CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Higher Diploma in Data Analytics for Business |
| **Assessment Title:** | Data Preparation & Visualisation 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.

* [1](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Table-of-Contents) [Table of Contents](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Table-of-Contents)
* Import libraries
* Importing The Data
* [Data Description and Exploratory Visualisations](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "3---Data-Description-and-Exploratory-Visualisations)
* [Missing Values](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "4--Missing-Values)
* [EDA & Outliers](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "5-EDA-&-Outliers)
* Cluster analysis
* Check For Normality
* Data Preprocessing & Encoding For LDA & PCA
* [Linear Discriminant Analysis](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "9---Linear-Discriminant-Analysis)
* Principal Component Analysis
* Feature Engineering- Binning
* [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)
* [Train Dummy Classifier](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "13--Train-Dummy-Classifier)
* [Machine Learing Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "14--Machine-Learing-Models)
* Cross Validation
* [Hyperparameter Tuning of Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "16---Hyperparameter-Tuning-of-Models)
* [18  17- Feature Importances](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "17--Feature-Importances)
* [19  18- Evaluation of Best Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "18--Evaluation-of-Best-Models)
* [20  19 - Feature Importances](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "19---Feature-Importances)
* [21  20 Statisical Modelling](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "20-Statisical-Modelling)
* [22  21 Descripitive Stats](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "21-Descripitive-Stats)
* [23  22 - Inferential Statistics](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "22---Inferential-Statistics)
* [24  23 - Appendix](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "23---Appendix)
* [25  import MinMaxScaler](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "import-MinMaxScaler)
* [26  Display rows and columns](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Display-rows-and-columns)

**Table of Figures**

Figure 1. Pairplot of Dataset 12

Figure 2. Correlation Matrix 13

Figure 3. Histogram & Boxplot Amounts Column 14

[Figure 4. Histogram & Boxplot By Day Column.](#_bookmark3) 14

[Figure 5. Histogram & Boxplot Month Column](#_bookmark4) 14

[Figure 6. Histogram & Boxplot Time of day Column. ...1](#_bookmark5)4

[Figure 7. Histogram & Boxplot Age Column ...10](#_bookmark6)

[Figure 8. Pie Plot 2020 v 2019 Fraudulent Transactions ...11](#_bookmark7)

Figure 9. Countplot Distribution of Fraudulent Transactions across Category ...12

[Figure 10. Countplot Distribution of Fraudulent Transactions across Gender ...13](#_bookmark8)

[Figure 11. Scatter plot Amount V Time of Day ...13](#_bookmark9)

[Figure 12. Kernel Density Plot ...14](#_bookmark10)

[Figure 13. Kernel Density Plot 50 euros excluded ...15](#_bookmark12)

[Figure 14. Algorithms Comparing Results of Models …16](#_bookmark13)

[Figure 15. Baseline Model Dummy Regressor .16](#_bookmark14)

[Figure 16.](#_bookmark15) Scatter Plot Predicted V Actual [18](#_bookmark15)

[Figure 17. Pie Plot under €50 18](#_bookmark16)

[Figure 18. Pie Plot under €100 19](#_bookmark17)

[Figure 19.](#_bookmark18) Pie Plot under €200 [20](#_bookmark18)

[Figure 20. Pie Plot under Median Amount €390 20](#_bookmark19)

[Figure 21. Residual Plot for GradientBoostingRegressor 21](#_bookmark20)

[Figure 22. Prediction Error for GradientBoostingRegressor . 21](#_bookmark21)

[Figure 23. Scatter Plot for Final GradientBoostingRegressor . 2](#_bookmark21)1

[Figure 24. Feature importance for Final GradientBoostingRegressor . 2](#_bookmark21)2

[Figure 25. Final Model Pie for GradientBoostingRegressor . 2](#_bookmark21)3

[Figure 26. Prediction Error for Final GradientBoostingRegressor . 2](#_bookmark21)3

### Data Preparation and Exploratory Data Analysis (EDA):

1. Characterization of the Data Set (0-10 marks):
   * Determine the size of the dataset.
   * Identify the number of attributes.
   * Check for missing values and assess the number of observations.
2. Data Preparation and EDA (0-20 marks):
   * Clean the data by handling missing values, duplicates, and outliers.
   * Rename variables if necessary for clarity.
   * Apply exploratory data analysis techniques such as histograms, box plots, and correlation matrices.
   * Provide a clear rationale for each step in your data preparation.
3. Encoding, Scaling, and Feature Engineering (0-30 marks):
   * Apply encoding techniques for categorical variables.
   * Scale numerical features if needed (e.g., using StandardScaler or MinMaxScaler).
   * Perform feature engineering based on your understanding of the data and problem.
4. Dimensionality Reduction (0-40 marks):
   * Apply Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA).
   * Visualize the results and compare the separation of classes.
   * Explain the differences between LDA and PCA and discuss their implications for classifying or clustering.

### Statistical Techniques:

1. Descriptive Statistical Analyses (0-30 marks):
   * Calculate measures of central tendency and dispersion.
   * Create frequency distributions.
   * Generate correlation matrices.
   * Summarize your findings.
2. Hypothesis Testing (0-40 marks):
   * Formulate and test hypotheses using appropriate statistical techniques (e.g., t-tests or ANOVA).
   * Use at least two statistical tests.
   * Summarize the findings.
3. Jupyter Notebook Results (0-10 marks):
   * Use a Jupyter notebook to produce result sets (e.g., scatter plots, regression models).
   * Provide a summary of your findings.
4. Communication of Results (0-20 marks):
   * Write a report with clear and concise explanations.
   * Use visualizations and appropriate statistical terminology.
   * Address stakeholders' needs.

### Machine Learning:

1. Choice of Machine Learning Approach (0-20 marks):
   * Justify the choice of supervised or unsupervised learning based on the dataset's characteristics.
   * Discuss the pros and cons of both approaches.
2. Feature Selection and Hyperparameter Tuning (0-30 marks):
   * Choose suitable features using feature selection methods.
   * Use hyperparameter tuning (e.g., GridSearchCV) to optimize machine learning models.
3. Training and Testing (0-30 marks):
   * Implement supervised learning with different splits.
   * Utilize k-fold cross-validation for unsupervised learning.
4. Comparison of ML Modeling Outcomes (0-20 marks):
   * Present a comparison of outcomes using tables or graphs.
   * Discuss the statistical understanding of the results.

### Submission Requirements:

1. Document Preparation (Min 3000 words / Max 4000 words):
   * Organize your report in a clear structure.
   * Clearly specify the number of words used.

