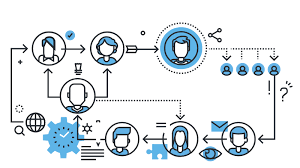
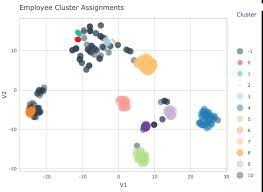
**BLOG ON HR ANALYTICS PROJECT**

**Machine Learning to Understand & Predict HR Attrition**

****

****

****

**by**

**Nikita Kumari**

**PROBLEM DEFINITION-**

Each year, numerous companies embark on the journey of recruiting new employees, investing substantial time and resources in their training and development. Moreover, these organizations often conduct internal training programs to enhance the skills and effectiveness of their existing workforce. The overarching goal of these initiatives is to bolster employee performance and productivity. However, where does HR Analytics fit into this equation? Is its sole purpose confined to optimizing employee performance?

HR Analytics, a pivotal component in the realm of workforce management, involves the application of analytical methodologies to the human resources domain. Its primary objective is to leverage data-driven insights to enhance employee performance, consequently yielding a higher return on investment for organizations. However, HR analytics transcends mere data collection on employee efficiency; its essence lies in delving deep into various HR processes, harnessing data, and making informed decisions to streamline these processes effectively.

Employee attrition refers to the process of workers leaving a company for voluntary or involuntary reasons, without being immediately replaced. Sometimes employee attrition is due to a hiring freeze, at other times, there are deeper issues at play. However, whatever the cause of employee attrition, there’s one inevitable result: the company’s workforce shrinks in size.

Attrition, a prevalent concern in human resources, refers to the gradual loss of employees over time. Elevated attrition rates pose significant challenges for companies, prompting HR professionals to take proactive measures in designing compensation programs, fostering conducive work cultures, and implementing motivation systems to retain top talent. How does attrition impact companies, and how can HR Analytics aid in its analysis? Let's delve into the first query here, and for the latter, we'll delve into the intricacies through a step-by-step examination of the process.

The ramifications of high employee attrition are manifold, with one of the primary concerns being its substantial cost to organizations. The expenses incurred in advertising job vacancies, conducting recruitment processes, managing paperwork, and facilitating new hire training sessions constitute significant financial burdens. Furthermore, frequent turnover impedes the accumulation of collective knowledge and experience within the organization over time. This becomes particularly worrisome for customer-facing businesses, as clientele often prefer interacting with familiar faces. The continuity of service is disrupted when new employees are constantly introduced, increasing the likelihood of errors and issues arising.

**IBM Dataset**

In this case study, we will leverage the IBM HR Analytics database, a fictional dataset meticulously crafted by IBM professionals, readily accessible for download from GitHub and Kaggle repositories. Alternatively, you can procure the dataset directly from my GitHub profile, available [here](link). Comprising 1470 rows and 35 features delineating each employee's background, characteristics, and a target variable, this dataset is designed to shed light on the intricacies of employee attrition. Attrition serves as the target variable to be predicted, rendering this case study a classification machine learning endeavor. Our objectives are twofold:

1. Uncover the pivotal factors contributing to employee attrition.

2. Develop a robust ML model for accurately predicting attrition.

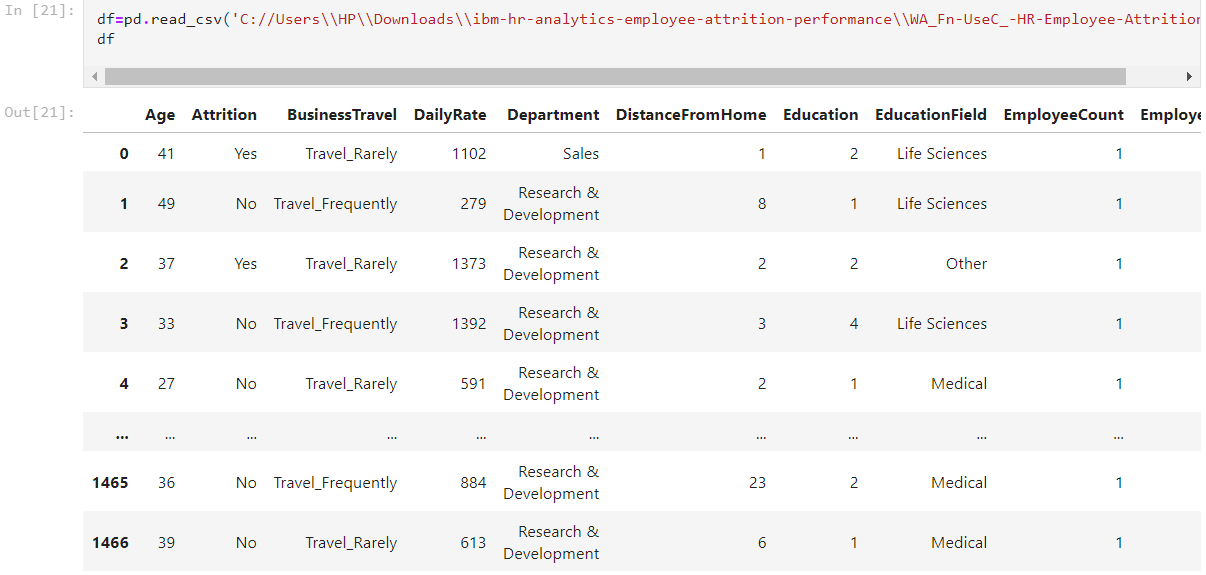
Dataset link - https://github.com/dsrscientist/IBM\_HR\_Attrition\_Rate\_Analytics

