# **DATA INTERPRETATION**

## **DOMAIN ANALYSIS:**

**1) EmpNumber:** Unique identifier for each employee in the dataset.

**2) Age:** Age of the employee, providing insight into workforce demographics and potential correlations with attrition.

**3) Gender:** Gender of the employee, which may impact workplace dynamics and attrition patterns.

**4) EducationBackground:** The educational background of the employee, influencing skillset and career trajectory.

**5) MaritalStatus:** Marital status of the employee, potentially affecting work-life balance and job satisfaction.

**6) EmpDepartment:** Department in which the employee works, indicating job role and organizational structure.

**7) EmpJobRole:** Specific job role of the employee within their department, reflecting responsibilities and career path.

**8) BusinessTravelFrequency:** Frequency of business travel for the employee, impacting lifestyle and job satisfaction.

**9) DistanceFromHome:** Distance of employee's residence from the workplace, influencing commuting stress and retention.

**10) EmpEducationLevel:** Level of education attained by the employee, reflecting qualifications and potential for advancement.

**11) EmpEnvironmentSatisfaction:** Employee satisfaction with the work environment, affecting morale and turnover.

**12) EmpHourlyRate:** Hourly wage of the employee, a factor in compensation satisfaction and retention.

**13) EmpJobInvolvement:** Level of involvement and engagement in the job role, affecting performance and attrition risk.

**14) EmpJobLevel:** Level of hierarchy within the organization, indicating seniority and career progression.

**15) EmpJobSatisfaction:** Satisfaction level with the job role, impacting employee morale and retention.

**16) NumCompaniesWorked:** Number of companies the employee has previously worked for, indicating job stability and turnover risk.

**17) OverTime:** Whether the employee works overtime, influencing work-life balance and burnout.

**18) EmpLastSalaryHikePercent:** Percentage of the employee's last salary hike, affecting compensation satisfaction and retention.

**19) EmpRelationshipSatisfaction:** Satisfaction with relationships at work, influencing job satisfaction and likelihood of turnover.

**20) TotalWorkExperienceInYears:** Total work experience of the employee, influencing skill level and career trajectory.

**21) TrainingTimesLastYear:** Number of training sessions attended by the employee last year, indicating investment in skill development and career growth.

**22) EmpWorkLifeBalance:** Employee's perceived balance between work and personal life, affecting job satisfaction and retention.

**23) ExperienceYearsAtThisCompany:** Years of experience at the current company, indicating loyalty and potential for promotion.

**24) ExperienceYearsInCurrentRole:** Years of experience in the current job role, influencing expertise and potential for advancement.

**25) YearsSinceLastPromotion:** Time since the employee's last promotion, impacting career progression and job satisfaction.

**26) YearsWithCurrManager**: Years of tenure with the current manager, affecting job satisfaction and retention.

**27) Attrition:**   indicates whether the employee has left the company or not.

28) PerformanceRating:  Target variable for the given problem.  this is the performance rating assigned to the employee, influencing career development and potential for retention.

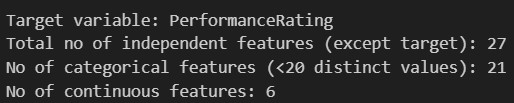
# **SUMMARY OF IMPORTANT ASPECTS OF THE MODEL**

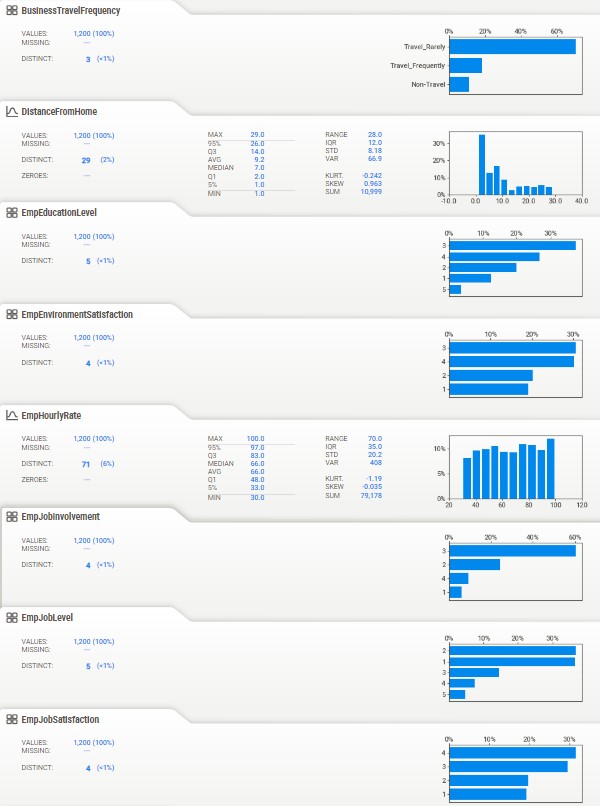
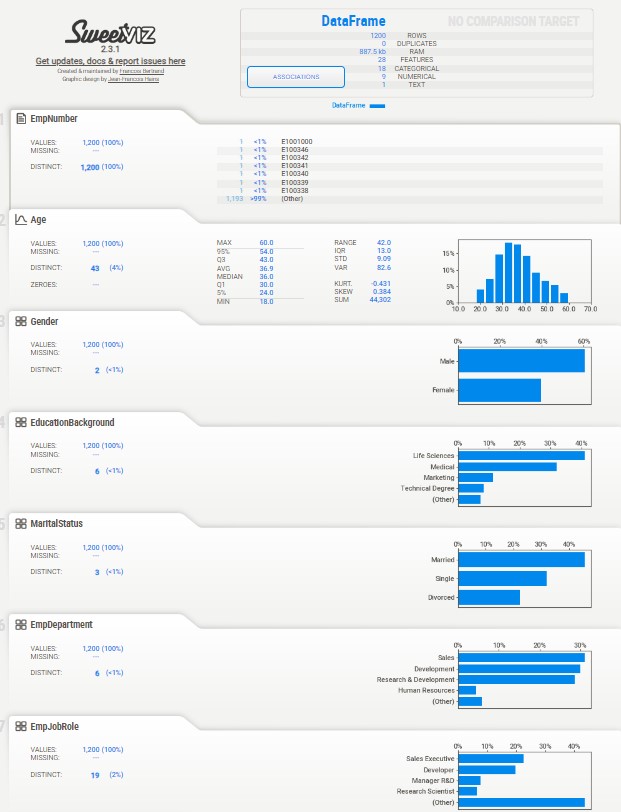
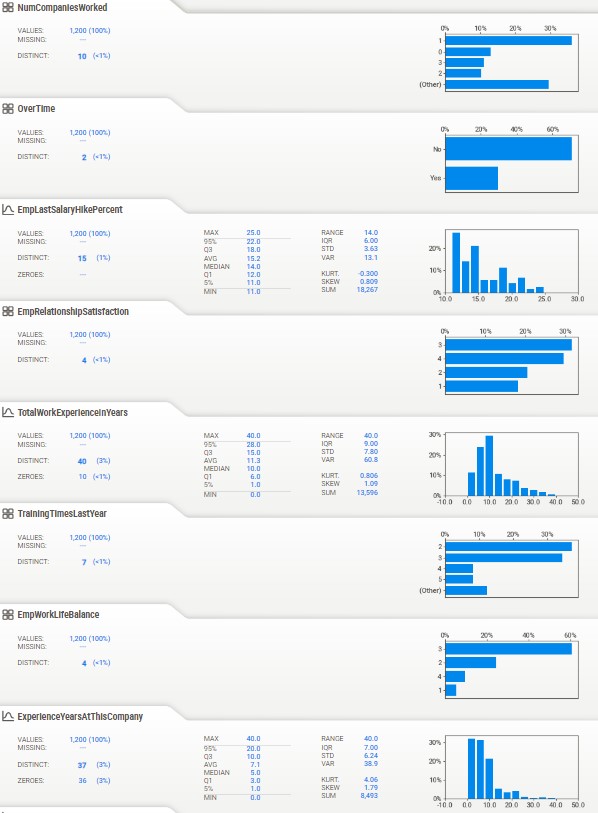
## **DATA ANALYSIS**

