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.

Contents

## Introduction ………………………………..5

## Credit Card Industry …………………………………………………………………………………………….6

Business Understanding 7

Data Understanding ………………………………………………………………………………………………...7

Libraries Used7

Overall Objectives8

Dataset Structure8

Initial Data Analysis9

Descriptive Statistics11

Inferential Statistics 17

Data Preparation & Data Preprocessing 17

Models First Run17

Dummy Regressor18

GradientBoosting Model Results18

Evaluation Plots20

Feature Importance21

Statistical Test22

[Flowchart of Data Preparation and Modeling](#_TOC_250008) 22

Conclusion22

[Challenges encountered](#_TOC_250005) 23

[Inclusion of strategies to overcome them 2](#_TOC_250004)3

Appendix- ….…………………………………………………………………………………………………...25

[References](#_TOC_250000) 26

**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.

## Introduction

As a data analyst, our task is to prepare and analyse the data set using appropriate data preparation,

statistical techniques and ML models. Our analysis will aim to identify any relationships or trends in the

data that can be used to improve employee satisfaction and productivity.

A study by the Center for American Progress on employee turnover showed that replacing a worker costs around 20% of their compensation. Increased turnover can result in significant expenditures that can be avoided by providing workplace flexibility and earned sick days at little or no cost.

Thus, most firms still find employee replacement expensive. The time spent interviewing and finding a replacement, sign-on bonuses, and the loss of productivity for several months while the new employee adjusts to the new workplace

## 

## (‘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 retention plan to boost employee satisfaction and productivity. Step-by-step systematic approach employing a strategy that can be utilized for many ML problems. 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). We investigate the probability of an employee leaving the organization (Y).

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 : v  
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 liabraries were used for data handling & analysis, data visualization, data preprocessing, 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

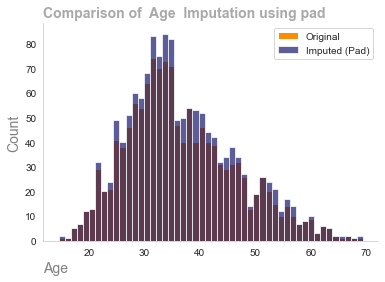
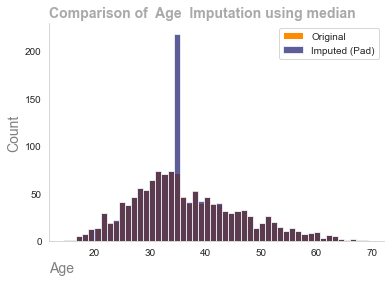
We have multiple imputation options available:

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

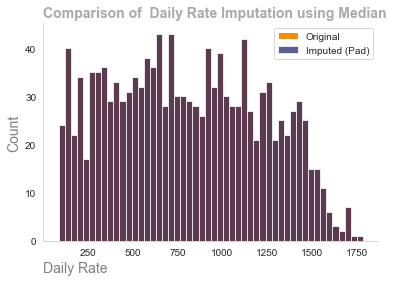
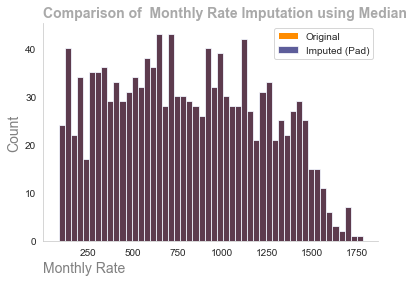
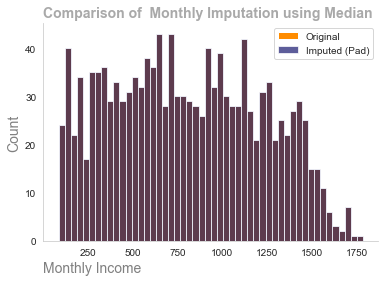
Each of the above models have their individual pros and cons.

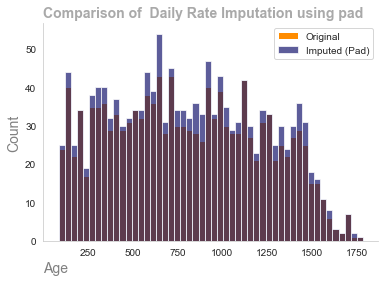
We will first deal with numerical values. Our strategy is use the the imputation method that is the simplest to use (Occam Razor Principle) while not been so simple as to introduce unwanted noise it the data.

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 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.