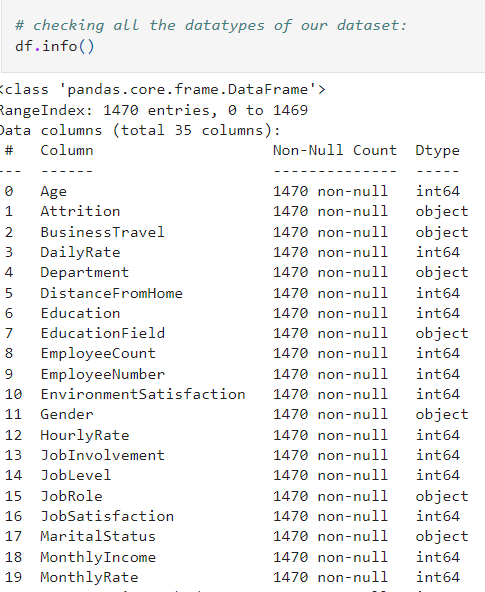
**DATA ANALYSIS-**

**Data Preparation: Load, Clean and Format-**

Let’s begin with importing libraries for EDA and dataset itself.

****

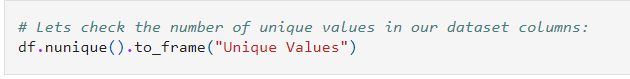


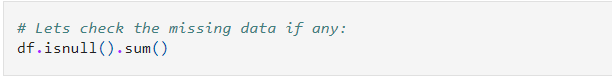
****

###### As per above information we can see that our dataset has 26 integer values and 9 object values.

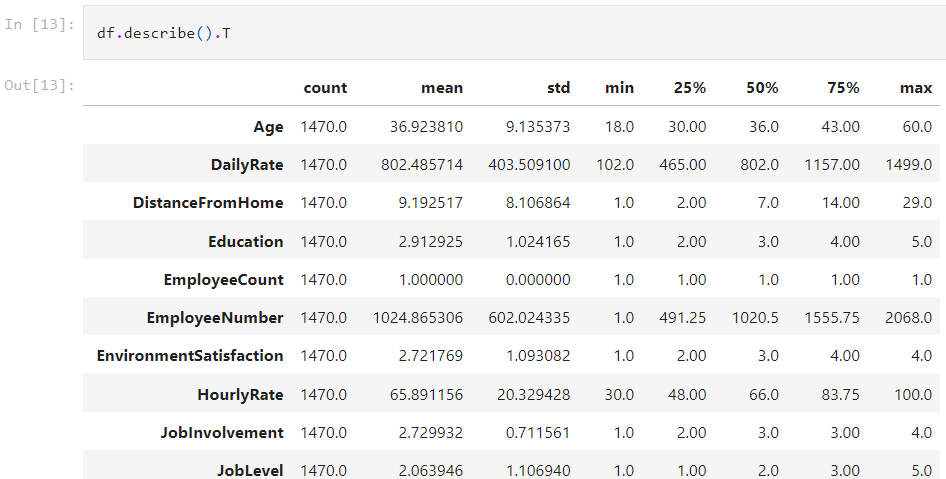
###### We will need to treat the object datatype columns so that our machine learning model can understand the data

**Data Integrity Check:** Dataset can have missing values, duplicated entries and whitespaces. Now we will perform this integrity check of dataset.





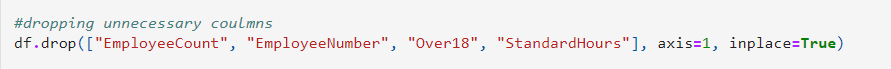
Fantastic news! The dataset is pristine, devoid of any missing entries, duplicates, or unwanted characters like whitespace, 'NA', or '-'. With this clean slate, we're all set to delve deeper into the data. Statistical parameters such as mean, median, and quantiles offer invaluable insights into the dataset. Now, it's time to unveil the statistical matrix of our dataset.



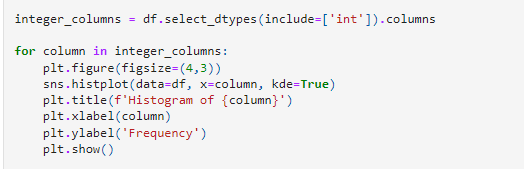
Using the 'describe' method in transpose format provides a clear overview of column details. The 'count' column reaffirms the absence of any missing data concerns within our dataset.

