**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Higher Diploma Data Analytics for Business   * Data Preparation & Visualisation * Statistical Techniques for Data Analytics * Machine Learning |
| **Assessment Title:** | Individual / Practical |
| **Lecturer Name:** | Dr. Muhammad Iqbal  David McQuaid  Marina Soledad Iantorno |
| **Student Full Name:** | Miroslava Slavikova |
| **Student Number:** | sba22382 |
| **Assessment Due Date:** | 05 Jan 2024 |
| **Date of Submission:** | 02 Jan 2024 |

**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. |

Attrition – title……

**Subject area:**

Contents

[Introduction 3](#_Toc154935847)

[Methodology 3](#_Toc154935848)

[Libraries 4](#_Toc154935849)

[Data Preparation 5](#_Toc154935850)

[Handling missing values 5](#_Toc154935851)

[Exploratory Data Analysis 7](#_Toc154935852)

[Data Visualisation 7](#_Toc154935853)

[Data Exploration 8](#_Toc154935854)

[Outliers 11](#_Toc154935855)

[Feature Engineering 13](#_Toc154935856)

[Statistical Techniques 15](#_Toc154935857)

[Correlation Matrix 16](#_Toc154935858)

[Data Modelling 19](#_Toc154935859)

[Hyperparameter Testing 19](#_Toc154935860)

[Machine Learning using Best Hyperparameters 19](#_Toc154935861)

[Analysis of Results 20](#_Toc154935862)

[Challenges encountered 21](#_Toc154935863)

[Conclusion 22](#_Toc154935864)

# Introduction

A company has gathered data on its employees with intention to uncover patterns and trends that will assist in improving employee satisfaction, increase productivity and gain an understanding of how to keep a healthy employee retention rate.

With the provided dataset that includes wide range of information about their employees, we are going to look at the variety factors impacting employee attrition. By understanding the reasons why employees leave, the company can revise or develop new strategies to improve talent retention and in the long term reduce their recruitment budget



Figure For illustration purpose only

# Methodology

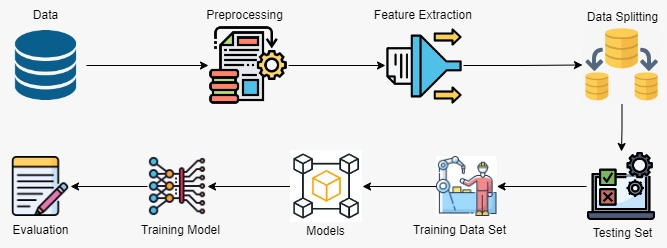
PCA/LDA

Hypothesis testing …..

Three different supervised machine learning models have been used:

* Decision Tree with accuracy score of
* KNN with accuracy score of
* Logistic Regression with accuracy score of

The results provided reasonable values, albeit not outstanding…



# Libraries

I have loaded all the necessary libraries, panda for data manipulation and analysis library, matplotlib.pyplot and seaborn for data visualisation, numpy for numerical computing, sklearn libraries for future scaling for ML models, PCA, encoding, training and testing sets, linear regression model for predictive modelling. Libraries have been imported and assigned the abbreviated formats. The abbreviated format makes recalling and use of these libraries more efficient. Lastly, I have uploaded a dataset "Employee\_Attrition.csv".

# Data Preparation

By using command df.head(), we get a quick overview, it shows first 5 rows of the DataFrame and we can look at the structure of the data, which is particularly useful when working with large datasets. When further inspecting the dataset, we have 1470 rows (observations) and 35 columns (features or variables). Additionally, the DataFrame contains following data types: 26 columns “float64(26)” representing numerical values with a decimal point and 9 columns “object(9)” representing string values.

