CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Higher Diploma in Data Analytics for Business |
| **Assessment Title:** | Data Preparation & Visualization Statistical Techniques for Data Analytics Machine Learning |
| **Lecturer Name:** | Muhammad Iqbal David McQuaid Marina Soledad Iantorno |
| **Student Full Name:** | Andrew Maher |
| **Student Number:** |  |
|  | sba22521 |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** | 5-12-24 |
| **Date of Submission:** | 5-12-24 |

*Andrew Maher*

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

[Table of Contents](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Table-of-Contents) Page

Import libraries...................................................................................................................................................................................................4

Importing The Data..............................................................................................................................................................................................4

Business Understanding.......................................................................................................................................................................................9

Data Understanding..............................................................................................................................................................................................9

Overall Objective.................................................................................................................................................................................................10

# Characterization of the Data.................................................................................................................................................................................10

[Missing Values](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "4--Missing-Values).....................................................................................................................................................................................................10

Data Preparation....................................................................................................................................................................................................18

Exploratory Data Analysis.....................................................................................................................................................................................25

[Outliers](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "5-EDA-&-Outliers)...................................................................................................................................................................................................................28

Cluster analysis......................................................................................................................................................................................................30

Check For Normality..............................................................................................................................................................................................31

[Linear Discriminant Analysis](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "9---Linear-Discriminant-Analysis).................................................................................................................................................................................32

Principal Component Analysis................................................................................................................................................................................32

Feature Engineering- Binning.................................................................................................................................................................................34

[Machine Models Data Preprocessing and Encoding](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "12---Machine-Models-Data-Preprocessing-and-Encoding)................................................................................................................................................37

[Train Dummy Classifier](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "13--Train-Dummy-Classifier)..........................................................................................................................................................................................37

[Machine Learing Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "14--Machine-Learing-Models)...........................................................................................................................................................................................37

Cross Validation..........................................................................................................................................................................................................40

[Hyperparameter Tuning of Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "16---Hyperparameter-Tuning-of-Models)............................................................................................................................................................................42

[Feature Importances](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "17--Feature-Importances)...................................................................................................................................................................................................42

[Evaluation of Best Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "18--Evaluation-of-Best-Models)..........................................................................................................................................................................................43

[Statisical Modelling](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "20-Statisical-Modelling).....................................................................................................................................................................................................46

[Descripitive Stats](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "21-Descripitive-Stats).........................................................................................................................................................................................................47

[Inferential Statistics](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "22---Inferential-Statistics).....................................................................................................................................................................................................50

[Appendix](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "23---Appendix)

References

**Table of Figures**

Figure 1. Replacing Employees 6

Figure 2. Age Imputation Histogram 11

Figure 3. Monthly Imputation Histogram 11

[Figure 4. Daily Rate Imputation Histogram](#_bookmark3) 12

[Figure 5. Gender Histogram](#_bookmark4) 13

[Figure 6. Business Travel Histogram. ...1](#_bookmark5)4

[Figure 7. Department Histogram Imputation ...10](#_bookmark6)

[Figure 8. Education Imputation Histogram ...1](#_bookmark7)5

Figure 9. Catplot Job Role ...16

[Figure 10. Catplot Marital Status ...1](#_bookmark8)6

[Figure 11. Catplot Attrition ...1](#_bookmark9)6

[Figure 12. Correlation Plot ...1](#_bookmark10)7

[Figure 13. Leavers Histogram ...1](#_bookmark12)8

[Figure 14. Eda Histograms ...1](#_bookmark13)9

[Figure 15. Violin Plots ....](#_bookmark14)21

[Figure 16.](#_bookmark15) Scatter Plots….22

[Figure 17. Employed/Left Tables](#_bookmark16) ….22

[Figure 18. Boxplots Outliers](#_bookmark17) ….24

[Figure 19.](#_bookmark18) Boxplots after Outlier Removal [….2](#_bookmark18)5

[Figure 20. Normality Tests QQ plots …..2](#_bookmark19)7

[Figure 21. Histplot Stock Performance 2](#_bookmark20)7

[Figure 22. Scatter Plot Cluster analysis. 2](#_bookmark21)9

[Figure 23. Scatter Plot for LDA.](#_bookmark21) 30

[Figure 24. Confusion Matrix LDA.](#_bookmark21) 30

[Figure 25. PCA Scatter plot.](#_bookmark21) 32

[Figure 26. Decision Tree Binning....](#_bookmark21) 34

[Figure 27. Confusion Matrix Gradient Boosting.](#_bookmark21) 38

[Figure 28. Confusion Matrix Logistic regression.](#_bookmark21) 38

[Figure 29. Confusion Matrix Gradient Boosting Fine Tuning.](#_bookmark21) 39

[Figure 30. Confusion Matrix Logistic Regression Fine Tuning.](#_bookmark21) 40

[Figure 31. Gradient Boosting Test set.](#_bookmark21) 41

[Figure 32. Logistic regression Test set.](#_bookmark21) 42

[Figure 33. Rock curve](#_bookmark21) 43

[Figure 34. Feature Importance's Histogram.](#_bookmark21) 46

[Figure 35. Correlation Matrix Top 10 Features.](#_bookmark21) 47

[Figure 36. Age Bar Chart.](#_bookmark21) 47

[Figure 37. Hourly Rate Bar Chart Employed v Leavers.](#_bookmark21) 48

[Figure 38. Monthly Rate Bar Chart.](#_bookmark21) 49

[Figure 39. Distance From Home Bar Chart.](#_bookmark21) 50

[Figure 40. Years Since Last Year Bar Chart.](#_bookmark21) 51

[F](#_bookmark21)

# Introduction

A Center for American Progress research on employee turnover found that replacing a person costs 20% of their pay. Free or low-cost workplace flexibility and earned sick days can reduce attrition and save money.

Thus, most companies find employee replacement costly. Interviewing and finding a successor, sign-on bonuses, and months of lost production while the new hire adjusts to the company As data analysts, we organize and evaluate data using statistical methods and machine learning algorithms. We will analyze data to find trends and linkages that can enhance employee satisfaction and productivity.

A Center for American Progress research on employee turnover found that replacing a person costs 20% of their pay. Free or low-cost workplace flexibility and earned sick days can reduce attrition and save money.

Thus, most companies find employee replacement costly. Interviewing and finding a successor, sign-on bonuses, and months of lost production while the new hire adjusts to the company

## 

## Figure 1

## (‘There Are Significant Business Costs to Replacing Employees’, 2012)

# Business Understanding

Understanding why and when employees depart might help retain them. The reasons people quit a job can help us create a plan to boost employee satisfaction and productivity. This project falls under "HR Analytics" or "People Analytics".

This report aims to tackle the problem statement:

* What are the factors behind former employees leaving the company?
* What are the key signs of employee turnover?
* What policies or tactics may be implemented based on the results to boost employee happiness and productivity?

Since we have data on previous employees, this is a normal supervised classification problem with a binary label: 0 (current employee), 1 (former employee).

# Data Understanding

The datasets used in this portfolio were created by [IBM](https://www.kaggle.com/datasets/rohitsahoo/employee) data scientists to analyze the factors that lead to employee attrition.

Age: Age of the Employee  
Attrition : Employee who stayed: 0 , Employee who leave: 1  
Business Travel: ‘Travel\_Rarely’ ‘Travel\_Frequently’ ‘Non-Travel’  
Daily Rate : Daily Rate of Employee  
Department : ’Sales’ ‘Research & Development’ ‘Human Resources’  
Disfranchisement : Distance from home from work for each Employee  
Education: 1:Below College, 2: College, 3:Bachelor, 4; Master, 5:Doctor  
Education Field: Life Sciences’ ‘Other’ ‘Medical’ ‘Marketing’ ‘Technical Degree’  
‘Human Resources’  
Environment Satisfaction : 1: Low, 2: Medium, 3: High, 4 :Very High  
Gender : Female or Male  
Hourly Rate : Hourly Rate of Employee  
Job Involvement: 1: Low, 2 :Medium, 3 :High, 4 :Very High  
Job Level  
Job Role : ‘Sales Executive’ ‘Research Scientist’ ‘Laboratory Technician’  
‘Manufacturing Director’ ‘Healthcare Representative’ ‘Manager’  
‘Sales Representative’ ‘Research Director’ ‘Human Resources’  
Job Satisfaction: 1: Low, 2 :Medium, 3 :High, 4 :Very High  
Marital Status : ‘Single’ ‘Married’ ‘Divorced’  
Monthly Income : Monthly income of Employee between 2094 and 26999.  
Monthly Rate  
Num Companies Worked : Number of Companies for the employee work before the current one.  
Over18 : ’Y’  
Over Time : ‘N’  
Percent Salary Hike : Percentage of Salary increase between %11-%25.  
Performance Rating : 1 :Low, 2 :Good, 3 :Excellent, 4 :Outstanding  
Relationship Satisfaction 1: Low, 2 :Medium, 3 :High, 4 :Very High  
Standard Hours : standard work hour for each employee: 80 Hours  
Stock Option Level : It categorized from 0 to 3 indicate the stock level of employee  
Total Working Years : Employee total working years and it varies between 0 to 40 years.  
Training Times Last Year : Employee training time in the last year.  
Work Life Balance 1-Bad, 2-Good, 3-Better, 4-Best  
Years At Company: Employee total working year at the company and it varies between 0 to 40 years.  
Years In Current Role : Employee current position at the company and it varies between 0 to 18 years.  
Years since Last Promotion: The time the employee get the last promotion and it varies between 0 to 15 years.  
Years With Curr Manager : The time for time employee working with current manager and it varies between 0 to 17 years.

The dataset consists of 35 features and 1470 attributes.

# Libraries Used

The libraries were used for data handling & analysis, data visualization, data prepossessing, data modeling, model helpers, performance metrics & Stats model.

This list is a sample of the libraries used:

* Pandas- Seaborn
* Sklearn- Sklearn Feature extraction-
* Sklearn Metrics Stats model

# Overall Objectives

The objective of the report is to develop features(reasons) why employees may leave the company. We will use these features as a proxy to identify areas where the company can look to improve productivity. To achieve this we will need to use the machine learning models as a risk management tool. Rather that a predictive tool. By doing list we can create features which we can then undertake statistical modeling to ascertain the validity of these features. From this we can create actionable suggestions to the company to improve satisfaction and productivity

# Characterization of the Data

* 1470 attributes x 35 features
* data types float64 & objects
* Missing values detected in dataset's
* 51,450 observations

# Missing Values

The dataset contains 147 missing values across the entire dataset. Machine learning models will not correctly if they are presented with missing values, any decision made with regards to imputation will have far reaching consequences when we present final models.

Common Types of Missing Data

* Missing Completely at random(Mcar) missingness is unrelated to any other variables making it easy to handle.
* Missing at Random(Mar) The probability of missing values depends on other variables making it more difficult to handle.
* Not Missing at Random. Missingness related to variable not observed making it more difficult to handle.

For the purpose of this project that the data is missing completely at random.

When dealing with missing values we need to consider the following

* Which imputation method can we use to minimize bias & noise in to the models
* How can we evaluate the chosen imputation method to insure the distribution of the data is maintained as far as possible.

Our Strategy

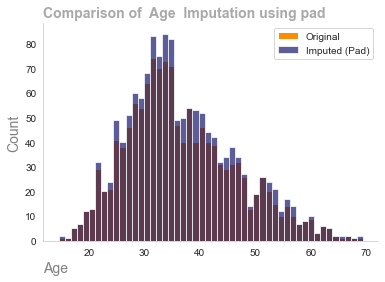
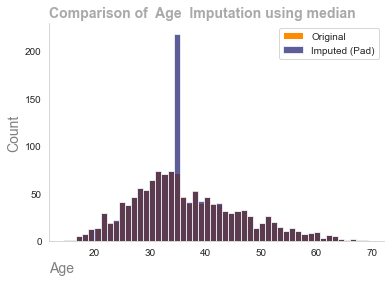
We have multiple imputation options available:

* Deletion Strategy
* Mean/ Median imputation
* Mode Imputation
* Interpolation (pad, fill, backward fill)
* Predictive Modeling(KNN)

Each of the above models have their individual pros and cons.