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

* [2  Import libraries](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Import-libraries)
* [3  Importing The Data](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Importing-The-Data)
* [4  - Data Description and Exploratory Visualisations](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "--Data-Description-and-Exploratory-Visualisations)
* [5  Missing Values](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Missing-Values)
* [6  EDA & Outliers](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "EDA-&-Outliers)
* [7  Cluster analysis](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Cluster-analysis)
* [8  Check For Normality](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Check-For-Normality)
* [9  Data Preprocessing & Encoding For LDA & PCA](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Data-Preprocessing-&-Encoding-For-LDA-&-PCA)
* [10  Linear Discriminant Analysis](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Linear-Discriminant-Analysis)
* [11  Principal Component Analysis](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Principal-Component-Analysis)
* [12  Feature Engineering- Binning](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Feature-Engineering--Binning)
* [13  Machine Models Data Preprocessing and Encoding](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Machine-Models-Data-Preprocessing-and-Encoding)
* [14  Train Dummy Classifier](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Train-Dummy-Classifier)
* [15  Machine Learing Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Machine-Learing-Models)
* [16  Cross Validation](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Cross-Validation)
* [17  Hyperparameter Tuning of Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Hyperparameter-Tuning-of-Models)
* [18  Feature Importances](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Feature-Importances)
* [19  Evaluation of Best Models](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Evaluation-of-Best-Models)
* [20  Feature Importances](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Feature-Importances)
* [21  Statisical Modelling](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Statisical-Modelling)
* [22  Descripitive Stats](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Descripitive-Stats)
* [23  Inferential Statistics](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Inferential-Statistics)
* [24  Appendix](http://localhost:8888/notebooks/SBA_22521_intregated_hdip_level_8_dm-ml_mi.ipynb" \l "Appendix)

# Introduction

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 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 use statistical models and machine learning models to identity the factors that lead to an Employee leaving the company, use this information to develop a set of proposals fro management to improve the satisfaction and productivity within the company

# Characterization of the Data

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

# Data Preparation & EDA

## 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

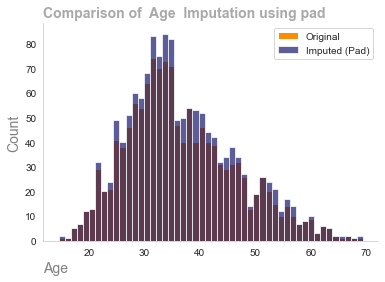
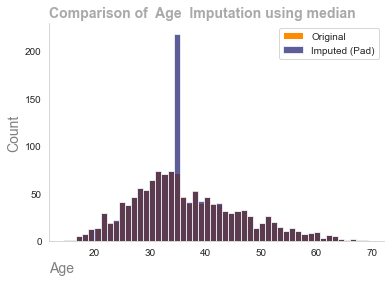
We have multiple imputation options available:

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

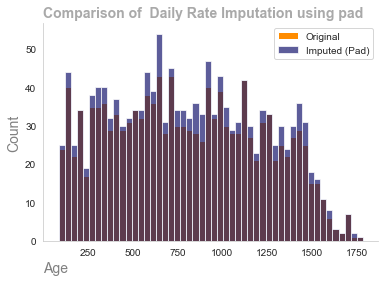
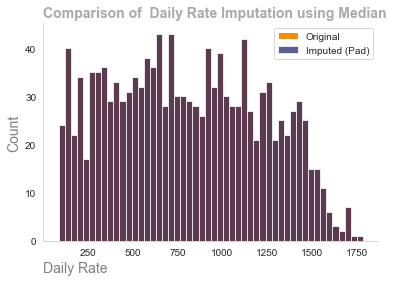
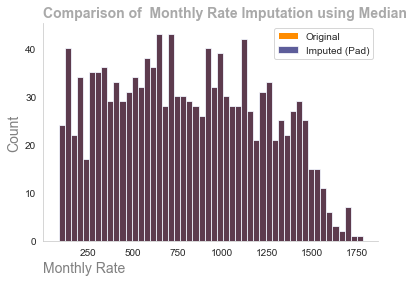
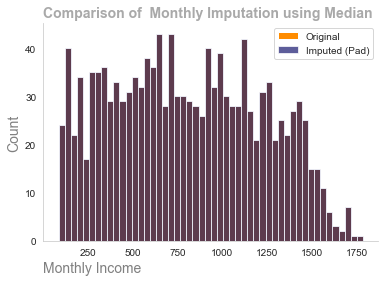
Each of the above models have their individual pros and cons.

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 been successful



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.

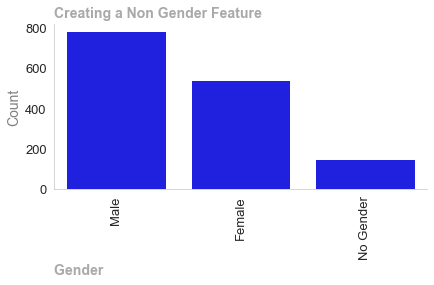
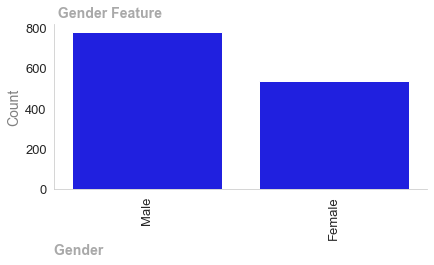
# Categorical Data

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.

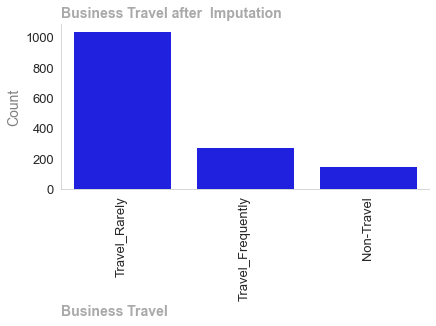
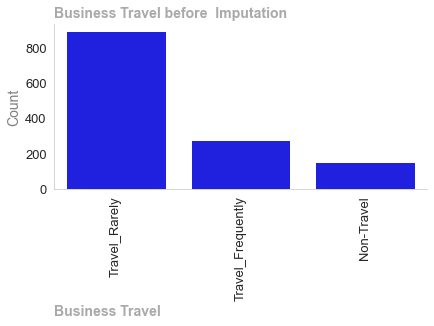
Our Strategy:

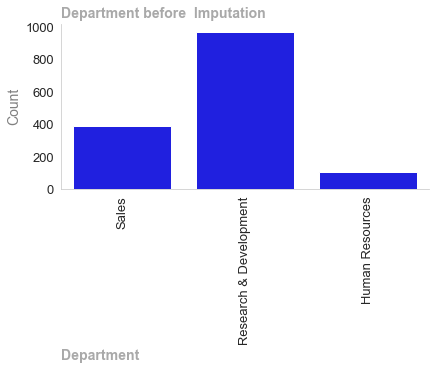
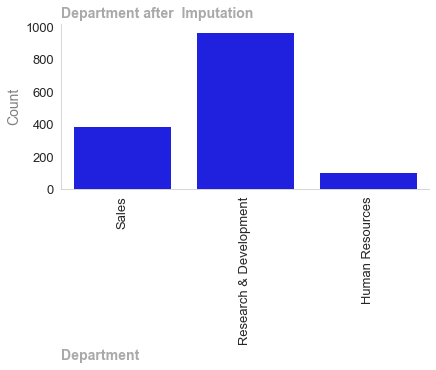
We have 8 features contain categorical database