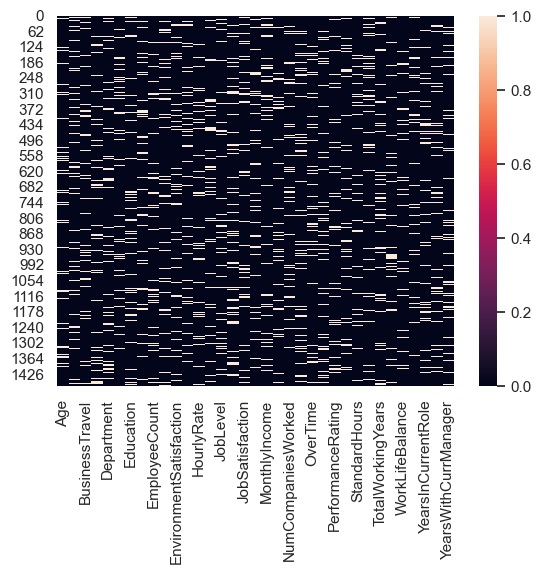
## Handling missing values

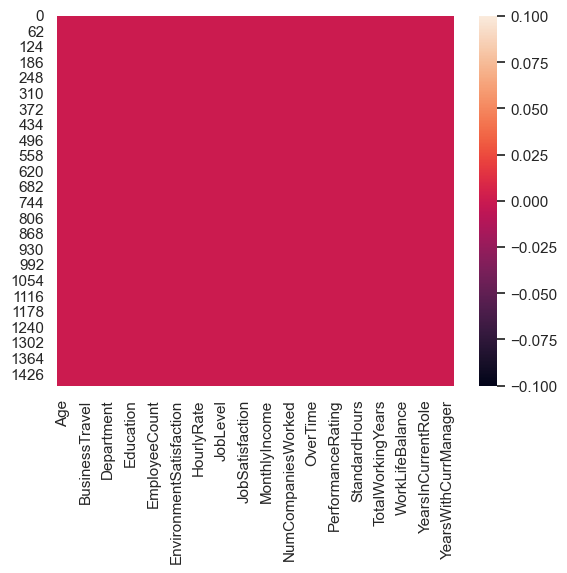
Missing data in dataset is a very common occurrence, therefore we need to understand, why is it missing and what exactly is missing. If we ignore the missing data, our results could lead into a bias result and reduce statistical power ultimately leading to invalid conclusions.

There are two possible options to deal with this, if the percentage of missing date is relatively low, we can perform imputation to replace the missing data. The second option is to remove data but it’s not inappropriate as it’ll lead into fewer observations resulting in inaccurate analysis and the data are lost.

From the heatmap below (fig. ), we can see that missing values are spread all across the dataset and when checking for Null Values, it’s apparent that 147 values are missing for each of the variables.

Missing data are typically filed with Mean or Median in case where the data is missing at random. The common method if we have to outliers is using Mean. Here, we impute missing values with the Mean which involves replacing NaN entries with the average value. Numeric are replaced with Mean (Mean of each column), however categorical are replaced with Mode, which represents the most common category, it is the category with the highest frequency.





The updated heatmap (fig. ) shows

that we have replaced the missing

values now.

# Exploratory Data Analysis

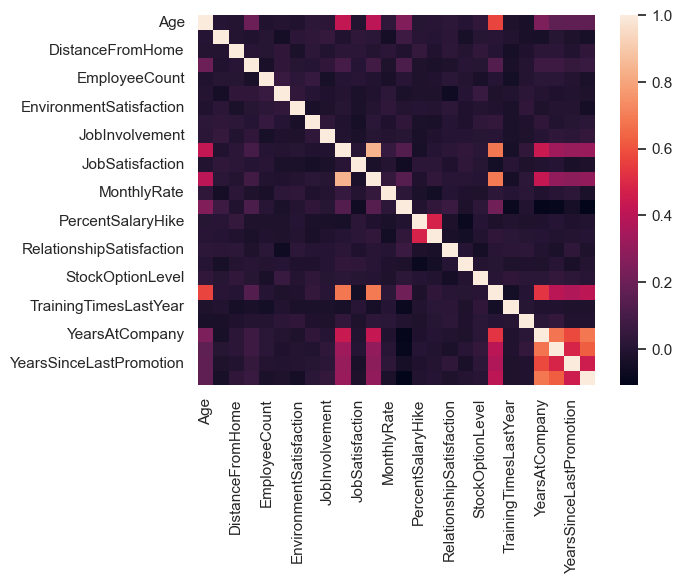
Exploratory Data Analysis (EDA)is a crucial process of performing an initial investigation of the data within the dataset. It helps us to gain better understanding of the dataset variables and the relationship between each of them. The main goal is to review the data, identify errors, missing values, understand the patterns within that might not be expected, find outliers or detect other anomalies. Python is a very useful and one of the most common data science tools to help us work on EDA.

## Data Visualisation

Data visualisations help us to better understand data, it’s a very effective technique to identify patterns, trends and possible outliers in our dataset. It helps us to present and communicate the data to wider audiences in a very effective way while delivering information they can understand and easily interpret.

The method “describe” helps us to display statistical values, to get a statistical summary and identify outliers for all our numerical columns. We obtained the following: count, mean, standard deviation, minimum, 25th percentile (first quartile), median (50th percentile), 75th percentile (third quartile), and maximum for each column.

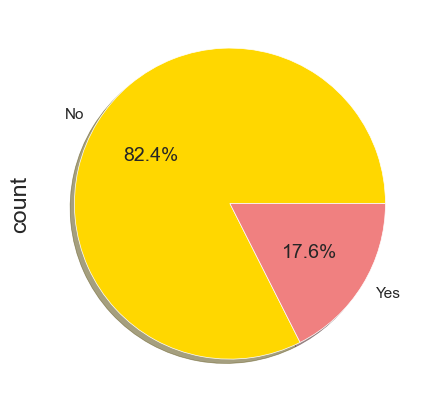
Correlation matrix (fig. ) is a very effective method to display correlation between multiple variables. It helps us to summarise the large data and easily identify patters, it shows clearly how the variables correlate with each other.