We will first deal with numerical values. Our strategy is use the imputation method that is the simplest to use (Occam Razor Principle) while not been so simple as to introduce unwanted noise it the data.

1. Delete all missing data. This method very simple to implement however we are only left with 1200 instances.
2. Mean. We can use the mean imputation this will not be useful from inspection out earlier histograms, many off the features have a tail to the would suggest that there are outliers in the database
3. Use interpolation pad method to by using values from neighboring data points. This method was successful for the majority of the features First we used the describe to get a summary of the data before we used the pad method. We then imputed the data ran the describe function again and compared the results.
4. As we can observe from the below graph the pad method did not work on the daily rate, monthly rate & monthly income features by overlaying the original distribution with the distribution we can observe if the median imputation method has been chosen has successful

Figure 2

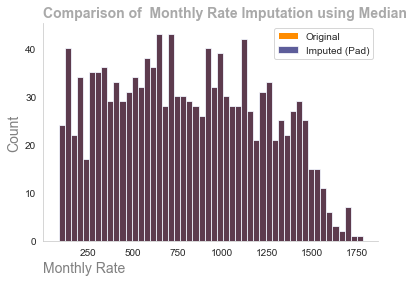
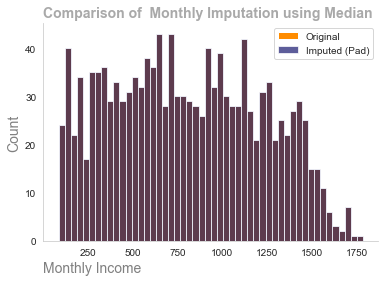
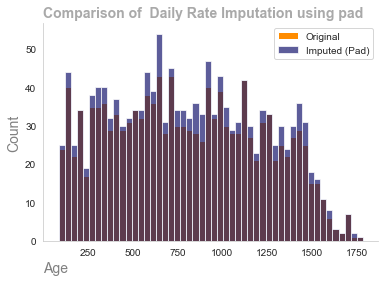
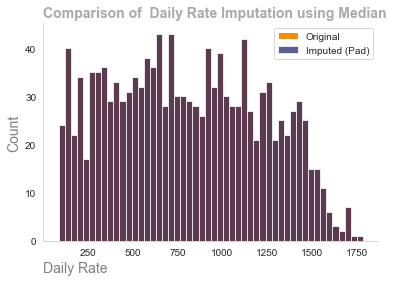


Figure 3



## Figure 4

We can observe from the visualizations the importance using different imputation methods. The critical part is to understand the nature of the missingness of the data and also understand how the data is distributed.

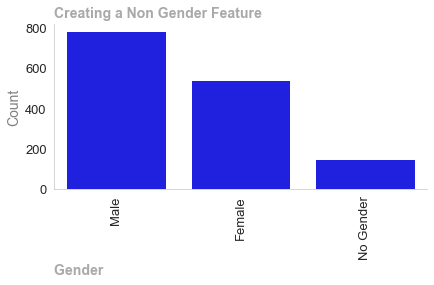
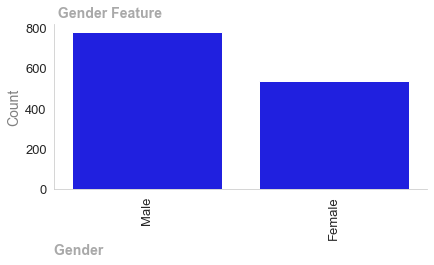
# Categorical Data

The most commonly method used for categorical data is the mode imputation (the most common occurring value), this method will work well with the majority of data however a problem may arise where there is only two classes such as gender (Male or Female) the majority class will become imbalanced due to the incorrect imputation method being applied this will have the introduce considerable noise in to the data making the statistical models & machine learning models sub optimal.

Our Strategy:

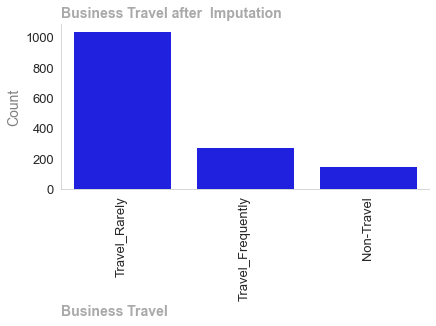
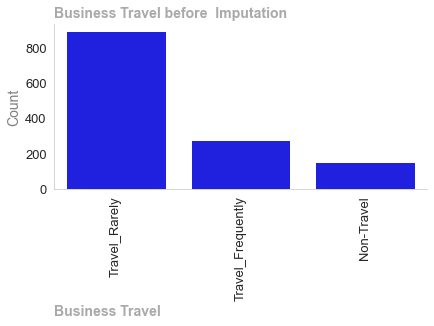
We have 8 features contain categorical database

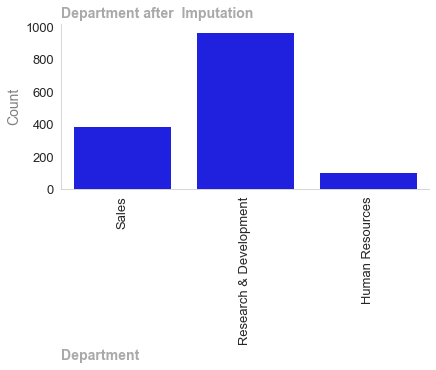
1. Assume mode will mode be the correct method of imputation
2. Evaluate results through the use of histograms to ensure the distribution of the original data has been maintained
3. If the distribution has not been maintained work on models such as predictive modeling.

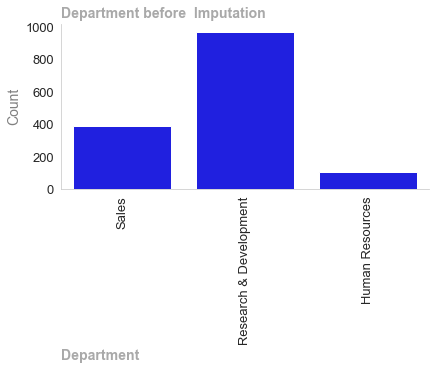


## Figure 5

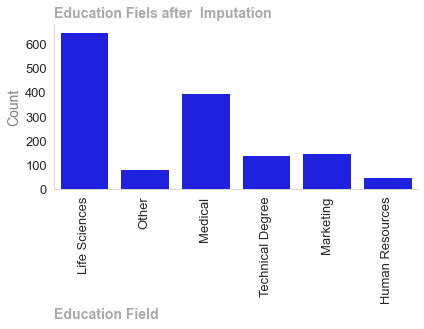
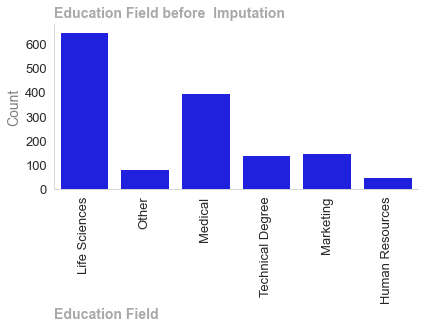
We can observe the problem we spoke about if we apply model imputation method to the gender feature the male column will increase exponentially. We need to consider another method which could be predictive modeling such as KNN. However as data analyst we need to be a ware of modern society. Was the missingness created random or was it created by a person stating they were non binary. If we simply impute we could we misrepresenting the data a valid action would be go back to the company and ask the question. Is this missingness due to people within there company wishing to be defined as non binary. Due to this uncertainty the safest method would be to introduce a third class as non gender. The benefits of this are we have the changed the distribution of the original data and by creating a third class it may help us gain insights into the changing demography of the company.



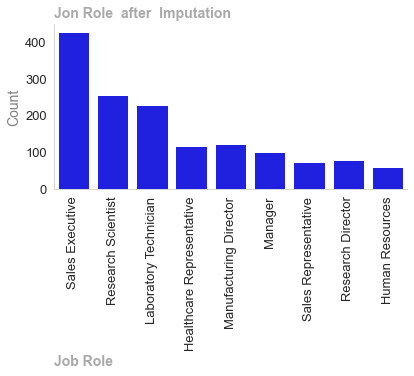


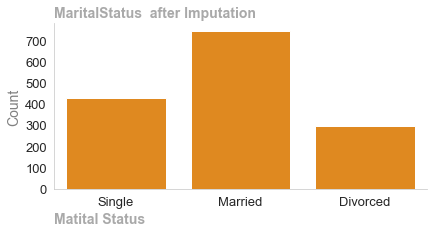
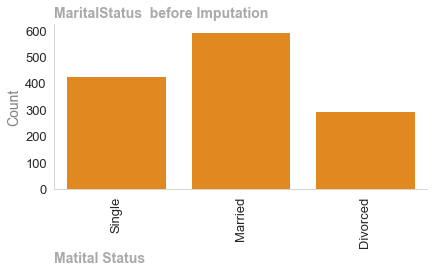


## Figure 7



## Figure 8





## Figure 10

We have used the mode imputation method on the remaining features as can be observed the distribution of the data has been maintained.

The attrition column which will be our target variable in the classification models we will run has two classed yes or no

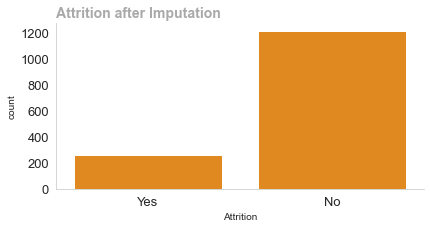
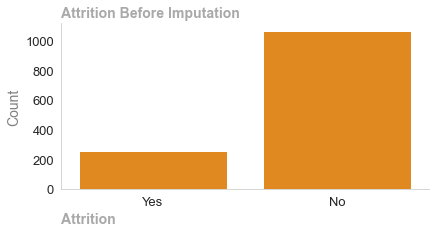


Figure 11

Here we observe where mode and not worked as it has more to imbalance to a unbalanced dataset. The course of action would be use a predictive model such as KNN that uses other features to calculate the missing values. However I could not get the code to work and have added my workings in the appendix of my jupyter notebook. The imputation method was to use mode imputation a method such as under sampling oversampling, smote to deal with the imbalanced attrition class.

# Data Preparation & EDA

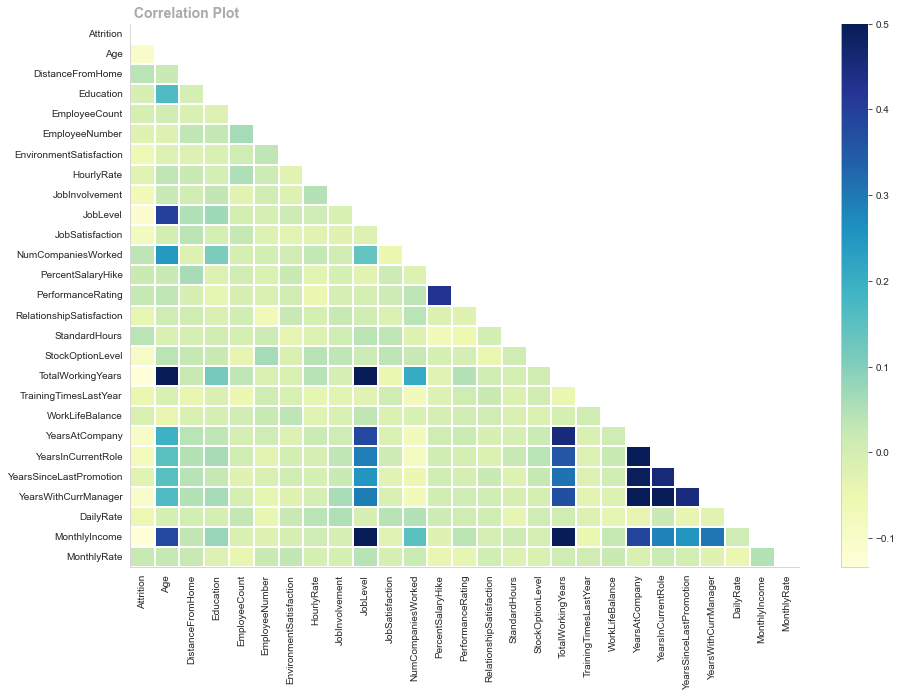
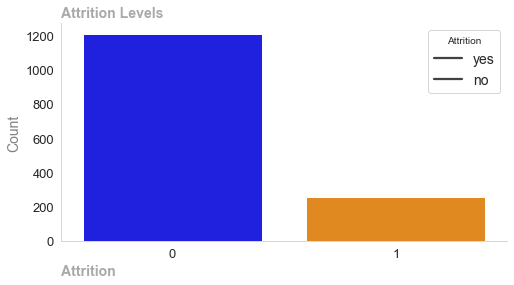
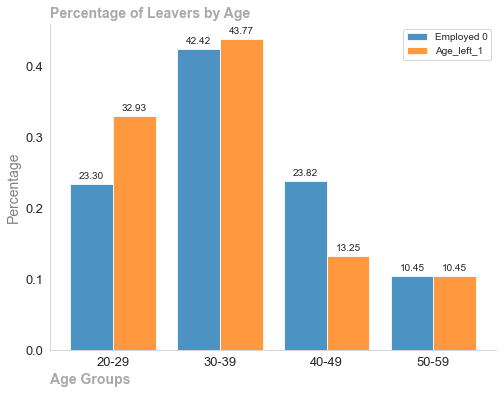


Figure 12

Monthly Income, Total Working Years, and Job Level are adversely connected with Attrition, but there are no positively correlated factors. A correlation plot without positive correlations shows a lack of linear linkage, not no link. Non-linear correlations may help us comprehend date relationships. Non-linear models may be useful if the correlation plot displays non-linear associations or if we suspect non-linear data patterns. Decision trees, random forests, SVNs, and neural networks model complex non-linear connections. This figure shows how feature engineering can reveal relationships in the data that were not obvious before.



