**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:** | David McQuaid  Marina Soledad Iantorno  Dr. Muhammad Iqbal |
| **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. |

**A Data-Driven Exploration of Employee Turnover Trend**

Contents

[Introduction 3](#_Toc155090579)

[Methodology 3](#_Toc155090580)

[Libraries 5](#_Toc155090581)

[Data Preparation 5](#_Toc155090582)

[Handling missing values 5](#_Toc155090583)

[Exploratory Data Analysis 7](#_Toc155090584)

[Data Visualisation 8](#_Toc155090585)

[Data Exploration 9](#_Toc155090586)

[Outliers 12](#_Toc155090587)

[Feature Engineering 14](#_Toc155090588)

[Principal Component Analysis 14](#_Toc155090589)

[Linear Discriminant Analysis 15](#_Toc155090590)

[Statistical Techniques 17](#_Toc155090591)

[Measures of Central Tendency 17](#_Toc155090592)

[Correlation Matrix 19](#_Toc155090593)

[Data Modelling & Hyperparameter Testing 21](#_Toc155090594)

[Machine Learning using Best Hyperparameters 22](#_Toc155090595)

[Analysis of Results 23](#_Toc155090596)

[Challenges encountered 24](#_Toc155090597)

[Conclusion 24](#_Toc155090598)

# 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 variety factors impacting employee attrition. By understanding the reasons why employees leave, the company can revise current policies and develop new strategies to improve talent retention and in the long term reduce their recruitment budget.



Figure For illustration purpose only

# 

# Methodology

In this study, we will explore the dataset and deep dive into it while using several analysis methodologies, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and hypothesis testing. We will also go over the results of three different supervised machine learning models: Decision Tree, K-Nearest Neighbors (KNN), and Logistic Regression.

**Dimensionality Reduction using PCA and LDA:**

* PCA is a dimensionality reduction approach that is used to convert a high-dimensional dataset into a lower-dimensional space while preserving as much of the original variability as feasible.
* LDA, like PCA, is a dimensionality reduction method that focuses on maximizing the separation between distinct classes in the dataset. It is especially effective for enhancing classification model performance.

**Hypothesis Testing:**

Based on the sample information, the testing of hypotheses was used to derive statistical inferences about the population. In validating hypotheses and coming to conclusions regarding the relationships between variables in the dataset.

**Machine Learning Models:**

Three different supervised machine learning models have been used:

* Decision Tree
* K-Nearest Neighbor
* Logistic Regression

**Methodology Steps (fig. 2)**

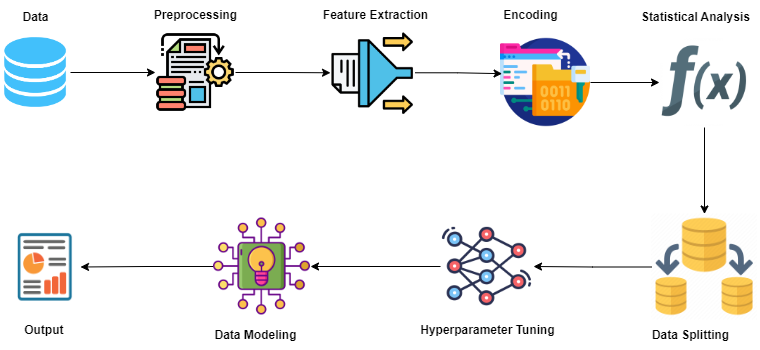


Figure Methodology steps

# Libraries

Libraries are very useful functions and they play important roles in developing machine learning, data visualisations, data manipulations applications and many more, without the need to write the codes from scratch.

For the purpose of our dataset analysis, the following libraries have been loaded and used - panda for data manipulation and analysis, 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, we have uploaded a dataset "Employee\_Attrition.csv", as provided by the client.