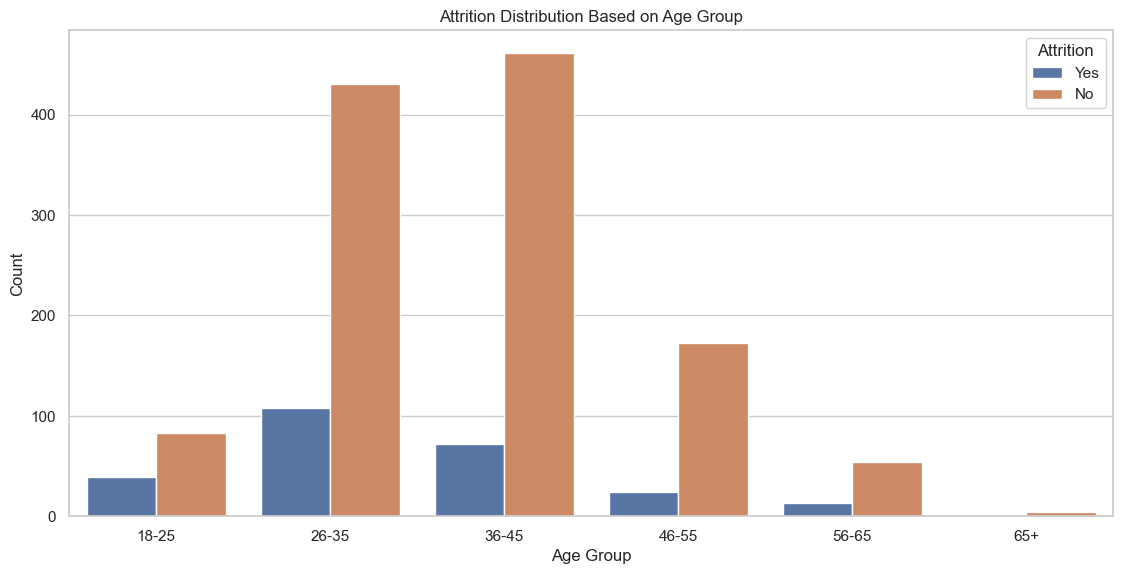
## Data Exploration

**Attrition**

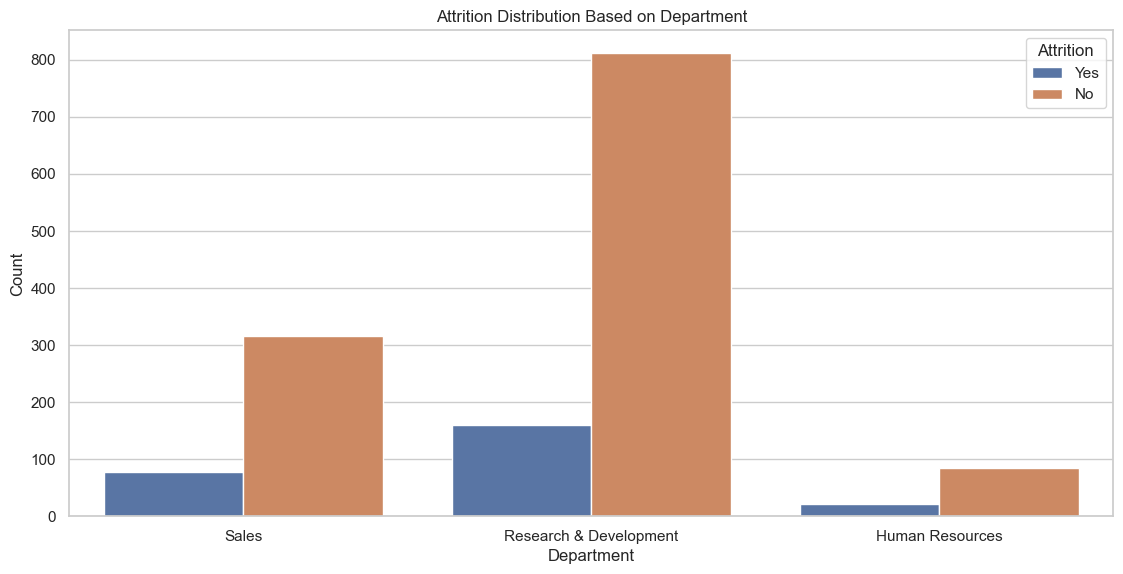
**As per the fig…. 82.4% is not attrite and 17.6% is not attire. Since the dataset does not have any tmeline, we are not able to deliver acurate results. But 10% is a market att art + ref**