## Figure 13

## 

Figure 14

## 

## Figure 14

## 

## 

Figure 15

## 

Figure 16

Employed Left

Age 37.190027 34.609336

DistanceFromHome 9.004461 9.893881

Education 2.924942 2.901589

EmployeeCount 0.999010 0.997946

EmployeeNumber 1030.559051 992.177597

EnvironmentSatisfaction 2.749307 2.560969

HourlyRate 65.729472 64.213450

JobInvolvement 2.765483 2.618081

JobLevel 2.091541 1.761240

JobSatisfaction 2.754480 2.511047

NumCompaniesWorked 2.604728 2.840930

PercentSalaryHike 15.283715 15.475213

PerformanceRating 3.142657 3.174729

RelationshipSatisfaction 2.744992 2.624496

StandardHours 80.079068 80.885349

StockOptionLevel 0.841568 0.633798

TotalWorkingYears 11.730730 9.130775

TrainingTimesLastYear 2.823380 2.673212

WorkLifeBalance 2.781650 2.753256

YearsAtCompany 6.817308 5.528411

YearsInCurrentRole 4.370313 3.649021

YearsSinceLastPromotion 1.939663 1.768886

YearsWithCurrManager 4.248773 3.281085

DailyRate 810.653894 750.652016

MonthlyIncome 6288.552492 4975.772016

MonthlyRate 14281.097178 14694.936783

# Exploratory Data Analytics

# The strongest positive correlation with the attrition features are: TotalWorkingYears, Age, PerformanceRating, Job Level, Monthly Income, Years in Current Role, Years since last promotion, years with curr manager.

# The dateset is imbalanced with the majority of observations being described as employed.

# 

# Age group 20-29 more likely to leave than an other group

# Majority of people who leave will do so before 4 year anniversary.

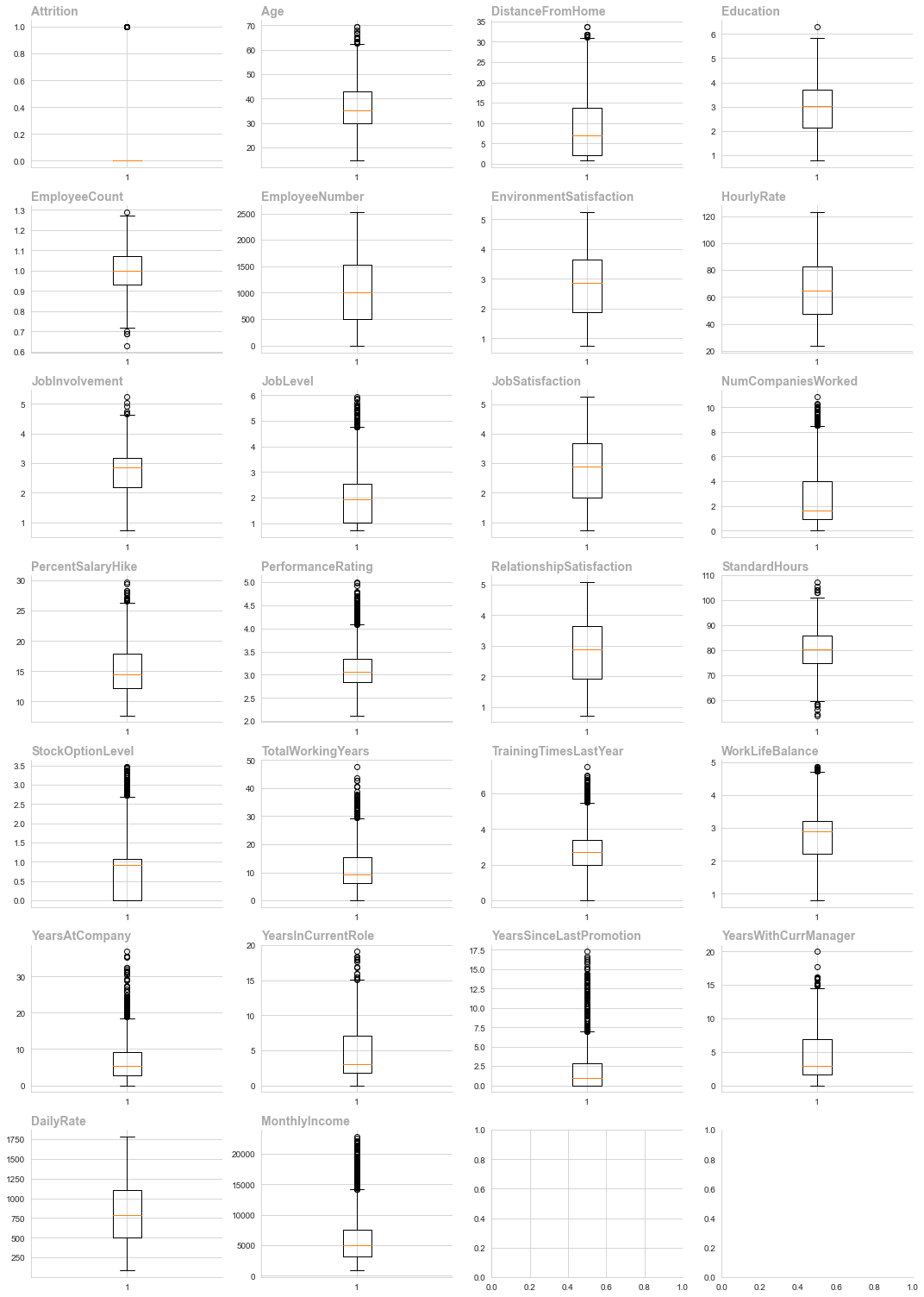
# People who live further away from their work show higher proportion of leavers compared to their counterparts.

* People who left did not receive any stock options
* People who left did had a lower JobSatisfaction, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobSatisfaction than people who remained
* Employees that have already worked at several companies previously (already “bounced” between workplaces) show higher proportion of leavers compared to their counterparts.

# 

# Outliers

## Outliers must be addressed to ensure statistical analysis integrity, model accuracy, data quality, and meaningful data exploration and interpretation. Data-driven insights are more credible when outliers are identified and addressed.



Most numerical columns in box-plots have outliers, which must be eliminated for machine learning to work in an optimal manner

Removing Outliers using IQR

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

Outliers (Previously) Outliers Count Column Lower Limit Upper Limit

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

True False 15 Age 10.3288 62.3787

True False 10 DistanceFromHome -15.1462 31.0037

True False 6 JobInvolvement 0.695 4.655

True False 50 JobLevel -1.19125 4.75875

True False 45 NumCompaniesWorked -3.58 8.5

True False 15 PercentSalaryHike 3.6625 26.4025

True False 58 TotalWorkingYears -7.9725 29.3675

True False 68 TrainingTimesLastYear -0.14125 5.48875

True False 12 WorkLifeBalance 0.71 4.71

True False 93 YearsAtCompany -6.9325 18.7275

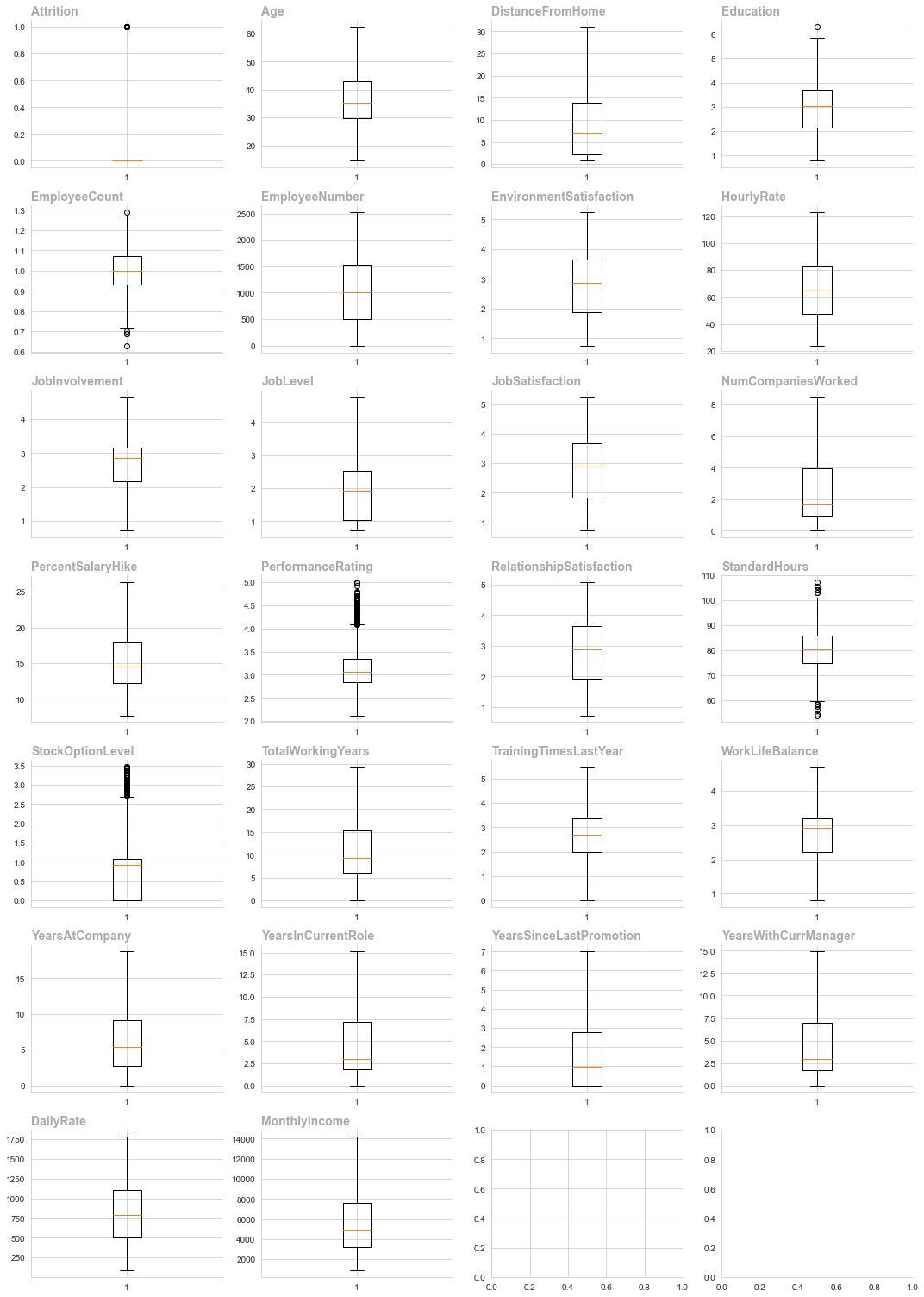
True False 15 YearsInCurrentRole -6.14625 15.1037

True False 153 YearsSinceLastPromotion -4.19625 6.99375

True False 14 YearsWithCurrManager -6.28625 14.9037

True False 135 MonthlyIncome -3438.8 14205.20

We used IQR to remove outliers. Standard Hours, Stock Option level, and performance rating are still outliers that must be addressed through log transformation before machine learning.