# 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.

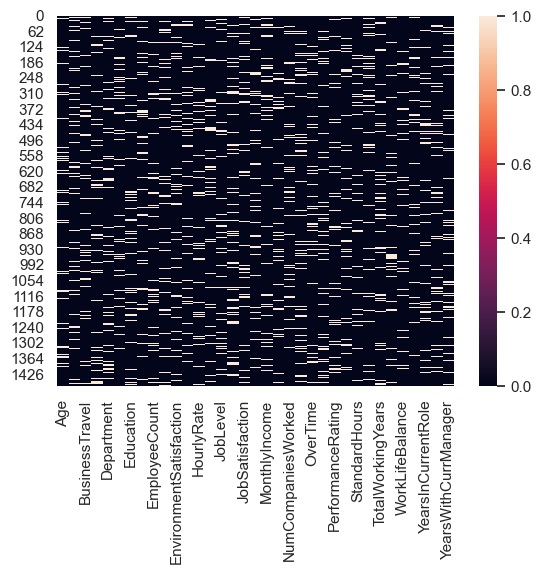
## Handling missing values

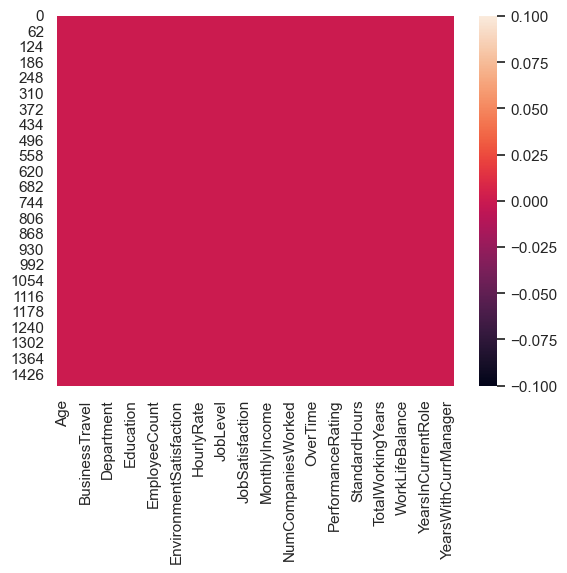
Missing data in dataset is a very common occurrence, therefore we need to understand, why is it missing and what exactly is missing. If we ignore the missing data, our results could lead into a bias result and reduce statistical power ultimately leading to invalid conclusions.

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. 3), 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.



Figure Heatmap of missing values

The updated heatmap (fig. 4) shows

that we have replaced the missing

values now.

Figure Heatmap shows missing values are replaced

# Exploratory Data Analysis

Exploratory Data Analysis (EDA)is a crucial process of performing an initial investigation of the data within the dataset. As a good practice, it is important to understand the data and obtain as many insights from it as feasible.

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.

The fundamental steps of EDA are following:

1. Data Cleaning
2. Descriptive Statistics
3. Data Visualization
4. Feature Engineering
5. Correlation and Relationships
6. Data Segmentation, if required
7. Hypothesis Testing
8. Conclusions

## 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. 5) 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.

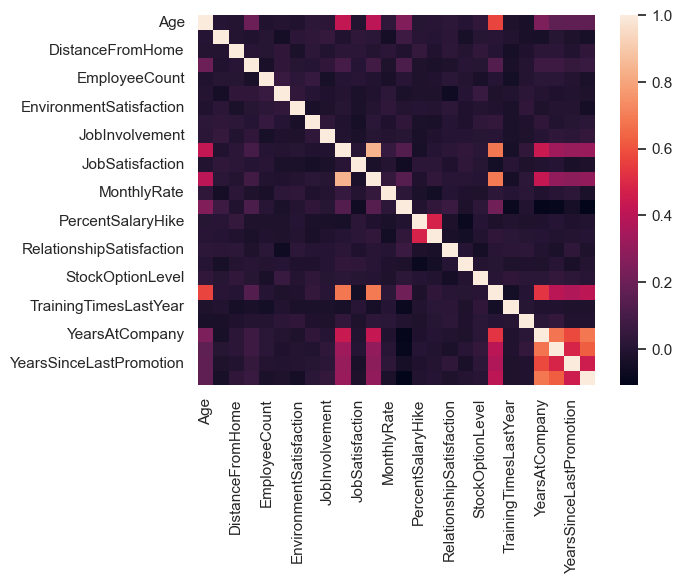


Figure Correlation Matrix

## Data Exploration

**Attrition**

According to the pie chart (fig. 6), the data shows that 82.4% employees are not attrite and 17.6% are attrite. The attrition rate is rather high, but since the dataset does not have any timeline, we are not able to deliver accurate results for a specific time, for example per annum. However, according to the market report, the current market employee turnover is expected at 11.1%.