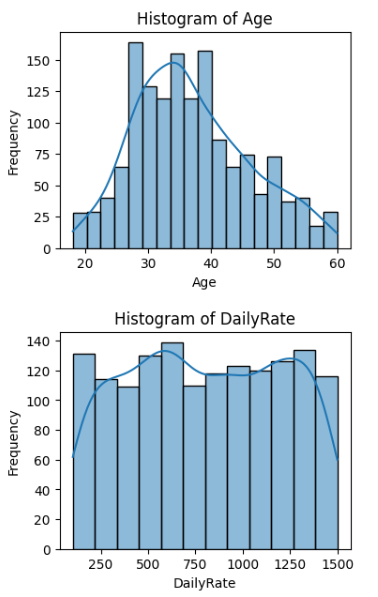
However, upon inspecting the 'min' column, we observe instances where columns contain zero values. Yet, it's conceivable that freshers within an organization might have these fields marked as zeros in their records.

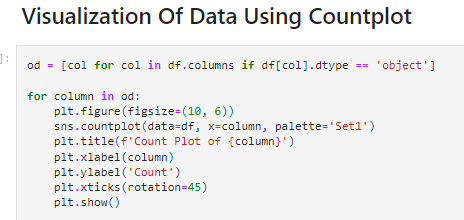
It's worth noting that only numerical data information is obtained here, with all object datatype columns being disregarded.