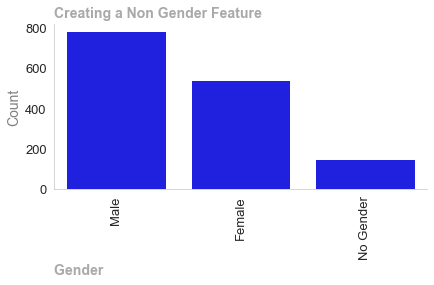
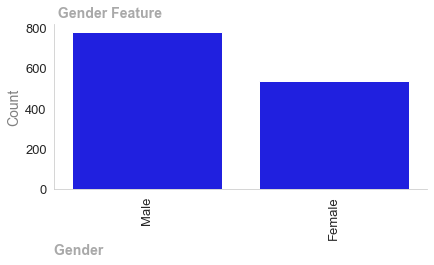
Categorical Data

The most commonly method used for categotical data is the mode imputaion (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.

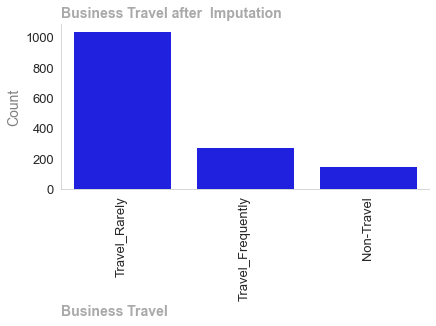
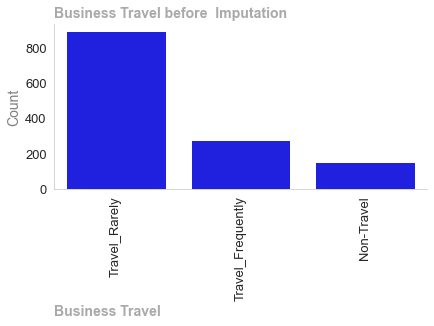
Our Strategy

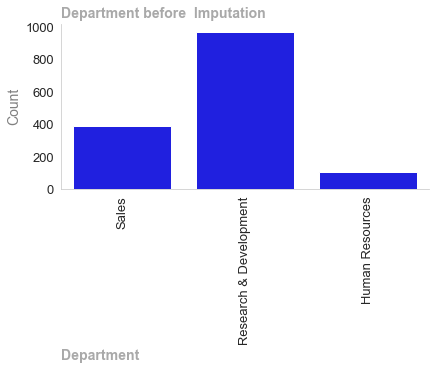
We have 8 features contain categorical database

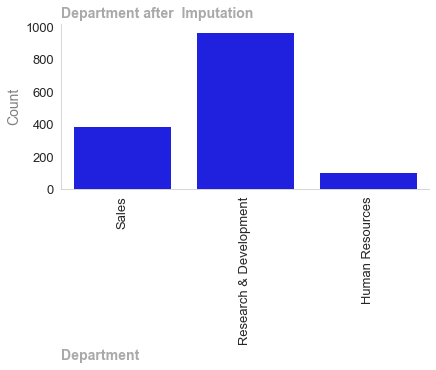
1. Assume mode will mode be the correct method of imputation
2. Evaluate results through the use of histograms to ensure the distribution of the original data has been maintained
3. If the distribution has not been maintained work on models such as predictive modeling.

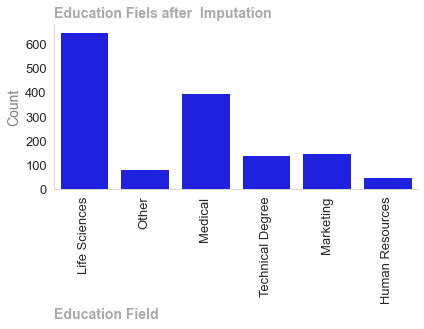
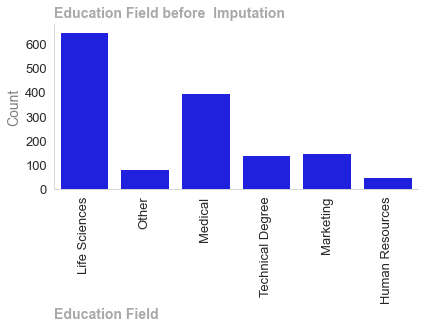


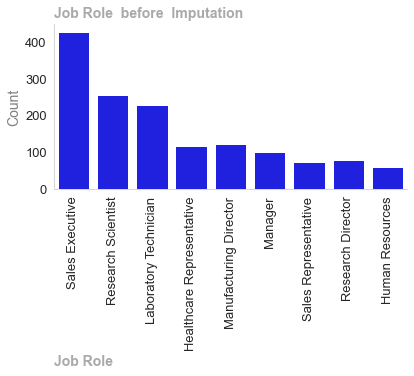
We can observe the problem we spoke about if we apply model imputation method to the gender feature the male column will increase exponentially. We need to consider another method which could predictive modeling such as KNN. However as data analyst we need to be a ware of modern society. Was the missingness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