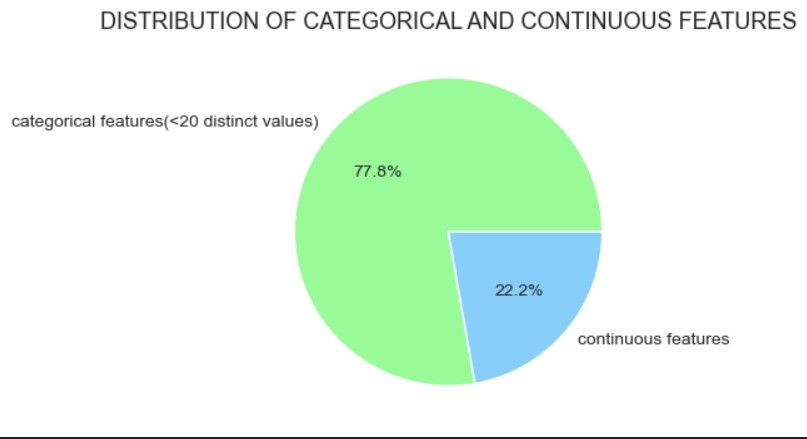
### **VISUALIZATION**

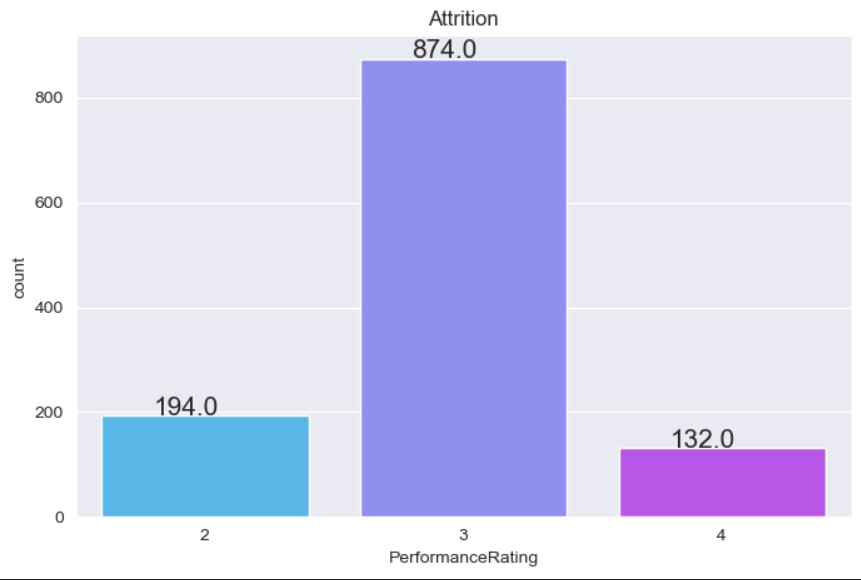
Used visualizations such as **histograms, count plots, scatter plots, pie charts, box plots, line plots, cat plots, pair plot and heatmaps** to effectively communicate analysis findings. Visual representations of data can make complex patterns and relationships more accessible and understandable.