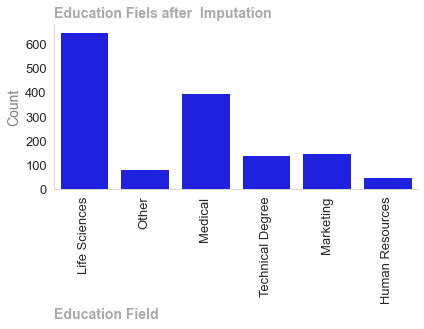
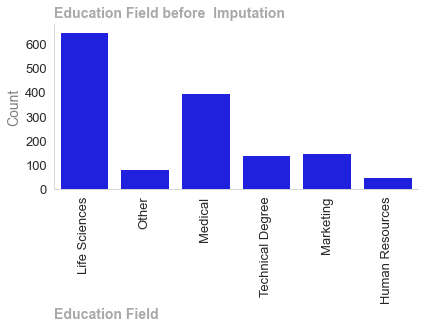
1. Assume mode will mode be the correct method of imputation
2. Evaluate results through the use of histograms to ensure the distribution of the original data has been maintained
3. If the distribution has not been maintained work on models such as predictive modeling.

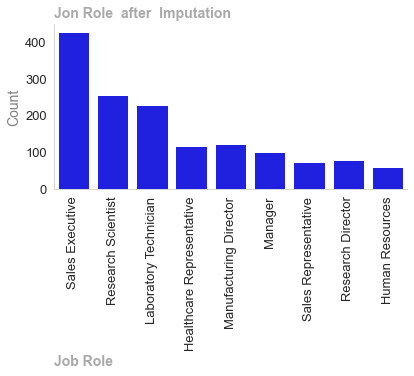
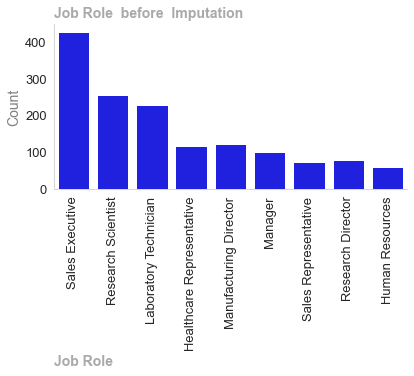


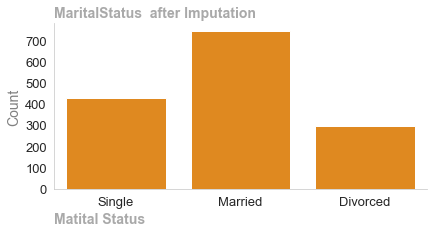
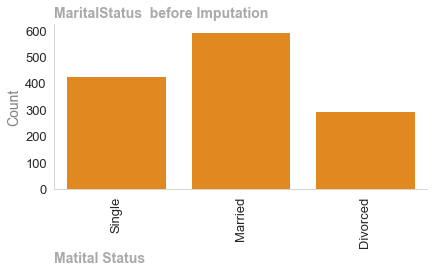
We can observe the problem we spoke about if we apply model imputation method to the gender feature the male column will increase exponentially. We need to consider another method which could be predictive modeling such as KNN. However as data analyst we need to be a ware of modern society. Was the missingness 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.





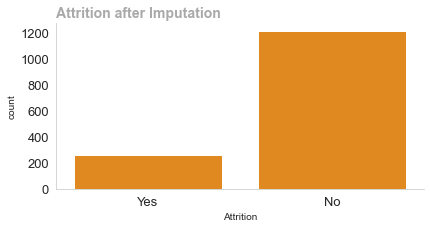
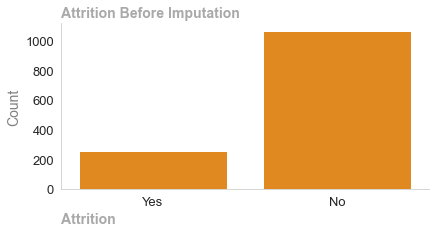






We have used the mode imputation method on the remaining features as can be observed the distribution of the data has been maintained.

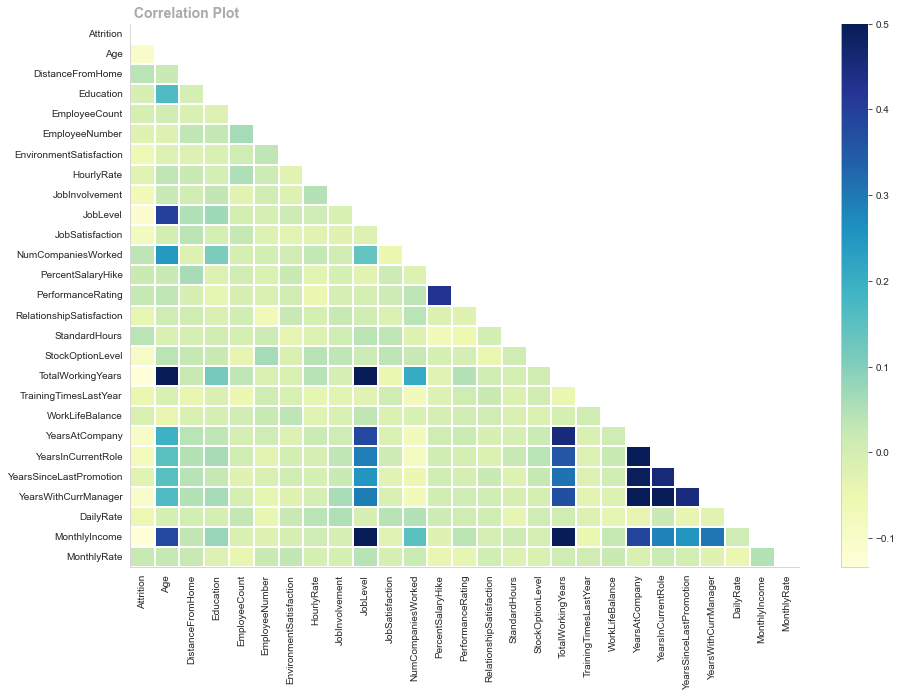
The attrition column which will be our target variable in the classification models we will run has two classed yes or no



