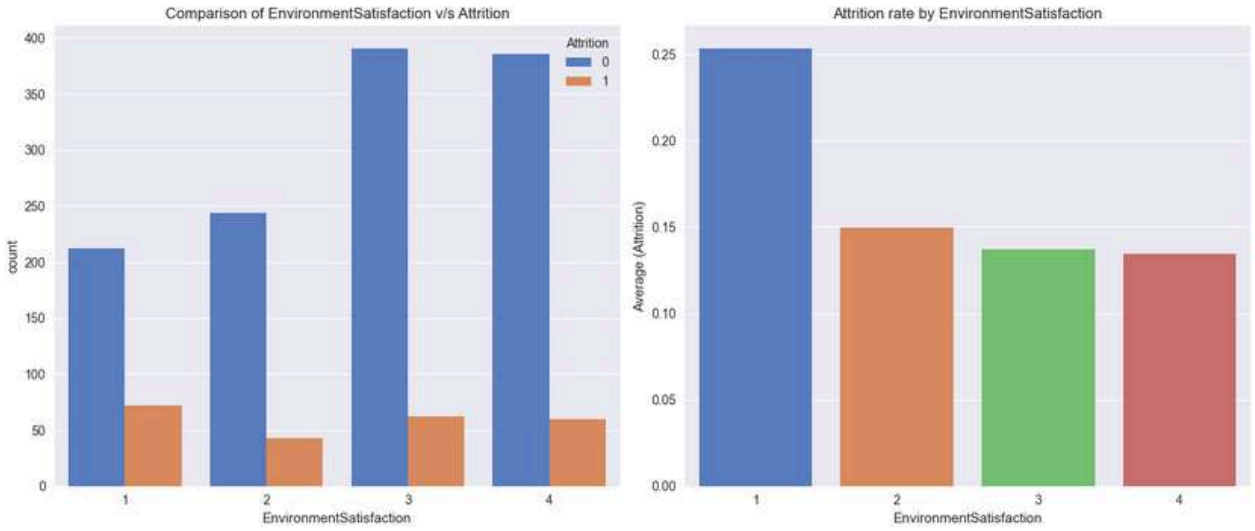
Attrition vs Education Field and Education Level:

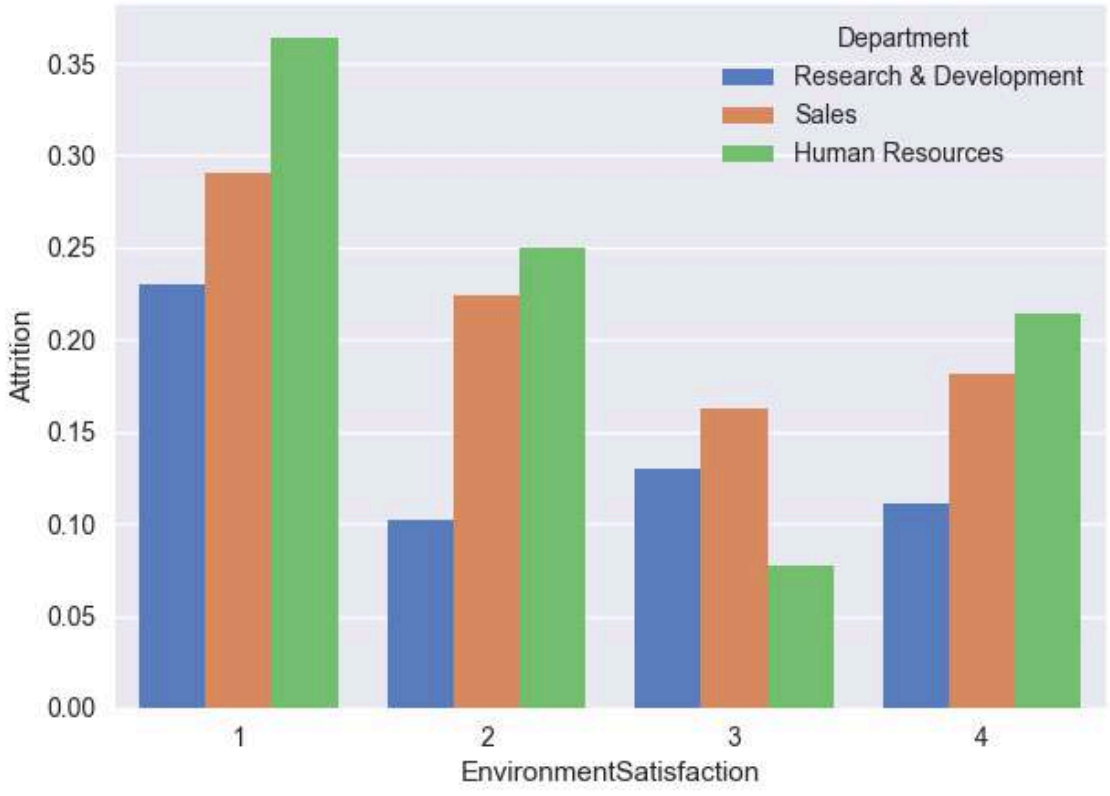


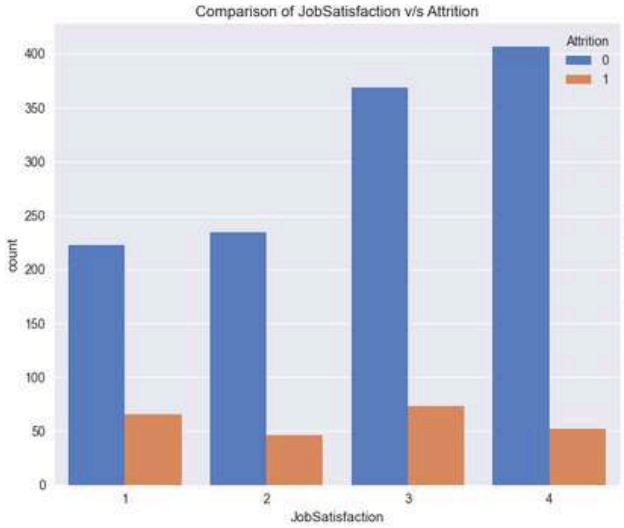


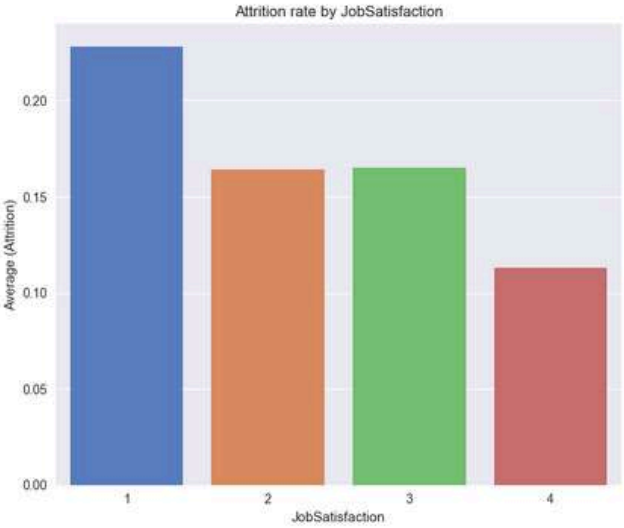
- Around 41% of total employees have Life Sciences Education Field, followed by Medical(around 31%) and Marketing(around 11%).
- Employees with HR and Technical Degree Education fields have the highest attrition levels with 26% and 24% respectively.
- Among Employees in the Human Resources and Technical Degree Education field, employees with the highest level of education (i.e. Doctor here) are more likely to leave the company.

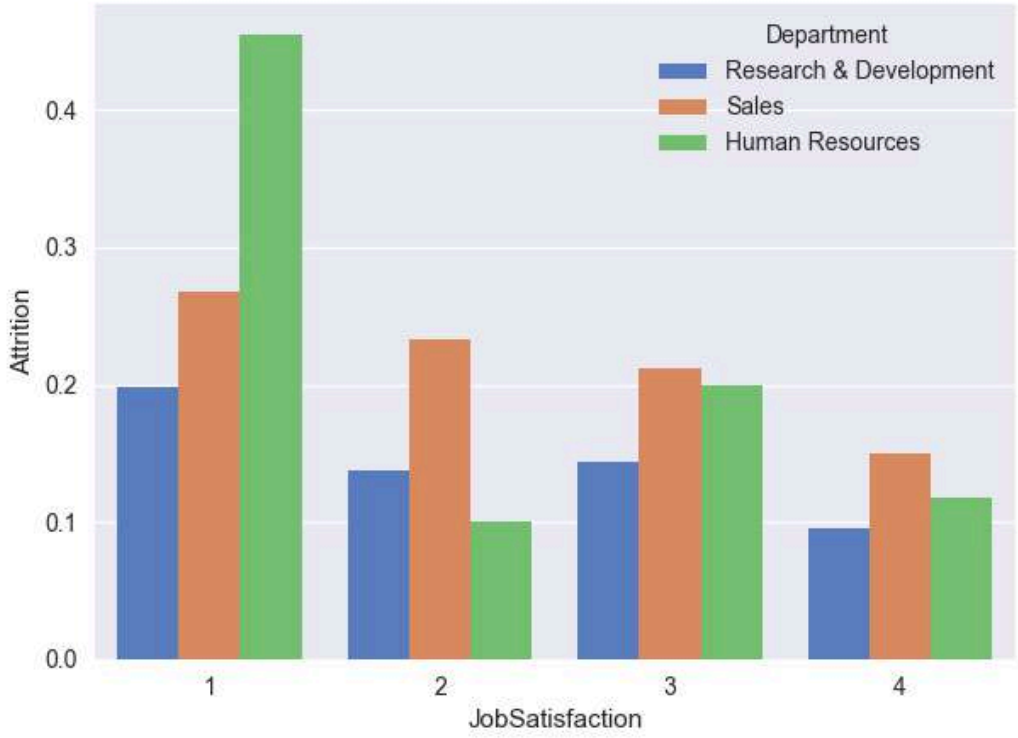
Attrition vs Environment Satisfaction, Job Satisfaction, Relationship Satisfaction:

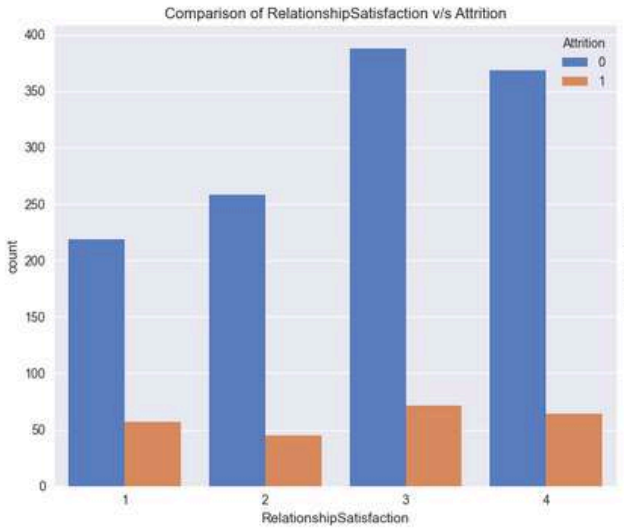


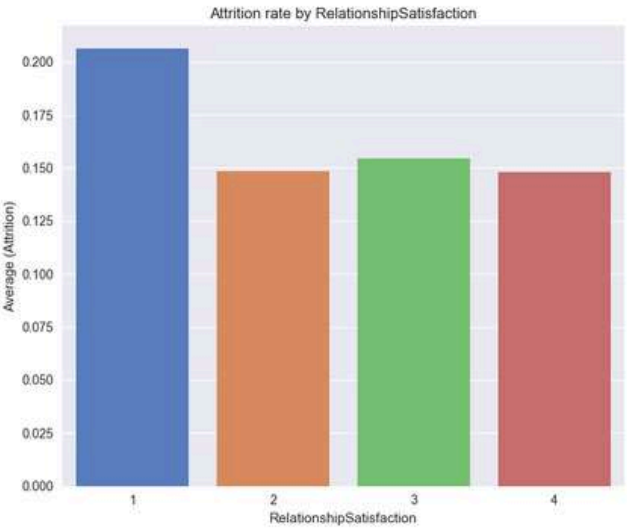



Comparison of JobSatisfaction v/s Attrition
Attrition
0
1
count
JobSatisfaction


Attrition rate by JobSatisfaction
Average (Attrition)
JobSatisfaction

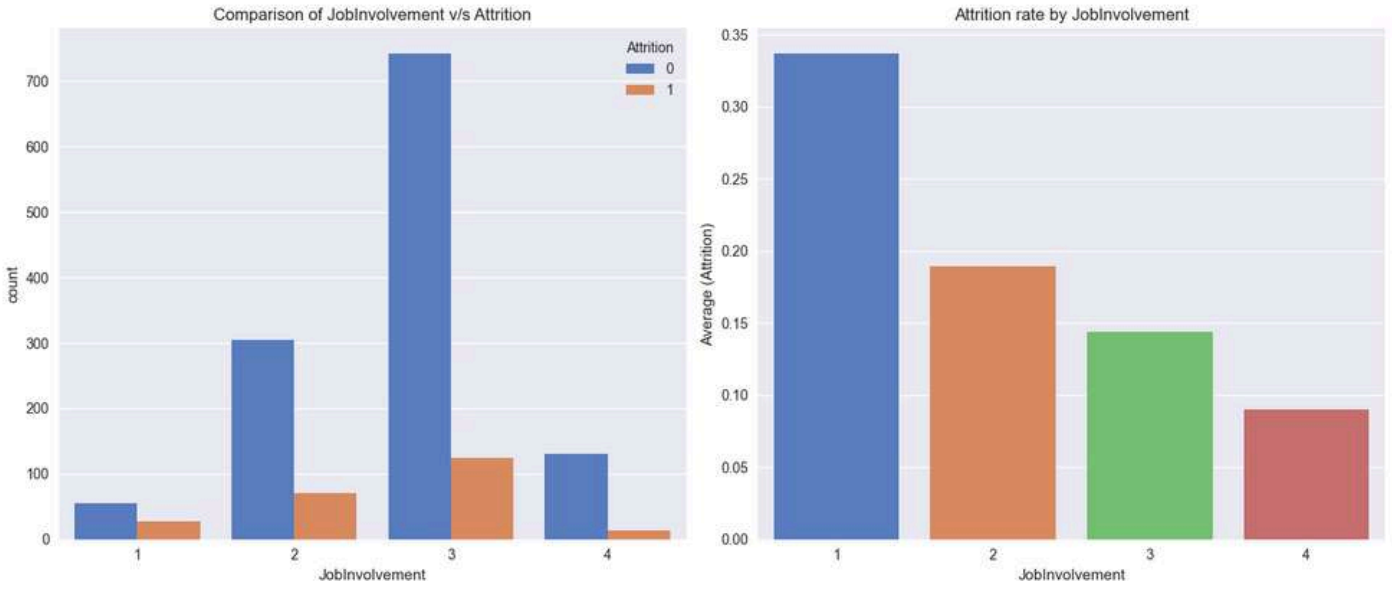

Department
Research & Development
Sales
Human Resources
Attrition
JobSatisfaction

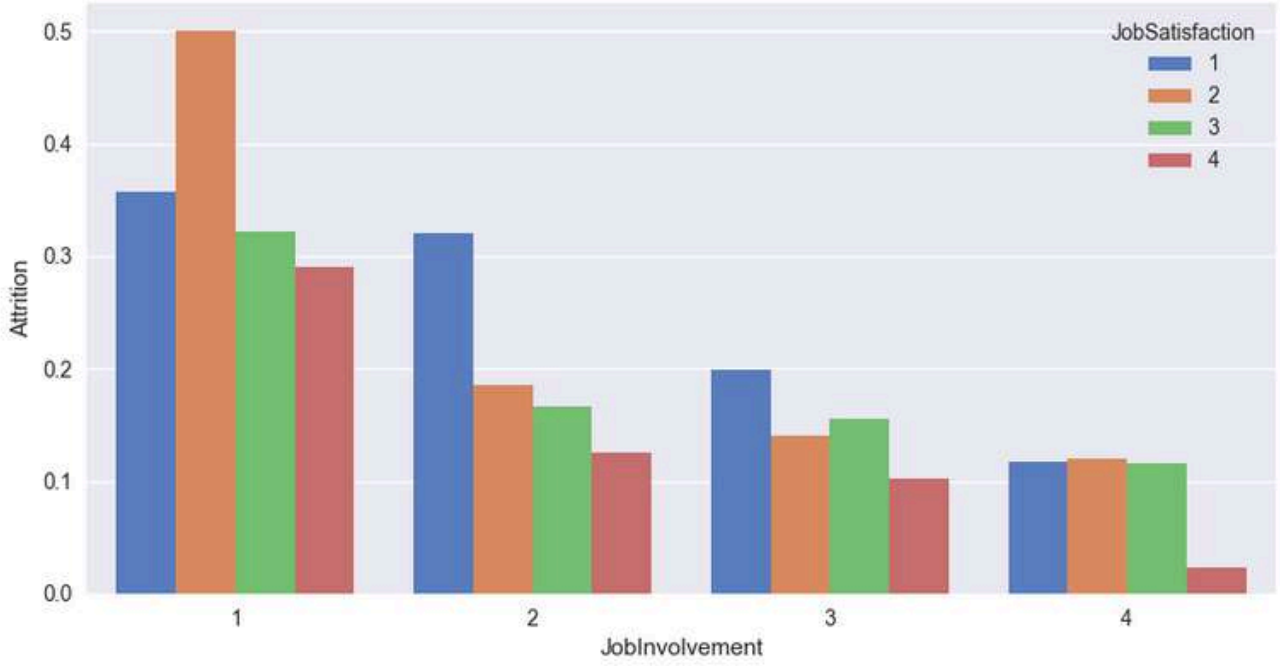

Comparison of RelationshipSatisfaction v/s Attrition
Attrition
0
1
count
RelationshipSatisfaction


Attrition rate by RelationshipSatisfaction
Average (Attrition)
RelationshipSatisfaction

- The distribution of ratings in EnvironmentSatisfaction, JobSatisfaction, and RelationshipSatisfaction are similar. (around 30%-30%-20%-20% for 4-3-2-1 ratings respectively. These are not exact values.)
- Employees who are not satisfied with their environment, jobs, and relationships, have high attrition levels (20-25%).
- Employees who are highly satisfied with Job, Environment, and Relationships have also left the company but attrition levels are lower relatively.
- Upon mapping the distribution of EnvironmentSatisfaction and JobSatisfaction with Department, it is seen that the employees who have the least EnvironmentSatisfaction and JobSatisfaction, the majority are from the HR Department.

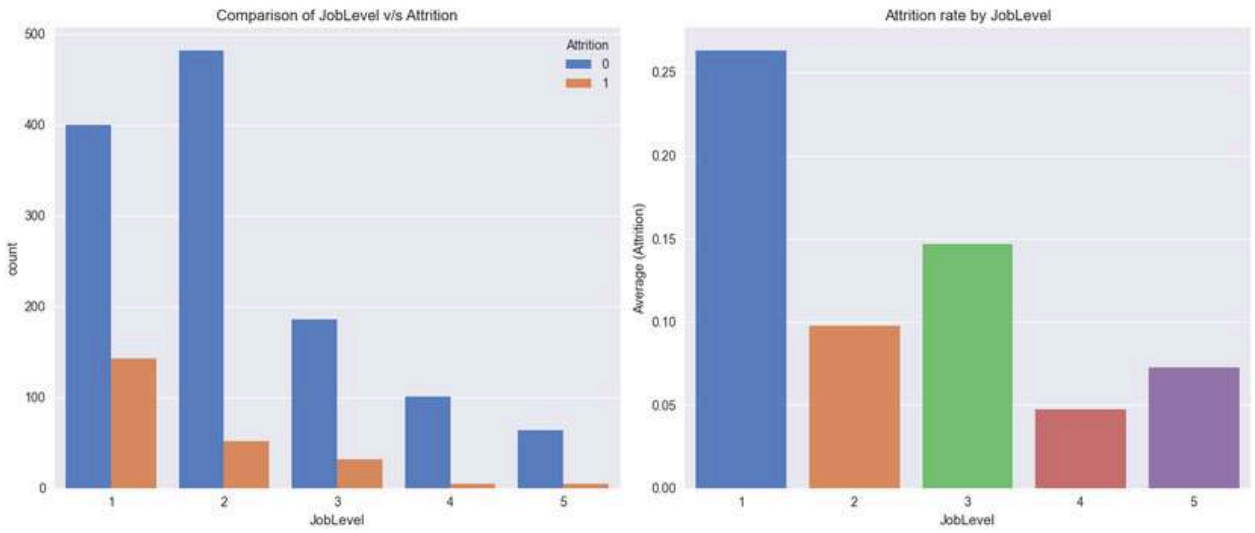
Attrition vs Job Involvement:





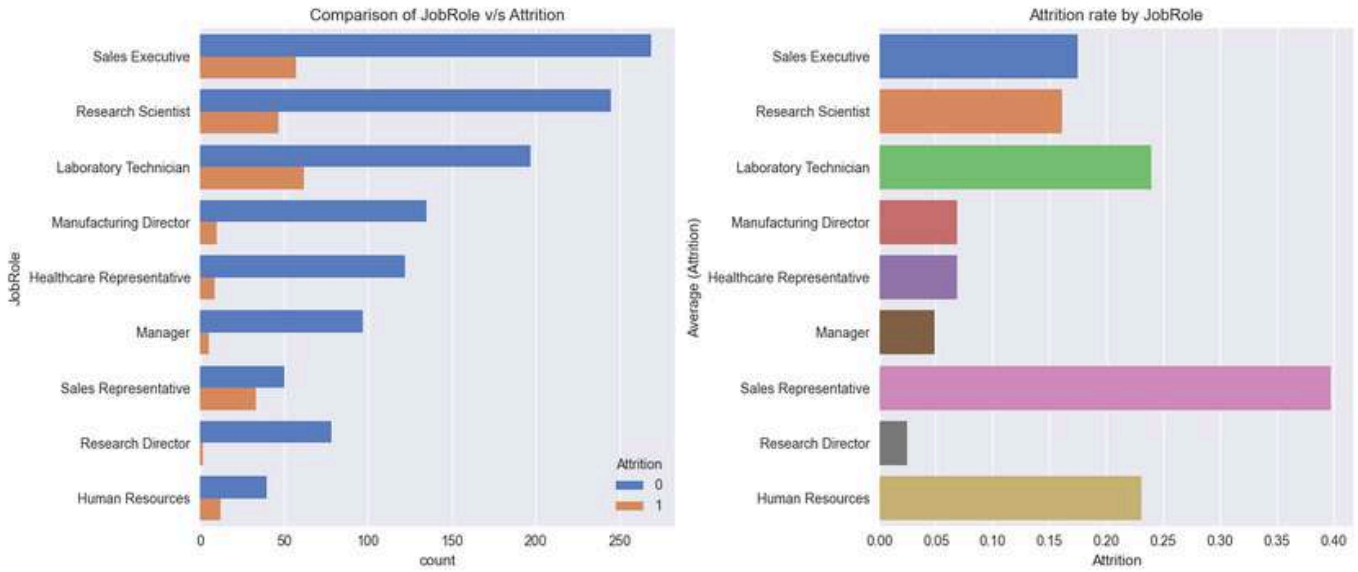
- Around 59% of total employees have high job involvement whereas 25% have medium job involvement.
- From the barplot above, we can see that around 33% of employees who have low job involvement have left the company.
- Even employees with high job involvement, have considerable attrition levels (around 10-15%).
- Maybe, Employees with high job involvement are not satisfied with the job.

Attrition vs Job Level:



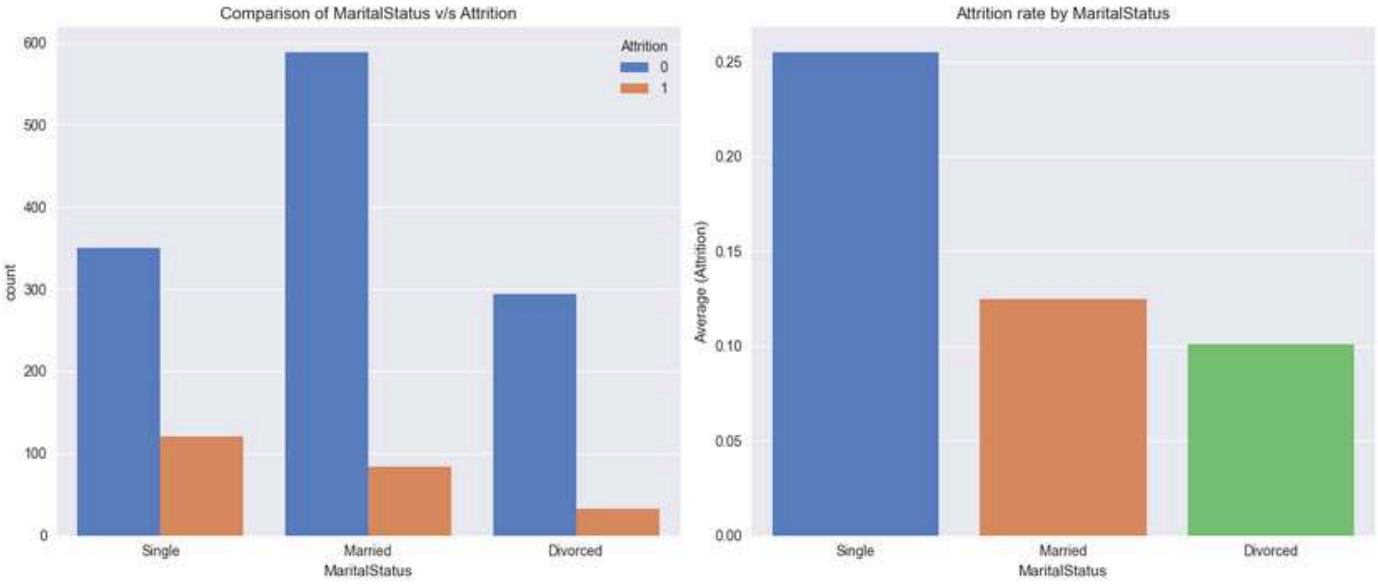
- Attrition Level is highest for employees with Job Level 1, it is around 26%.
- For other higher job levels, the trend is quite irregular.
- No significant relation was found after the first job level between Attrition and Job Level.

Attrition vs Job Role:



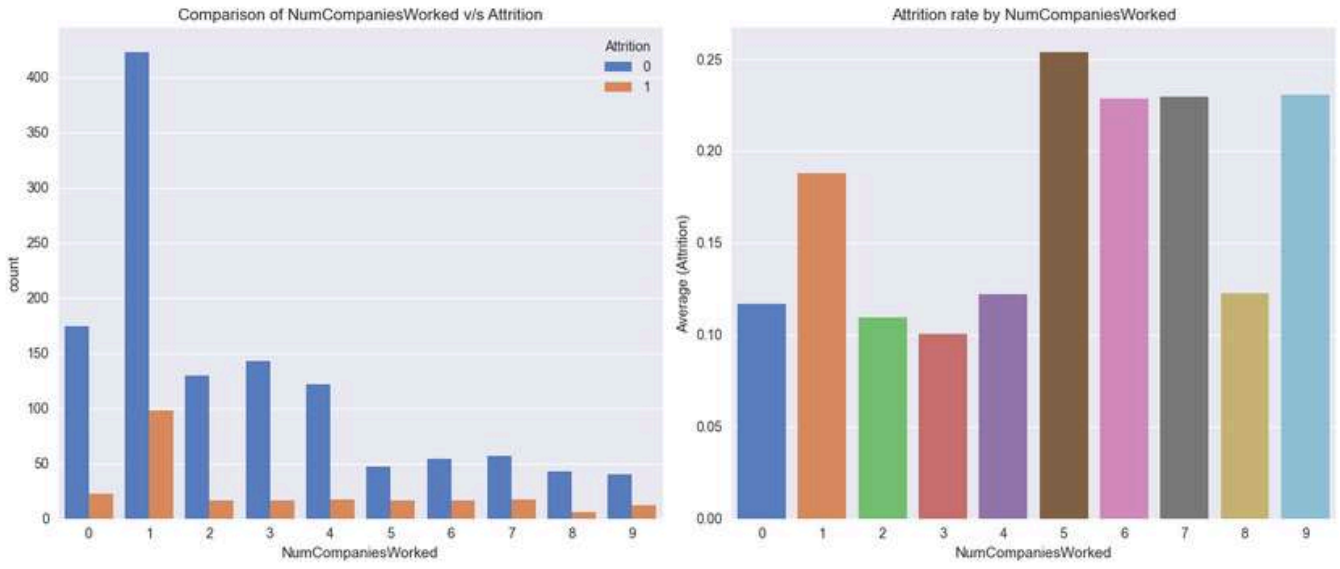
- There is a large number of Sales Executives (22% of the total) in the company, followed by Research Scientists (20%), then Laboratory Technicians (18%).
- Employees working in the Sales department, especially Sales Representatives are most likely to leave the company followed by Laboratory Technician, then Human Resources. Their attrition levels are 40%, 24%, and 23% respectively.

Attrition vs Marital Status:



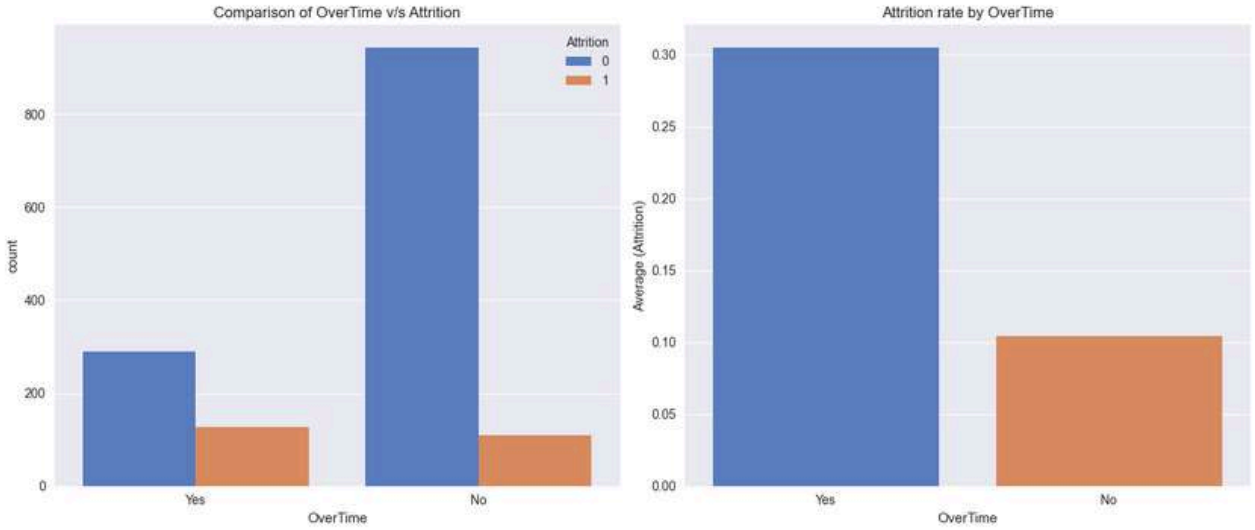
- Single Employees are more likely to leave the company with an attrition level of 25%.
- There is no significant difference in attrition levels between Married and Divorced employees.