.

Summary of Dealing with Outliers

Summary of How outliers were dealt and why

* Visual inspection (box plots), statistical methods (Z-score or IQR)
* Understanding Outliers: Determine if outliers are errors, oddities, or important data points. Consider domain knowledge to assess relevance.
* Handling Outliers: Choose a context-specific outlier strategy. Alternatives include elimination, transformation, and robust modeling.
* Assessment of Impact: Consider how outliers affect your study or modeling. Compare your model's performance with and without outliers to see how they affect outcomes.
* Domain Consultation: Domain specialists can advise on outlier importance in the context. Domain knowledge helps decide whether to keep or treat outliers.
* Process iteration: Iterating on outliers is common. Try multiple methods, evaluate their results, and adjust your strategy based on your analytical goals.
* Model chosen Inter Quartile Range(IQR) We have observed that our data was skewed to the right from earlier analysis and this has also been confirmed by visual inspection of box-plots. We have chosen the IQR method for the following reasons:

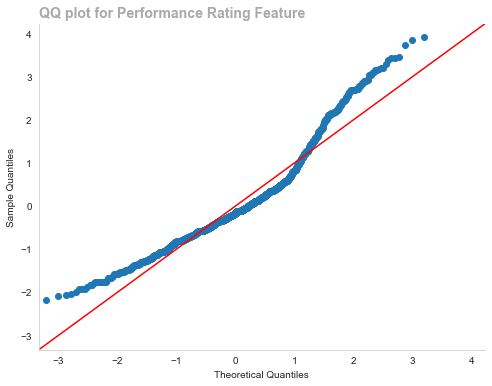
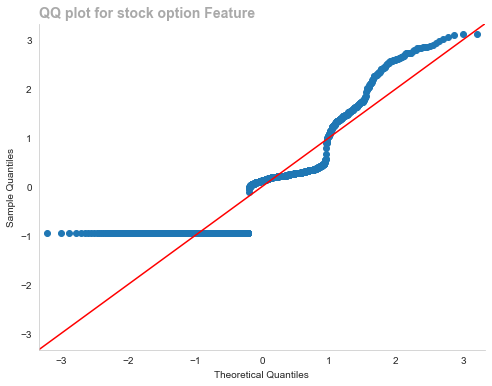
Ideal for outlier datasets, IQR is resilient to extreme values. Without considering extreme values, it determines the range between Q1 and Q3.

IQR is beneficial for skewed or non-normal data. It is less sensitive to extreme values than mean and standard deviation, which are highly influenced by outliers.

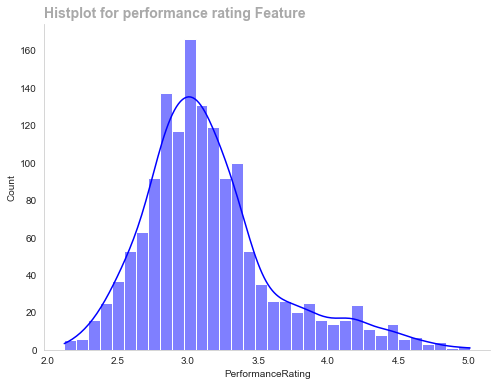
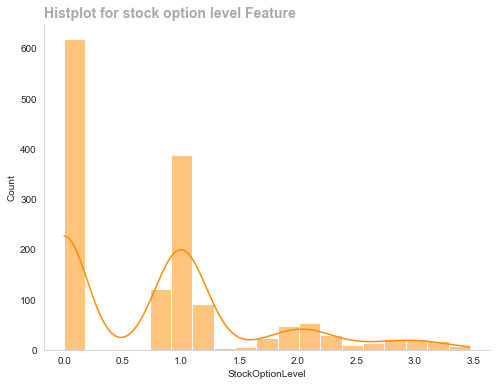
A non-parametric measure like IQR doesn't assume a data distribution. This makes it suitable for many datasets, including ones with non-standard distributions such as ours.

Box plots with IQR show data spread. This visual technique can broaden outlier awareness.

# Check For Normality



## Figure 20



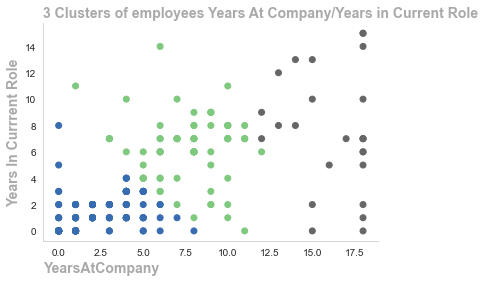
## Figure 21

Statistical analysis requires checking data normalcy for numerous reasons. Validating the assumption of a normal distribution for statistical tests like t-tests and ANOVA is essential for dependability. If data is regularly distributed, parametric tests are more effective. Normal distributions simplify statistical inference by making outcomes interpret able.

Even if the population distribution is not normal, the Central Limit Theorem suggests that the sample mean distribution is. To use this theorem, normality checks verify the sample size is sufficient. Another benefit of normalcy screening is detecting outliers and skewness, which can affect statistical analysis.

Regression analysis relies on normality checks to verify residual distribution. This validates regression results. Inference accuracy improves with regularly distributed data, including confidence intervals and hypothesis tests. Normality is not always required, especially with large sample sizes, however screening for it improves statistical studies.

# Cluster Analysis



## Figure 22

Cluster analysis is a sophisticated EDA tool that helps explore and comprehend complicated datasets, find patterns, generate hypotheses, andimprove decision-making by revealing data structures. Cluster analysis shows that youthful employees leave the organization within 5 years

# Data Preprocessing for Linear Discriminant Analysis & PrinciplaComponent Analysis

Preprocessing LDA or PCA requires encoding and scaling. Encoding translates all variables into numeric notation and scaling addresses variable scale concerns, making these dimensionality reduction strategies stable, efficient, and interpret able.

# Linear Discriminant Analysis

# Dimensionality reduction methods Linear Discriminant Analysis(LDA) and Principal Component Analysis(PCA) have different objectives assumptions on data classification and groupings

LDA Objectives:

* LDA is developed for supervised learning to maximize dataset class separation. This separation forms decision boundaries where new data can be positioned
* Class Separation: LDA maximizes between class variance to within class variation(IE employed & leavers), if finds a projection that maximizes class mean distance while minimizing data point spread within each class
* Class Distribution LDA assumes multivariate normal distribution for Attrition class, this is why we used SMOTE to balance our class using synthetic data. It assumes all classes have the same covariance matrix.
* LDA finds characteristics that maximize class separation it excels at categorizing data as can be seen from our visualizations.

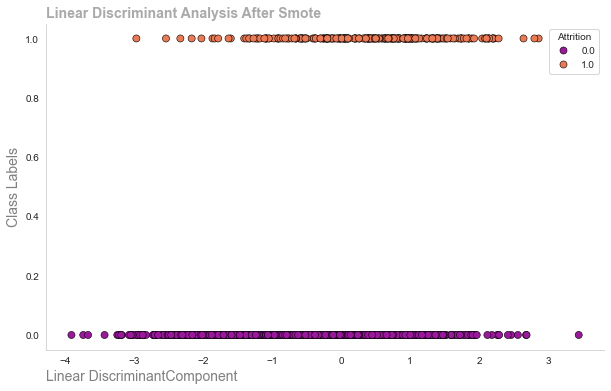
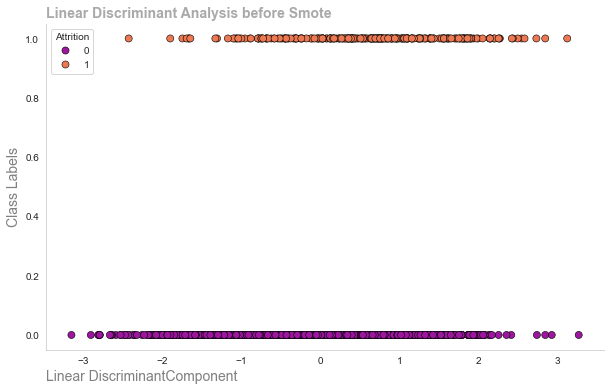


Figure 24

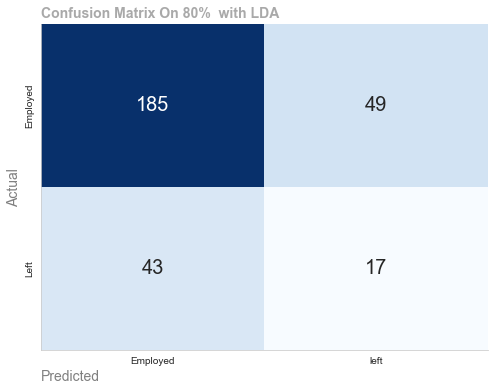
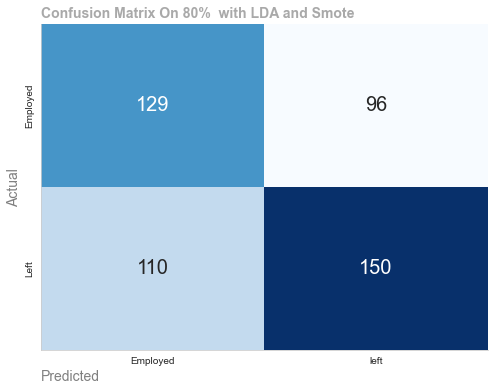


Figure 23

Principal component analysis PCA:

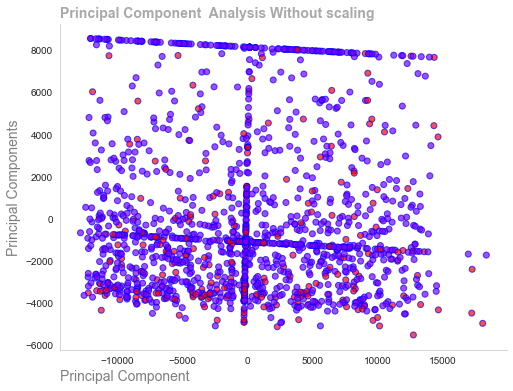
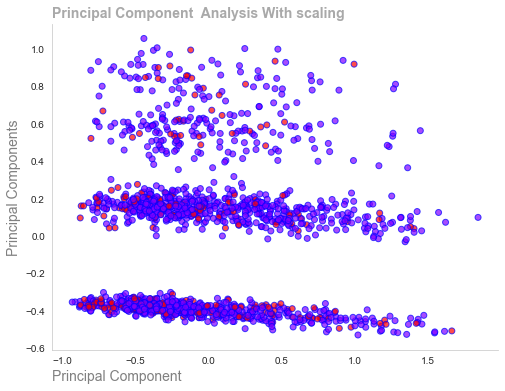
* PCA is an unsupervised method that maximizes dataset variance regardless of class labels.
* Eigenvalue Maximization: PCA finds orthogonal axes(principal components that maximize data variance.
* PCA reduces dimensionality without considering class labels

Implications for Clustering or Classification:

* PCA reduces data dimensionality during preprocessing. This helps capture the most important components across the dataset
* PCA does not include class labels, yet its reduced dimensional data can be used later in clustering techniques.
* LDA feature extraction is used for classification problems, especially where class separability is important, it focuses on decision boundaries which help improve classification accuracy.

PCA for Dimensionality Reduction:

* Data is commonly dedimensionalized using PCA before classifiers or clustering algorithms
* The curse of dimensionality is mitigated and computational complexity is reduced



## Figure 25

Combining LDA/PCA

* LDA and Pca are sometimes used together. LDA improves class separability and PCA reduces dimensionality

In conclusion, LDA & PCA achieve different objectives and assumptions and LDA can be used for supervised classification algorithms where PCA can be used for unsupervised algorithms such as clustering.

# Feature Engineering Binning

Feature engineering is a critical aspect of Data Analytics, it involves creating new features or modifying current ones to improve the performance of machine learning models and enhance the interpret ability of the models

Improved model performance

* Relevance: Create features that are relevant to the problem at hand.