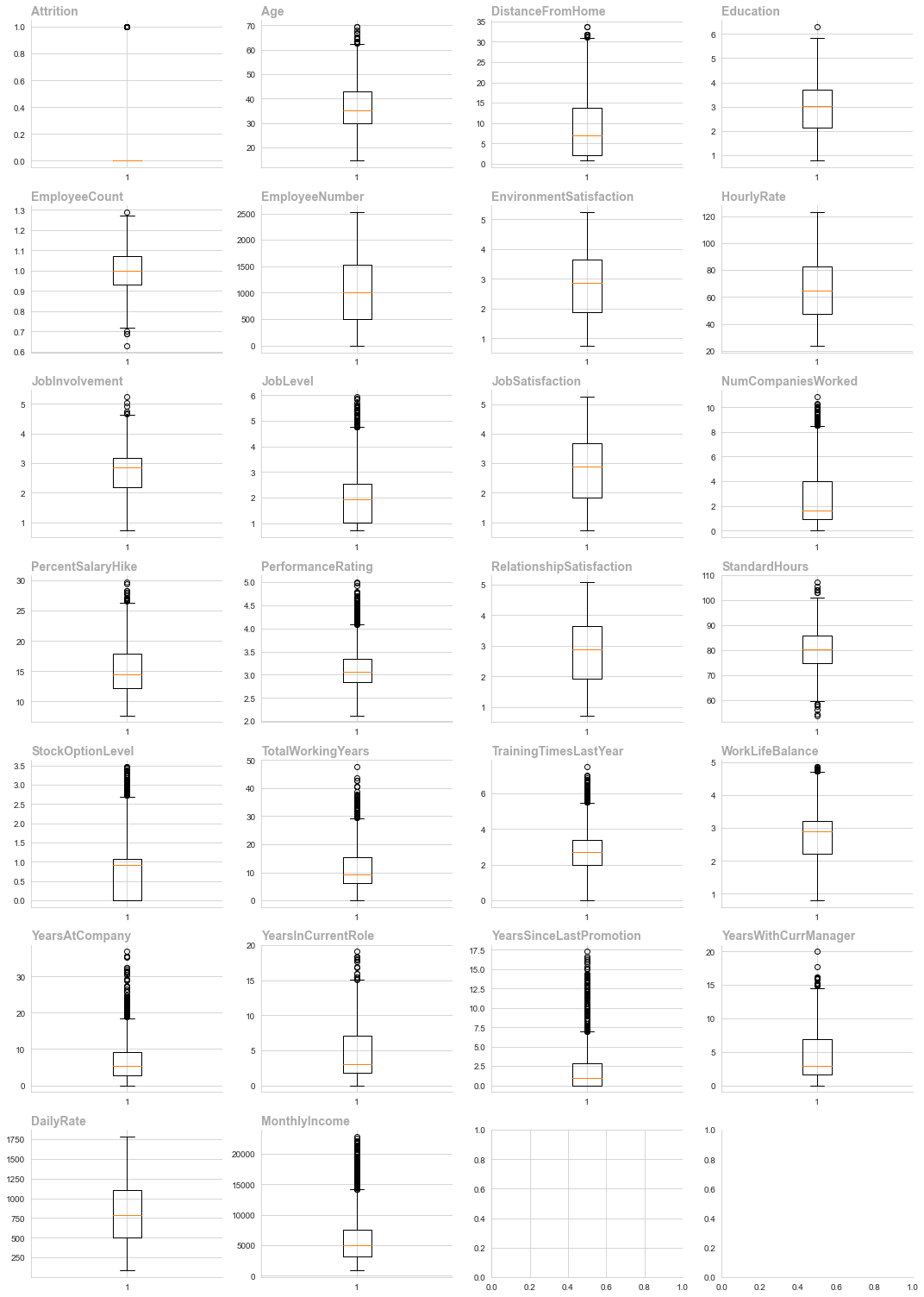
Here we observe where mode and not worked as it has more to imbalance to a unbalanced dataset. The course of action would be use a predictive model such as KNN that uses other features to calculate the missing values. However I could not get the code to work and have added my workings in the appendix 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.

1518 Words

# Data Anaysis Plots



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, SVMs, 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.



Most numerical columns in boxplots have outliers, which must be eliminated for machine learning to work.