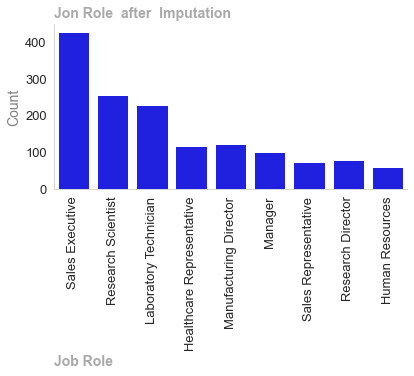


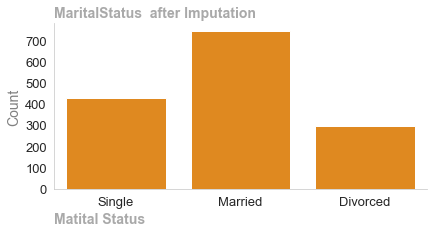
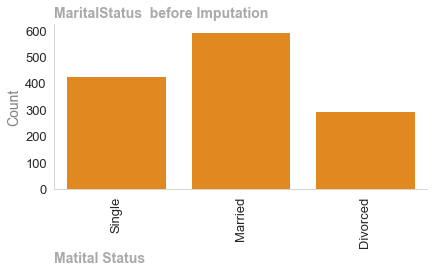






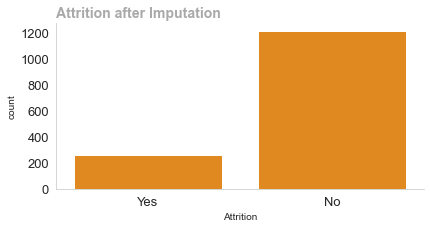
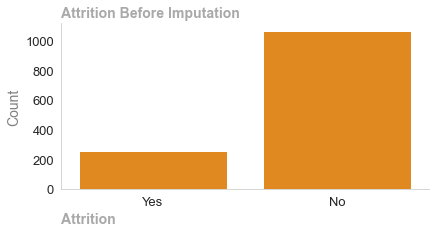






We have used the mode imputation method on the remaining features as can be observed the distribution of the data has been maintained.

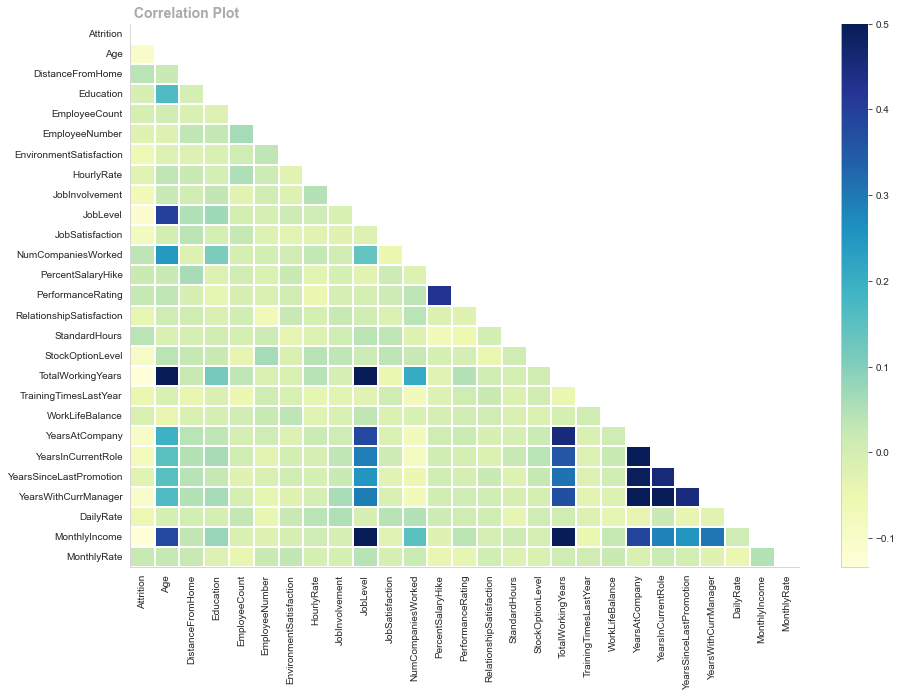
The attrition column which will be our target variable in the classification models we will run has two classed yes or no