#### **UNIVARIATE ANALYSIS:**

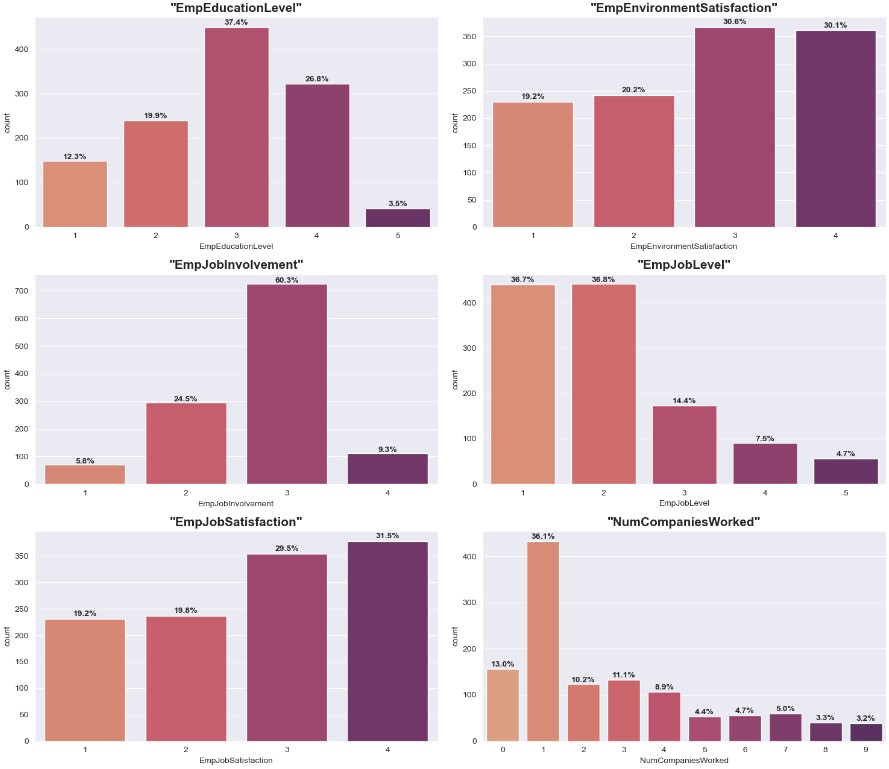


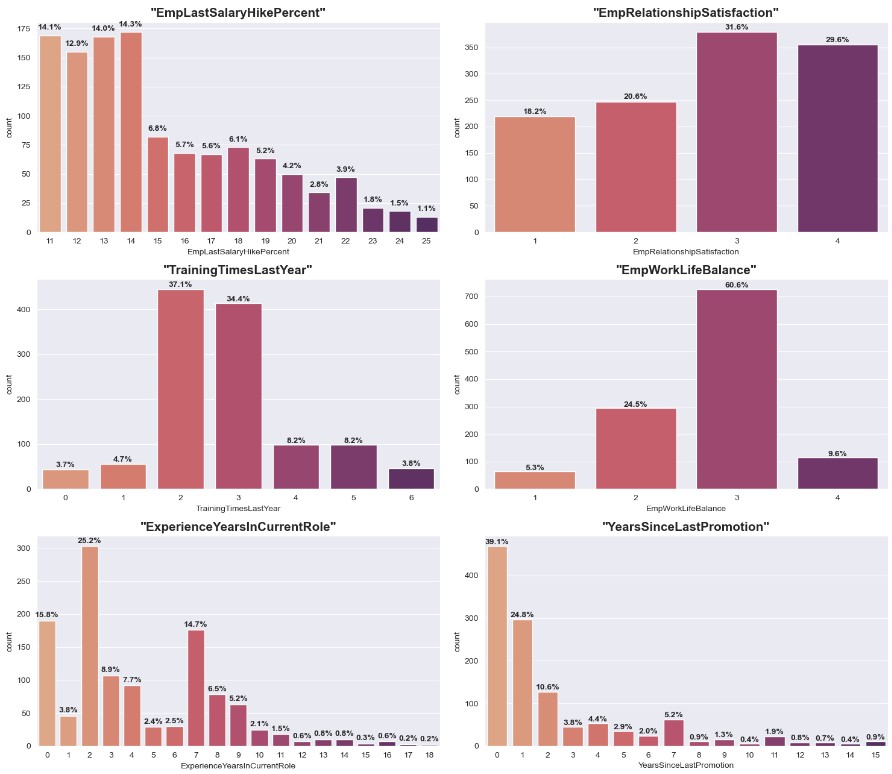






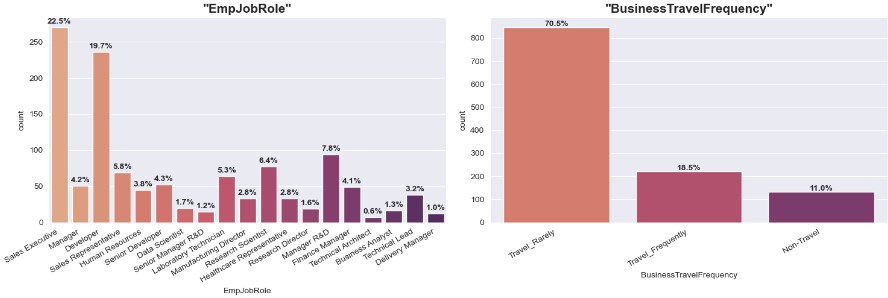
**Discrete features:**

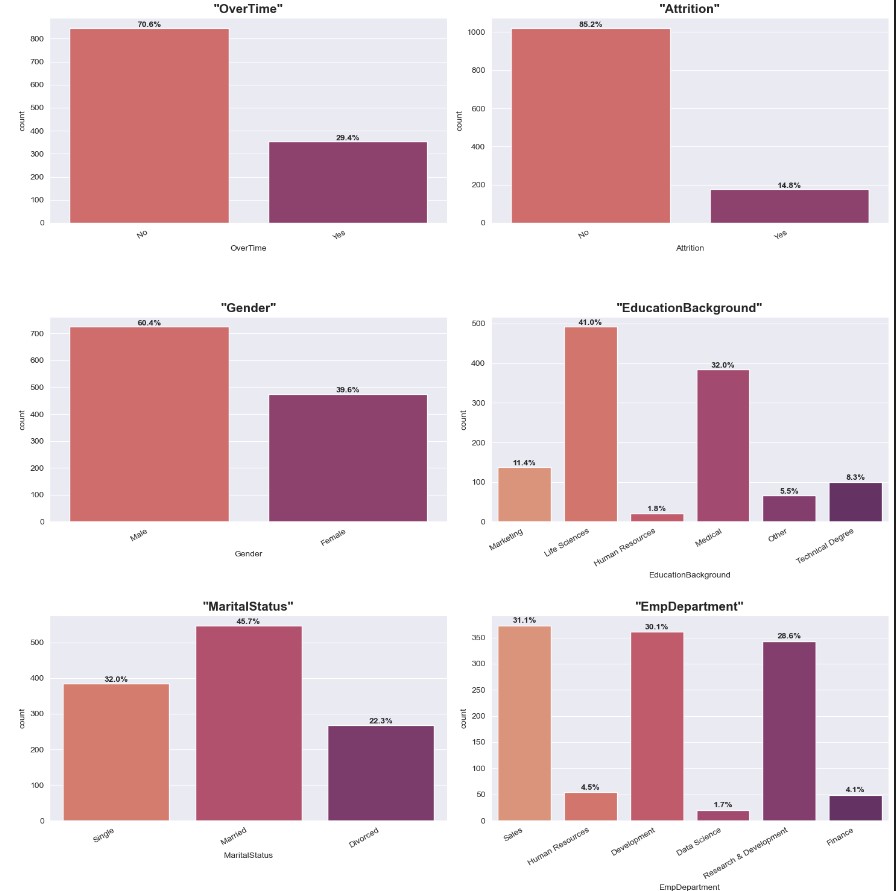




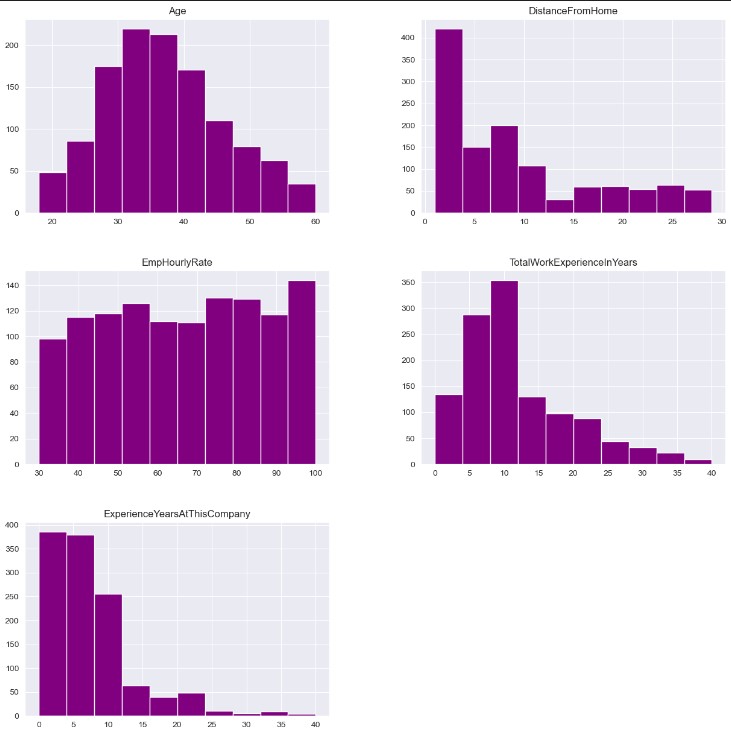


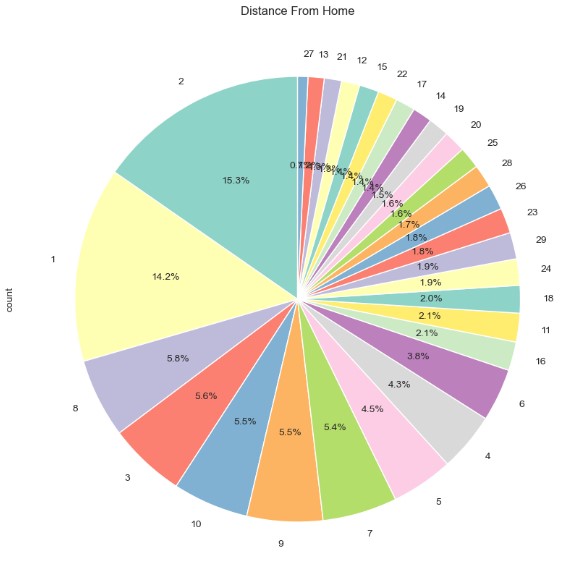
**Binary and nominal features:**





**Continuous features:**



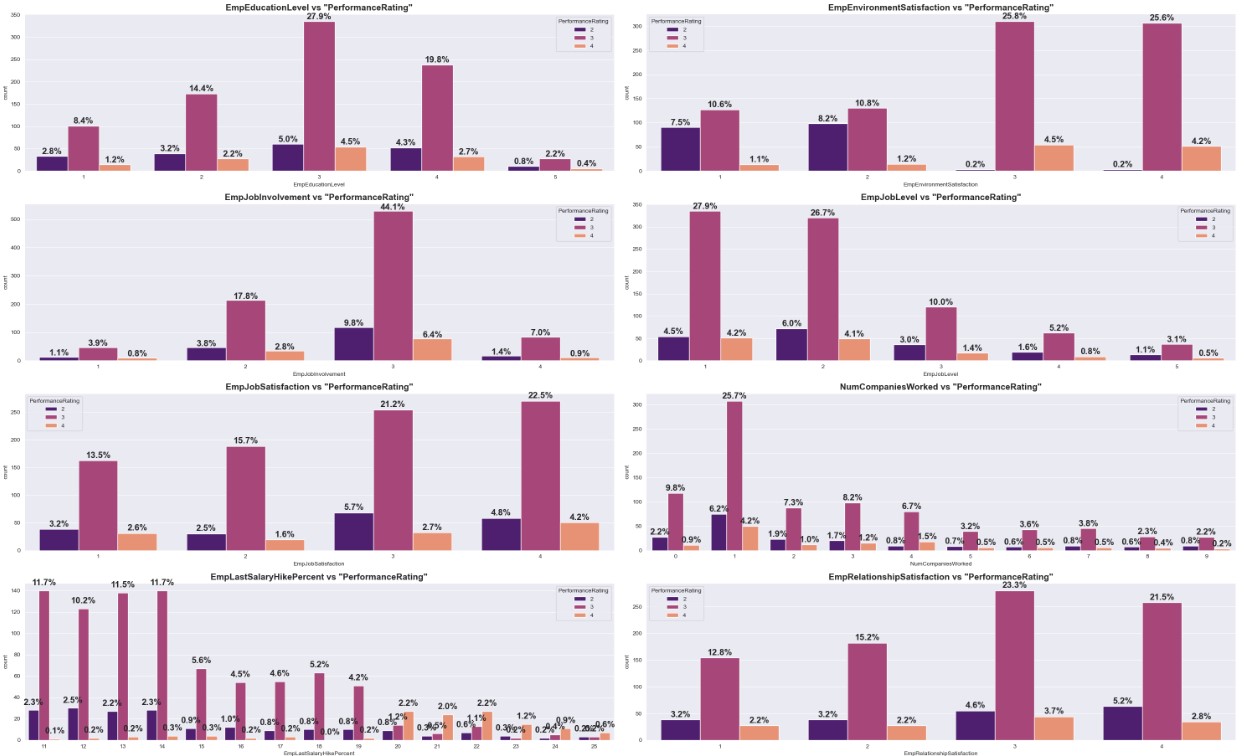


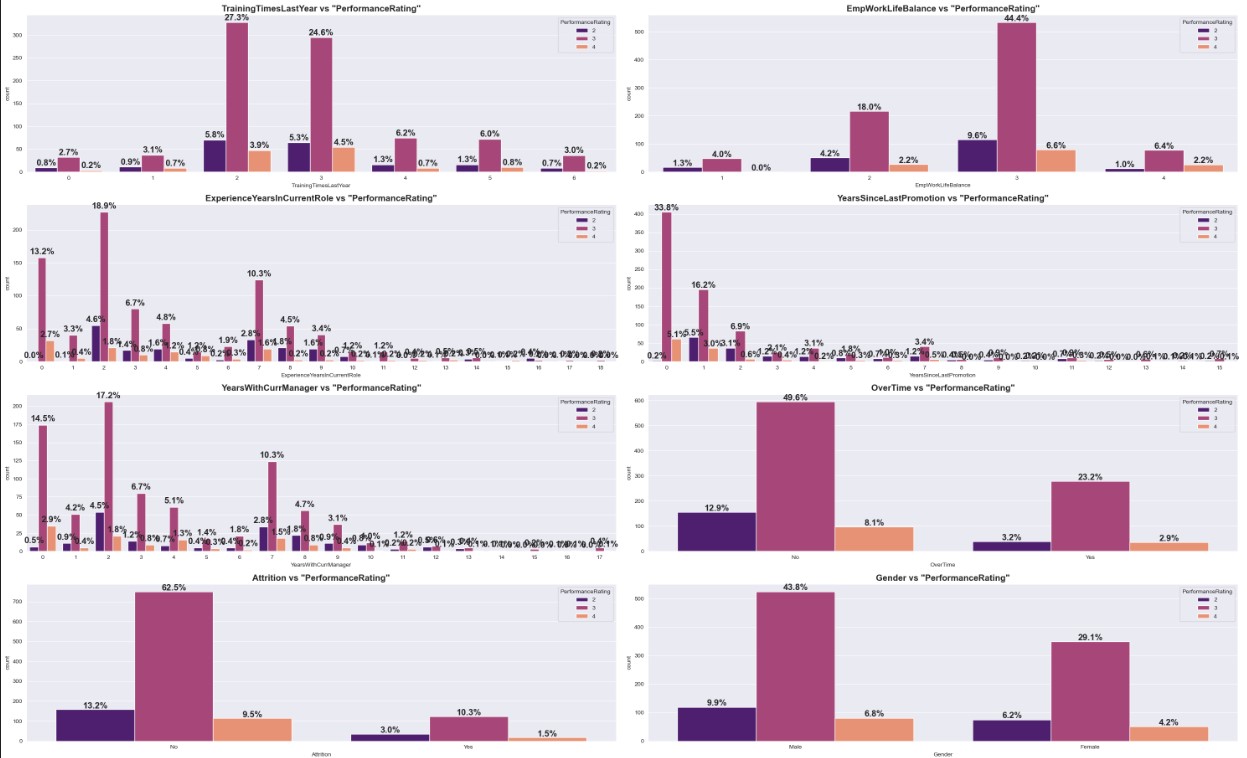
**Insights from univariate analysis:**

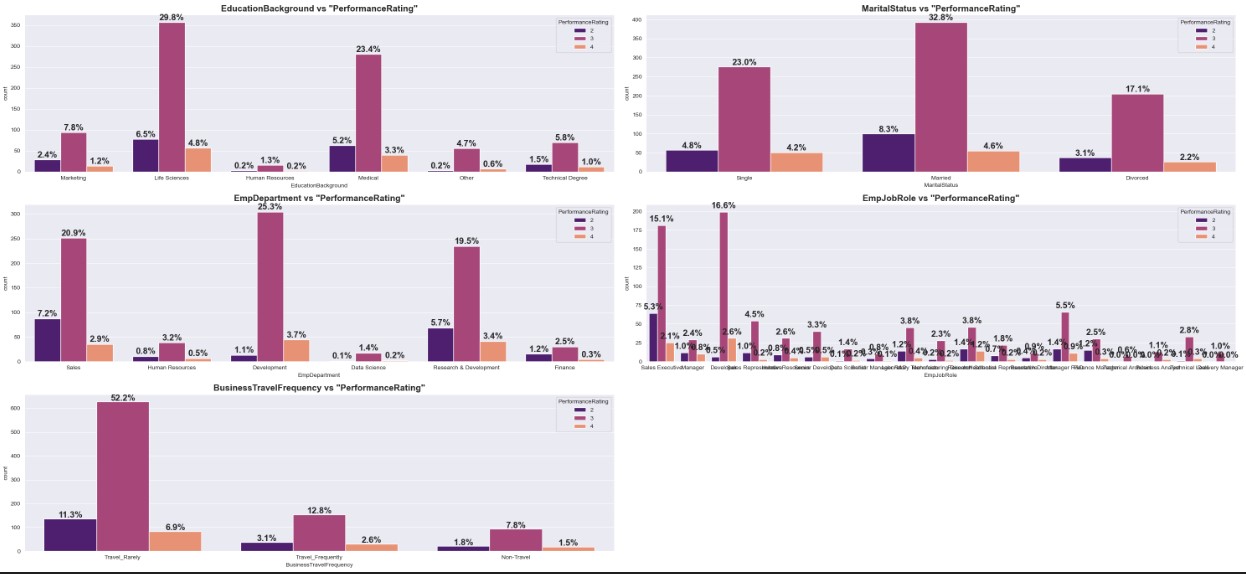
* Employees between the age group 30-35 are the majority in number.
* 70% of the people travel rarely, 20% travel frequently, rest do not travel.