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 14205.2

Removing Outliers using 3 Standard Deviation

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

Outlier (Previously) Outliers Count Column Lower Limit Upper Limit

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

False False Age 7.29665 66.1775

False False DistanceFromHome -15.4958 33.8169

False False JobInvolvement 0.445604 5.03362

False False JobLevel -1.2005 5.26764

False False NumCompaniesWorked -4.75161 10.044

False False PercentSalaryHike 3.2585 27.3761

False False TotalWorkingYears -11.1607 33.7095

False False TrainingTimesLastYear -0.98982 6.58387

False False WorkLifeBalance 0.497852 5.05548

False False YearsAtCompany -8.75277 21.935

True True 15 YearsInCurrentRole -6.6146 15.102

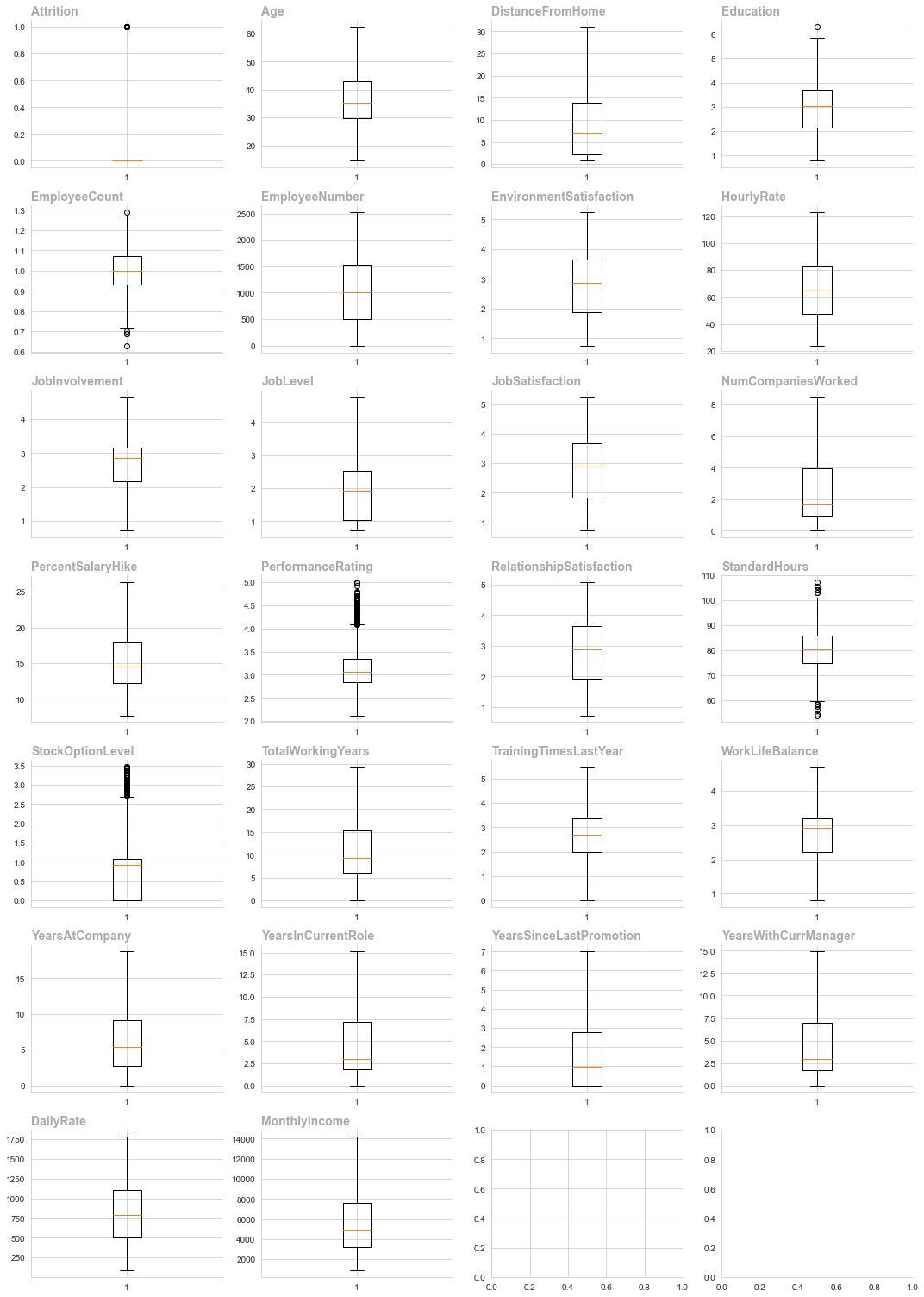
False False YearsSinceLastPromotion -5.33531 9.15469

True True 14 YearsWithCurrManager -6.61229 14.7702

False False MonthlyIncome -5102.13 17218.4

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

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



# Feature Engineering

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 interpretability of the models

Improved model performance

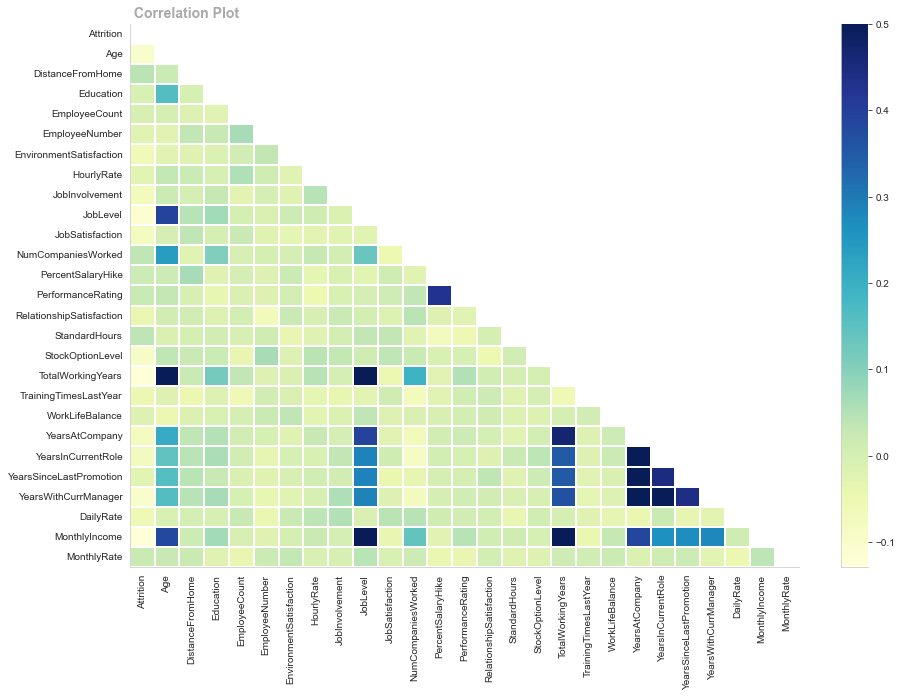
* Relevance: Create features that are relevant to the problem at hand.

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

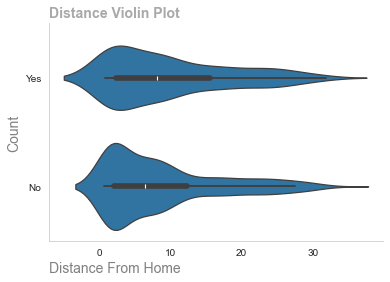
* Handling Non Linearity: Introduce features that can help capture non linear data improving model performance.
* Creating composite Features: Combine multiple features into one feature which can simplify the model and improve interpretability.
* Addressing Data Skewness: Apply transformation to features 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 conxext of data analytics more data is always more prefarable and we need to be aware that we not mutate the date to the point where it is unusable.