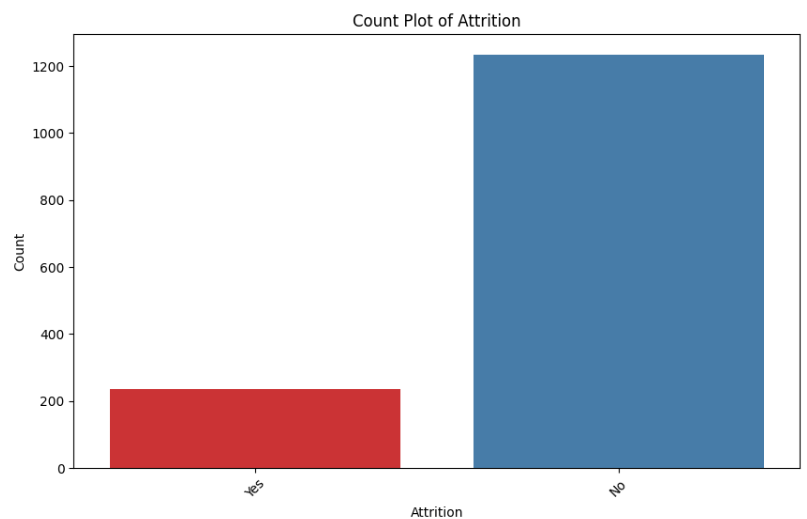


**DATA VISUALIZATION AND ANALYSIS-**



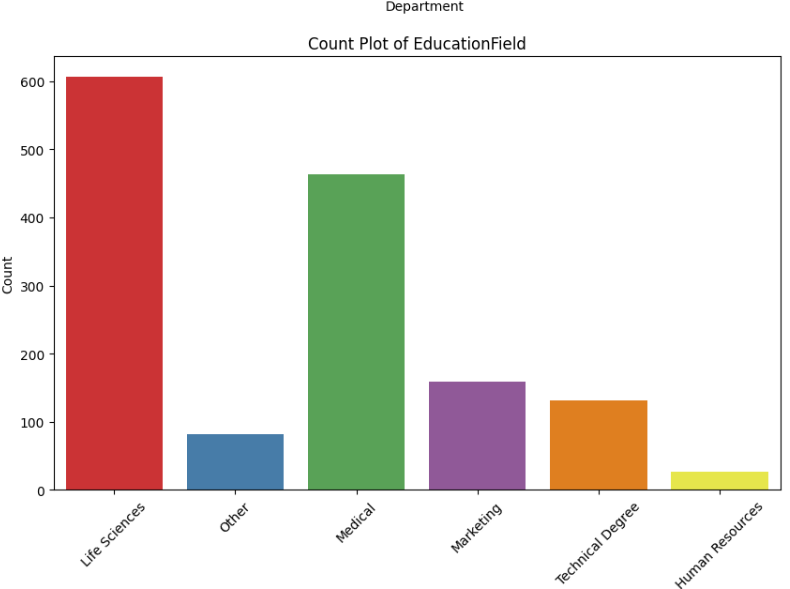


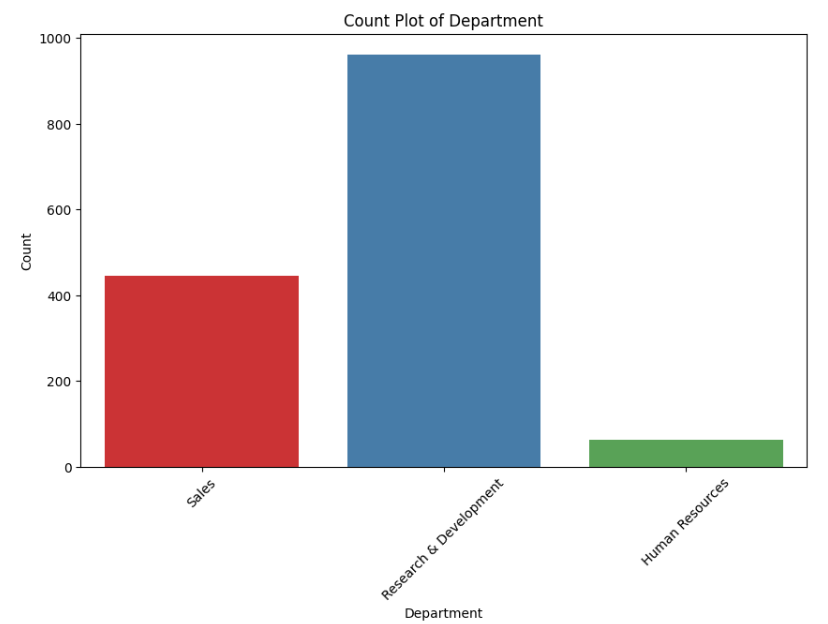




Out of the total workforce, 83.88% (1237 employees) opted to remain with the organization, whereas 16.12% (237 employees) chose to leave. This distribution suggests an imbalance, with a greater number of employees staying compared to those leaving.

Within the dataset, features such as education, department, education field, job role, and job satisfaction are interconnected. Misalignment between job roles and educational backgrounds can contribute to attrition.

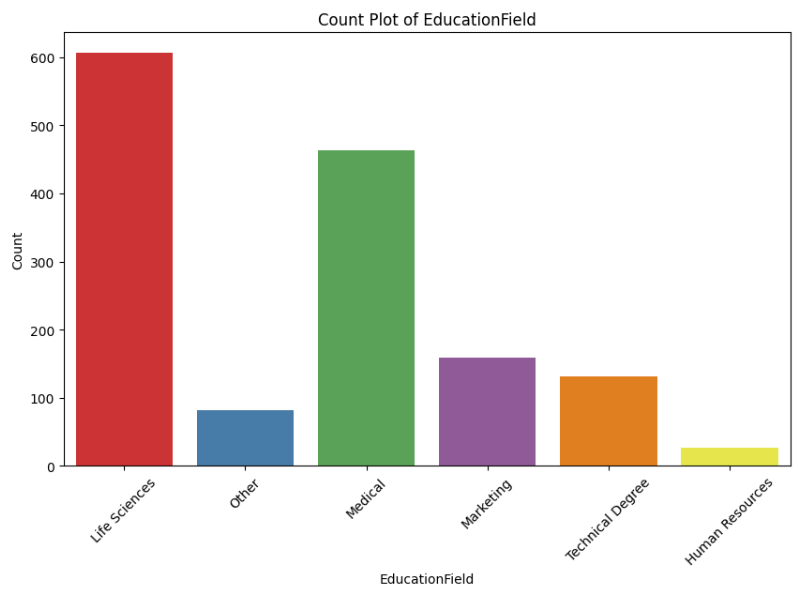


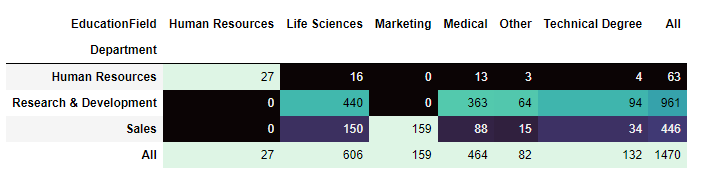


Key Insights on Department and Education Level of employee in each Department

1. 65.37% of Employees work inside Research & Development Department. Out of Total 961 Employee number of employees with education level of Bachelors, Masters, Doctor are 379, 255 and 30 respectively.
2. Only 63 Employee work in HR department.

Employees distribution as per eduction field-





Attrition is predominant with approximately 1200 "No" responses and fewer "Yes" responses, around 200.

Regarding Business Travel, the majority, around 1000 individuals, travel rarely, followed by approximately 300 frequent travelers, and less than 200 non-travelers.

Within departments, Research and Development house the most employees, nearly 1000, followed by Sales with approximately 400, and Human Resources with about 50.

In terms of education, Life Sciences boast the highest number of individuals, with over 600, while Human Resources have significantly fewer, around 50.

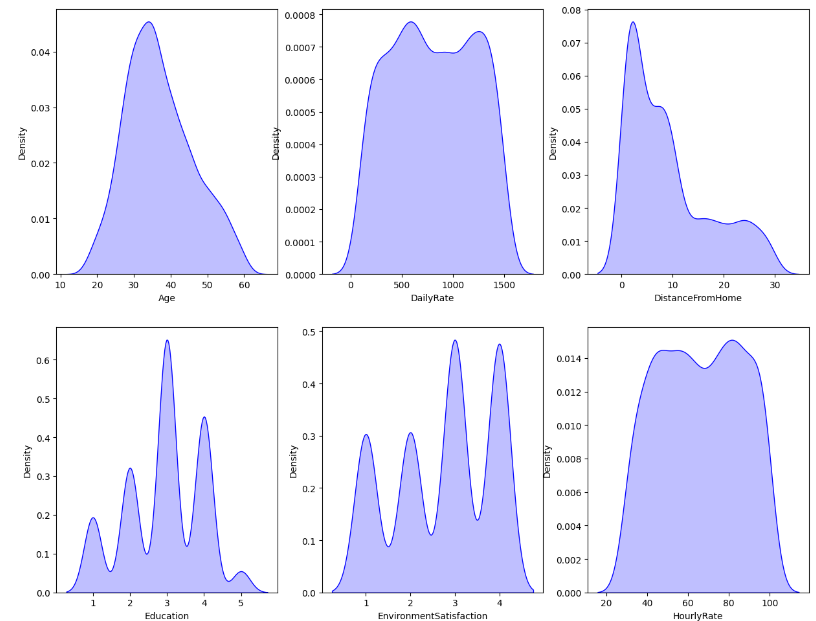
Gender distribution shows that males outnumber females, with over 800 males compared to around 600 females.