**As per the below data (fig ….), we can see that attrition based on age is highest for age from 26 to 45 …**

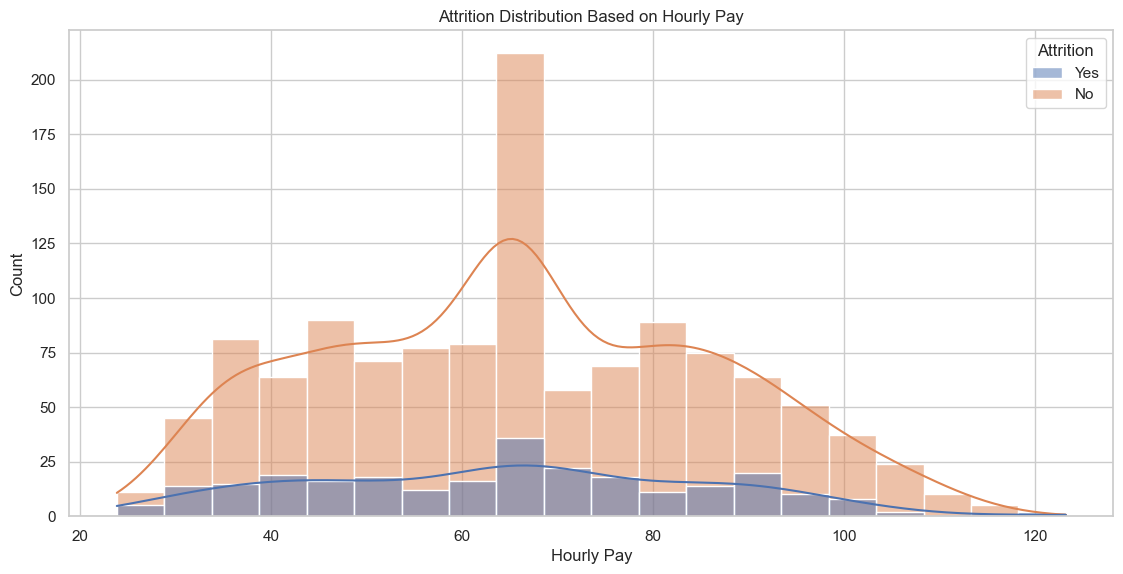


**Attrition per department …….**

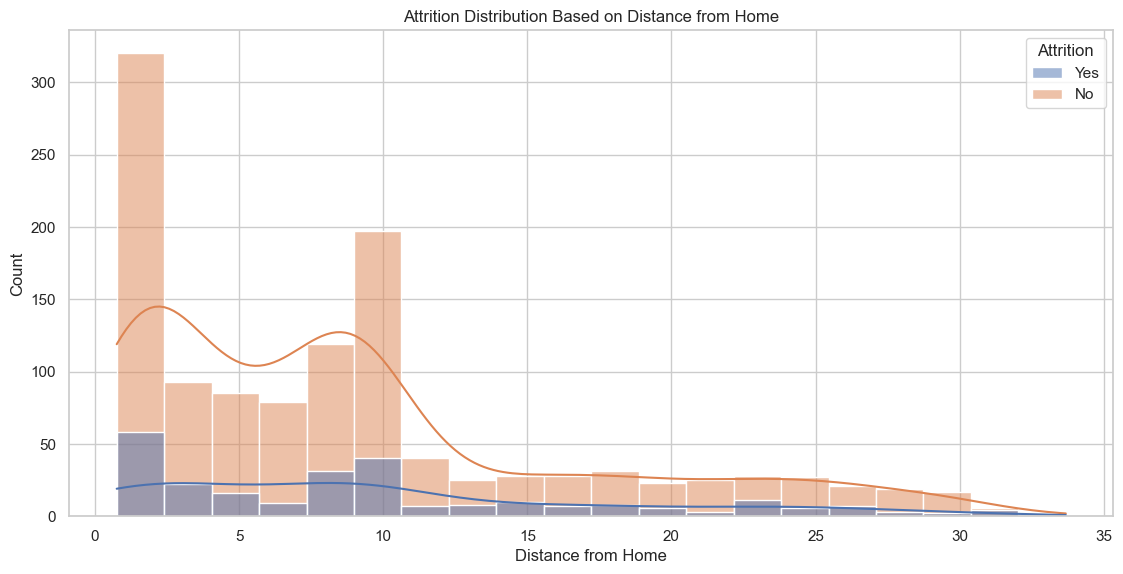


**Attrition per hourly pay**

**We can see that attrition is decreasing when the pay is increasing, so the highest earners are likely to stay. The average salary is likely to stay.**