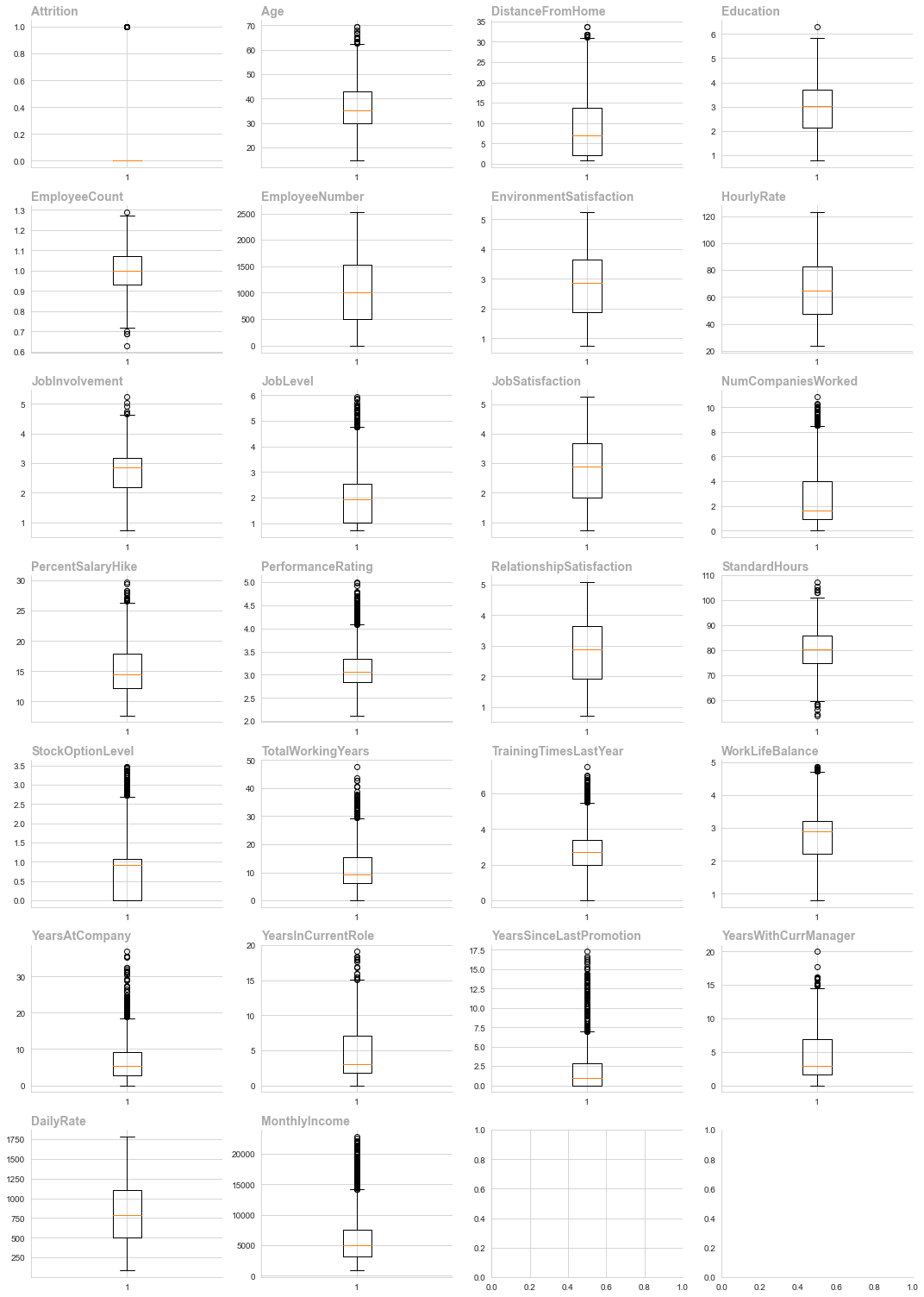
Here we observe where mode and not worked as it has more to imbalance to a unbalanced dataset. The course of action would be use a predictive model such as KNN that uses other features to calculate the missing values. However I could not get the code to work and have added my workings in the appendix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



As shown above, Monthly Income,Total Working Years and Job Level are negatively correlated to Attrition; while there are no positively correlated features to the target variable Attrition. a correlation plot without positive correlations does not indicate no link; it merely indicates a lack of linear association. To understand the date's relationships, we may need to examine non-linear correlations. Non-linear models may be useful when the correlation plot shows non-linear associations or if we suspect non-linear patterns in the data. Decision trees, random forests, SVMs, and neural networks can be used to model complex non-linear connections. This plot also guides as the need to use feature engineering to reveal relationships that are not apparent in the original data