“In 2022, retention was already a hot topic. Paul Dooley, CEO of Typetec, told SiliconRepublic.com that [his approach](https://www.siliconrepublic.com/business/typetec-paul-dooley-leaders-insights) to tacking the problem has involved the IT company switching to a four-day working week. And Careers editor Jenny Darmody looked into the phenomenon of ‘[stay interviews](https://www.siliconrepublic.com/careers/staff-retention-stay-interviews-employee-engagement)’ for staff retention.

For the coming year, talent acquisition, retention and recruitment will remain the top priorities for businesses, according to [Adare Human Resource Management’s](https://www.adarehrm.ie/) HR Barometer Report.

According to the report, the expected average employee turnover of 11.1pc in 2023 demonstrates a return of confidence in the labour market from the upheaval felt in 2022.

The report recommends businesses move away from reactionary measures and lean more towards longer-term planning to retain valued employees. (O’Dea, 2023)”

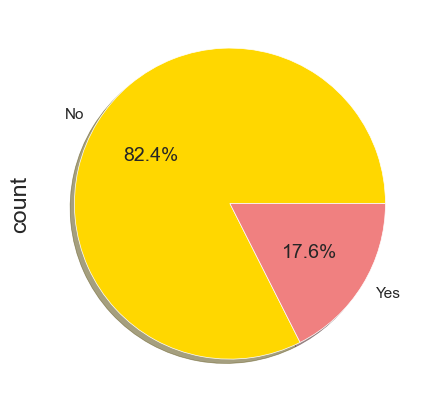


Figure Employee Attrition pie chart

**Attrition per age**

As per the below data (fig. 7), we can see that attrition based on age is highest between the ages from 26 to 45 years old.

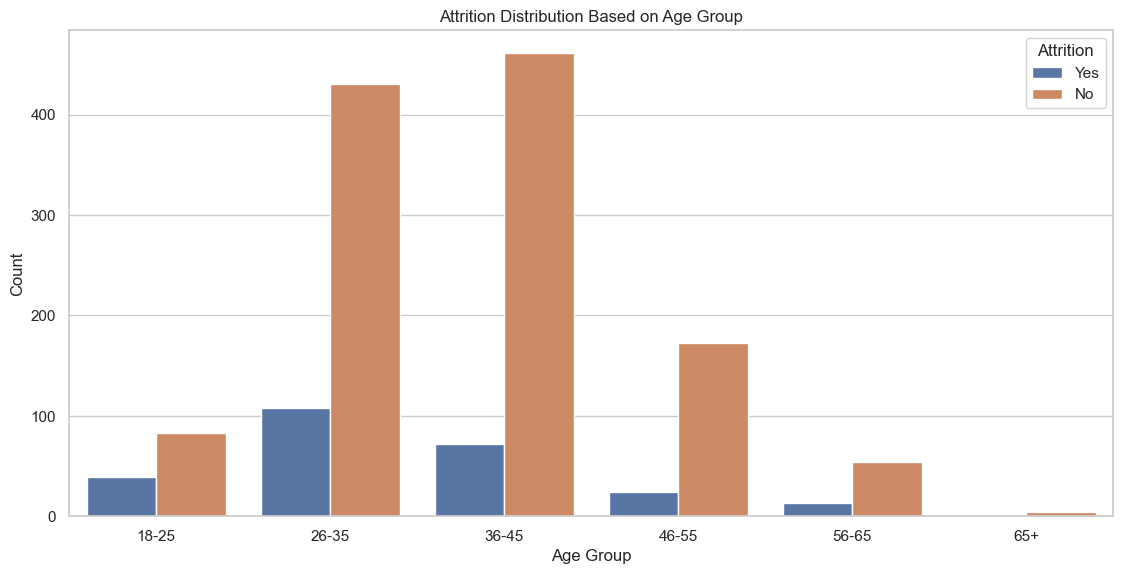


Figure Attrition per age

**Attrition per department**

As per the below data (fig. 8), we can see that attrition based on the company’s department is highest for the Research & Development but the department also has the largest number of employees working in the department.