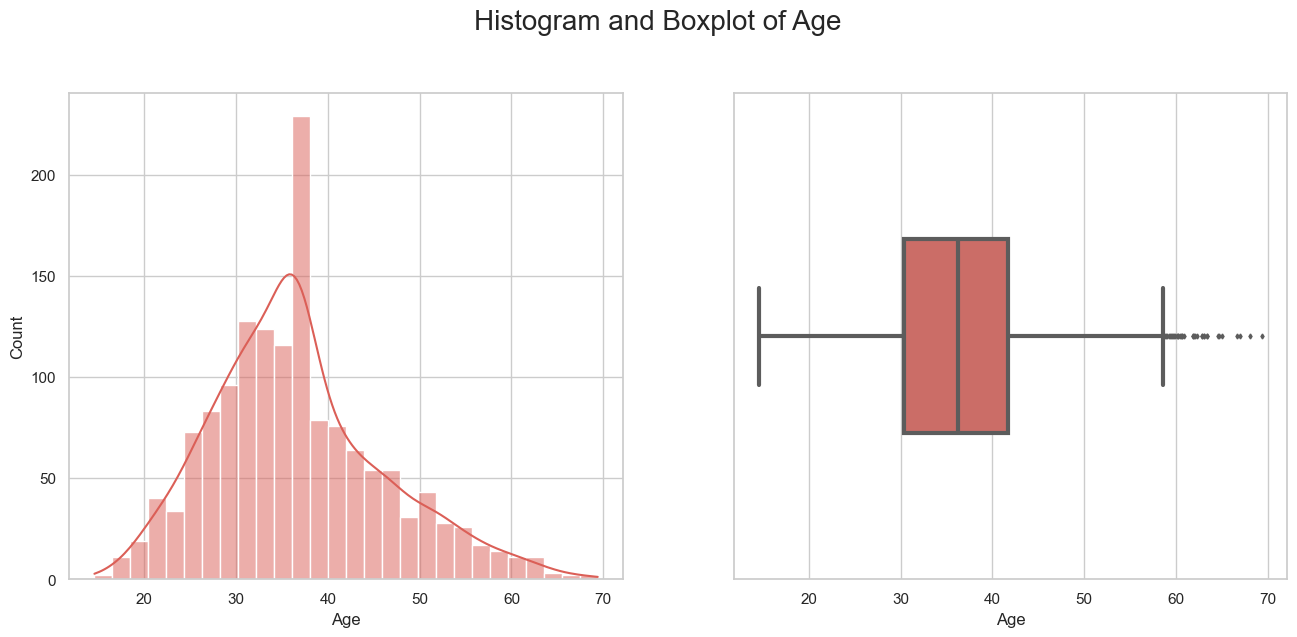


**Distance from home shows us that employees are likely to stay if they live close to the office.**

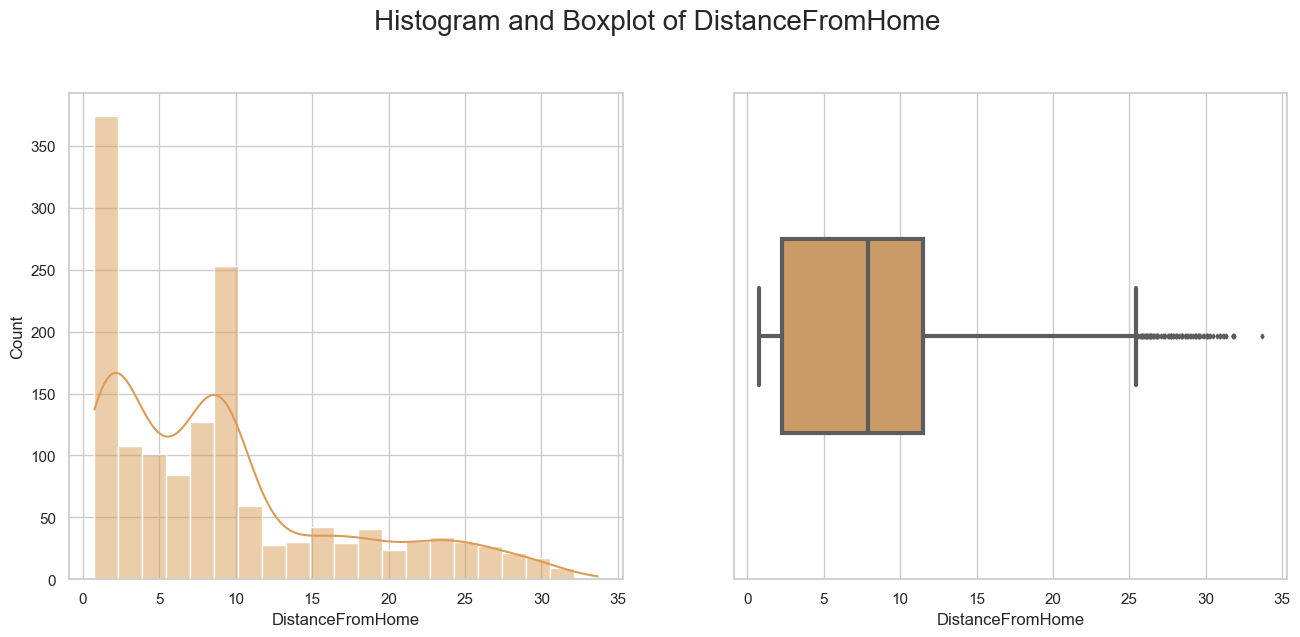


## Outliers

Checked outliers, that majority employees are 30-40 yo Mean age is 36



We can see that there is an outlier on the distance from home, but we don’t have to drop it, because based on our correlation results, I found out it won’t have any major effect the data.



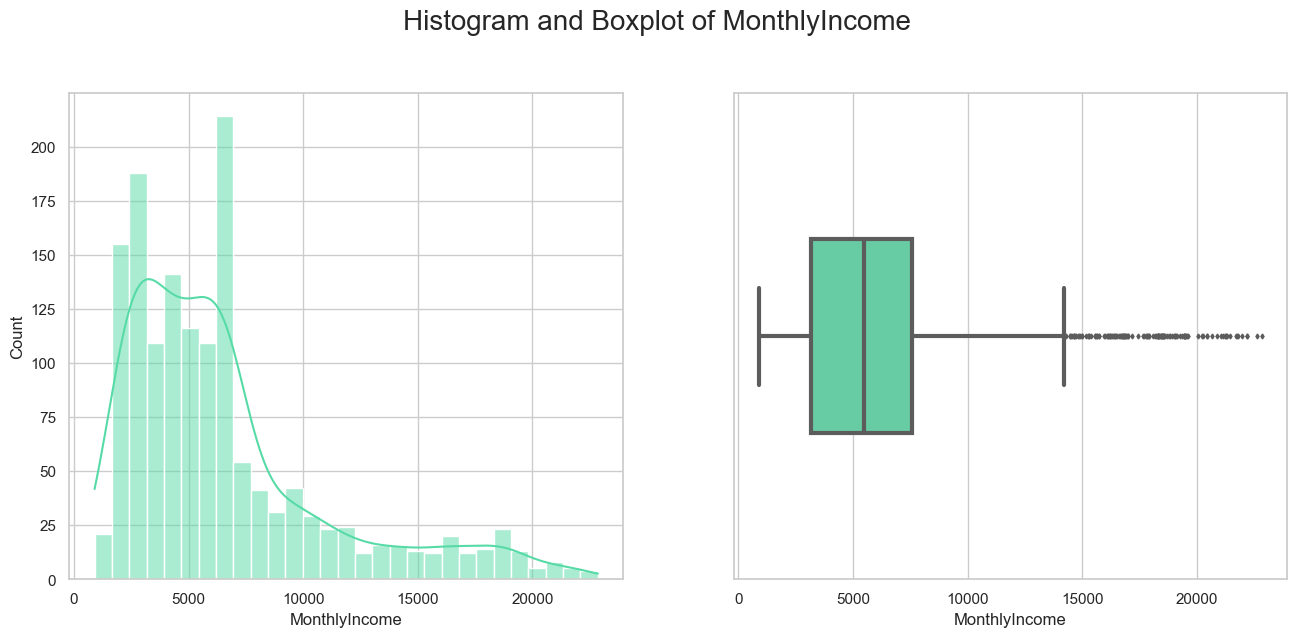
From the Histogram (fig. ) we can conclude, there are no outliers in the monthly income variable, we can conclude that the range of monthly income is following:

MonthlyIncome

Mean 6544.72

Median 5495.90

Mode 6544.72



# Feature Engineering

We performed encoding for feature engineering using the standard scaler…..

**Principal Component Analysis**

Principal Component Analysis (PCA), is often used when dealing with large datasets. Its goal is to transform a large dataset of variables into smaller ones, while retaining as much of the original information as possible within the large dataset. It is unsupervised learning technique that examines the relationships between variables.

It is taking all factors of the dataset and combining into one. They’re correlated with each other and ranked from the most important to least important. The new factors that we produce by inducing PCA are called principal components.

It is particularly useful for data preprocessing, noise reduction or feature extraction.

Some of the key disadvantage is the sensitivity to the scale of the data. If not scaled correctly, the PCA won’t work well. There is also a possibility of information loss, therefore, we need to select the most prominent components that we wish to retain.  Lastly, the PCA can be overfitted if too many components are used or the model is trained on a smaller dataset.

**Linear Discriminant Analysis**

Linear Discriminant Analysis (LDA) is a supervised learning algorithm. The main goal is to remove the redundant and dependent features by changing the dataset to a lower dimensional space. In atoner words, we reduce the dimensions, for example variables, while retaining majority of the data within the dataset.

**PCA versus LDA**

* Both PCA and LDA are used in case of dimensionality reduction
* PCA performs better with a smaller sample sizes
* LDA has better performance on multi-class classification tasks
* PCA is unsupervised technique that finds principal components that increase the variance in our data
* whereas LDA is a supervised technique that projects features in higher dimensional  space to a lower dimensional space. This is to avoid the curse of dimensionality.

PCA/ KNN

We use StandardScaler to encode categorical values into numerical

Results

LDA/ KNN



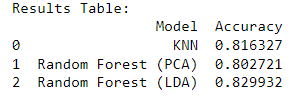
We reduced the noise …….how it’s done by LDA/PCA

In the comparison between Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), it has been observed that the accuracy achieved through PCA is higher than that of LDA. This outcome can be attributed to the inherent characteristics and goals of each dimensionality reduction technique.

PCA is primarily designed for unsupervised dimensionality reduction, aiming to capture the maximum variance in the data without considering class labels.

LDA is a supervised technique that considers class labels during dimensionality reduction. It aims to find the linear combinations of features that maximize the separation between different classes.

LDA/ Random Forest



Why is PCA is working better with KNN and LDA with RandomForest …

We use 3 models, PCA, LDA, KNN and we see that the best accuracy is derived y Random Forest LDA with accuracy score of 0.83.