Noise Reduction Feature engineering can help remove noise from the database.

* Handling Non Linearity: Introduce features that can help capture non linear data improving model performance.
* Creating composite Features: Combine multiple features into one feature which can simplify the model and improve interpret ability.
* Addressing Data Skewness: Apply transformation to features can help address issues which may arise due to the skews of the data.
* Dimensionality Reduction: Choosing the most important features can help remove the dimensionality of the data making the model more efficient and less prone to over fitting.

In the context of data analytics more data is always more preferable and we need to be aware that we not mutate the date to the point where it is unusable.

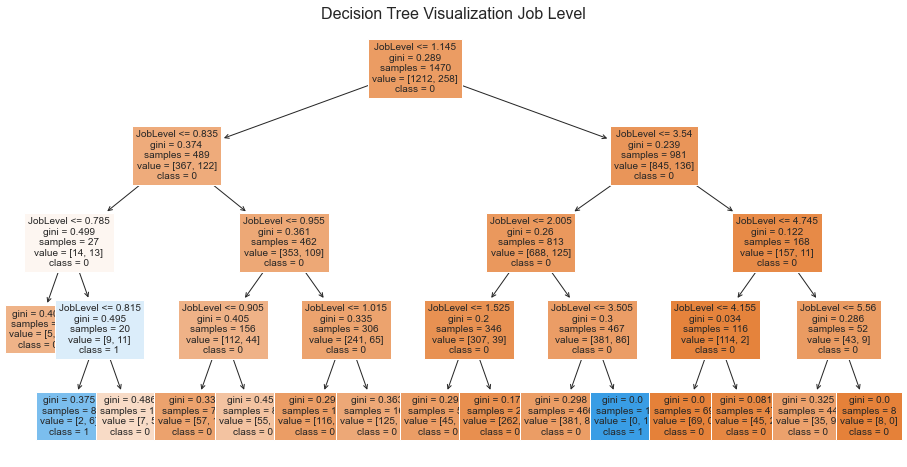


Figure 26

We used Figure binning to reduce the complexity of the data on a series of the features. Bu using a decision tree classifier we could decide the appropriate bins for the features. This process improves the AUC score from 62.00 to 64.8. We could used other techniques such as arranging features into classes medium high low etc however this method allows us to systematically decide the correct bins. Every numerical feature was tested if it improved the model it was kept otherwise disregarded.

# Machine Models Data Preprocessing & Encoding

Scaler: StandardScaler

Accuracy: 0.8401, Precision: 0.6098, Recall: 0.5769

Scaler: MinMaxScaler

Accuracy: 0.8401, Precision: 0.6098, Recall: 0.5769

Scaler: RobustScaler

Accuracy: 0.8401, Precision: 0.6098, Recall: 0.5769

Scaler: MaxAbsScaler

Accuracy: 0.8401, Precision: 0.6098, Recall: 0.5769

Scaling is necessary to prevent variable values from affecting machine learning models or statistical methods. Stability, convergence, and interpret ability of analytical methods are improved We used a dummy classifier to test different scaling methods for our models to ascertain the best one. As can be observed the scalers returned all the same scores

Categorical variable data preprocessing requires encoding. It ensures algorithm compatibility and prevents misinterpretations from non-numeric input by include vital information in the analysis. We used a mixture of label encoding where features to reduce the size of the model and thus reduce its complexity. For the other features we used one hot encoding however this led to 48 columns with many 0s which is not ideal if the model performs badly this may be the cause.

# Train Dummy Classifier

Accuracy Dummy Classifier: 0.68

Classification Report Dummy classifier:

precision recall f1-score support

0.0 0.82 0.77 0.80 242

1.0 0.18 0.23 0.20 52

accuracy 0.68 294

macro avg 0.50 0.50 0.50 294

weighted avg 0.71 0.68 0.69 294

# Results from the dummy classifier which we use as a baseline this allows us to quantify the performance of our models

# Machine Learning Models

Machine Learning Models. Supervised Learning

Supervised learning works best with labeled input-output pairs so the algorithm cab learn from the input features to predict output. This is applicable in classification and regression. As our dataset contains binary labels(yes, no) as such we have a supervised classification with two class binary labels.

Superior predictive accuracy is achieved by training supervised learning modes to predict future events using labeled data. The algorithm learns labeled data patterns and relationships within the features on the data using statistical calculations such as distance between features and then produces well defined outputs. However supervised learning relies on labeled data for training, which can be costly and time consuming depending on the size of the dataset. The model can only predict outputs it has observed making it unsuitable for emerging classes.

Unsupervised Learning

Unsupervised learning works without labels. The algorithm finds data patterns and structures without instructions. Cluster Algorithms such as KNN work on distanced between data points to form decision boundaries where new data can be fitted. Unsupervised learning works in cases where labels are difficult or expensive. It can highlight patterns structures and linkages that a labeled dataset may miss. However unsupervised learning results are evaluated subjectively and based on metrics such as centroids (distance from the center) as there are no labels. In some instances due to no labels unsupervised models may not be as precise as supervised tasks in some tasks

Results of Models with 20% & 30% split

|  | index | Accuracy | Precision | Recall | F1-Score | AUC-ROC | AUC-PR | Training Time (s) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Naive Bayes 20% split | 0.691 | 0.691 | 0.765 | 0.726 | 0.725 | 0.709 | 0.001 |
| 1 | AdaBoost 20% split | 0.676 | 0.665 | 0.800 | 0.726 | 0.719 | 0.694 | 0.217 |
| 2 | LDA 20% split | 0.689 | 0.717 | 0.692 | 0.705 | 0.723 | 0.689 | 0.002 |
| 3 | SVM 20% split | 0.682 | 0.704 | 0.704 | 0.704 | 0.718 | 0.698 | 1.019 |
| 4 | CatBoost 20% split | 0.682 | 0.707 | 0.696 | 0.702 | 0.722 | 0.697 | 4.239 |
| 5 | Gradient Boosting Classifier 20% split | 0.676 | 0.700 | 0.692 | 0.696 | 0.718 | 0.697 | 0.485 |
| 6 | Logistic Regression 20% split | 0.680 | 0.713 | 0.677 | 0.694 | 0.723 | 0.689 | 0.005 |
| 7 | XGBoostn 20% split | 0.602 | 0.638 | 0.596 | 0.616 | 0.677 | 0.681 | 0.194 |
| 8 | Random Forest 20% split | 0.575 | 0.610 | 0.577 | 0.593 | 0.621 | 0.628 | 1.301 |
| 9 | Descision Trees 20% split | 0.575 | 0.610 | 0.577 | 0.593 | 0.575 | 0.579 | 0.016 |
| 10 | k-NN 20% split | 0.577 | 0.616 | 0.562 | 0.588 | 0.628 | 0.616 | 0.002 |

|  | index | Accuracy | Precision | Recall | F1-Score | AUC-ROC | AUC-PR | Training Time (s) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Naive Bayes 30% split | 0.691 | 0.691 | 0.765 | 0.726 | 0.725 | 0.709 | 0.002 |
| 1 | AdaBoost 30% split | 0.676 | 0.665 | 0.800 | 0.726 | 0.719 | 0.694 | 0.237 |
| 2 | LDA 30% split | 0.689 | 0.717 | 0.692 | 0.705 | 0.723 | 0.689 | 0.002 |
| 3 | SVM 40% Split | 0.682 | 0.704 | 0.704 | 0.704 | 0.718 | 0.698 | 1.104 |
| 4 | CatBoost 30% split | 0.682 | 0.707 | 0.696 | 0.702 | 0.722 | 0.697 | 4.236 |
| 5 | Gradient Boosting Classifier 30% split | 0.676 | 0.700 | 0.692 | 0.696 | 0.718 | 0.697 | 0.525 |
| 6 | Logistic Regression 30% split | 0.680 | 0.713 | 0.677 | 0.694 | 0.723 | 0.689 | 0.004 |
| 7 | XGBoost 30% Split | 0.602 | 0.638 | 0.596 | 0.616 | 0.677 | 0.681 | 0.202 |
| 8 | Random Forest 30% Split | 0.575 | 0.610 | 0.577 | 0.593 | 0.619 | 0.631 | 1.309 |
| 9 | Descision Trees 30% split | 0.575 | 0.610 | 0.577 | 0.593 | 0.575 | 0.579 | 0.019 |
| 10 | k-NN 30% split | 0.577 | 0.616 | 0.562 | 0.588 | 0.628 | 0.616 | 0.002 |

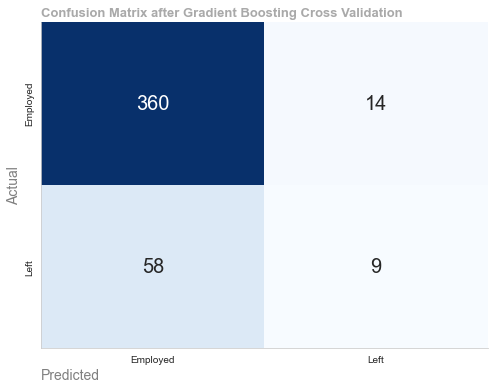
AUC-ROC is a versatile statistic that evaluates a classification model's performance, especially under class imbalance, threshold variability, and model comparison situations. Three logistic regression models perform well on the 30% & 20% split. Gradient boosting, naive Bayes. On training times, logistic regression and naive Bayes make sense. We must employ machine learning models to manage risk for our problem. Our purpose is to measure company churn. We then want to use these statistics to identify ways a company might boost productivity and satisfaction. The logistic regression model and gradient boosting boosting model will be used to generate feature importance. We can then conduct statistical modeling. on these importances to verify them and use them to create a data-driven solution.

# Cross Validation Best Models

Cross\_Validation mean Accuracy: 0.8115

Cross\_Validation mean Precision: 0.5214

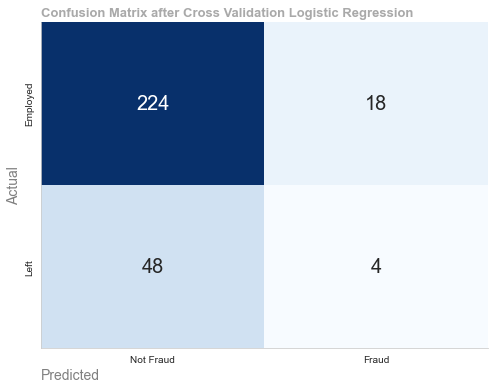
Cross\_Validation mean Recall: 0.1312



## Cross\_Validation mean Accuracy: 0.8291

Cross\_Validation mean Precision: 0.5452

Cross\_Validation mean Recall: 0.1458



## Figure 27 & 28

## Hyperparameter Tuning

Accuracy after hyperparameter training: 0.85

Classification Report after hyperparameter training:

precision recall f1-score support

0.0 0.85 0.99 0.92 374

1.0 0.44 0.06 0.11 67

accuracy 0.85 441

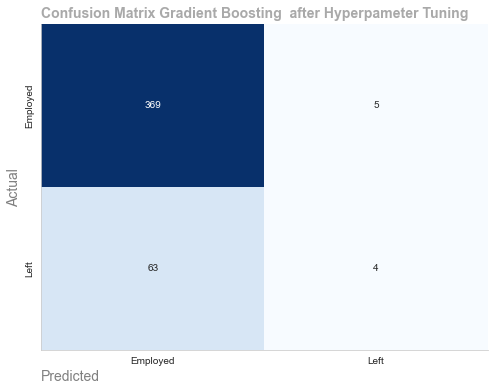
macro avg 0.65 0.52 0.51 441

weighted avg 0.79 0.85 0.79 441

Precision after hyperparameter training: 0.44

Recall after hyperparameter training: 0.06

F1 Score after hyperparameter training: 0.11



## Figure 29

Accuracy after hyperparameter training: 0.88

Classification Report after hyperparameter training:

precision recall f1-score support

0.0 0.87 1.00 0.93 242

1.0 1.00 0.33 0.49 52

accuracy 0.88 294

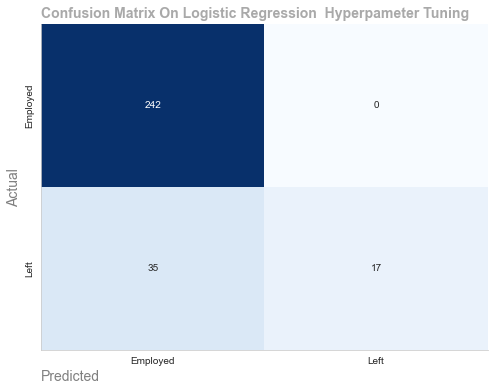
macro avg 0.94 0.66 0.71 294

weighted avg 0.90 0.88 0.85 294

Precision after hyperparameter training: 1.00

Recall after hyperparameter training: 0.33

F1 Score after hyperparameter training: 0.40



## Figure 30

## Model performance can vary greatly with hyperparameter settings. Hyperparameter tuning identifies settings that improve predicted accuracy and generalization to new data. Model creation requires hyperparameter tuning to optimize machine learning models for predictive accuracy, generalization, and practical deployment. It involves methodical exploration of hyperparameter space to discover settings that match data properties and modeling. goals.