Sales Executives comprise over 300 individuals in their job role, whereas Human Resources have significantly fewer, around 50.

Married individuals are the most numerous, with approximately 650, followed by singles and divorced individuals, each around 450.

About 400 individuals work overtime, while approximately 1000 do not.

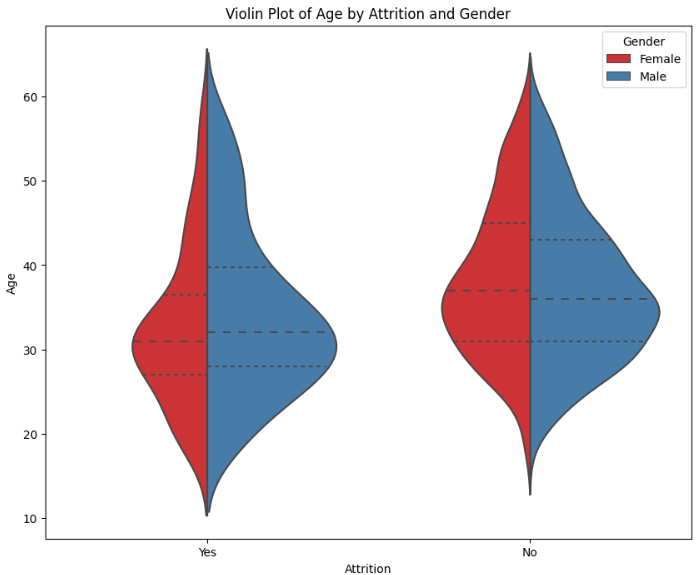
## **Visualization of data using Distribution plot**



Some skewness is present in few columns that we will need to handle.

1. DistanceFromHome
2. MonthlyIncome
3. NumCompaniesWorked
4. PercentSalaryHike
5. TotalWorkingYears
6. TrainingTimesLastYear
7. YearsAtCompany
8. YearsInCurrentRole
9. YearsSinceLastPromotion
10. YearsWithCurrManager

## **Visualization of Data using Violinplot**

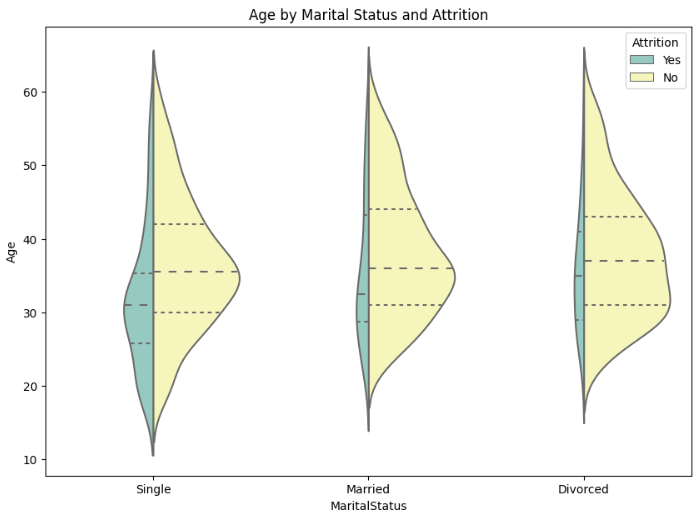




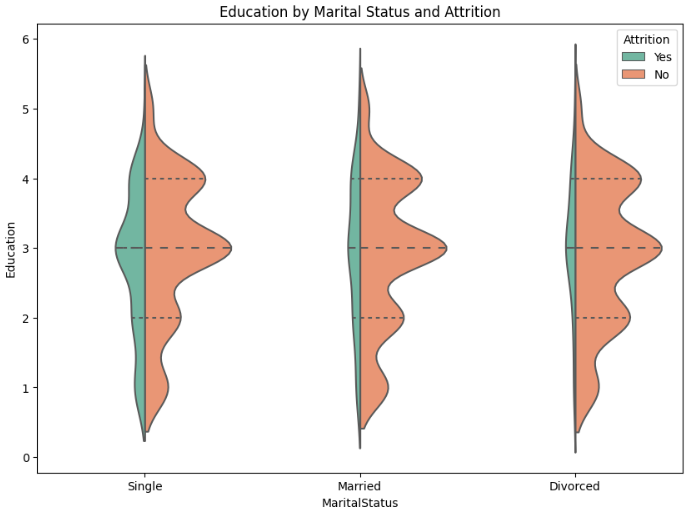
In the above plot we can see that the Attrition peaks for both male and female employees the monthly iwhen ncome is less than 5000