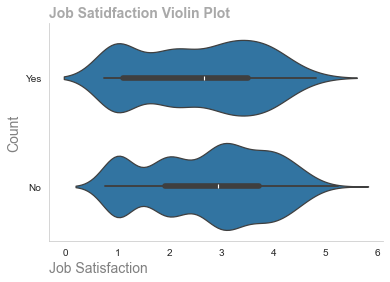
# EDA



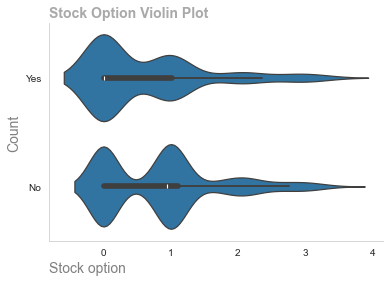


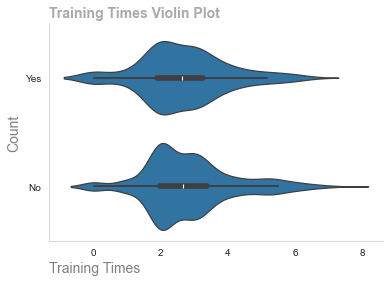


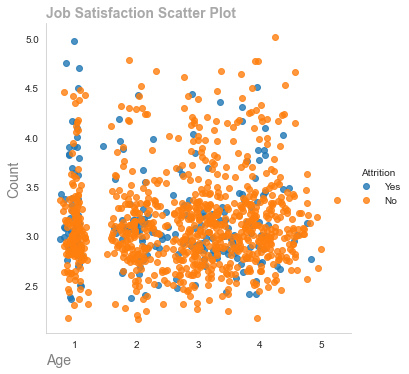


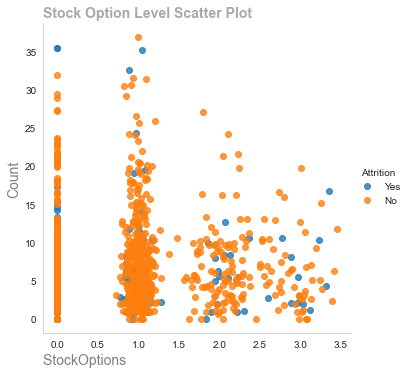












Employed Left

Age 37.190027 34.609336

Distance From Home 9.004461 9.893881

Education 2.924942 2.901589

Employee Count 0.999010 0.997946

Employee Number 1030.559051 992.177597

Environment Satisfaction 2.749307 2.560969

Hourly Rate 65.729472 64.213450

Job Involvement 2.765483 2.618081

Job Level 2.091541 1.761240

Job Satisfaction 2.754480 2.511047

Num Companies Worked 2.604728 2.840930

Percent Salary Hike 15.283715 15.475213

Performance Rating 3.142657 3.174729

Relationship Satisfaction 2.744992 2.624496

Standard Hours 80.079068 80.885349

StockOption Level 0.841568 0.633798

Total Working Years 11.730730 9.130775

Training Times Last Year 2.823380 2.673212

Work Life Balance 2.781650 2.753256

YearsAtCompany 6.817308 5.528411

Years In CurrentRole 4.370313 3.649021

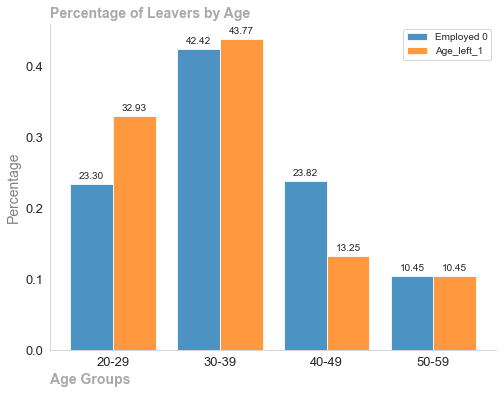
Years Since Last Promotion 1.939663 1.768886

Years With Curr Manager 4.248773 3.281085

Daily Rate 810.653894 750.652016

Monthly Income 6288.552492 4975.772016

Monthly Rate 14281.097178 14694.936783

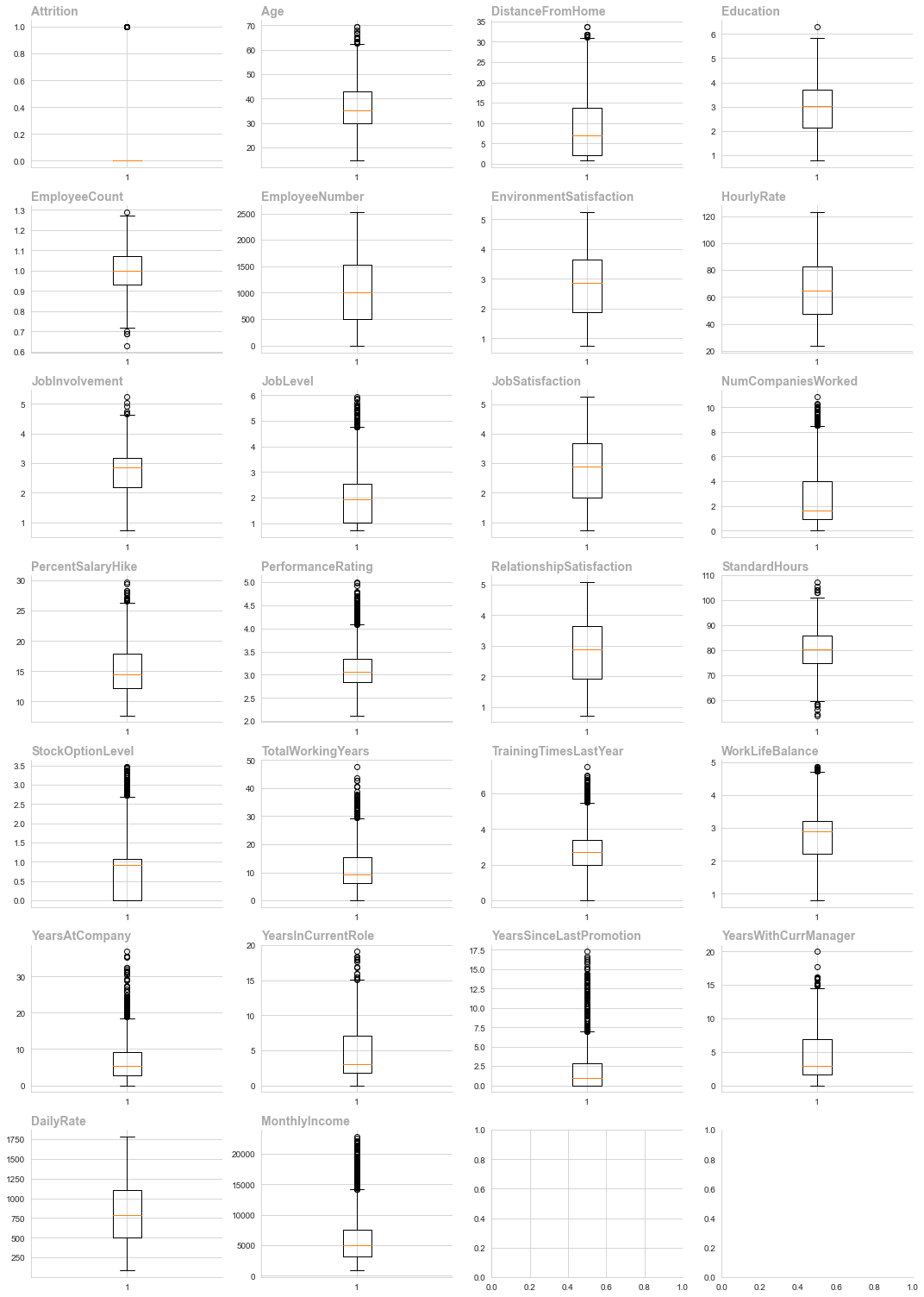


# EDA Takeaway

* People under the age of 29 are more likely to leave than stay
* People who leave are travel further than people who stay
* People who leave have a lower Job satisfaction & Environment satisfaction than people who stay
* People who leave have not been awarded stock options.