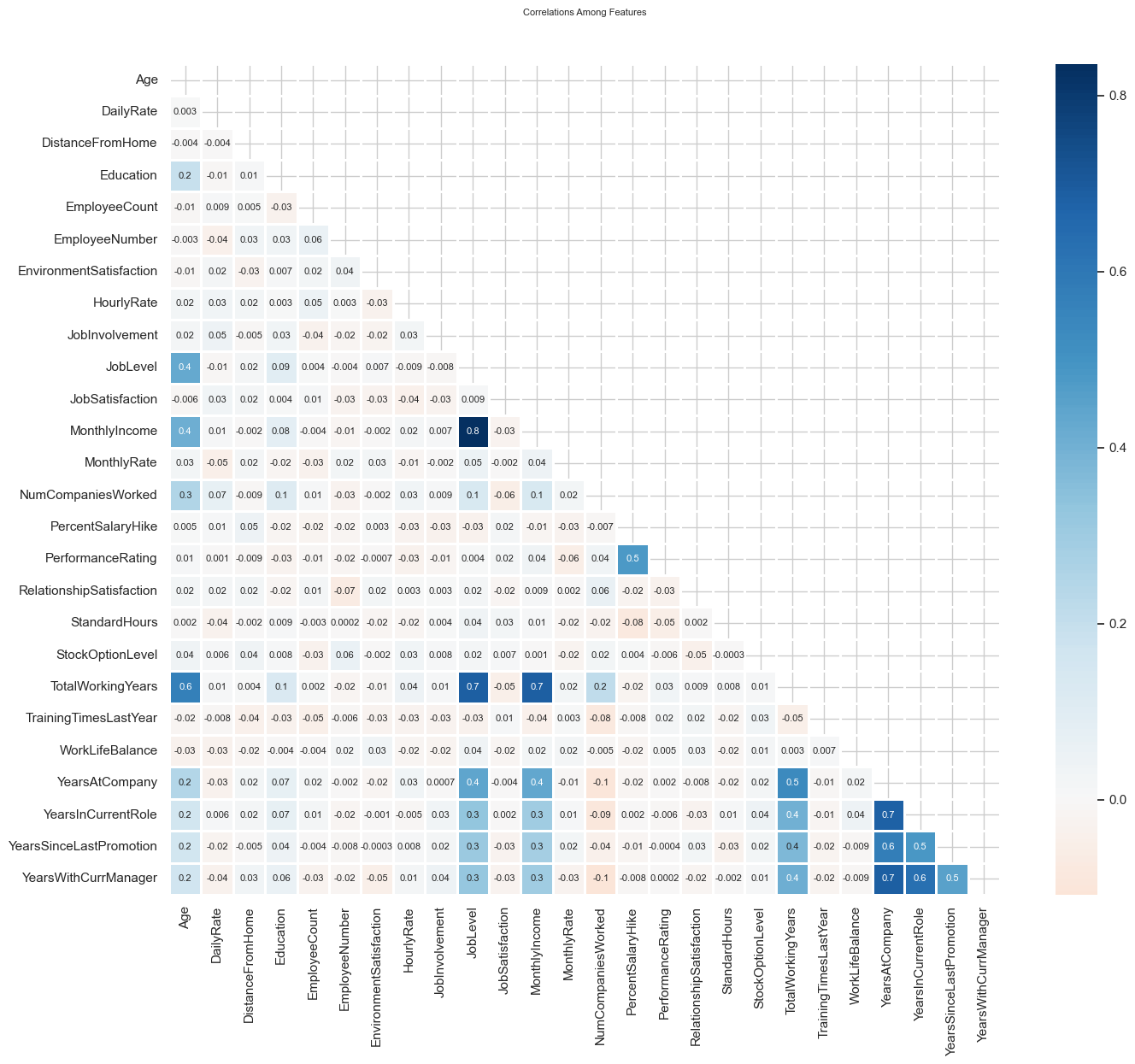
KNN works on the probabilities

# Statistical Techniques

**Measures of Central Tendency:**

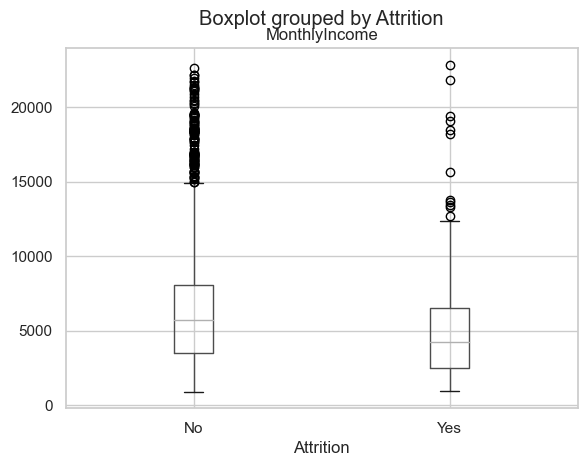


Correlation Matrix **Findings…. What is correlated**

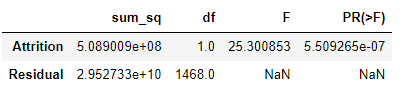


Hypothesis Testing

We use 2 variables to identify how the monthly income affects the attrition and ,..

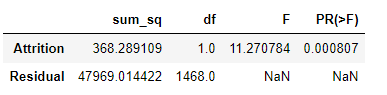


Annova



The p-value for 'Attrition' is very low (5.509265e-07), suggesting that there is a significant effect of 'Attrition' on 'MonthlyIncome.'

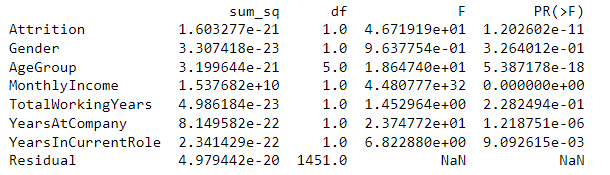
Year is the company does not affect



Hypothesis test where

Dep. Variable: MonthlyIncome

Anova table where obtain good p-values, which is higher than 0.05, anything below BAD



# Data Modelling

## Hyperparameter Testing

Selection of Hyperparameters

I used the n estimators

'n\_estimators': [50, 100, 200]

'max\_depth': [None, 10, 20],

Models decided:

Best Hyperparameters: {'max\_depth': 10, 'min\_samples\_leaf': 4, 'min\_samples\_split': 2, 'n\_estimators': 100}