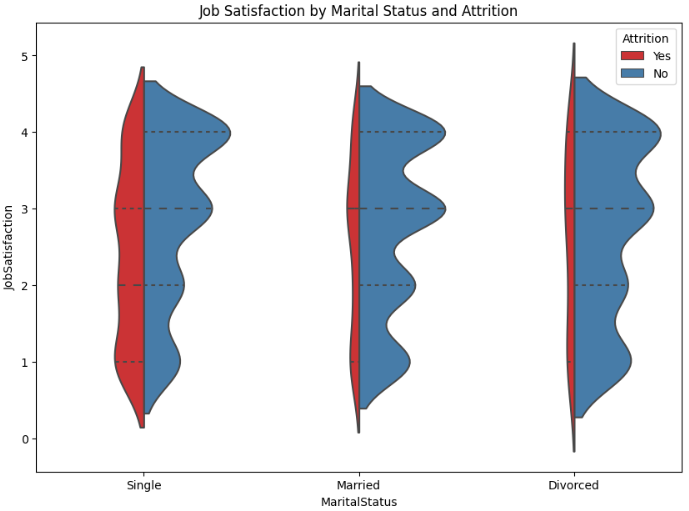
In the above plot we see that the Attrition for both make and female employees happen when they do not see prootions happening after years of gaining experience.



In the above violin plot we can see that the Attrition rate is quite less in employees when they are married or divorced as compared to when they are single and have lesser responsibilities to deal with at their age



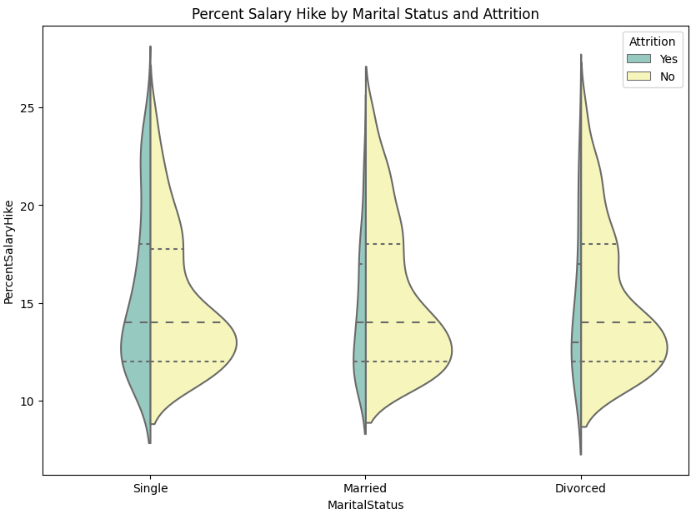
In the above plot we notice that once again employees who are married or divorced and with good education choose stability in life rather than the one's who are single and are okay to take risks and oppotunities in life.



In the above voilin part we can see that the job satisfaction part for singles is not that great compared to employees who are married or divorced may be due to the year of experience difference that makes a huge gap in pay scale. But we do notice stability and lesser attrition rate amongst employees who are married or divorced.



In the above plot we can see that Work Life balance maintained by singles are quite less therefore there are attritions observed as they have to gather lots of skills and experience to get better in their career.



In the above violin plot we can see that the Percent Salary Hike plays a major role when it come to Attrition amongst the Singles as comapred to their married or divorced counterparts.

**FEATURE ENGINEERING- DATA PRE-PROCESSING-**

Feature Engineering is very important step in building Machine Learning model. Some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. In Feature engineering can be done for various reason. **Some of them are mention below:**

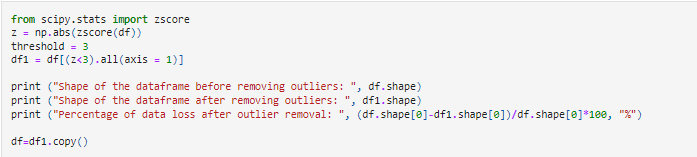
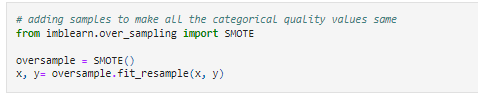
There are Varity of techniques use to achieve above mention means as per n**Feature Importance**: An estimate of the usefulness of a feature

1. **Feature Extraction**: The automatic construction of new features from raw data (Dimensionality reduction Technique like PCA)
2. **Feature Selection**: From many features to a few that are useful
3. **Feature Construction**: The manual construction of new features from raw data (For example, construction of new column for month out date - mm/dd/yy)