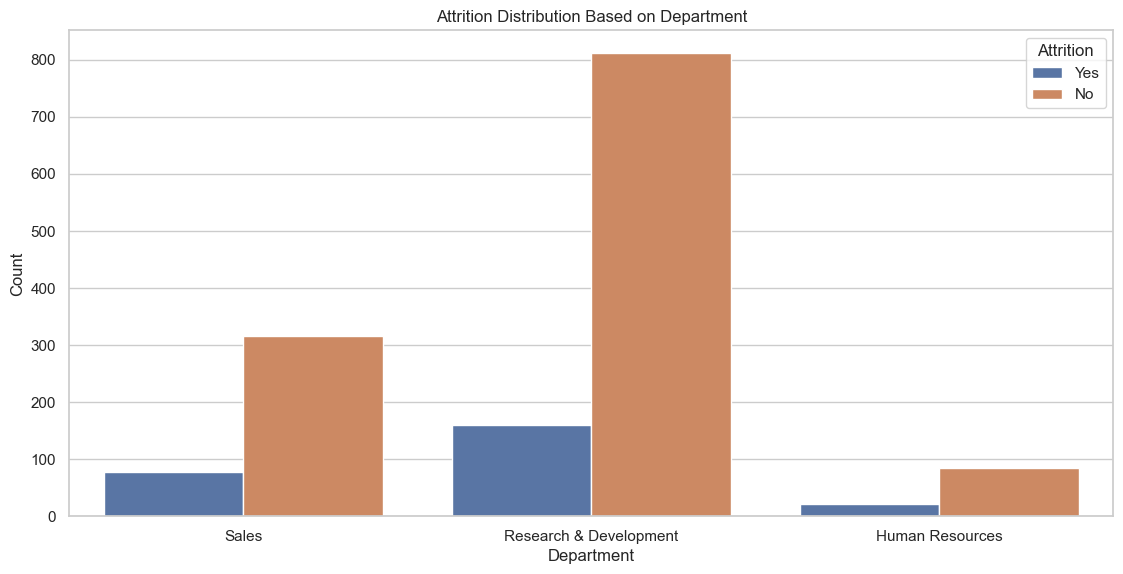


Figure Attrition per department

**Attrition per hourly pay**

As per the data (fig. 9), we can conclude that attrition is decreasing when the pay is increasing, therefore the highest earners are likely to stay. The average salary earners are likely to stay as well.