Outlier handling pros and downsides must be addressed based on data and analytic goals. 2 traits showed outliers after IQR. Continued investigation was needed. Stock option levels are excessive, with few employees owning them and many not. Log transformation altered this feature. Box cox was better, but I couldn't fix the code and had to move on owing to time constraints. However, these outliers may boost corporate satisfaction and productivity. Remember that not all outliers are the same and may assist us understand the problem.

# Outliers



We can oberve from the boxplots thst there are outliers cotained with the dataset. These will need to be dealt with prior to using LDA as this feature extraction tevhnique expectd to be presented with normal distribution.

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

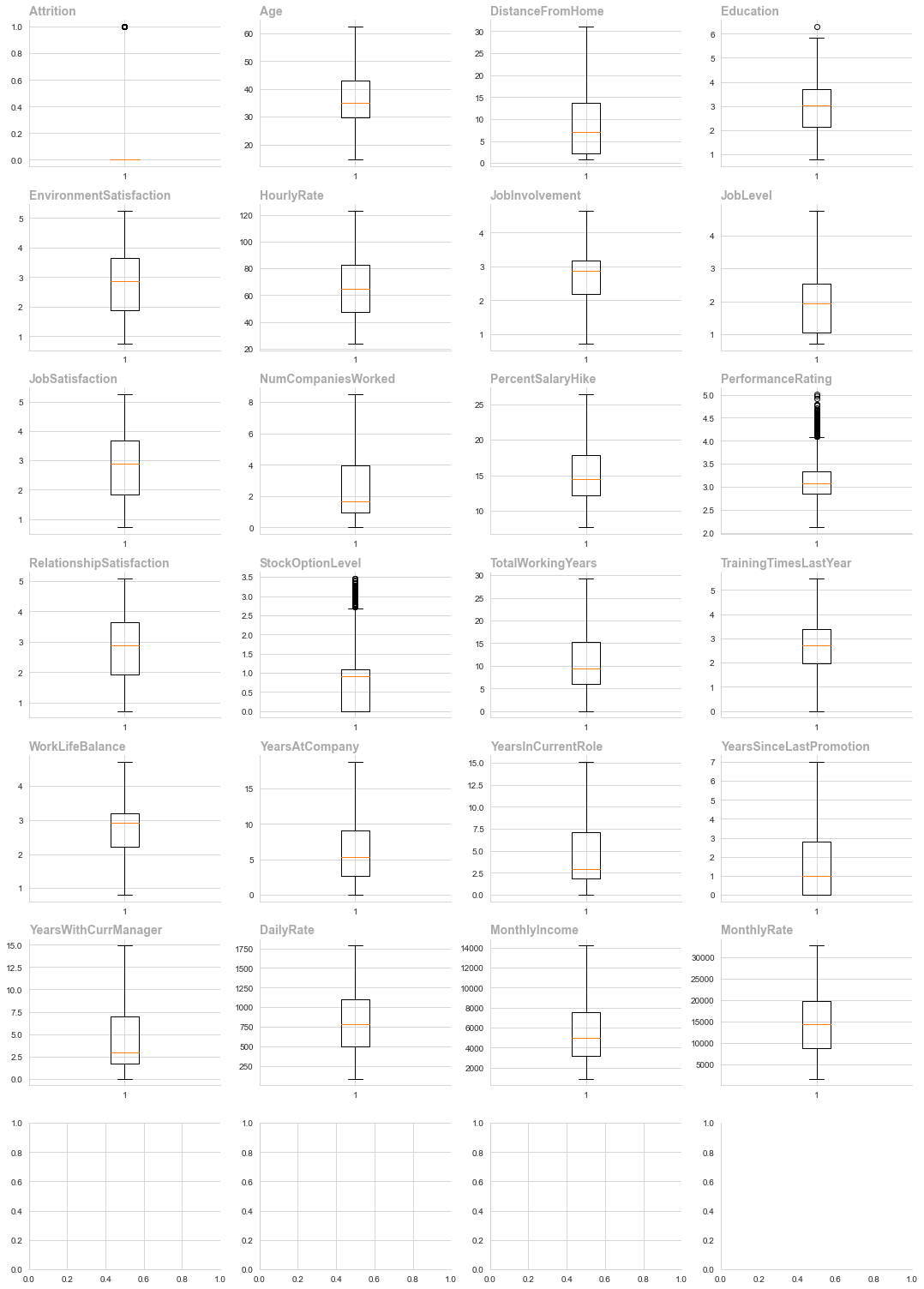
True False 15 YearsInCurrentRole -6.14625 15.1037

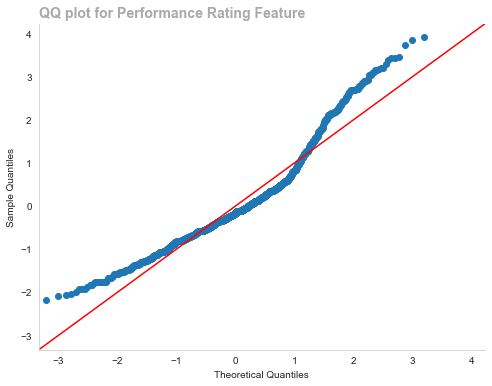
True False 153 YearsSinceLastPromotion -4.19625 6.99375

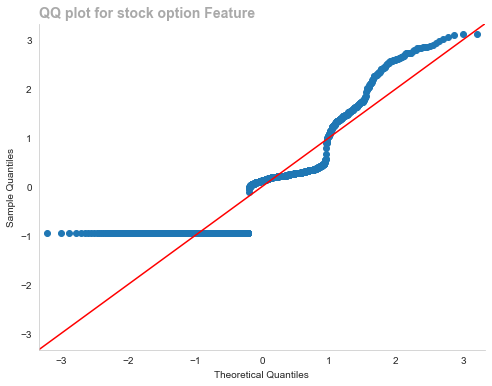
True False 14 YearsWithCurrManager -6.28625 14.9037

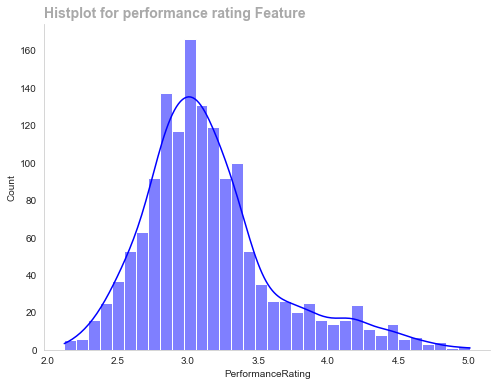
True False 135 MonthlyIncome -3438.8 14205.2