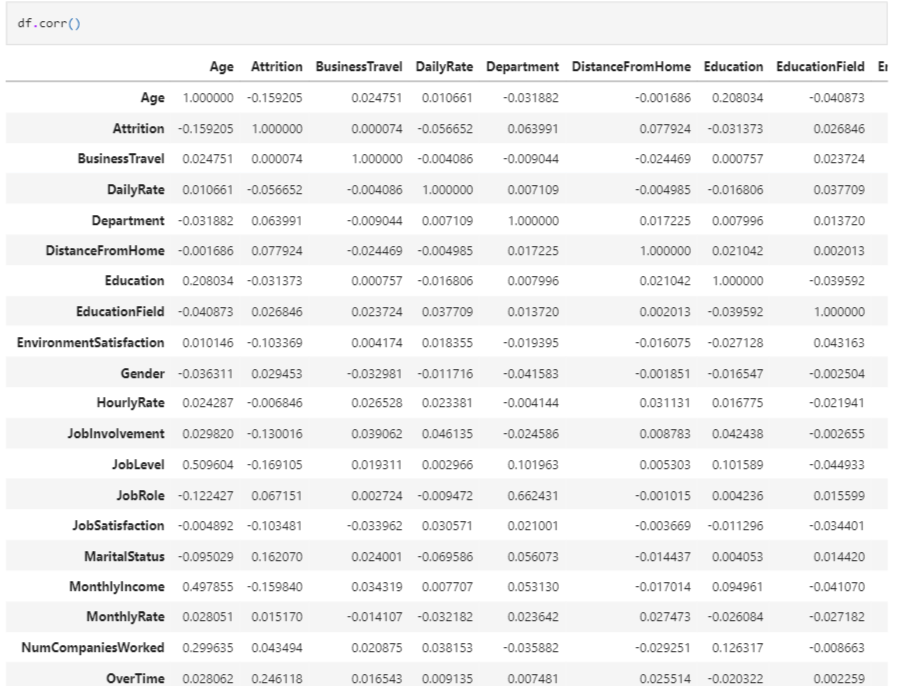
eed of dataset. Some of Techniques important are as below:

* Handling missing values
* Handling imbalanced data using SMOTE
* Outliers’ detection and removal using Z-score, IQR
* Scaling of data using Standard Scalar or Minmax Scalar
* Binning whenever needed
* Encoding categorical data using one hot encoding, label / ordinal encoding
* Skewness correction using Boxcox or yeo-Johnson method
* Handling Multicollinearity among feature using variance inflation factor
* Feature selection Techniques:
* Correlation Matrix with Heatmap
* Univariate Selection – SelectKBest
* ExtraTreesClassifier method

In this case study we will use some of the mention feature engineering Techniques one by one.

1. **Dropping unnecessary features -**Feature like ‘Over18’, ‘StandardHours’ contain single unique value. Features like EmployeeCount, EmployeeNumber are irrelevant from ML model building perspective. We will drop these features.
2. **Encoding Categorical & Ordinal Features-**Label Encoding is employed over target variable ‘Attrition’ while Ordinal encoding employ for rest categorical features.
3. **Outliers’ detection and removal-**Machine learning algorithms are sensitive to the range and distribution of attribute values. Data outliers can spoil and mislead the training process resulting in longer training times, less accurate models and ultimately poorer results. Outliers can be seen in boxplot of numerical feature. We did not added boxplot here as it will make this article length, I left it to reader to further investigate. 
4. **Handling imbalanced data using SMOTE-** This two-class dataset is imbalanced (84% vs 16%). As a result, there is a possibility that the model built might be biased towards to the majority and over-represented class. We can resolve this by Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class. 
5. **Scaling of data using Standard Scalar-**

****

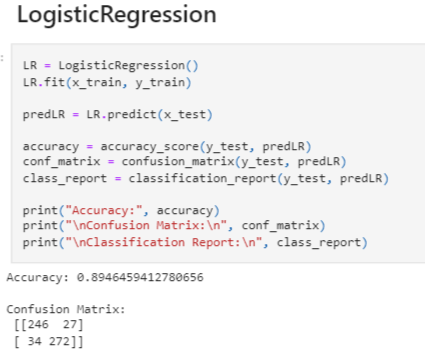
1. **Correlation - **

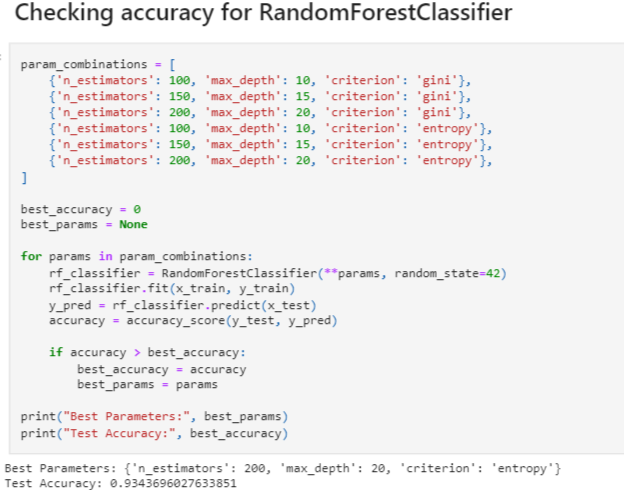
**MACHINE LEARNING MODEL BUILDING-**

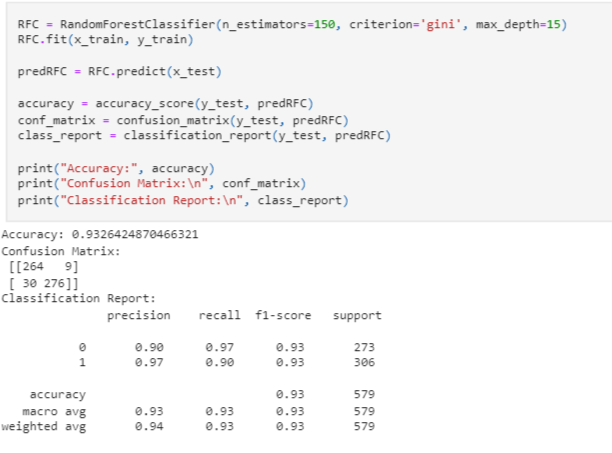
In this section we will build Supervised learning ML model-based classification algorithm. As objective is to predict attrition in ‘Yes’ or ‘No’ leads to fall problem in domain of classification algorithm. train\_test\_split used to split data with size of 0.33