We can observe from the boxplots that we have outliers in the majority of the numerical columns these need to be dealt with as our machine learning will not correctly well with skewed data

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

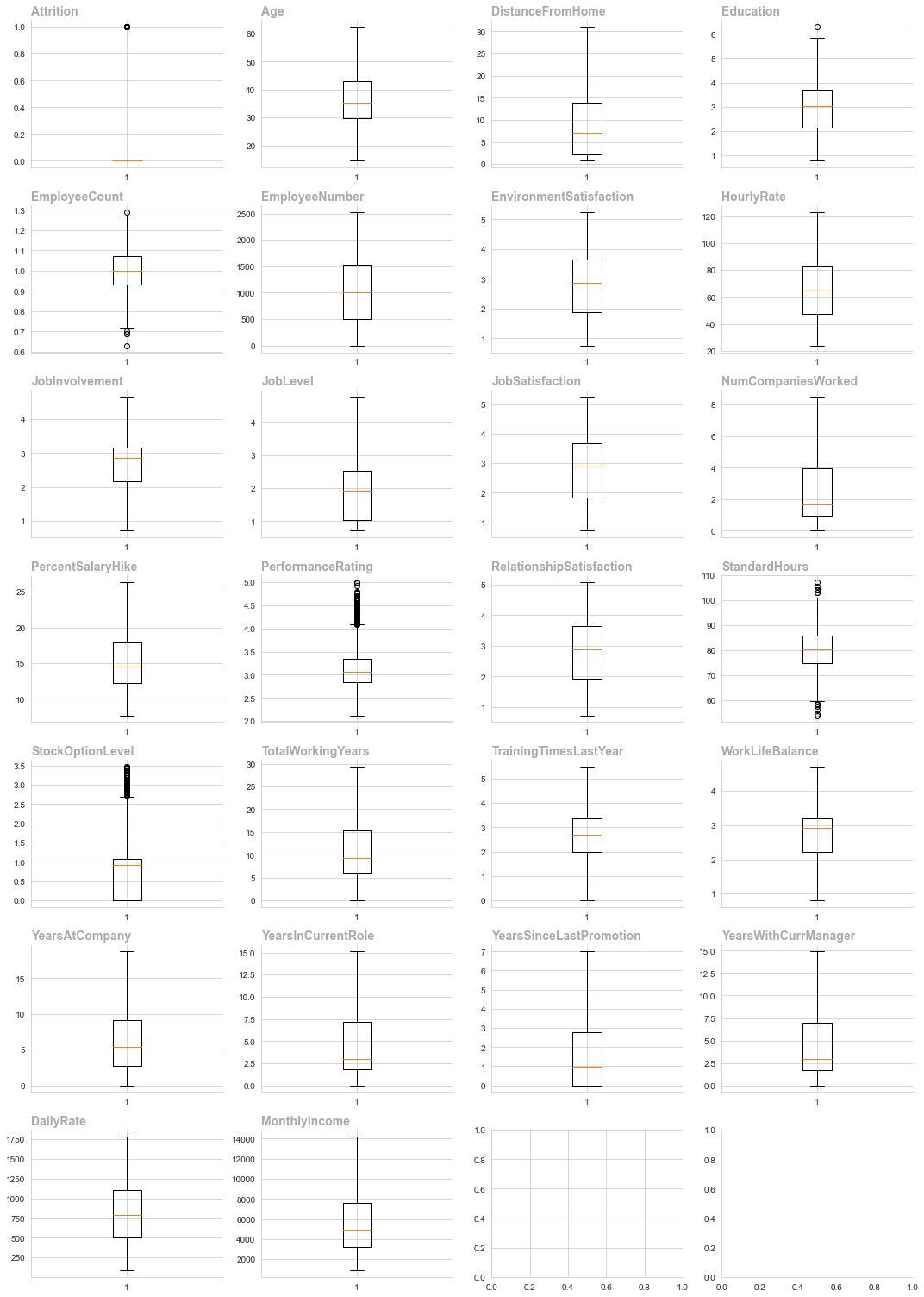
False False YearsSinceLastPromotion -5.33531 9.15469

True True 14 YearsWithCurrManager -6.61229 14.7702

False False MonthlyIncome -5102.13 17218.4

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

The tables summarise the outlier removal process for IQR nominal & 3 standard deviation, including the count of outliers and their IQR method-based range. Outlier elimination reduces extreme results and strengthens statistical studies and machine learning models



We can observe that we have dealt with the outliers by the use of IQR method . There are still 3 columns with outliers, Standard Hours,Stock Option level & performance rating we will need to deal with these 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.+