Attrition vs Number of Companies Worked:



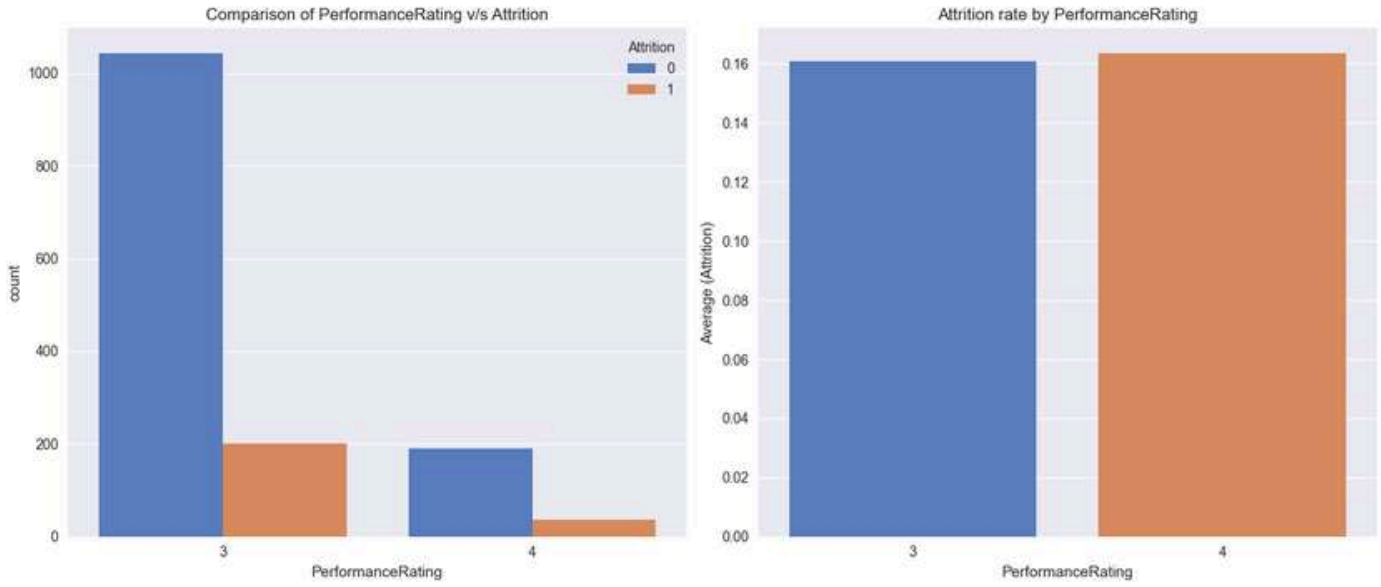
- The majority of employees have worked only for 1 company.
- No significant relationship between the attrition levels and number of companies worked.
- A significant amount of employees have left the company irrespective of the number of companies they have worked with.

Attrition vs Overtime:



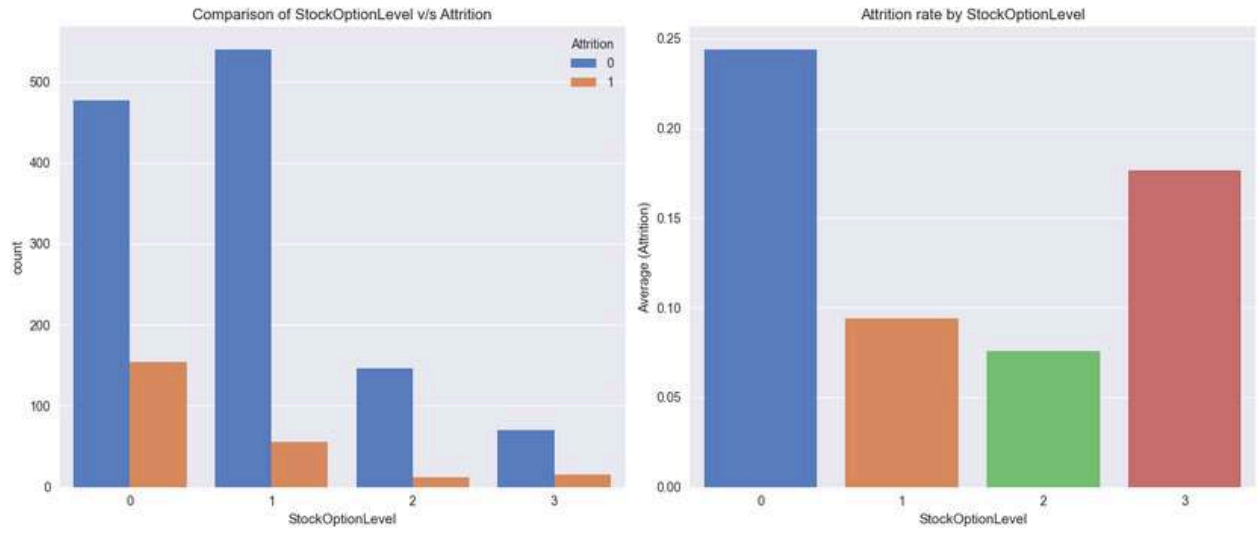
- Only 28% of total employees have worked overtime, remaining 72% of employees have no experience of overtime.
- Despite having no experience of overtime, employees of this category have also left the company with an attrition level of 10%.
- There must be some other reason for leaving the company apart from OverTime for these employees.

Attrition vs Performance Rating:



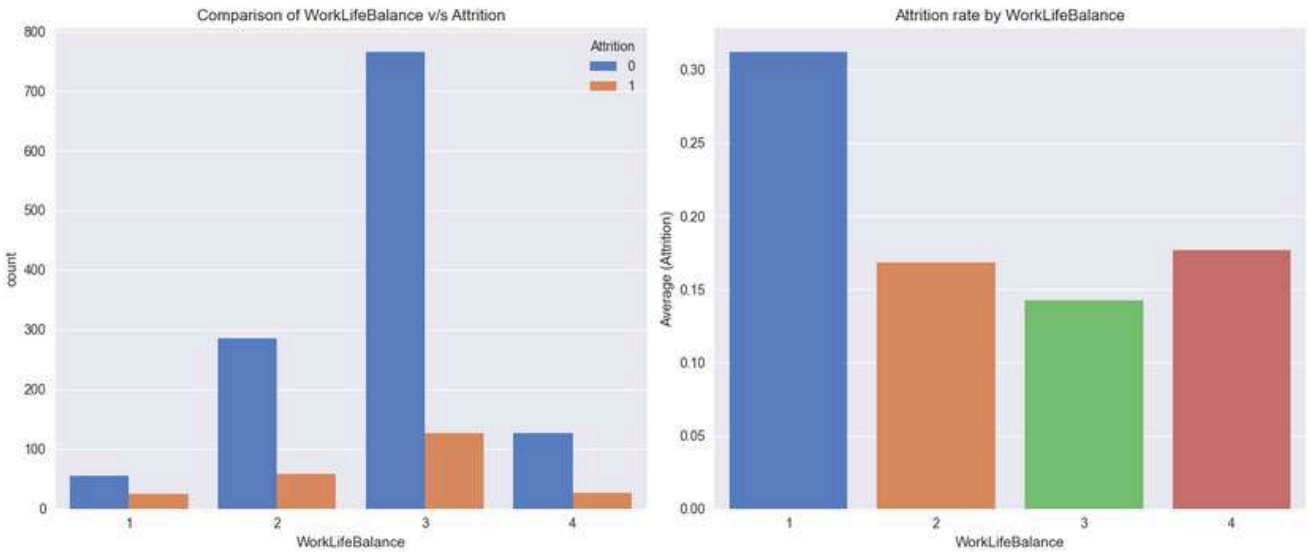
- All employees have either excellent(85%) or outstanding (15%) performance ratings.
- Whether it's excellent or outstanding, the employees of both categories have left the company with nearly similar attrition levels (16%).
- Performance Rating has no significant impact on Attrition.

Attrition vs Stock Option Level:



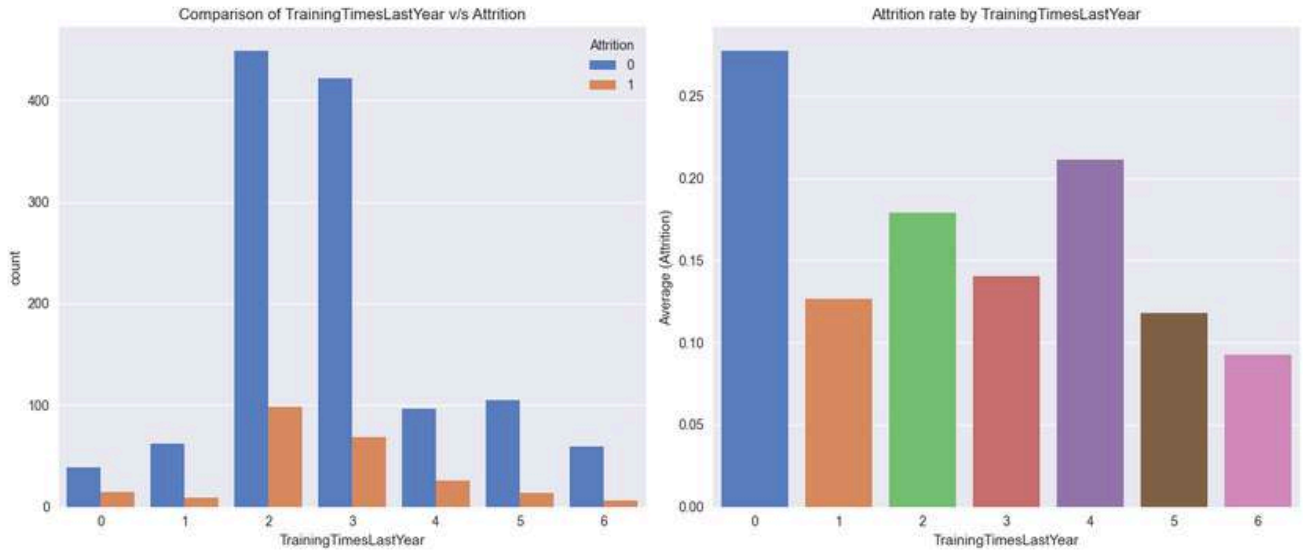
- The majority of Employees have 0 and 1 Stock Options Levels.
- Employees with lower stock option levels are more likely to leave the company with an attrition level of 24%.
- Even employees with the highest stock option level left the company with an attrition level of 18%.