# Evaluation Metrics

Accuracy test set: 0.88

Classification Report on test set Gradient Boosting:

precision recall f1-score support

0.0 0.88 0.99 0.93 374

1.0 0.89 0.25 0.40 67

accuracy 0.88 441

macro avg 0.89 0.62 0.67 441

weighted avg 0.88 0.88 0.85 441

Precision test set: 0.89

Recall test set: 0.25

F1 Score on test set: 0.40

# 

## Figure 30

## Accuracy test set: 0.83

Classification Report on Logistic Regression test set:

precision recall f1-score support

0.0 0.84 0.98 0.90 242

1.0 0.57 0.15 0.24 52

accuracy 0.83 294

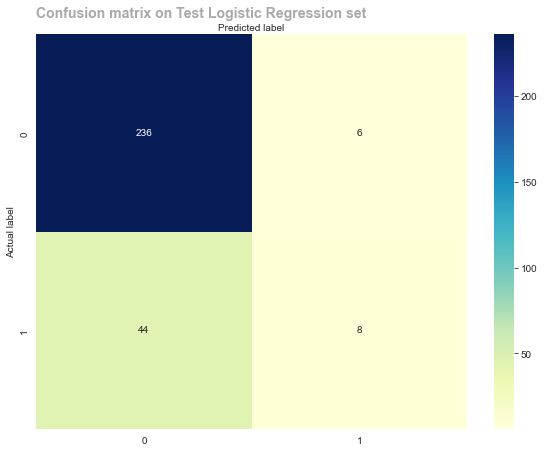
macro avg 0.71 0.56 0.57 294

weighted avg 0.79 0.83 0.79 294

Precision test set: 0.57

Recall test set: 0.15

F1 Score on test set: 0.10

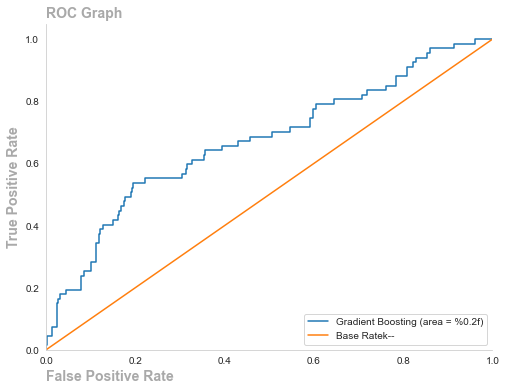


## Figure 32

Accuracy of Gradient Boosting Classifier on test set: 88.21

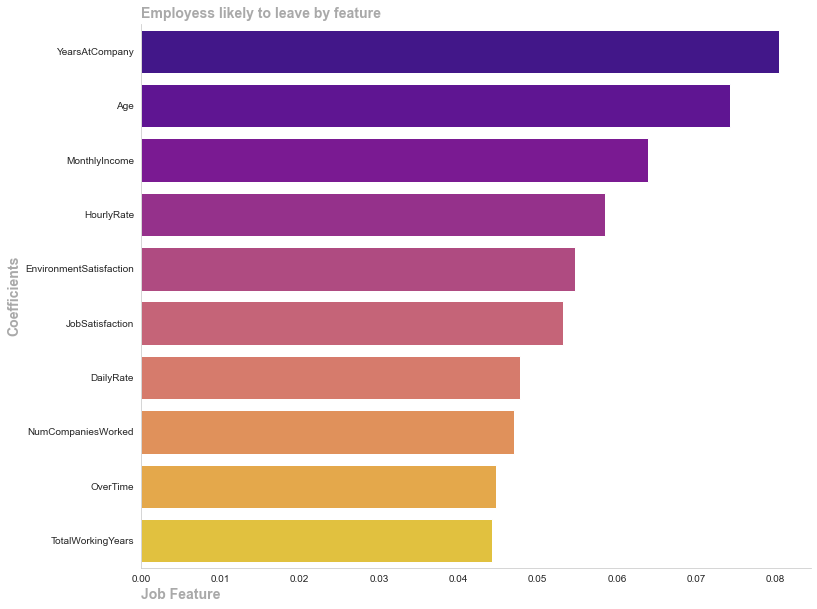
Accuracy of Logistic Regression on test set: 82.99

The gradient boosting model performs best after cross validation and hyperparameter tuning.



**Figure 32**

# Feature Importance's



## Figure 34

# 

# Statistical Modeling

# Statistical modeling using descriptive and inferential statistics helps get data insights.

# Descriptive statistics summarize and describe key elements of a dataset. Descriptive statistics help analyze employee churn variables' central tendency, variability, and distribution. The main descriptive statistics measures are:

# The average value of a variable is its mean.

# The median divides a dataset into two equal half.

# Standard Deviation shows value dispersion around the mean.

# Max and Min Values: Display the dataset's range.

# Divide data into quartiles or percentage intervals.

# Descriptive statistics can show trends and patterns for each within the data.

# Inferential Statistics: Inferential statistics use sample data to make predictions about a population. Inferential statistics can draw conclusions about the total workforce from a subset. Common methods:

# Hypothesis testing: Testing correlations or variable differences.

# Finding the degree and direction of variable correlations using regression analysis.

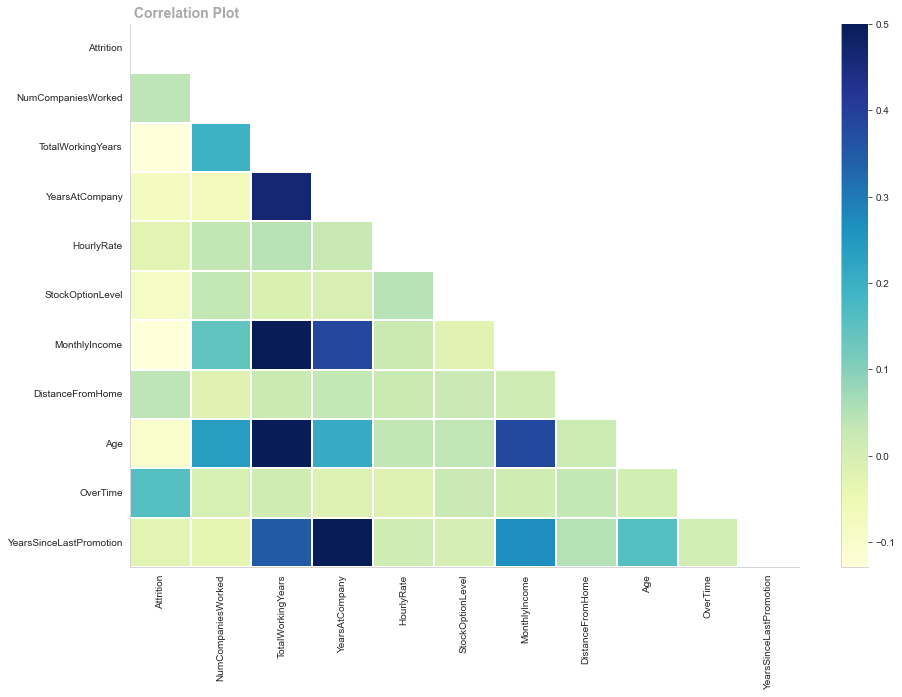
# ANOVA: Comparing data groups.

# Correlation Analysis: Assessing linear data connections.

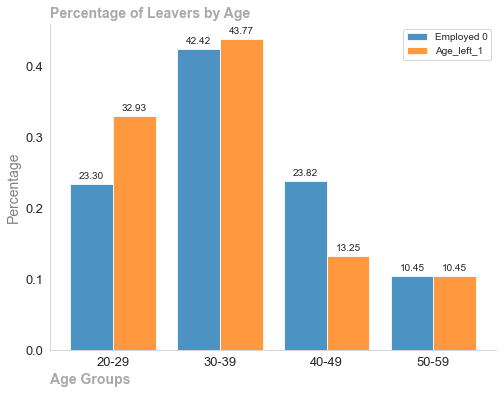
# We may validate assumptions, discover and make informed decisions to address organizational difficulties such as low satisfaction by using inferential statistics.

# In conclusion, using descriptive and inferential statistics in our analysis helps interpret patterns, draw meaningful conclusions, and make data-driven decisions to improve productivity.

|  | count | mean | std | min | 25% | 50% | 75% | max |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Attrition | 1470.0 | 0.18 | 0.38 | 0.0 | 0.00 | 0.0 | 0.0 | 1.0 |
| NumCompaniesWorked | 1470.0 | 2.24 | 2.47 | 0.0 | 0.00 | 1.0 | 3.0 | 10.0 |
| TotalWorkingYears | 1470.0 | 10.99 | 8.04 | 0.0 | 6.00 | 9.0 | 15.0 | 47.0 |
| YearsAtCompany | 1470.0 | 6.43 | 6.05 | 0.0 | 2.00 | 5.0 | 9.0 | 36.0 |
| HourlyRate | 1470.0 | 64.97 | 21.34 | 23.0 | 47.00 | 64.0 | 82.0 | 123.0 |
| StockOptionLevel | 1470.0 | 0.53 | 0.77 | 0.0 | 0.00 | 0.0 | 1.0 | 3.0 |
| MonthlyIncome | 1470.0 | 6385.50 | 4546.83 | 891.0 | 3177.25 | 4957.0 | 7588.0 | 22858.0 |
| DistanceFromHome | 1470.0 | 8.69 | 8.24 | 0.0 | 2.00 | 6.0 | 13.0 | 33.0 |
| Age | 1470.0 | 36.27 | 9.91 | 14.0 | 29.00 | 35.0 | 42.0 | 69.0 |
| OverTime | 1470.0 | 0.27 | 0.44 | 0.0 | 0.00 | 0.0 | 1.0 | 1.0 |
| YearsSinceLastPromotion | 1470.0 | 1.93 | 3.16 | 0.0 | 0.00 | 0.0 | 2.0 | 17.0 |