* Majority of the employees belong to research and development.
* Almost 35% of the people are nearer to the office i.e. the distance from their home is lesser than or equal to 10.
* More than 60% of the people have educational qualification of 2 and 4.
* Majority (40%) of the people are from life science field and 30% are from medical field.
* 60% of the people are almost satisfied with environment condition of the office with more than 3 and 4 ratings.
* Gender count: 60% male, 40% female.
* Almost 40% of the people have partial involvement in job and 20% have good involvement.
* More than 45% employees seem to be satisfied with their job.
* 50% of the people are married, 30% single and the rest are divorced.
* 40% of the employee have work experience of less than 10 years.
* 15% of the people have worked for less than 1 company which implies that they are freshers.
* 30% of the people have worked for more than 5 companies.
* 80% of the people have average work rating.

#### **BIVARIATE ANALYSIS:**

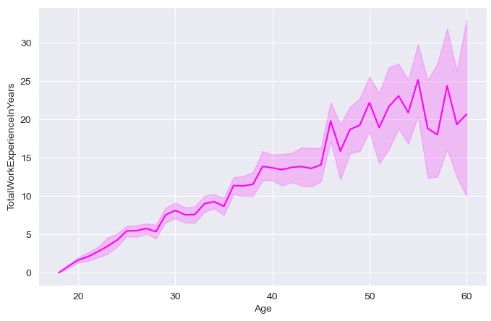
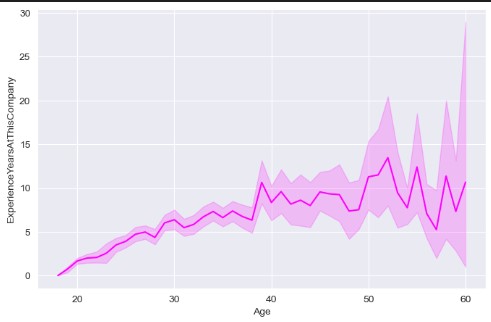
**Categorical features Vs target variable:**

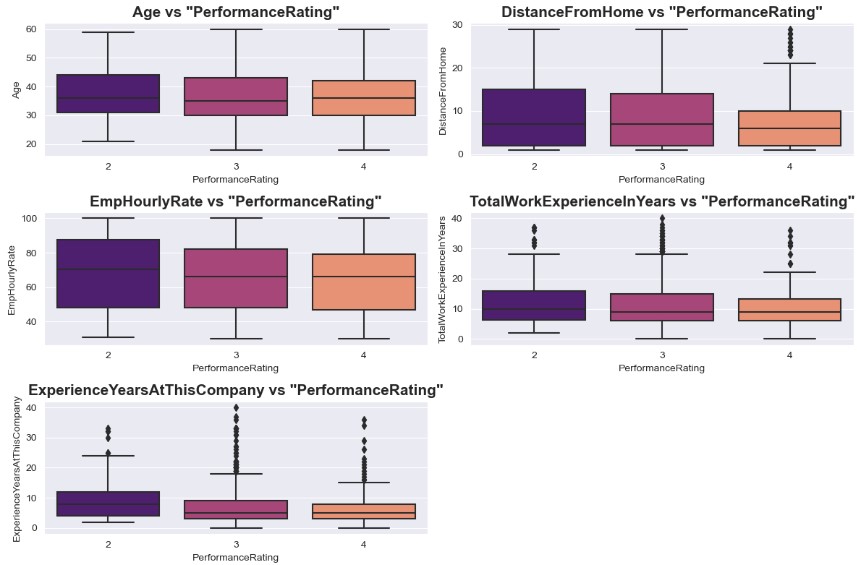


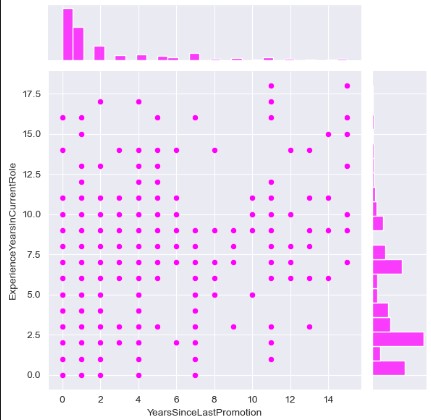
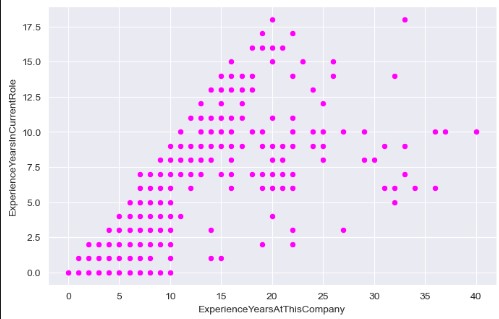


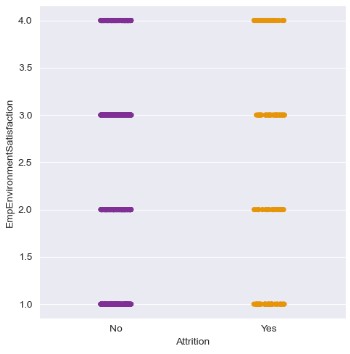


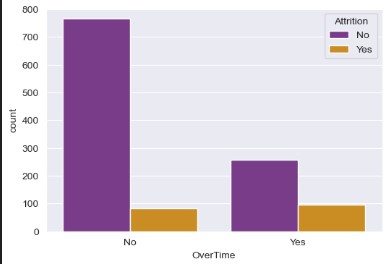
**Continuous features Vs target variable:**