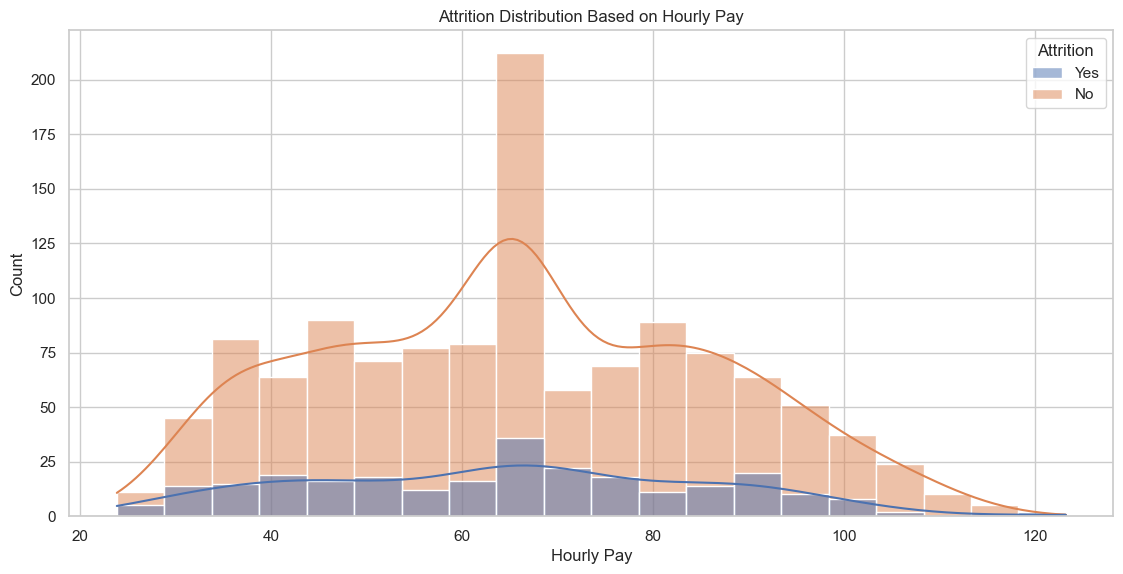


Figure Attrition per hourly pay

**Attrition based on distance from home**

Distance from home (fig. 10) shows us that employees are likely to stay if they live close to the office**.**

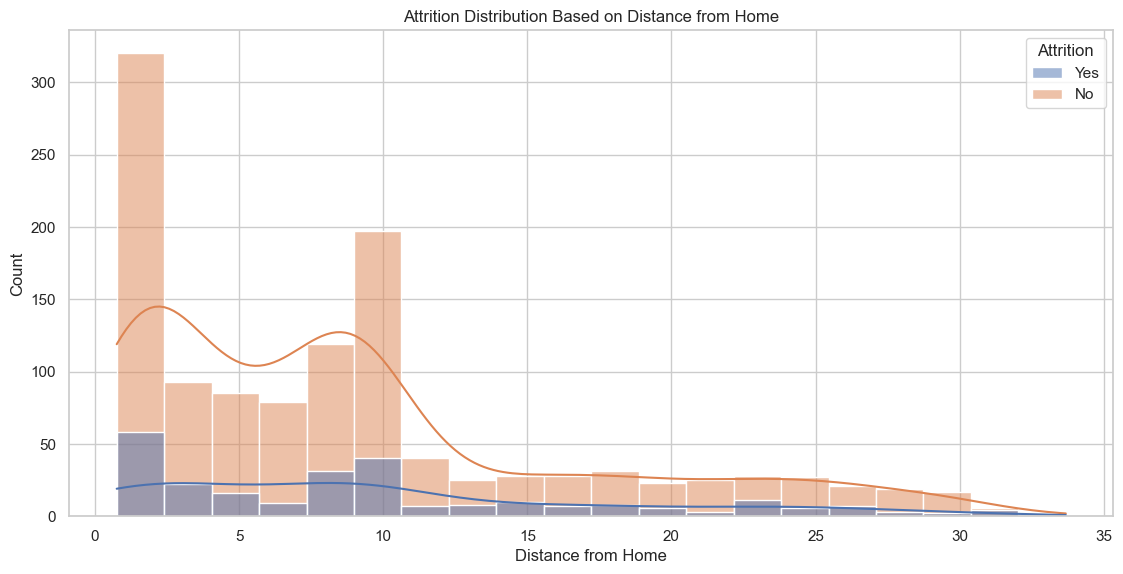


Figure Attrition based on the distance from home

## Outliers

There’re no outliers in our data, the data shows the majority employees are between 30 to 40 years old (fig. 11), Mean age is 36.