Grid search cv

This means that the model correctly predicted the target variable 'Attrition' for about 80.27% of the instances in the test data. (NOTES: overfitting ……70 is bad, 95 is overfitted) too many variables, with one or 2 is easier

## Machine Learning using Best Hyperparameters

KNN

Random Forest PCA

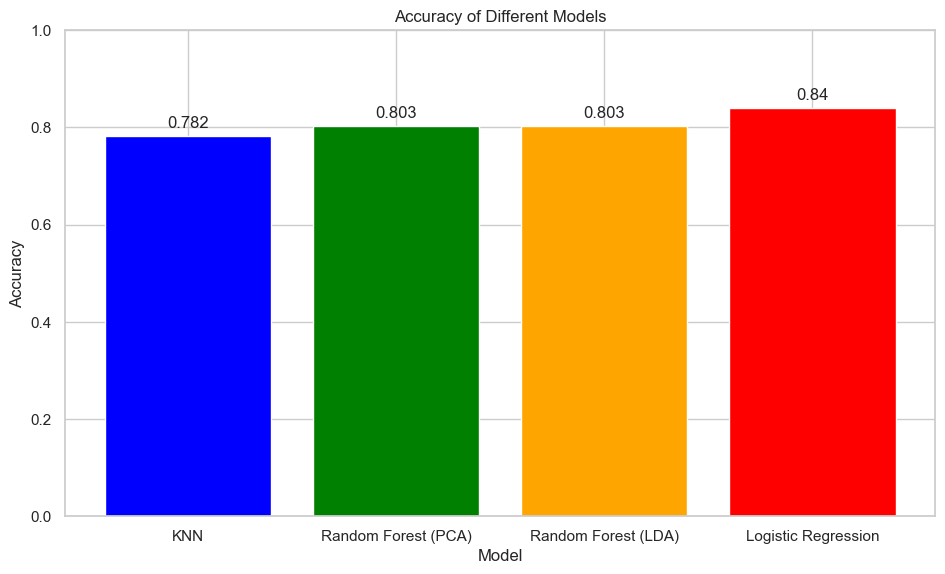
Randon Forest LDA

K-fold using split 10 (least time to run and most accurate)

, 20, 30 (30 will take more optimization)

## Analysis of Results

**We used the 3 models… present results**



# Challenges encountered

I have encountered numerous challenges while working on this dataset. Larger volume of data provides more information for the training and might lead to better performance.

Additionally, I’ve identified a lot of missing vales, and identifying the best performing model with better accuracy score.

Admittedly, the project overall was a challenging task, but generally speaking, I really enjoyed learning and attempting to understand such a complex subject of study as Data Analysis and Machine Learning.

**Milestones**

Researching while working on this project helped me to gain better understanding of visualisation techniques and how to interpret the various graphs and charts, gradually gaining confidence in this subject. I've also developed a better understanding of GitHub, including version control and how to create repositories.

## Conclusion

Employee turnover peaked in 2021 at 18%, up from 8% in 2020. This was the highest percentage recorded, since the HR Barometer Report was established in 2017. It was projected that the turnover will reach 18.2% for the year 2022. As reported by 56% employers, the main cause of the increase is an attractive renumeration, followed by the key factors, like better work-life balance and career advancement.

Commenting on increasing salaries, Sarah Fagan, Managing Director of Adare Human Resource Management said: “While the market is seeing increases in salary, total reward is not just about the financial benefits. It should incorporate the intrinsic value brought through your business culture. It recognises and shares successes of employees, it balances paid benefits with those of a non-financial value. And, most importantly, it considers the employee experience as an asset that requires the right investment.” (Fagan, 2022)

At the same time, the expense of hiring has significantly increased to as high as €7,491 per employee, going up from €4,215 in 2021 and €6,895 in 2020. For large companies, with more than 250 workers, the cost of new hire soars to €14,690 per person.

Over the past year, 39% of businesses have considered flexing working arrangements, that includes hybrid or work from home models, to be crucial indicative that will lead to higher employee retention.

**GitHub link:**

**References & Bibliography:**

[**https://www.siliconrepublic.com/careers/irish-workers-concern-2023-tech-trends#:~:text=According%20to%20the%20report%2C%20the,planning%20to%20retain%20valued%20employees**](https://www.siliconrepublic.com/careers/irish-workers-concern-2023-tech-trends#:~:text=According%20to%20the%20report%2C%20the,planning%20to%20retain%20valued%20employees)**.**

[**https://www.fingalchamber.ie/news/details/employee-turnover-more-than-doubles-in-12-months-as-employers-battle-to-keep-staff**](https://www.fingalchamber.ie/news/details/employee-turnover-more-than-doubles-in-12-months-as-employers-battle-to-keep-staff)

[**https://www.cipd.org/globalassets/media/knowledge/knowledge-hub/reports/2023-pdfs/2023-hr-practices-ireland-report-8238.pdf**](https://www.cipd.org/globalassets/media/knowledge/knowledge-hub/reports/2023-pdfs/2023-hr-practices-ireland-report-8238.pdf)

<https://www.kaggle.com/>

<https://pypi.org/project/fasteda/>

<https://towardsdatascience.com/how-to-clean-your-data-in-python-8f178638b98d>

<https://realpython.com/>

<https://www.geeksforgeeks.org/>

<https://pandas.pydata.org/>