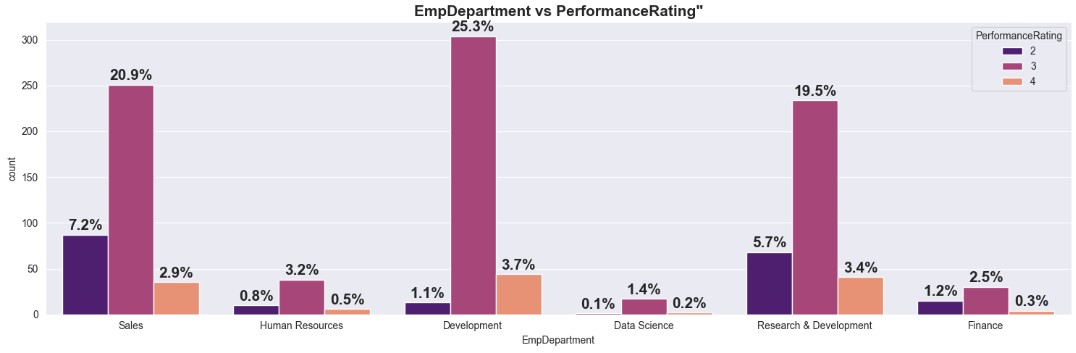


**Insights from bivariate analysis:**

* One common insight from all the above graphs seems to be that the "3" category of Performance Rating dominates in all the categorical input features.
* This represents the previous insights as similar to that of univariate analysis and when comparing the categories of each input feature, most of the employees in the dominating category have moderate performance rating of 3.
* For instance, in the **Employees education level**, the dominating category is 3rd level. Among them, most of the employees have shown moderate performance rating than the rest of the categories.
* However, the trend is different in few features.
* Even though the **employees have work environment satisfaction of 4th level, they show less performance rating** compared to the other categories. This goes in a similar way to that of employee’s job satisfaction.
* Since **the target variable is imbalanced**, this visualization may not reveal the true trend of the data distribution.
* Age does not significantly affect the performance rating as shown in the boxplot.
* Employees with travel **distance of more than 12 miles** are more probably low performing employees. Greater the distance from Home, lower the performance rating of the employees.
* **Hourly rate** of employees shows **significant different trend** with respect to the performance.
* **Total work experience** as well as **Experience at this company** shows similar trend as that of hourly rate.
* **Total work experience:**  Less than 15 years experienced employees have high performance rating compared to employees with >15 years of experience.
* **Experience years at this company**: experience of >8 years have poor performance rating whereas employees with **< 8 years of experience at this company** show high performance rating.
* This work experience trends could be due to their increasing age along with experience, **but due to the presence of many outliers**, these features might not capture the underlying pattern of the data.
* People who travel more are more expected to leave the job.
* People who do not do overtime less likely to leave the job.
* People from job roles of sales executive, developer, sales representative, laboratory technician, research scientist and manager of R&D are more likely to leave the job.

#### **DEPARTMENT WISE PERFORMANCE ANALYSIS:**

The analysis assesses the performance of employees across different departments to identify any disparities or trends.



**Insights:**

* Development department dominates the other departments in number.
* In terms of performance rating, employees who belong to development department perform well with the dominating 25% of rating 2.
* Hence departments such as **development, sales and Research & development** constitute the high performing employees.

### **FEATURE SELECTION/ FEATURE ENGINEERING:**

Feature selection techniques determine the relative importance of different features in predicting employee performance.

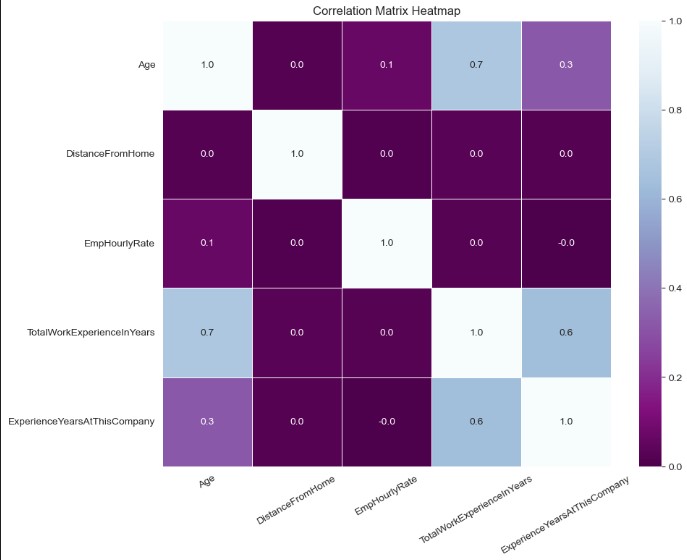
These techniques:

1. Ensure that visual analysis results are rigorously validated through appropriate methodologies and techniques.
2. Continuously monitor and evaluate analysis outcomes to ensure consistency and validity over time.

#### **THE TECHNIQUES USED IN THIS PROJECT ARE:**

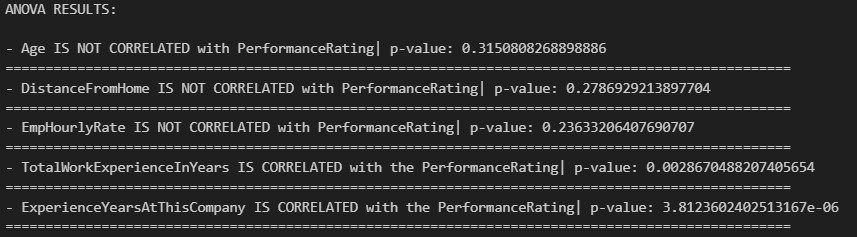
#### **Correlation analysis:**

Pearson’s correlation coefficient is calculated to find out the correlation coefficients between two continuous variables.



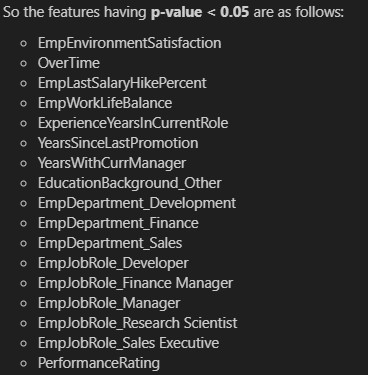
#### **ANOVA Test:**

ANOVA test is performed between categorical and continuous variable to find out the statistical significance of the features.



#### **CHI SQUARE Test:**

Chi square test is performed between two categorical features to assess the feature relevance for the predictions.



#### **INSIGHTS:**

**IMPORTANT FEATURES SELECTED:**

There are a total of 19 columns being selected after feature selection techniques performed as above. They are as follows (they also include encoded extra columns):

* TotalWorkExperienceInYears
* ExperienceYearsAtThisCompany
* EmpEnvironmentSatisfaction
* OverTime
* EmpLastSalaryHikePercent
* EmpWorkLifeBalance
* ExperienceYearsInCurrentRole
* YearsSinceLastPromotion
* YearsWithCurrManager
* EducationBackground\_Other
* EmpDepartment\_Development
* EmpDepartment\_Finance
* EmpDepartment\_Sales
* EmpJobRole\_Developer
* EmpJobRole\_Finance Manager
* EmpJobRole\_Manager
* EmpJobRole\_Research Scientist
* EmpJobRole\_Sales Executive
* PerformanceRating

These features are selected after performing correlation analysis, ANOVA test and chi-square test to fetch the most correlated input features for predictive modelling.

**THE ABOVE FEATURE SELECTION TECHNIQUES ARE USED DUE TO:**

The dataset contains the datatypes such as:

* **Input features:** continuous and categorical
* **Target variable:** categorical

So, three relevant tests are selected to assess the importance of all the features.

They are:

* Pearson’s correlation coefficient: Between continuous🡨🡪continuous
* ANOVA test: Between categorical🡨🡪continuous
* Chi-square test: Between categorical 🡨🡪categorical

**IMPORTANT FEATURE TRANSFORMATIONS:**

The final features selected also include the encoded features (one-hot encoding) hence some of them would be different form the features of the raw dataset.