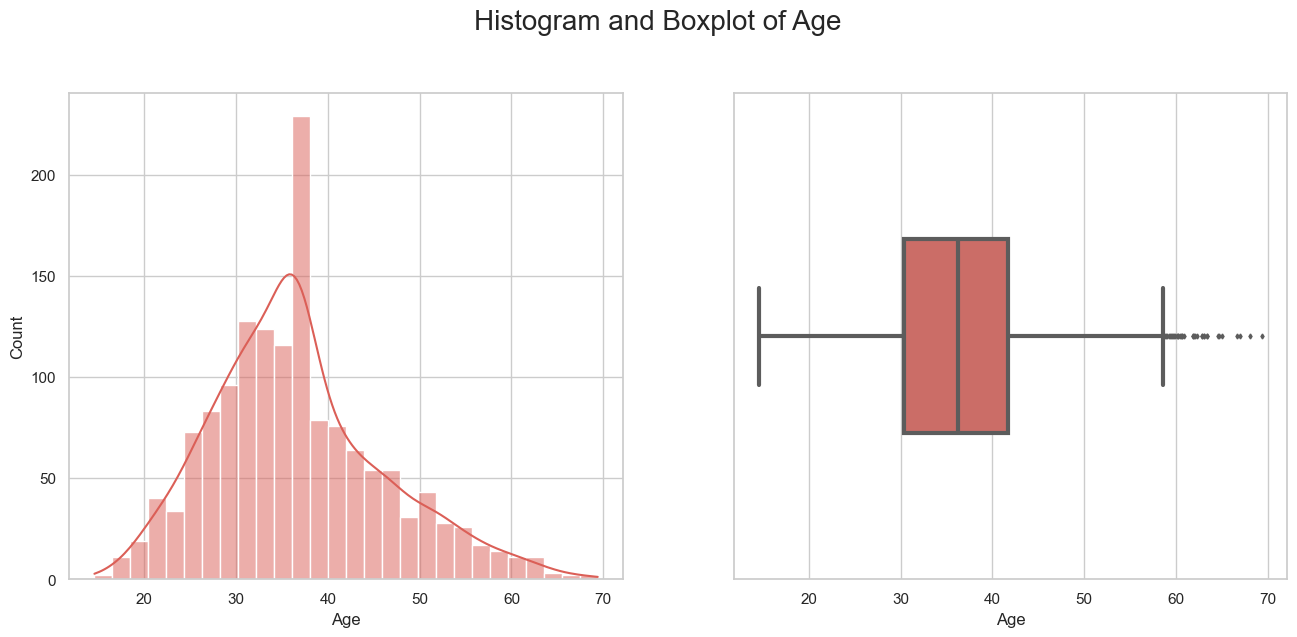


Figure Outliers based on the age of employees

From the below data visualisation (fig. 12), we can conclude that there is an outlier on the distance from home, but we don’t have to drop it. As per our findings based on the correlation results, it won’t have any major effect the data.