Attrition vs Work Life Balance:



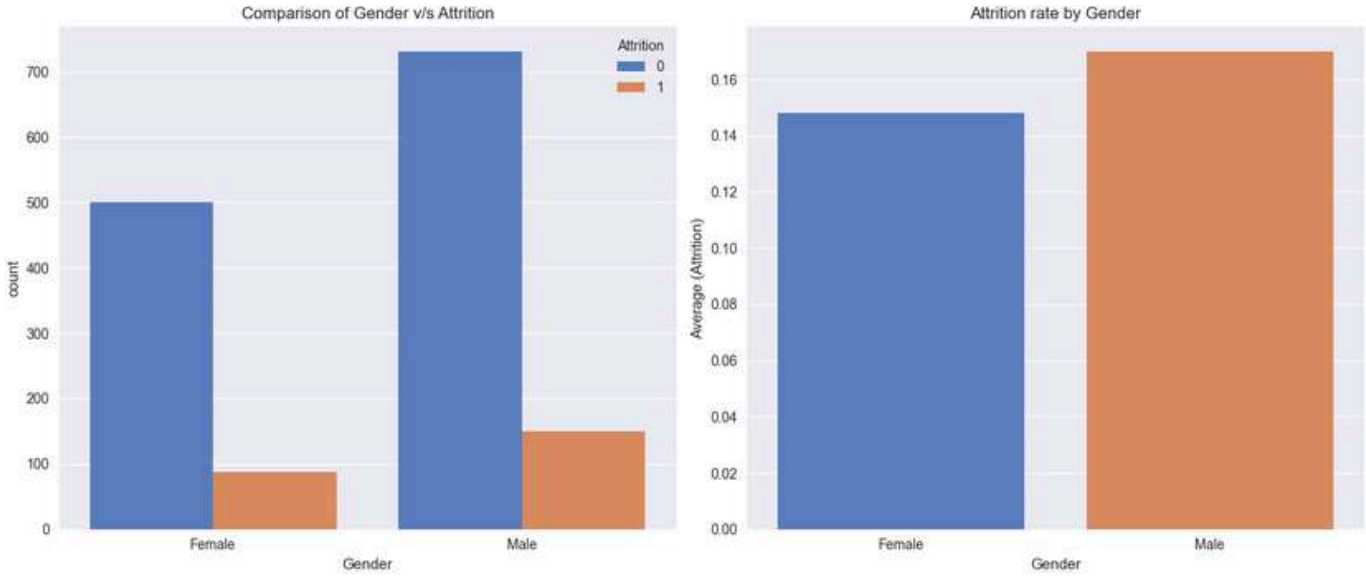
- Employees with Bad Work-Life Balance are more likely to leave the company with an attrition level above 30%.
- The majority of employees have a better work-life balance (around 60%).
- Employees with good, better, and best work-life balance also left the company but with a lower attrition rate (around 14-16%).

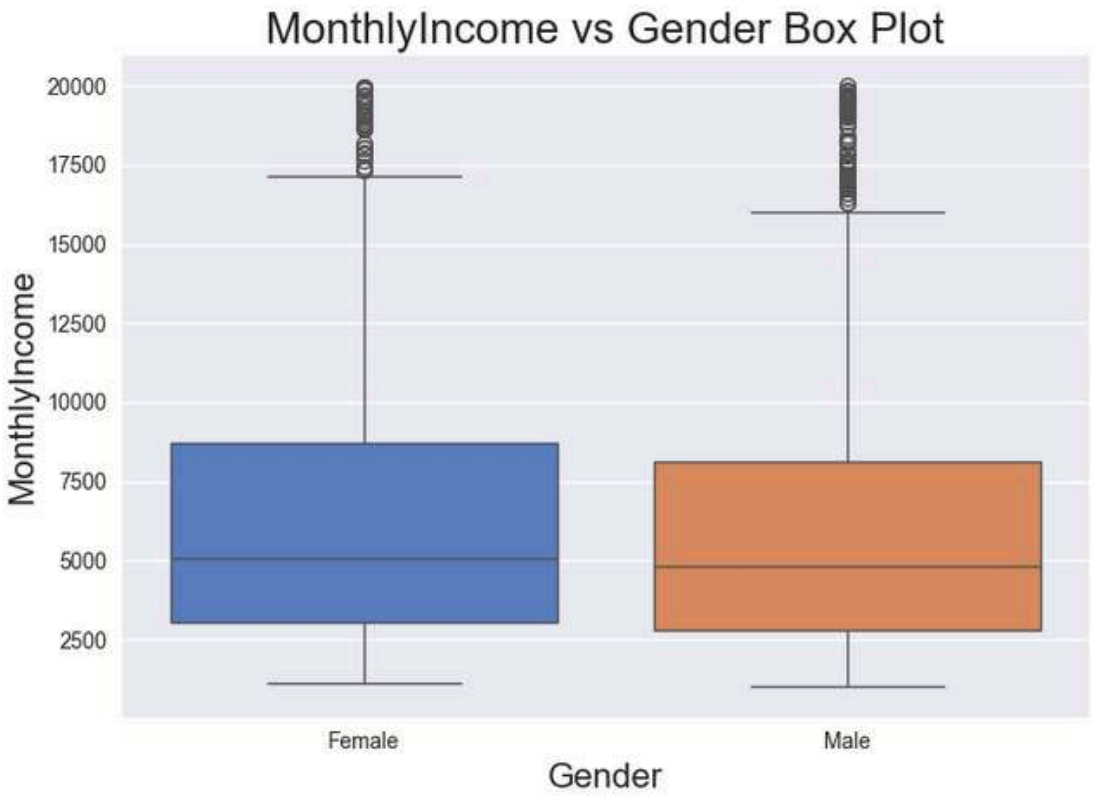
Attrition vs Training Times Last Year:



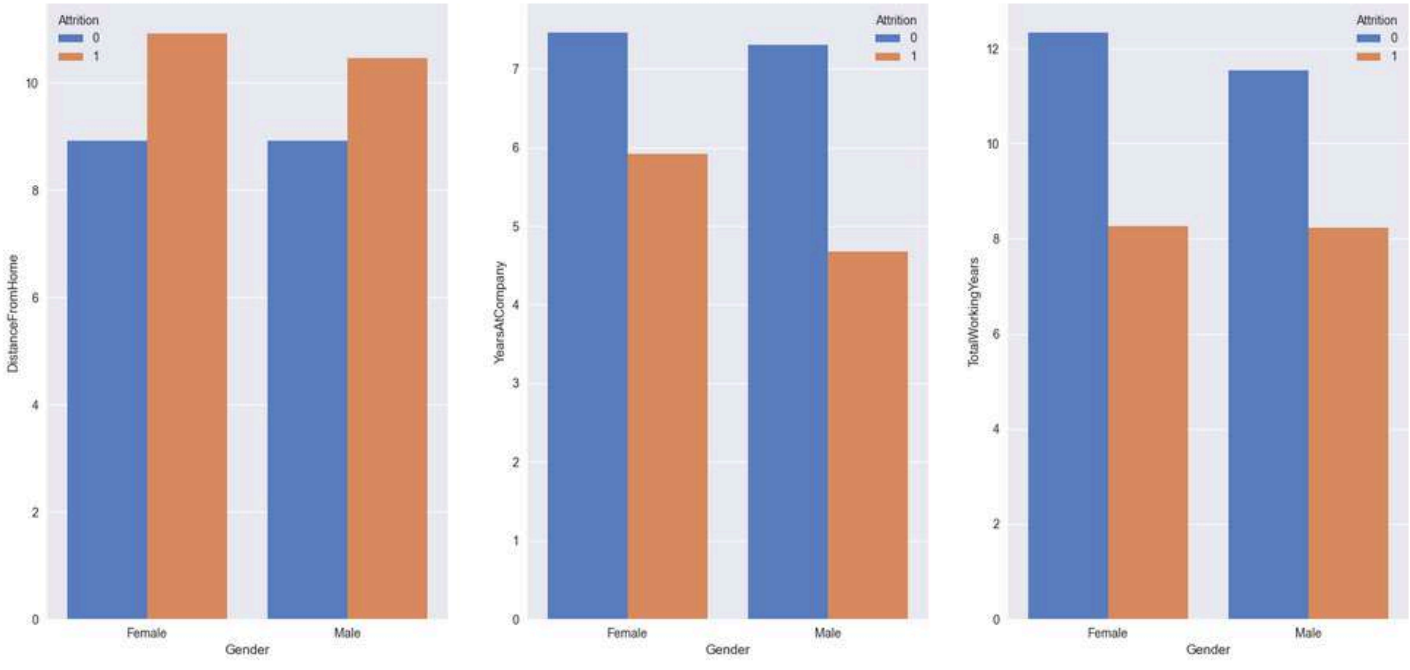
- Majority of employees have gone through training 2-3 times last year.
- Employees with no training last year are more prone to attrition.