***The selected original independent features for modelling:***

* TotalWorkExperienceInYears
* ExperienceYearsAtThisCompany
* EmpEnvironmentSatisfaction
* OverTime
* EmpLastSalaryHikePercent
* EmpWorkLifeBalance
* ExperienceYearsInCurrentRole
* YearsSinceLastPromotion
* YearsWithCurrManager
* EducationBackground🡪 *(After one-hot encoding)🡪*EducationBackground\_Other
* EmpDepartment🡪 *After one-hot encoding)🡪* EmpDepartment\_Development, EmpDepartment\_Finance, EmpDepartment\_Sales
* EmpJobRole🡪 *After one-hot encoding)🡪* EmpJobRole\_Developer, EmpJobRole\_Finance Manager, EmpJobRole\_Manager, EmpJobRole\_Research Scientist, EmpJobRole\_Sales Executive

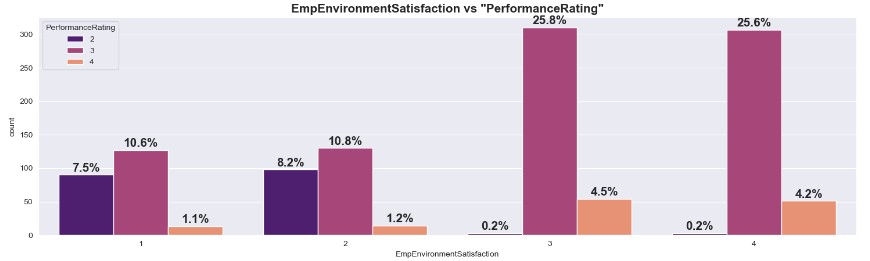
#### **TOP 3 FACTORS AFFECTING THE EMPLOYEE’S PERFORMANCE:**

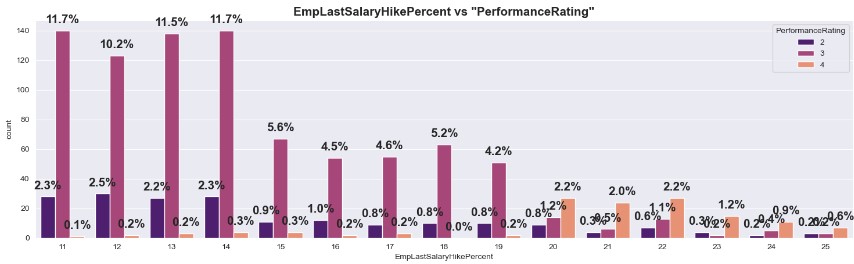
From the data analysis and the feature engineering, there are some of the factors which proved crucial in predicting the employee’s performance.

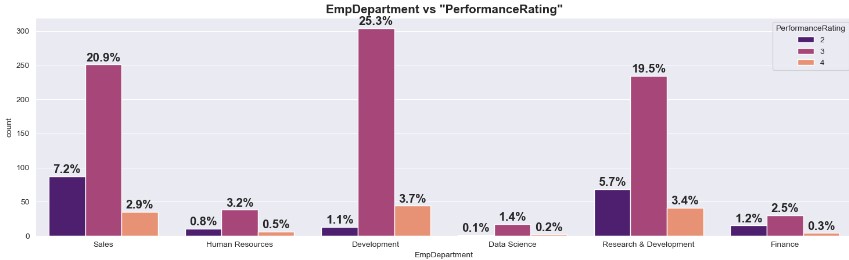
**Among them the top three factors are as factors:**

1. EmpEnvironmentSatisfaction
2. EmpLastSalaryHikePercent
3. EmpDepartment\_Development

**The visual analysis of these three factors are as follows:**







**Insights:**

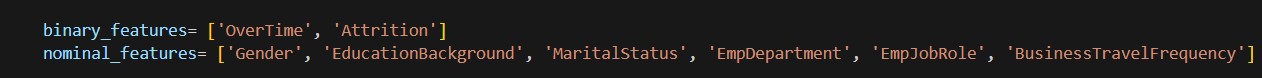
* The **work environment satisfaction** has positive relationship with the target. As the work environment satisfaction increases, the performance rating of the employees also increases.
* The feature “**EmpLastSalaryHikePercent**” holds negative relationship with the target. The performance rating is quite high consistently for a fixed range of salary hike percent of 11-14%. After this range, the performance decreases. This could attribute to other factors such as various departments and their working strength, limited top level positions etc.
* The **employee’s department of development** shows high performance rating compared to other departments.

### **ANALYSIS SUMMARY:**

* **Employees Travel:** The workers who travel a lot are more likely to have poor performance.
* **Department:**  The workers from job roles such as Manager, healthcare representative, Research director, business analyst more likely to stay than the workers of other job roles.
* **Educational Background:** The workers with Human Resources and Technical Degree are more likely to quit then employees from other fields of educations.
* **Marital Status:** The workers who have Single marital status are more likely to quit the Married, and Divorced.
* **Over Time:** More the overtime, more the attrition rate.

## **PREPROCESSING TCHNIQUES USED**

### **ENCODING TECHNIQUES:**



These features are the categorical types to be encoded in the dataset.

**The techniques utilized are as follows:**

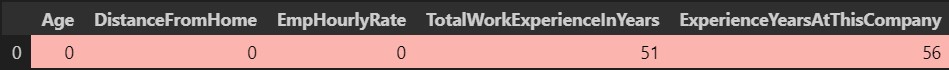
* **For binary features:**

Manual encoding is done, typically mapping is used ***(“No”🡪0, “Yes”🡪 1)***

* **For nominal features:**

One-hot encoding is done which assigns binary value to a category with the rest of the categories as zero in a categorical feature.

### **OUTLIERS HANDLING:**



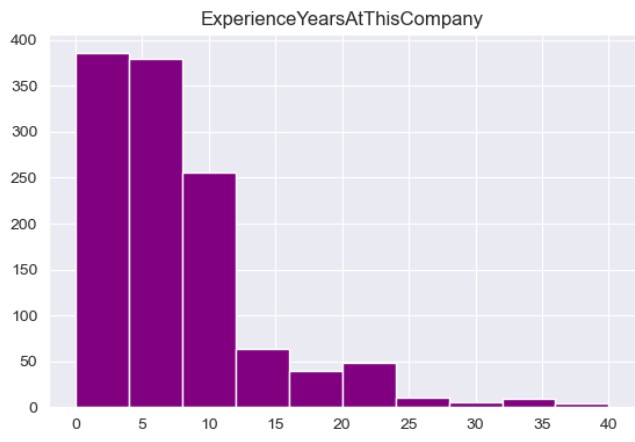
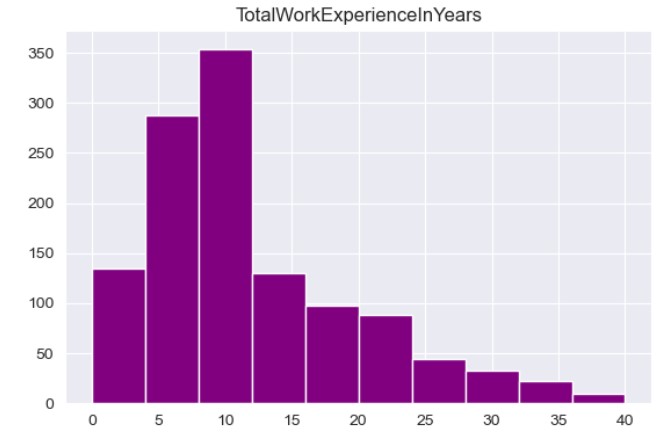
The above image is the output from the source code which shows the number of outliers in the continuous independent features present in the dataset.