## Figure 35

Age

Mean Years at Company is:6.427210884353742

Median Years at Company is:5.0

Standard Deviation is:6.045646176270098

Median Years at company is:5.0

Skew is:1.70086148136926

kurtosis is:3.488393990403337

Age Percentiles

[0.25, 0.5, 0.75]th Percentile: 2.0

[0.25, 0.5, 0.75]th Percentile: 5.0

[0.25, 0.5, 0.75]th Percentile: 9.0

Attrition 0 1 All

Age\_binned

20-29 0.233076 0.329317 0.250000

30-39 0.424165 0.437751 0.426554

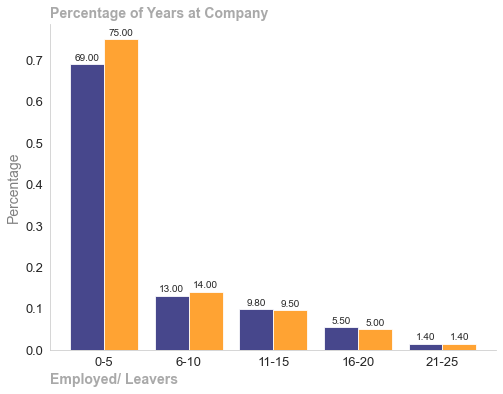
40-49 0.238218 0.132530 0.219633

50-59 0.104542 0.100402 0.103814

All 1.000000 1.000000 1.000000

Figure 35

**Years at Company**

****

Frequency Table for YearsAtCompany

YearsAtCompany

0-5 543

6-10 109

11-15 74

16-20 39

Attrition 0 1 All

YearsAtCompany

0-5 461 82 543

6-10 93 16 109

11-15 66 8 74

16-20 37 2 39

21-25 10 1 11

All 667 109 776

Attrition 0 1 All

YearsAtCompany

0-5 0.691154 0.752294 0.699742

6-10 0.139430 0.146789 0.140464

11-15 0.098951 0.073394 0.095361

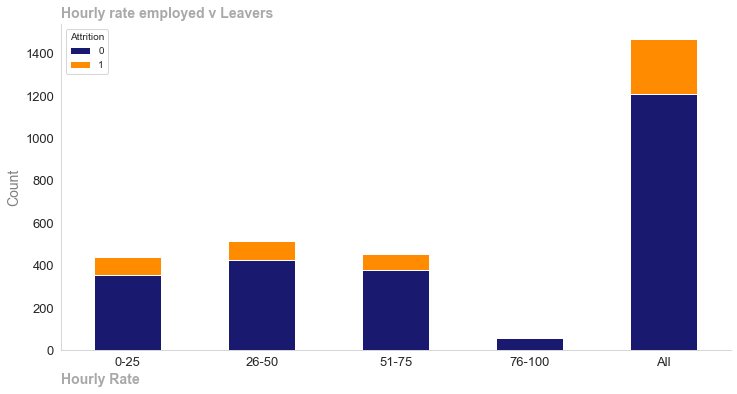
16-20 0.055472 0.018349 0.050258

21-25 0.014993 0.009174 0.014175

All 1.000000 1.000000 1.000000

## Figure 36

**Hourly Rate**



Frequency Table for HourlyRate

HourlyRate

0-25 440

26-50 512

51-75 453

76-100 62

Mean HourlyRate at Company is:64.97414965986394

Median HourlyRate at Company is:64.0

Standard Deviation is:21.343171870449364

Median HourlyRate at company is:64.0

Skew is:0.11440218104277446

kurtosis is:-0.9606806500063194

HourlyRate

[0.25, 0.5, 0.75]th Percentile: 47.0

[0.25, 0.5, 0.75]th Percentile: 64.0

[0.25, 0.5, 0.75]th Percentile: 82.0

Attrition 0 1 All

HourlyRate

0-25 354 86 440

26-50 424 88 512

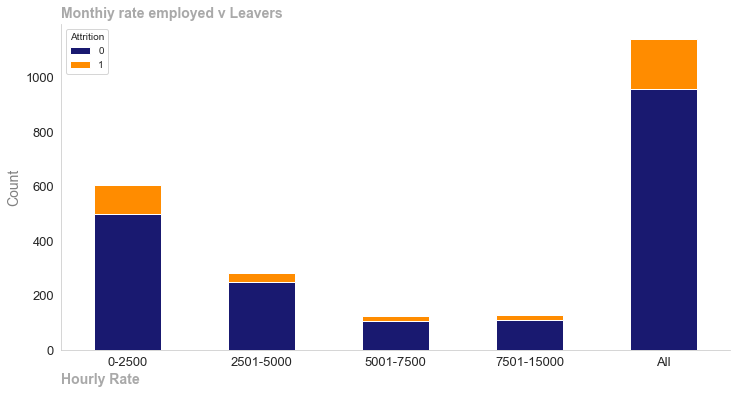
51-75 377 76 453

76-100 55 7 62

All 1210 257 1467

## Figure 37

Monthly Income

Frequency Table for Monthly Income

MonthlyIncome

0-2500 604

2501-5000 283

5001-7500 124

7501-15000 129

Name: count, dtype: int64

Mean MonthlyIncome at Company is:6385.496598639455

Median MonthlyIncomeat Company is:4957.0

Standard Deviation is:4546.827621203466

Median MonthlyIncome at company is:4957.0

Skew is:1.562916997702472

kurtosis is:1.842826615527795

Figure 38

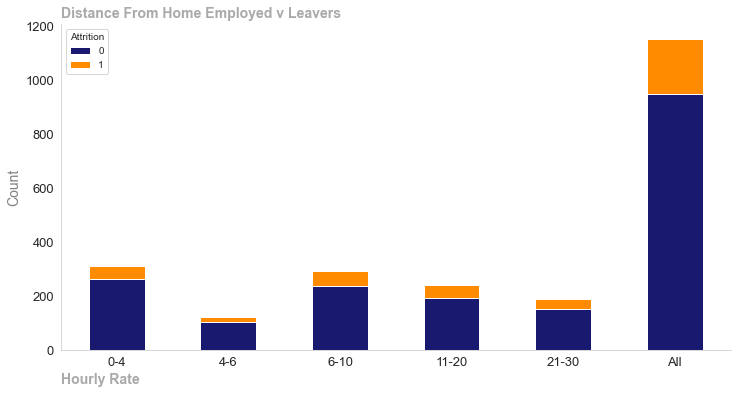
MonthlyIncome

[0.25, 0.5, 0.75]th Percentile: 3177.25

[0.25, 0.5, 0.75]th Percentile: 4957.0

[0.25, 0.5, 0.75]th Percentile: 7588.0

Distance From Home

Mean DistanceFromHome is:8.68639455782313

Median DistanceFromHome is:6.0

Standard Deviation is:8.237639125393605

Median DistanceFromHome is:6.0

Skew is:1.0082165558109475

kurtosis is:-0.06325832026833877

Attrition 0 1 All

DistanceFromHome

0-4 0.277134 0.226601 0.268229

4-6 0.107482 0.093596 0.105035

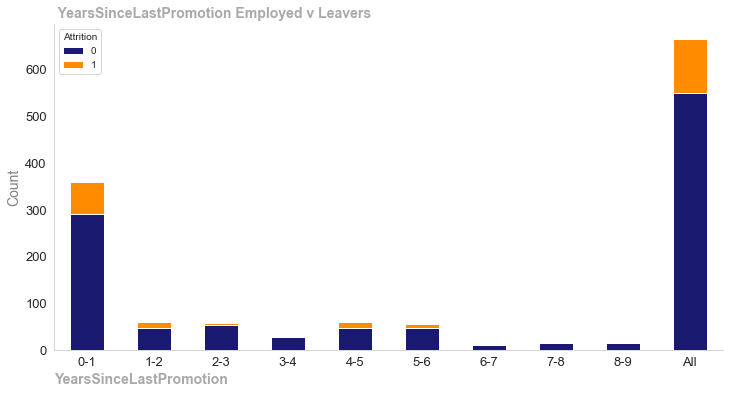
6-10 0.250790 0.270936 0.254340

11-20 0.203372 0.226601 0.207465

21-30 0.161222 0.182266 0.164931

All 1.000000 1.000000 1.000000

## Figure 40

Years since last promotion

Mean YearsSinceLastPromotion is:1.9346938775510205

Median YearsSinceLastPromotion is:0.0

Standard Deviation is:3.1620333881675293

Median YearsSinceLastPromotion is:0.0

Skew is:2.0636092744314474

kurtosis is:3.996706272723284

Age Percentiles

[0.25, 0.5, 0.75]th Percentile: 0.0

[0.25, 0.5, 0.75]th Percentile: 0.0

[0.25, 0.5, 0.75]th Percentile: 2.0

Attrition 0 1 All

YearsSinceLastPromotion

0-1 0.529197 0.586207 0.539157

1-2 0.083942 0.120690 0.090361

2-3 0.096715 0.043103 0.087349

3-4 0.049270 0.017241 0.043675

4-5 0.083942 0.112069 0.088855

5-6 0.083942 0.077586 0.082831

6-7 0.020073 0.008621 0.018072

7-8 0.027372 0.017241 0.025602

8-9 0.025547 0.017241 0.024096

All 1.000000 1.000000 1.000000

## Figure 41

Descriptive statistics summarize and describe key elements of a dataset. Descriptive statistics help analyze employee central tendency, variability, and distribution. The main descriptive statistics measures are:

The average value of a variable is its mean.

The median divides a dataset into two equal half.

Standard Deviation shows value dispersion around the mean.

Max and Min Values: Display the dataset's range.

Divide data into quartiles or percentage intervals.

Descriptive statistics can show trends and patterns for each employee churn feature.

Inferential Statistics: Inferential statistics use sample data to make predictions about a population. Inferential statistics can draw conclusions about the total workforce from a subset. Common methods:

Hypothesis testing: Testing correlations or variable differences.

Finding the degree and direction of variable correlations using regression analysis.

ANOVA: Comparing data groups.

Correlation Analysis: Assessing linear data connections.

You may validate assumptions, discover and make informed decisions to address organizational difficulties such as low satisfaction by using inferential statistics.

In conclusion, using descriptive and inferential statistics in our analysis helps interpret patterns, draw meaningful conclusions, and make data-driven decisions to improve productivity.

Measures of Central Tendency & Dispersion

Mean

* The mean level of NumCompaniesWorked is 2.24
* The mean level of TotalWorkingYears is 10.99
* The mean level of YearsAtCompany is 6.43
* The mean level of HourlyRate is 64.97
* The mean level of StockOptionLevel is..... .53
* The mean level of MonthlyIncome is 6385.50
* The mean level of DistanceFromHome is 8.69
* The mean level of Age is 36.27
* The mean level of YearsSinceLastPromotion is 1.93

Median

* The median level of NumCompaniesWorked is 1.00 indicating th middle of the data
* The mean level of TotalWorkingYears is 11 indicating th middle of the data
* The mean level of YearsAtCompany is 5.80 indicating th middle of the data
* The mean level of HourlyRate is 64.80 indicating th middle of the data
* The mean level of StockOptionLevel is 0 indicating th middle of the data
* The mean level of MonthlyIncome is 4957 indicating th middle of the data
* The mean level of DistanceFromHome is 6 indicating th middle of the data
* The mean level of Age is 35 indicating th middle of the data
* The mean level of YearsSinceLastPromotion is 0 indicating th middle of the data

Standard Deviation

* The Standard deviation of NumCompaniesWorked is 2.47 indicating a high variability
* The mean level of YearsAtCompany is 6.04 indicating a high variability
* The mean level of HourlyRate is 21.39 indicating a high variability
* The mean level of StockOptionLevel is .78 indicating a high variability
* The mean level of MonthlyIncome is 4546 indicating a high variability
* The mean level of DistanceFromHome is 8.23 indicating a high variability
* The mean level of Age is 9.9 indicating th a high variability
* The mean level of YearsSinceLastPromotion is 3.12 a high variability

Frequency Tables

* The majority of employees spend less than 5 years at the company
* Hourly rates spread evenly across employees
* Monthly Income majority at the lower end of the scale
* Majority of Employees under the age of 40
* Majority of employess travel less than 10 km to work

Correlation Matrices

# The strongest positive correlation with the attrition features are: TotalWorkingYears, Age, Monthly Income, Years Company

# Summary of Key Findings

# Using attrition levels as a proxy for satisfaction and productivity, descriptive statistics have shown areas to address. Our descriptive statistics suggest that firm remuneration packages are limited to service. Standard deviations demonstrate our features' considerable variability. Example: Monthly income has larger variability than daily rate, suggesting some employees earn end-of-month bonuses while others do not. This may diminish employee satisfaction, which may explain why most depart after 5 years.

# Utilizable knowledge The corporation may want to tie bonuses, stock options, etc. to KPIs rather than service. This could make the workplace more fair and productive.

Inferential Statistics

Inferential statistics uses samples from populations, but to ensure validity and inference, we must utilize tests like ttest or anova.T-Test on Num of companies worked The number of companies worked variable is useful for understanding various aspects of employee behavior & organizational dynamics. To give us context According to Staista the average number a companies a person works for in Ireland is 4.

Hypothesis

* Null Hypothesis (H0) The population Mean (Mu) is equal to 2.66
* Alternative Hypothesis (H1) The population of the Mean(Mu) is not equal to 2.66

T-Test Output

T-Statistics: 0.10127528990040056

p-value: 0.919345751795404

Interpretation

The t-statistic is the ratio of the difference between the sample mean and the hypothesized population to the standard error of the mean. In this case the t-statistic is close to zero 0,101, suggesting that the sample mean is not significantly different from the hypothesized population mean.