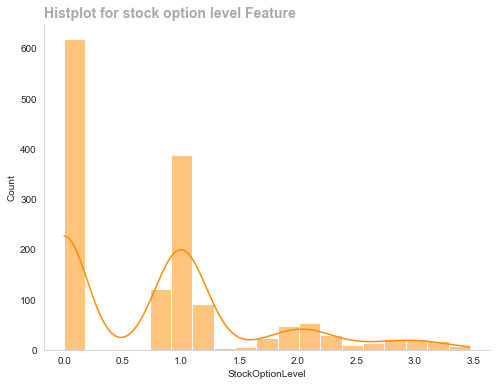
======================= ========== ======= ======================= ============= =============

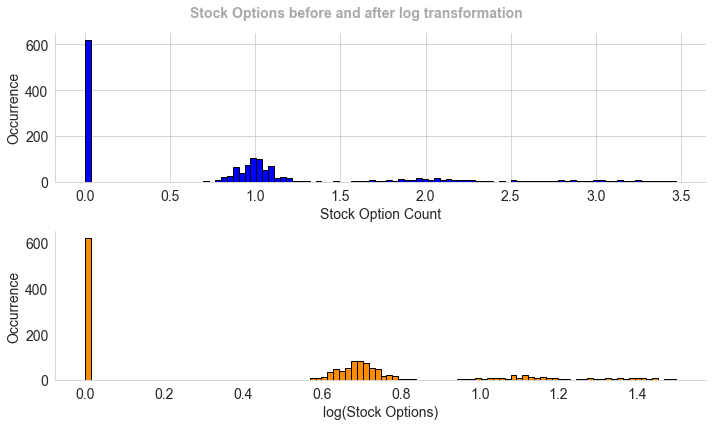












Summary of Dealing with Outliers

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 modelling.
* Assessment of Impact:Consider how outliers affect your study or modelling. 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 confimed by visual inspection of boxplots. We haave 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.

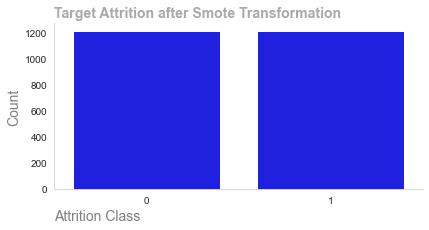
IQR can identify and remove outliers without affecting the dataset. This is crucial if outliers are anomalies, not errors.

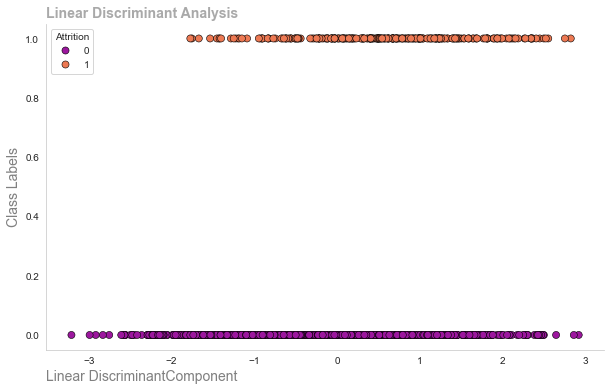
In LDA, importance:

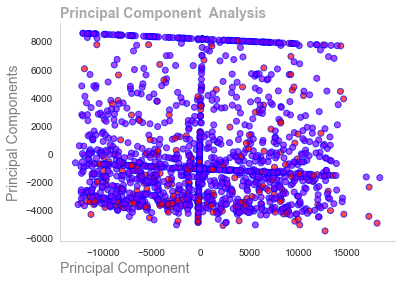
* Keeping Classes Separate:In Linear Discriminant Analysis, outliers can influence class parameter estimation. Maintaining class separation and accurate categorization requires handling outliers.
* Maximum Discrimination:LDA minimises within-class variance and maximises class mean distance. These measures can be greatly affected by outliers, reducing model discrimination.
* Sensitivity of ModelData dispersion affects LDA. Outliers can cause inaccurate Gaussian distribution assumptions, affecting model dependability.
* Dealing with Outliers Cons:Information Loss:Outlier removal or transformation may lose crucial information. Outliers can reveal insights or signify rare but notable events.
* Model biases:Rare events or unique patterns may cause extreme values. If these patterns are important to the analysis, removing them may bias the model.
* Subjectivity:Outlier handling is subjective and depends on the methodology. With no single solution, multiple approaches may yield different results.
* Risk of Overfitting:Aggressive outlier treatment can cause overfitting, when the model becomes overly suited to the training data and performs badly on new data.
* Assumption Error:Some outlier-handling approaches presume data distribution features. If these assumptions are violated, the approach may be ineffective.

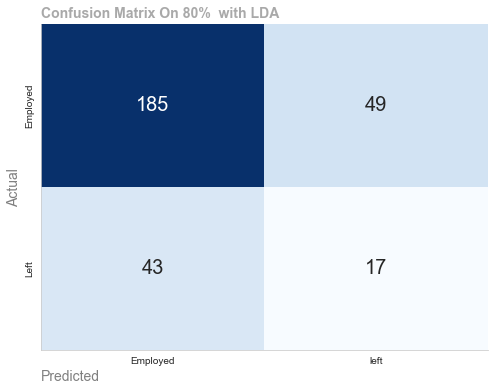
Based on data and analytic aims, outlier handling approaches' merits and cons must be considered. Two characteristics had outliers after IQR. More research was needed. The stock option level feature has extreme values, with few employees holding stock options and many not. This feature was transformed using log transformation. Box cox was better, but I couldn't get the code to work and had to move on due to time limits. However, these outliers may help us improve business satisfaction and productivity. It's vital to remember that not all outliers are the same and may some may help us with our analysis (word count 2168)

Linear Discriminant Analysis









Dimenaionality reduuction methods Linear Discriminant Analysis(LDA) and Principal Component Analysis(PCA) have diderent objectives assumptions on data classificaation and groupings

LDA Objectives:

* LDA is developed for supervised learning to maximise dataset class seperation. This sepreation forms decision boundariies where new data can be postioned
* Class Separation: LDA maximises between class variance to within class varietion(ie employed & leavers), if finda a projection that maximises class mean distance while minimising data point spread within each class
* Class Distributiion LDA assumes multivariate normal distribution for Attrition class, this is why we used SMOTE to balance our class using snthetic data. It assumes all classes have the same covariance matrix.
* LDA finds characteristics that maximise clas separation it excels at categorising data as can be seen from our visualizations.

Principal component analysis PCA:

* PCA is an unsuvervised method that maximises dataset variance regardless of class labels.
* Eigenvalue Maximisation: PCA finds orthogonal axes(principal components that maximise 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.

Implications for Clustering or Classifying:

* LDA feature extraction is used for classification problems, especially where class separability is important, it focuses on decision boundaries which help improbe cllassification accuracy.

PCA for Dimensionality Reduction:

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

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 se used for supervised classification algorithms where PCA can be used for unsupervised algorithms such as clustering.(2453 Words)

Model Preparation

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 centre) 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

.