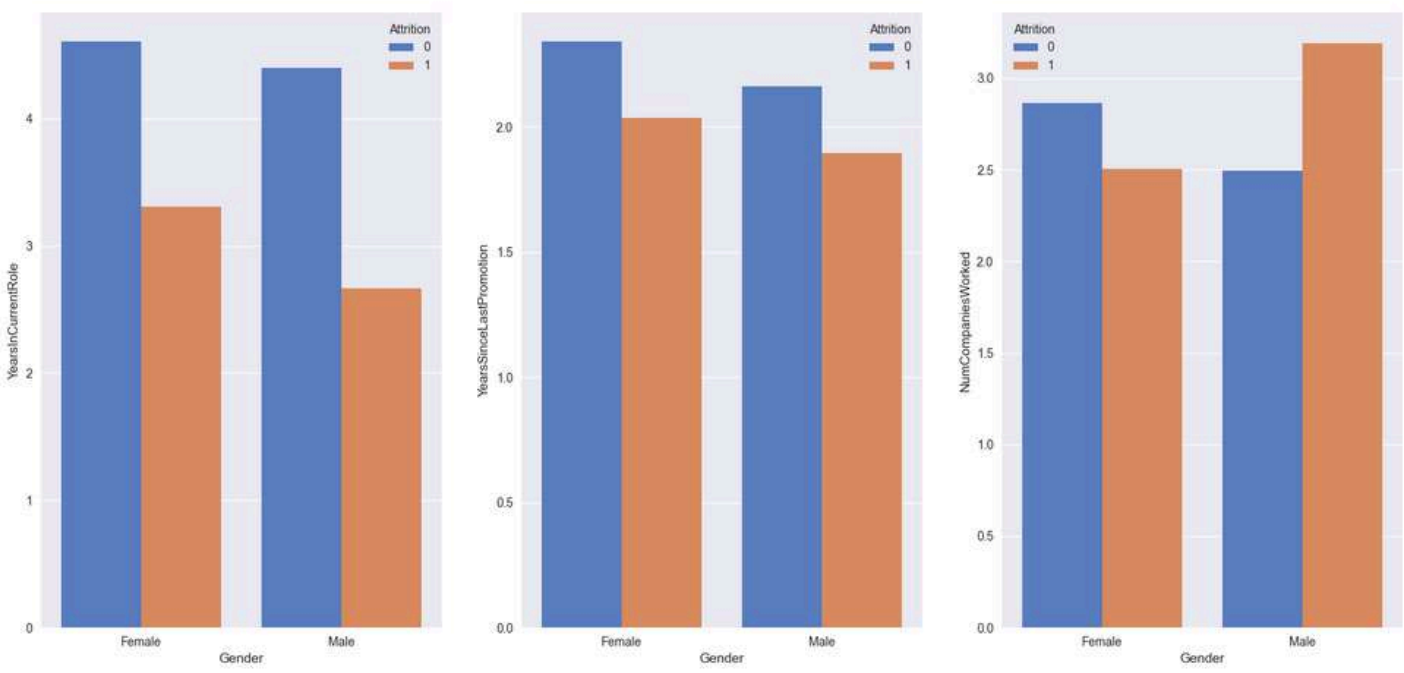
Attrition vs Gender:





- Monthly Income distribution for Males and Females is almost similar, so the attrition rate of Males and Females is almost the same around 15-16%.
- Gender is not a strong indicator of attrition.

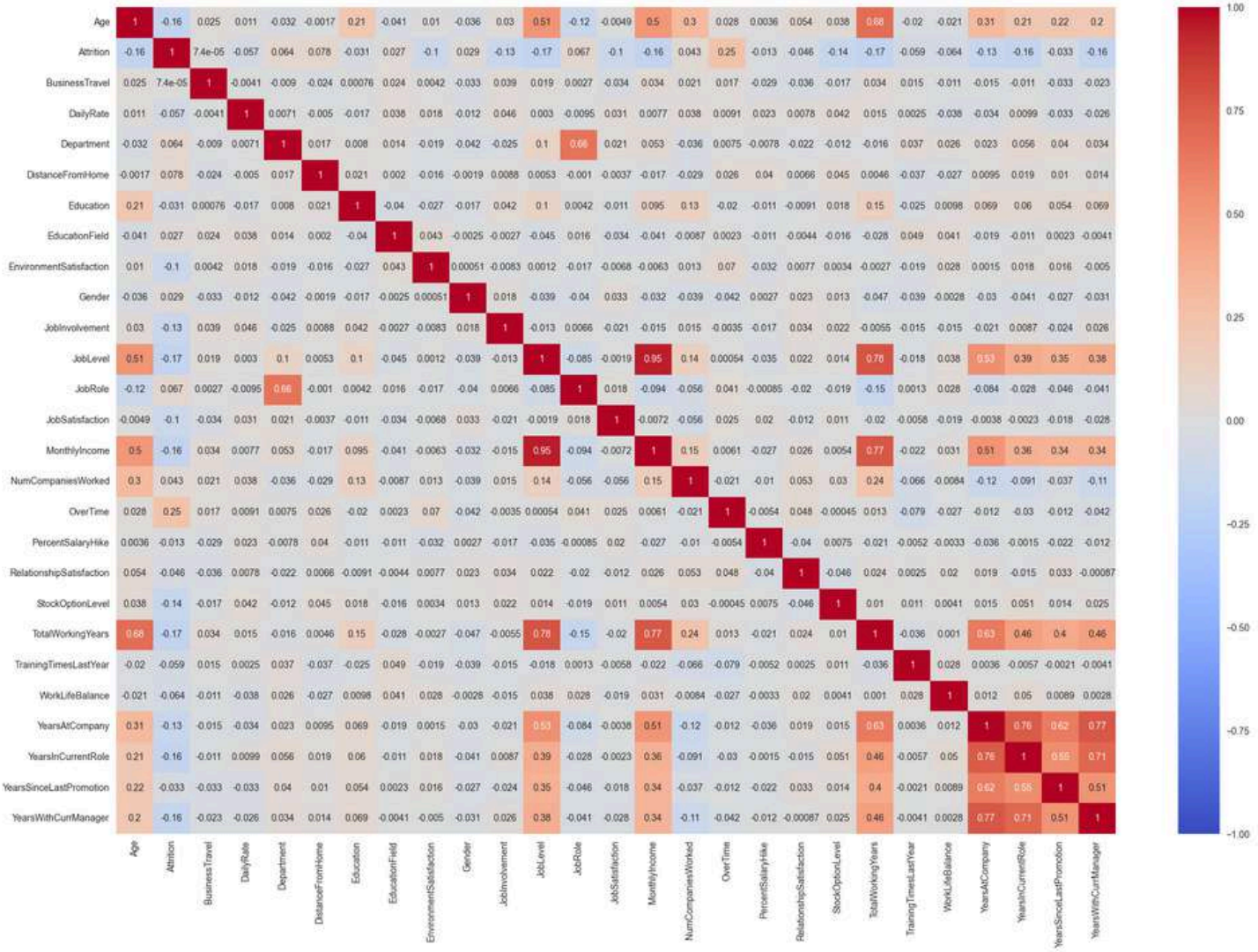




- Distance from home matters more to women employees than the male employees.
- Female employees are spending more years in one company compared to the men employees.
- Female employees spending more years in their current company are more prone to leave the company.
- Among men employees who worked for a large number of companies and men employees who worked for a relatively lesser number of companies, the former are more likely to leave the company than the latter.
- Among women employees who worked for a large number of companies and women employees who worked for a relatively lesser number of companies, the latter are more likely to leave the company than the former.
- Among women employees who worked for a large number of companies and women employees who worked for a relatively lesser number of companies, the latter are more likely to leave the company than the former.

- From the analysis, features that are most insignificant to attrition are: **HourlyRate**, **MonthlyRate**, **MaritalStatus**, **PerformanceRating**
- There are other variables also that are not significant to the attrition but these are extreme. So, we are going to drop them.
- Next, we need to see how every feature depends on one another through the Correlation Matrix. Especially the correlation of each relevant feature column with Attrition.

Correlation Matrix



Correlation Results:

- Monthly income is highly correlated with Job level.
- Job level is highly correlated with Total working years.
- Monthly income is highly correlated with Total working years.
- Age is also positively correlated with the Total working years.
- As expected Job Level is negatively correlated with Age.
- Job Role negatively correlated with Total working years.
- It seems like some employees with no experience of training last year are working overtime.
- Also, some employees, who are not satisfied with their jobs have worked for more companies than the employees who are satisfied.

Predictive Modelling

Machine Learning Models

I have used three different machine learning algorithms for creating the models and comparing the model's performances.

- Logistic Regression
- Random Forest
- Support Vector Machine

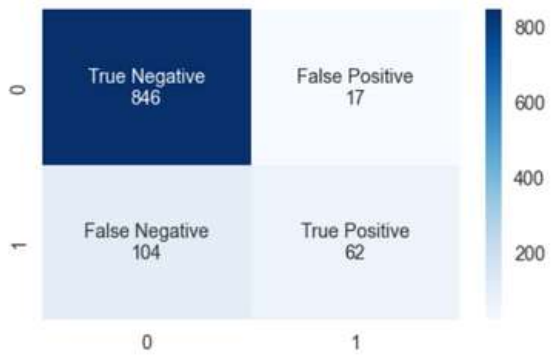
The major problem faced was, we had 1223 'No' and 237 'Yes' in Attrition, which is highly imbalanced data. So I used stratified sampling while splitting the data into training and test data based on a proportion of attrition in overall data.

- After reviewing the results of all three models, it's evident that the confusion matrix highlights a lot of misclassifications, indicating that a lot of instances were misclassified.
- This is mainly because the data is heavily imbalanced, with one class having more instances than the other.
- Logistic Regression, Random Forest, and SVM models tend to favor the majority class due to this data imbalance. As a result, features from the minority class may get overlooked, leading to higher misclassification rates for them.
- Accuracy measures might not reflect the actual model performance accurately in such cases.
- Instead, it's better to use metrics like precision, recall, and ROC curve analysis for a more precise evaluation.