The p-value (0.919) is greater than the 0.05 significance level commonly used. On this basis we fail to reject the null hypothesis

Conclusion

Based on the p\_value at a 5% significance level we do not have enough evidence to reject the null hypothesis. Therefore we do have sufficient evidence to conclude that the population mean is different from 2.66. The average number of businesses worked is 4, yet the null hypothesis cannot be rejected. Average number of firms worked: 2.66. This may reveal workforce mobility and career trajectory, and one interpretation may be age. If the average number of companies worked is low, like 2.66, it may indicate that the company has longer-tenured employees or fewer career transitions. An older workforce that has been with the organization longer may explain this.

Monthly Income

Hypotheses

* Null Hypothesis(H0): The average monthly income is the same for employees who have left the company and those who have stayed
* Alternative Hypothesis(H1): The average monthly income is not the same for employees who have left the company and those who have stayed

T-Test Output

T-statistic: -4.982194038913557

P-value): 7.031910768031407e-07

Interpretation

* Cas2 1 (H1 mu y =! mu n:)
* At a 5% significance level we reject the null hypothesis. The average monthly is not the same for employees who have left the company and those who are still employed
* Case 2 (H1: mu y < mu n):
* Dividing the p-value by 2 the result is still less than alpha therefore we reject the null hypotheses. At a 5% significance level, we can say that the monthly income of employees who leave the company is less than those who don’t.

Daily Rate

Hypotheses

* Null Hypothesis(H0): The average Daily Rate is the same for employees who have left the company and those who have stayed
* Alternative Hypothesis(H1): The Daily Rate is not the same for employees who have left the company and those who have stayed

T-Test Output

T-statistic: -1.036720196429877

P-value): 0.300037013267801

Interpretation

* First case (H1: mu y =! mu n) --> There is not enough evidence to suggest significant difference in hourly rate between employees who left the company and those who stayed.

Case 2(H1: mu y < mu n) --> At a 5% significance level, we fail to reject the null hypotheses, indication that there is no statistically significant difference in hourly rates between employees who have left the company and those who stayed

Inferential statistics the difference between daily rate and monthly income within the company. It would be reasonable to assume the monthly income would derive from the daily rate. The fact we accept the null hypothesis for the daily rate but reject it for the monthly rate would suggest that the company is making extra payments to employees. This is supported by a recent ABC newspaper article which stated the there is a gap of 42.9% between management \_level jobs and those considered to be non supervisory.(*New study highlights the pay gap between managers and employees*, no date)

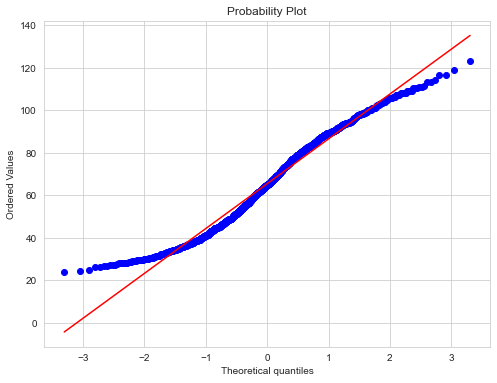
Anova Test

If there are statistically significant discrepancies in average hourly rates between departments, ANOVA might reveal them. This circumstance justifies an ANOVA:

* ANOVA compares three or more groups. In this example, you have numerous departments (groups) like Sales, Marketing, etc. and want to determine if average hourly rates differ significantly.

Efficiency in Testing:

* ANOVA lets you test for departmental differences in one analysis instead of many t-tests. This improves efficiency and reduces experiment-wide errors.



Now we will check it performing Shapiro Wilk test for the 3 categories

#H0: data comes from normal distribution

#H1: data does not come from normal distribution

#We use alpha (significance level 5%)

We start with Friends

stats.shapiro(df\_stat.HourlyRate[df\_stat['Department'] =="Sales"])

ShapiroResult(statistic=0.9732151627540588, pvalue=1.2233080042278743e-06)

We can Observe from the above p- value we can reject the null hypotheses that the data comes from a normal Distribution

stats.shapiro(df\_stat.HourlyRate[df\_stat['Department'] =="Research & Development"])

ShapiroResult(statistic=0.9748362898826599, pvalue=6.375634829108856e-12)

#Now we will check it performing Shapiro Wilk test for the 3 categories

#H0: data comes from normal distribution

#H1: data does not come from normal distribution

#We use alpha (significance level 5%)

We can Observe from the above p- value we can reject the null hypotheses that the data comes from a normal Distribution

Age

Statistics 0.974389910697937 p-value 1.6567358819736828e-15

The null hypothesis can be rejected

DistanceFromHome

Statistics 0.862089991569519 p-value 4.562174390777917e-34

The null hypothesis can be rejected

Education

Statistics 0.9777684211730957 p-value 2.681671074436509e-14

The null hypothesis can be rejected

EnvironmentSatisfaction

Statistics 0.9537271857261658 p-value 3.766294288931367e-21

The null hypothesis can be rejected

HourlyRate

Statistics 0.9750291109085083 p-value 2.749492663709982e-15

# Apply log10 transformation to the entire Data Frame

df\_log\_transformed = np.log10(df\_num.replace(0, 1))

We have applied a log transformation to the datset to normalize the data

#Now we will check it performing Shapiro Wilk test for the 3 categories

#H0: data comes from normal distribution

#H1: data does not come from normal distribution

#We use alpha (significance level 5%)

stats.shapiro(df\_stat.HourlyRate\_log[df\_stat['Department'] =="Human Resources"])

ShapiroResult(statistic=0.9634526371955872, pvalue=0.005139612127095461)

We can Observe from the above p- value we can reject the null hypotheses that the data comes from a normal Distribution

The null hypothesis is rejected statistically since the p-value is less than 0.05. Thus, we reject the normal distribution hypothesis and accept the alternative hypostheis the data in not normally distributed . To conduct an anova test the data need to have a normalized distribution and as a result we can not proceed any further with the anova test

Conclusion

The initial data preparation process includes addressing missing values by employing imputation techniques such as median, padding, and mode. Exploratory data analysis revealed trends, while outlier detection and cluster analysis yielded insights into potential affecting elements. Conducting normality tests and employing sophisticated approaches such as Linear Discriminant Analysis and Principal Component Analysis improved our comprehension of the features.  
  
The purpose of feature engineering, which includes binning, is to enhance the performance of the model. The data underwent encoding to prepare it for use by machine learning models, and a Dummy Classifier was employed as a reference point for assessment. Thorough cross-validation and hyperparameter adjustment determined the best models. The feature importances offered a clear understanding of the significant aspects of the reasons why people leave the company. These features were used as a proxy people unhappiness. By anaslying these features it will allow for a data driven approach to improve satisfaction nd productuitvy with the company  
  
The investigation progressed to the stage of statistical modelling, where descriptive statistics were used to summarise important metrics, providing a concise overview of the dataset. Inferential statistics, such as t-tests and ANOVA, examined and analysed correlations and disparities. This comprehensive approach guarantees a strong comprehension of the factors contributing to low productivity and satisfaction levels within the company, allowing for well-informed decision-making regarding initiatives to improve satisfaction and productutivity and ensure organisational stability. The integration of exploratory, machine learning, and statistical studies provides a strong basis for strategic HR interventions aimed at enhancing satisfaction and productivity among employees.

Communication To Stakeholders

Improve productivity and satisfaction plan

Distance From Home Employees who travel further to work are more likely to leave the company, efforts should be made to support the employees with access to working from home.

Age employees in a relatively young age bracket are more likely to leave the company, efforts should be made to articulate the long term vision off the company and they will be able to grow and be successful within this environment.

Total working years the more experienced employees are more likely to leave leading to a knowledge gap they should be placed in a high risk group with the company having succession plans in place.

Company use regular interviews with staff to asses job satisfaction etc these metrics will be measured against the averages and if an employee has a high figure they would be considered high risk

Company must adopt a remuneration plan that is linked to performance and productivity. When Key performance indicators are met monthly bonuses, stock options woulds be used to reward employees which will increase satisfaction and productivity

References

*New study highlights the pay gap between managers and employees* (no date). Available at: https://www.abc15.com/news/in-depth/new-study-highlights-the-pay-gap-between-managers-and-employees (Accessed: 7 January 2024).

Hamza b,2018.Building an Employee Churn Model in Python to Develop a Strategic Retention Plan..Available at: <https://medium.com/towards-data-science/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d> [Accesed] 14-12-2023]

Anurag Dhadse,2021.Removing Outliers. Understanding How and What behind the Magic...Available at: <https://medium.com/analytics-vidhya/removing-outliers-understanding-how-and-what-behind-the-magic-18a78ab480ff> [Accesed] 22-12-2023]

Zach,2020.How to Perform a Box-Cox Transformation in Python..Available at: <https://www.statology.org/box-cox-transformation-python/> [Accesed] 22-12-2023]

Saloni Mishra,2020.Methods for Normality Test with Application in Python...Available at: <https://medium.com/towards-data-science/methods-for-normality-test-with-application-in-python-bb91b49ed0f5> [Accesed] 22-12-2023]

#Geron, A. (2019) Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. 2nd ed. Sebastopol, CA: O’Reilly Media.

meghan Natarajan,2023.Intuition behind AUC and Precision-Recall Curves..Available at: <https://medium.com/@megha.natarajan/intuition-behind-auc-and-precision-recall-curves-7a33d2d45e02> [Accesed] 21-11-2023]

Xiao,2022.A Practical Guide To Exploratory Data Analysis..Available at: <https://medium.com/mlearning-ai/a-practical-guide-to-exploratory-data-anal> ysis-fabbac2bcad [Accesed] 03-11-2023]

unknown,unknown.Mathematics - Statistics & Probability mastery for Data Science, Machine Learning..Available at: [https://www.udemy.com/course/statistics-probability-for-data-science/learn/lecture/35290462#overview](https://www.udemy.com/course/statistics-probability-for-data-science/learn/lecture/35290462" \l "overview) [Accesed] 1-11-2023]

#Farzad Mahmoodinobar,2023.XGBoost: Intro, Step-by-Step Implementation, and Performance Comparison...Available at: <https://medium.com/towards-data-science/xgboost-intro-step-by-step-implementation-and-performance-comparison-6018dfa21>

Nussbaumer Knaflic, C. (2015). Storytelling with data (C. N. Knaflic, Ed.). John Wiley & Sons

Michelangiolo Mazzeschi,2021.Exploring Dataset With Data Visualization: Wine Classification Using PCA..Available at: <https://medium.com/towards-artificial-intelligence/exploring-dataset-with-data-visualization-wine-classification-using-pca-d133427183fb> [Accesed] 03-11-2023]

unknown,2018.Predicting Employee Churn in Python..Available at: <https://www.datacamp.com/tutorial/predicting-employee-churn-python> [Accesed] 03-11-2023]

satyam Kumar,2021.Essential guide to perform Feature Binning using a Decision Tree Model..Available at: <https://medium.com/towards-data-science/essential-guide-to-perform-feature-binning-using-a-decision-tree-model-90bcc66d61f9> [Accesed] 05-01-2024]

col jung,2020.Predict Employee Churn with Machine Learning using Python..Available at: <https://medium.com/towards-data-science/will-your-employee-leave-a-machine-learning-model-8484c2a6663e> [Accesed] 05-01-2024]

ayse bat,2020. Employee churn[Accesed] 05-01-2024]..Available at: <https://medium.com/@aaysbt/employee-churn-analysis-be94751e4df5> [Accesed] 05-01-2024]

‘There Are Significant Business Costs to Replacing Employees’ (2012) *Center for American Progress*, 16 November. Available at: <https://www.americanprogress.org/article/there-are-significant-business-costs-to-replacing-employees/> (Accessed: 18 December 2023).

Ph.D, E.S. (2023) ‘The Problem with Employee Churn Modelling’, *Medium*, 12 November. Available at: <https://eric-sandosham.medium.com/the-problem-with-employee-churn-modelling-51f10a26cf58> (Accessed: 24 December 2023).

*Average job tenures in Europe 2022* (no date) *Statista*. Available at: <https://www.statista.com/statistics/1209552/average-time-spent-with-one-employer-in-europe/> (Accessed: 3 January 2024).

.

.