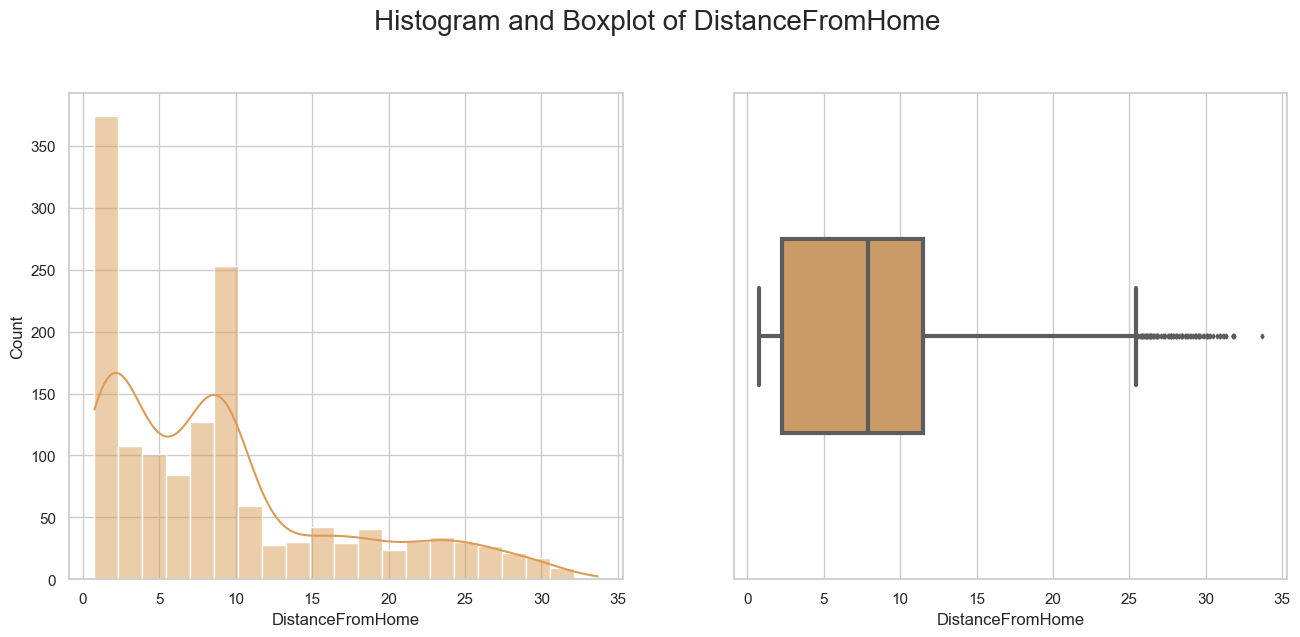


Figure Outliers based on the distance from home

From the Histogram (fig. 13), we can conclude that there are no outliers in the monthly income variable, and that the range of monthly income is following:

MonthlyIncome

Mean 6544.72

Median 5495.90

Mode 6544.72

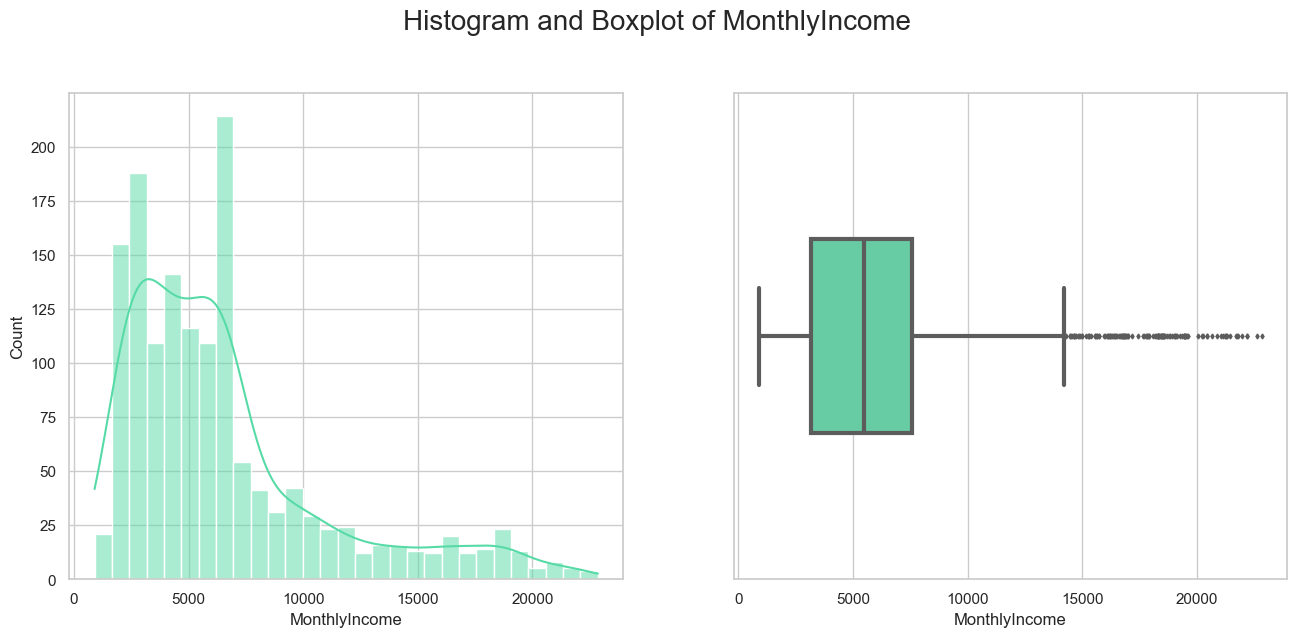


Figure Histogram based on a monthly income

# Feature Engineering

## 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 another words, we reduce the dimensions, for example variables, while retaining majority of the data within the dataset to ensure our results are not impacted or biased.