Logistic Regression

Model Results and Evaluation:

TRAINING RESULTS:
==========
CONFUSION MATRIX:

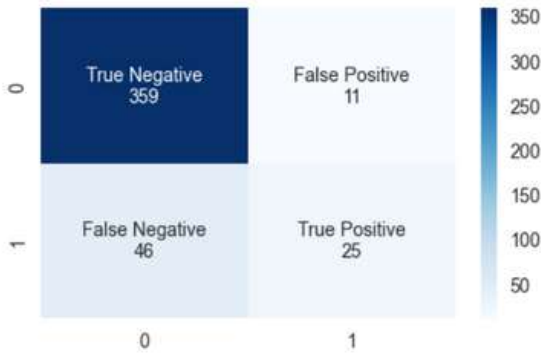


ACCURACY SCORE:
0.8824

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.890526	0.784810	0.88241	0.837668	0.873472
recall	0.980301	0.373494	0.88241	0.676898	0.882410
f1-score	0.933260	0.506122	0.88241	0.719691	0.864353
support	863.000000	166.000000	0.88241	1029.000000	1029.000000

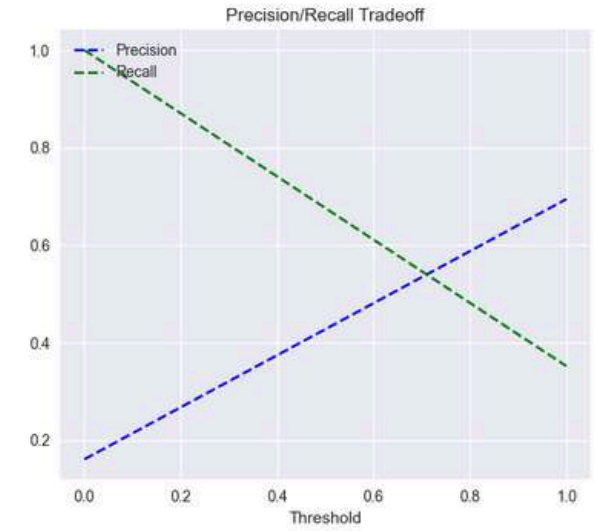
TESTING RESULTS:
==========
CONFUSION MATRIX:

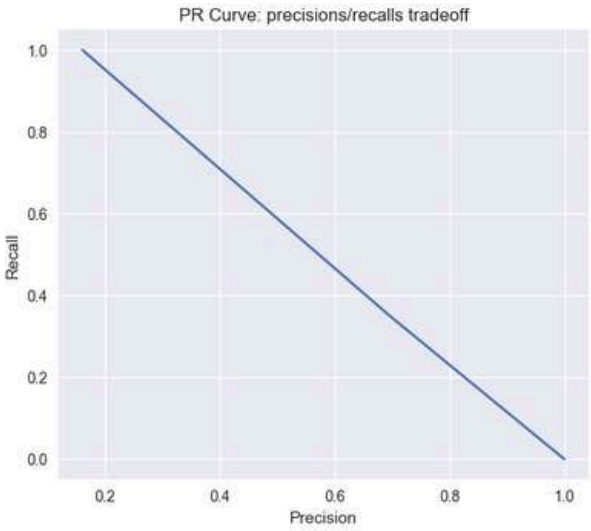


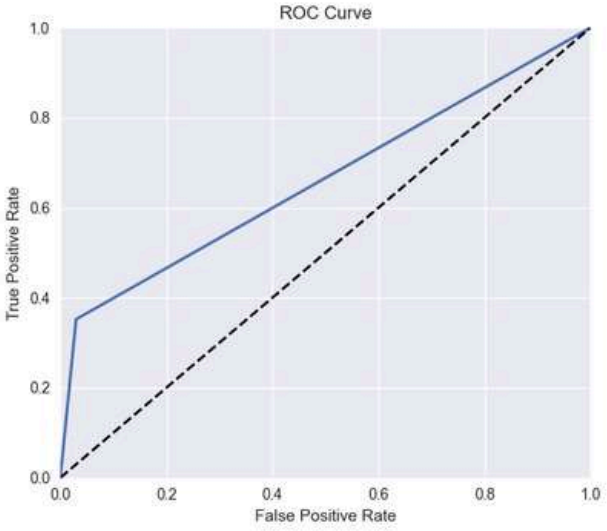
ACCURACY SCORE:
0.8707

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.886420	0.694444	0.870748	0.790432	0.855512
recall	0.970270	0.352113	0.870748	0.661191	0.870748
f1-score	0.926452	0.467290	0.870748	0.696871	0.852528
support	370.000000	71.000000	0.870748	441.000000	441.000000







Random Forest

Random Forest is an ensemble learning algorithm. It constructs multiple decision trees during training and merges them to get a more accurate and stable prediction. Random Forest avoids overfitting by randomly selecting subsets of features and data samples for each tree.
Pruning in Random Forest is implicitly done by limiting the tree depth or the number of trees in the forest through hyperparameters like max_depth or n_estimators.

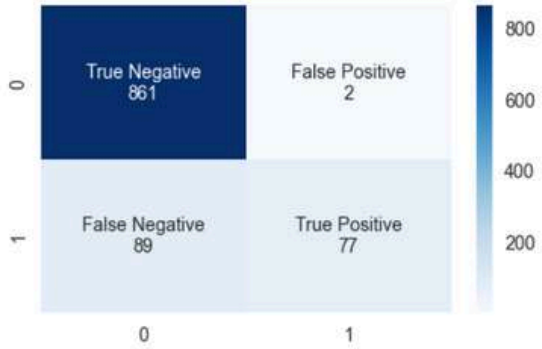
Here, I have used the K fold cross-validation technique to assess how the results of a model will generalize to an independent test data set. I used k=5 i.e. 5-fold cross-validation.

By using grid search best parameters were found to be -

```
{'bootstrap': True,
 'max_depth': 15,
 'max_features': 'sqrt',
 'min_samples_leaf': 4,
 'min_samples_split': 2,
 'n_estimators': 100}
```

Model Results and Evaluation:

TRAINING RESULTS:
==========
CONFUSION MATRIX:

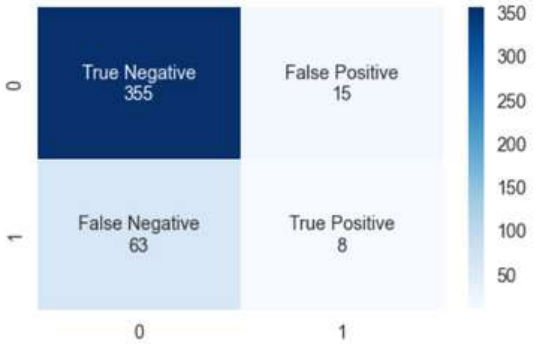


ACCURACY SCORE:
0.9116

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.906316	0.974684	0.911565	0.940500	0.917345
recall	0.997683	0.463855	0.911565	0.730769	0.911565
f1-score	0.949807	0.628571	0.911565	0.789189	0.897985
support	863.000000	166.000000	0.911565	1029.000000	1029.000000

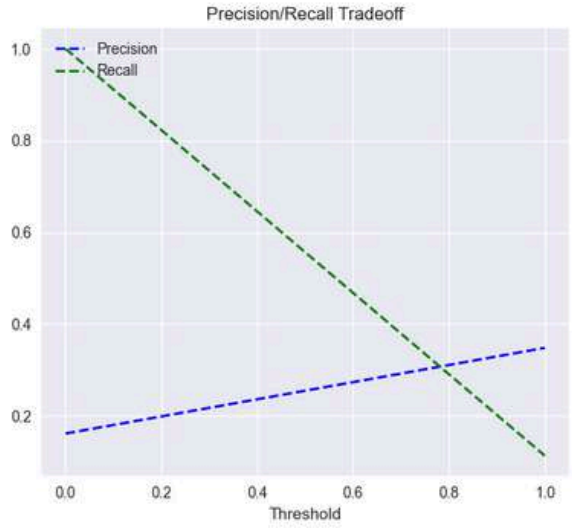
TESTING RESULTS:
==========
CONFUSION MATRIX:

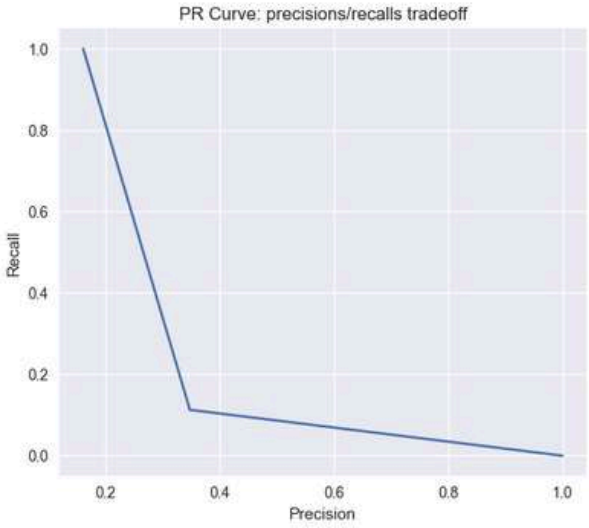


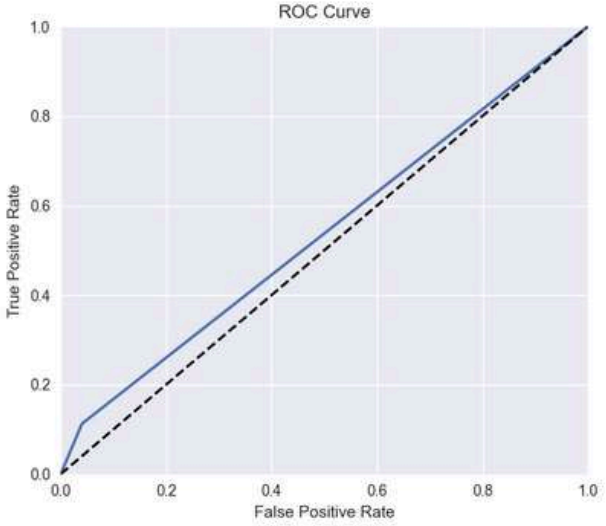
ACCURACY SCORE:
0.8231

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.849282	0.347826	0.823129	0.598554	0.768549
recall	0.959459	0.112676	0.823129	0.536068	0.823129
f1-score	0.901015	0.170213	0.823129	0.535614	0.783358
support	370.000000	71.000000	0.823129	441.000000	441.000000







- The ROC curve is a simple plot that shows the trade-off between the true positive rate and the false positive rate of a classifier for various choices of the probability threshold.
- From the ROC Curve, we have a choice to make depending on the value we place on true positive and tolerance for false positive rate. If we wish to find more people who are leaving, we could increase the true positive rate by adjusting the probability cut-off for classification. However, doing so would also increase the false positive rate. we need to find the optimum value of the cut-off for classification.

Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm for classification and regression. It finds the best hyperplane to separate classes, maximizing the margin. Regularization parameters like C control overfitting, while pruning hyperparameters like kernel and gamma fine-tune model complexity. SVM handles non-linear boundaries using the kernel trick.

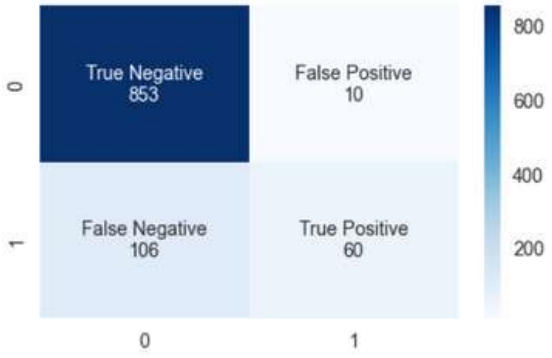
Here also, I have used the K-fold cross-validation technique. I used k=3 i.e. 3-fold cross-validation.

By using grid search best parameters were found to be -

```
{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
```

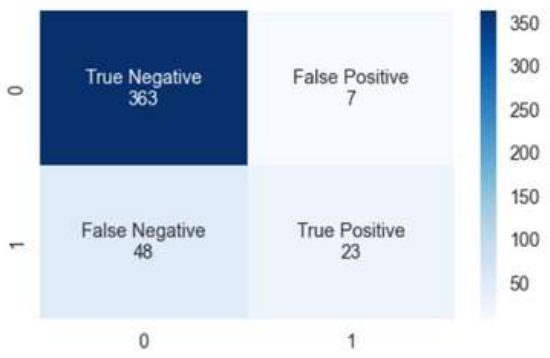
Model Results and Evaluation:

```
TRAINING RESULTS:
==============================
CONFUSION MATRIX:
```

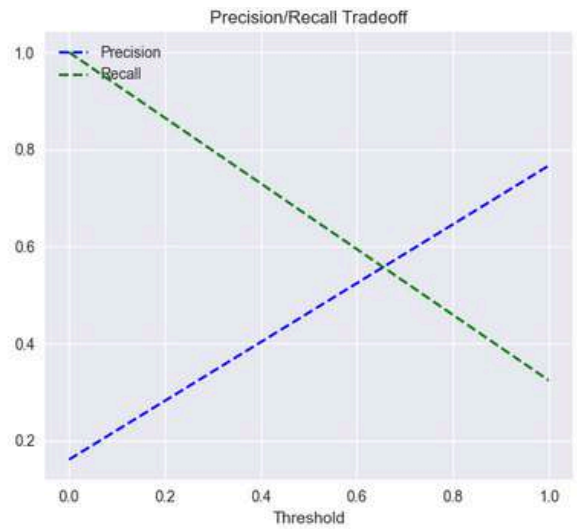


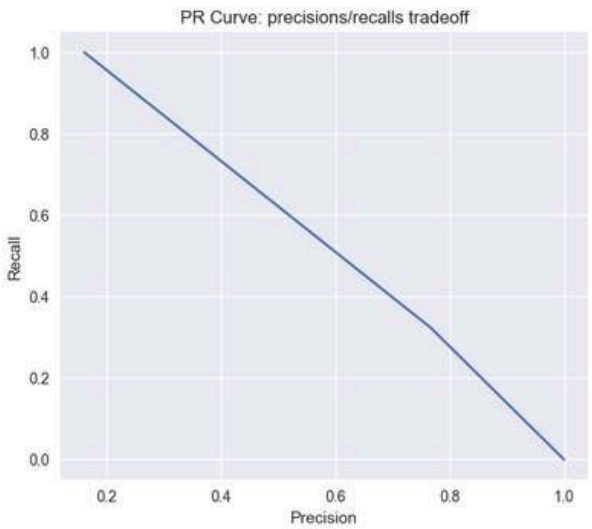

```
ACCURACY SCORE:
0.8873

CLASSIFICATION REPORT:
                   0           1    accuracy    macro avg  weighted avg
precision   0.889468    0.857143    0.887269     0.873306      0.884253
recall      0.988413    0.361446    0.887269     0.674929      0.887269
f1-score    0.936334    0.508475    0.887269     0.722404      0.867311
support   863.000000  166.000000    0.887269  1029.000000   1029.000000
```

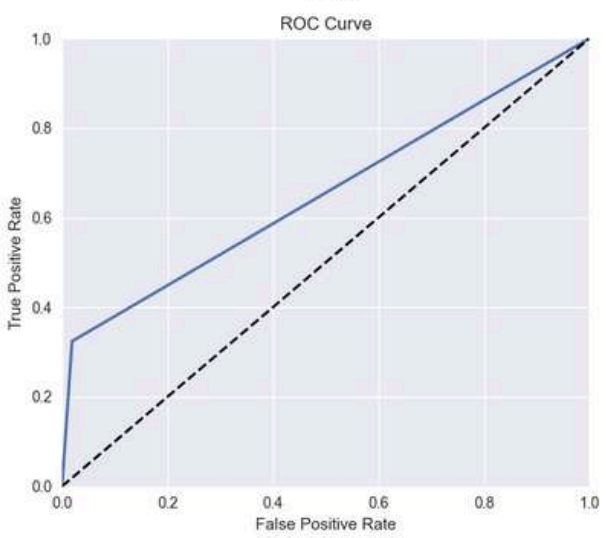
```
TESTING RESULTS:
==============================
CONFUSION MATRIX:
```




```
ACCURACY SCORE:
0.8753

CLASSIFICATION REPORT:
                   0           1    accuracy    macro avg  weighted avg
precision   0.883212    0.766667    0.875283     0.824939      0.864448
recall      0.981081    0.323944    0.875283     0.652512      0.875283
f1-score    0.929577    0.455446    0.875283     0.692512      0.853243
support   370.000000   71.000000    0.875283   441.000000    441.000000
```

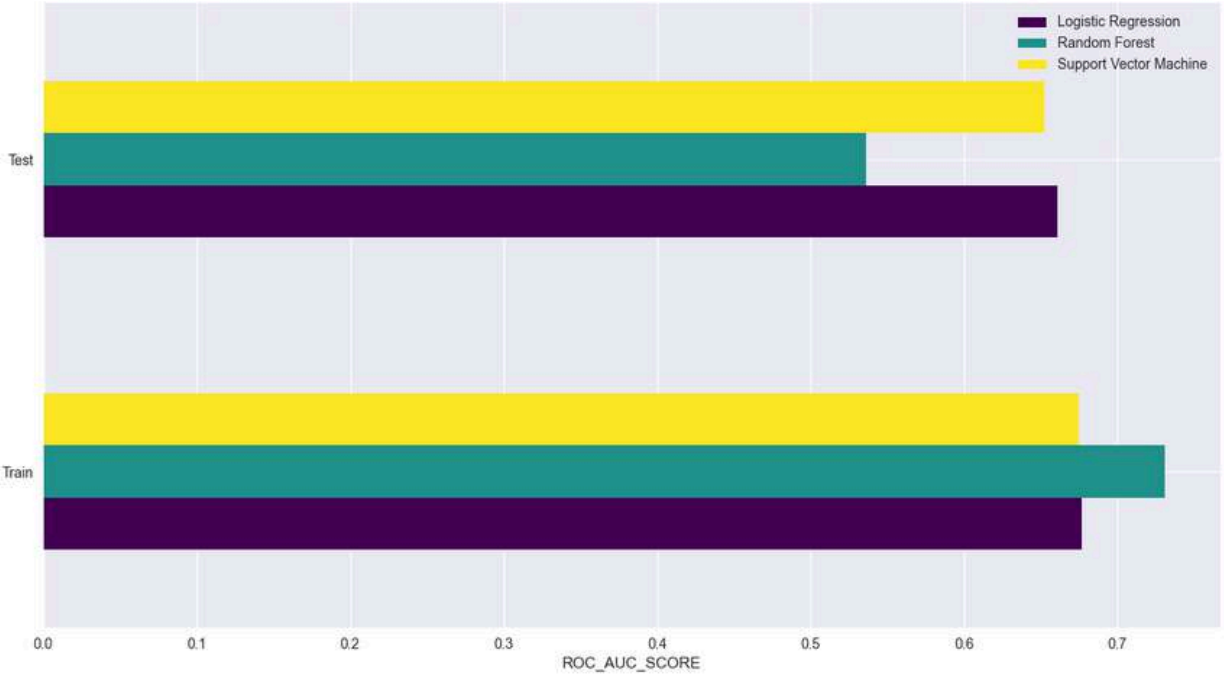






Comparison of Model's Performances

```
LOGISTIC REGRESSION:

Train ROC_AUC_SCORE: 0.6768976252635106
Test ROC_AUC_SCORE: 0.6611914731633042

================================================

RANDOM FOREST:

Train ROC_AUC_SCORE: 0.7307689622918091
Test ROC_AUC_SCORE: 0.5360677578987438

================================================

SUPPORT VECTOR MACHINE:

Train ROC_AUC_SCORE: 0.6749291488084436
Test ROC_AUC_SCORE: 0.652512371526456
```



Suggested Actions to reduce employee attrition:

Making work better:

- **I**nstead of worrying about every employee who might leave, HR should focus on improving the workplace environment.
- This could involve offering options like allowing employees to work from home, giving them flexibility in their working hours, or ensuring that their workspace is comfortable and ergonomic.
- By providing these choices, employees are more likely to feel happier and find it easier to balance their work and personal life.

Paying fairly and giving perks:

- To retain valuable employees, it's essential to offer fair compensation and additional benefits.
- This means ensuring that employees are paid competitively for their work and providing perks such as flexible working hours or travel discounts.
- By offering these incentives, employees are more likely to feel valued and motivated to stay with the company.

Keeping employees interested:

- Another important aspect of retaining employees is keeping them engaged and interested in their work.
- This can be achieved by providing opportunities for employees to learn new skills and develop professionally.
- By investing in their growth and development, employees are more likely to feel challenged and fulfilled in their roles, reducing the risk of them becoming bored or dissatisfied and considering leaving the company.