**Features with outliers are as follows:**

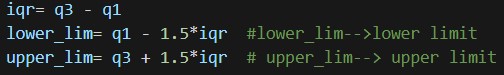
* TotalWorkExperienceInYears
* ExperienceYearsAtThisCompany

**Technique used to handle the outliers:**

**Interquartile method (IQR)** is used to find the outliers since these features are having skewed distribution.



**The formulae used in the IQR method are as follows:**



*# iqr🡪 inter-quartile range*

*# q1🡪 25th percentile*

*# q3🡪 75th percentile*

After finding the outliers, they are replaced by the **median value** of the corresponding feature.

### **SCALING:**

* There are two types of scaling in common practice.
* They are minmax scaler and the standard scalerMethods.
* However, in this project, **MinMaxScaler** is utilized from **scikit library**.

**MinMaxScaler:**

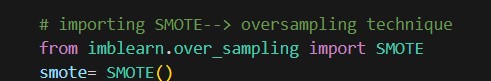
scales the features to a specified range, typically between 0 and 1.

**The continuous features that are scaled using MinMaxScaler are as follows:**

1. TotalWorkExperienceInYears
2. ExperienceYearsAtThisCompany

### **BALANCING THE DATASET:**

* Oversampling technique is used to balance the classes of the target variable.
* **SMOTE function** is utilized from **imblearn library**, which is one of the oversampling techniques, to create synthetic sample from under-represented minority class.
* This function is applied on the training data to make the model learn and predict in an unbiased manner.



**Classes of the target of the training sample before and after oversampling:**



## **ALGORITHMS AND TRAINING METHODS USED**

### **Algorithms used in this project:**

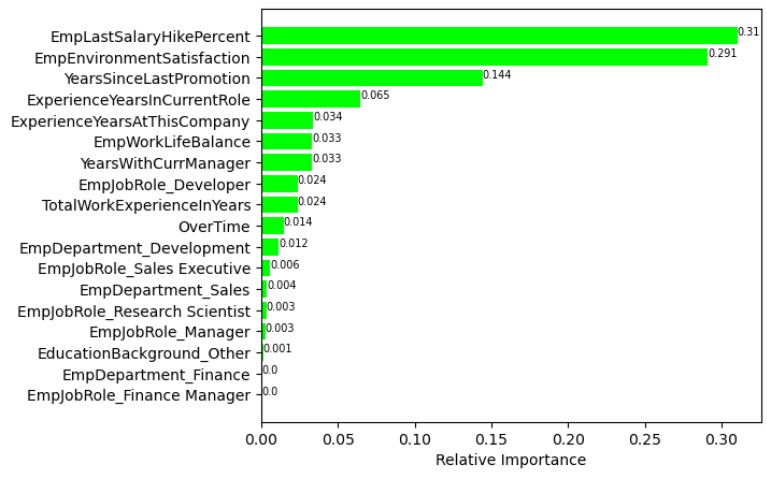
* Decision Tree classifier
* Random Forest classifier
* Support Vector Machine
* Extreme Gradient Boosting

### **Model Training Methods:**

#### **DECISION TREE CLASSIFIER:**

* This algorithm works on the principle of if-else condition and make decisions by partitioning the input features into regions and assigning a class label to each region.
* It is preferred in this dataset due to its flexibility and easy model interpretability.
* Since the given dataset is medium-sized, decision tree can handle it well.





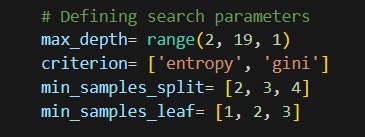
From the **feature importances** arrived at by this model, it is clear that the top three relevant features are:

* EmpLastSalaryHikePercent
* EmpEnvironmentSatisfaction
* YearsSinceLastPromotion

#### **TUNED DECISION TREE CLASSIFIER:**

This model refers to a decision tree algorithm that has been optimized for better performance.

The hyperparameters tuned in this model are as follows:

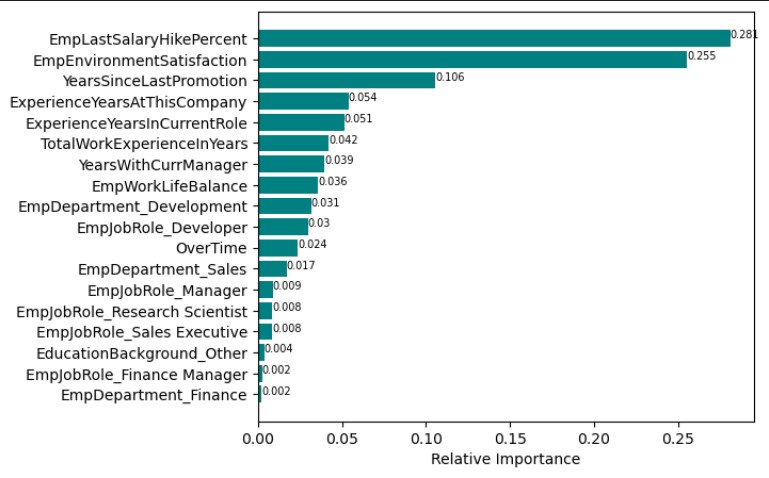


Cross validation technique used is **Grid Search CV** which exhaustively searches through a specified subset of hyperparameter combinations.

#### **RANDOM FOREST CLASSIFIER:**

* Random forest is an ensembling learning method that functions by constructing multiple decision trees during training and outputting the majority of the classes (classification) or the average prediction (regression) of the individual trees.
* This model is used here due to it’s ability to handle High-dimensional data, imbalanced datasets and capability of capturing complex relationships in the dataset.
* Moreover, it overcomes the overfitting issue of the decision tree model.

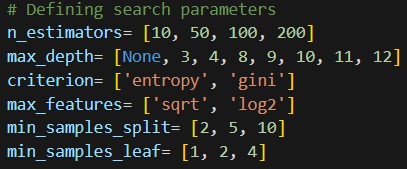




The feature importances in this model also gives the same top three features as the relevant features similar to that of the decision tree classifier.

#### **TUNED RANDOM FOREST CLASSIFIER:**

The hyperparameters tuned in this model are as follows:



Cross validation technique used is **Grid Search CV** which exhaustively searches through a specified subset of hyperparameter combinations.

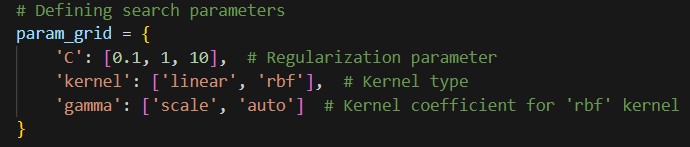
1. **SUPPORT VECTOR MACHINE:**

* A **Support Vector classifier** (SVC) works by finding the hyperplane that best separates the classes in the feature space.
* This model is trained in this project due to suitability to handle high-dimensional data.



1. **TUNED SUPPORT VECTOR CLASSFIER:**

The hyperparameters tuned in this model are as follows:



Cross validation technique used is **Grid Search CV** which exhaustively searches through a specified subset of hyperparameter combinations.

1. **TUNED XG BOOSTING:**

* XGBoost (eXtreme Gradient Boosting) is an efficient implementation of gradient boosting which builds a series of weak learners, with each new learner correcting errors made by the previous ones.
* This contains in-built regularization techniques and pruning to prevent overfitting and control the model complexity.
* This model is highly scalable due to parallelization and distributed computing.