First we will build base model using logistic regression algorithim. Best random state is investigated using for loop for random state

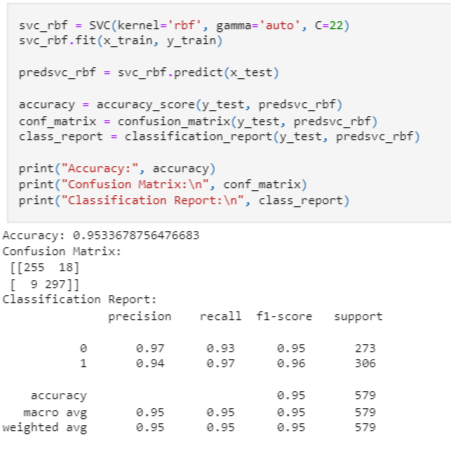




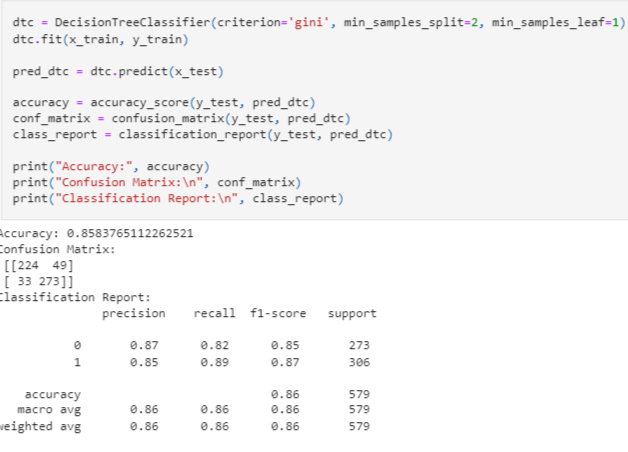




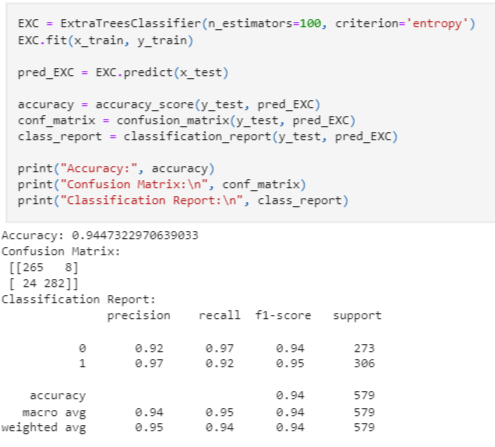




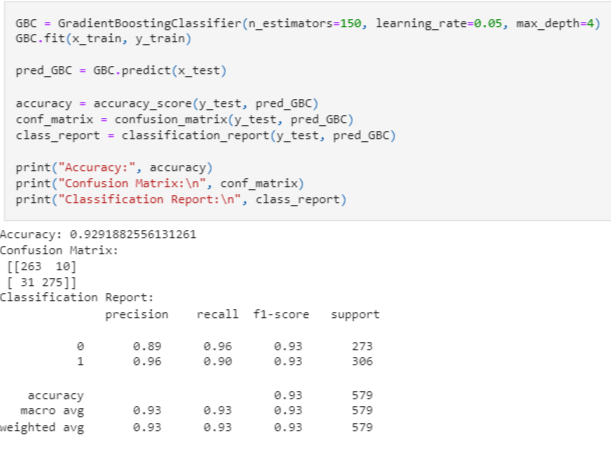








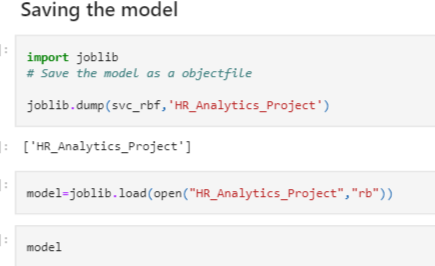








At last, we will save final model with joblib library, so it can be deploy on cloud platform.



**CONCLUSIONS-**

1. Attrition rate is high in age group of 29 to 33. HR need to keep eye over need & expectation of this age group from company.
2. Percentage of attrition is high in Sales Representative, Laboratory Technician
3. 16 % attrition rate among Research Scientist and no company afford to lose them.
4. Almost 50% employs in sales department from different education background. There is possibility of dissatisfaction among them as attrition high among these.
5. Different feature engineering techniques like balancing data, outliers’ removal, label encoding, feature selection & PCA are perform on data.