**Pre-Processing Technique**

The Standard Scaler is used as a pre-processing approach to standardize numerical features within a dataset before applying machine learning algorithms. Standardization entails changing numerical numbers to have a mean of 0 and a standard deviation of 1.

Results



As a result, the dataset becomes more streamlined (fig. 14), which may improve the performance of machine learning models, especially when dealing with high-dimensional data that contains noise or less useful features.

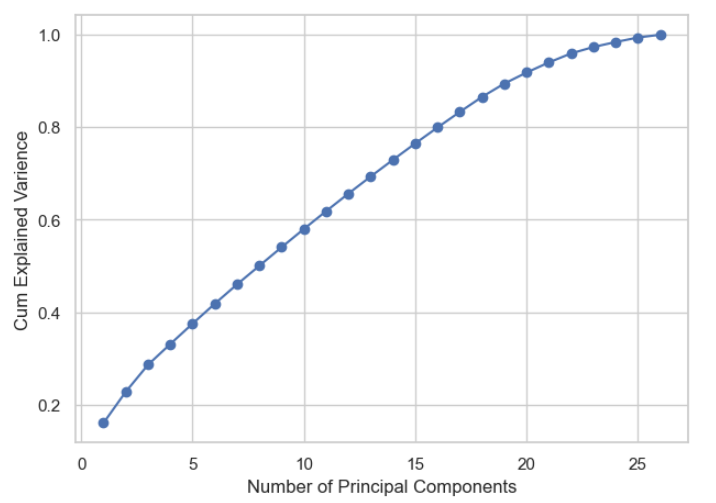
****

Figure New Streamlined Dataset

**PCA versus LDA:**

* Both PCA and LDA are used in case of dimensionality reduction
* PCA performs better with a smaller sample size
* 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 feature in higher dimensional space to a lower dimensional space. This is to avoid the curse of dimensionality.

To summarize, 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.

**Machine Learning Models**

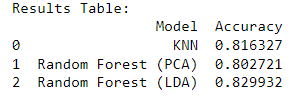
Here, we performed test using 3 machine learning models, the K-Nearest Neighbour model, Random Forest (PCA) and Random Forest (LDA).

K-Nearest Neighbour, commonly abbreviated as KNN, is supervised machine learning algorithm based on principle of the nearest neighbour number to an unknow variable that needs to be predicted. The new unknown variable is denoted by the symbol “K”. In principle, this method of classification is relatively simple, it works by locating the K nearest points in the training dataset. We find a value “K”, calculate the distance and count the points, check the labels and recount obtaining the probability.

We will also have a look at the Randon decision forest as a Machine Learning Method. This algorithm functions by constructing multitudes of decision trees during the training time. It is a popular machine learning algorithm that can be used for both classification (categorical target variable) and regression (numeric target variable) problems.

Let's briefly explain how Randon Forest Classification method works here. By using different and random subsets of the data, the multiple decision trees are created and predictions are made by calculating each prediction for each decision tree. The result is derived from the most commonly occurred and the more trees there are, it will be more robust, therefore delivering more accurate results.

Here, we used 3 different models, Random Forest PCA, Random Forest LDA, KNN and we can conclude that the best accuracy is derived from the Random Forest LDA with accuracy score of 0.83.



# 

# Statistical Techniques

## Measures of Central Tendency

A measure of central tendency is a summary that describe the entire data set with a single value that represents the middle of the distribution (fig. 15). In another words, it helps us to find the middle, or the average of the dataset.

There three main measures of central tendency are following:

* mode, which is the most frequently occurring value
* median, the middle number of the score for a dataset arranged in order of magnitude, less affected by skewed data and outliers
* mean, also called average, is the sum of all values divided by the total number of values.



Figure Measures of Central Tendency

Correlation Matrix

Correlation matrix shows us a correlation or relationships between variables, where each cell in the matrix table shows correlation between two variables. It is commonly used for data summary and as a base for further, more advance analysis.

From the matrix (fig 16), we can conclude that age is strongly correlated with job level, monthly income and total working years. There is also a relationship with years in the current role and years in the company.

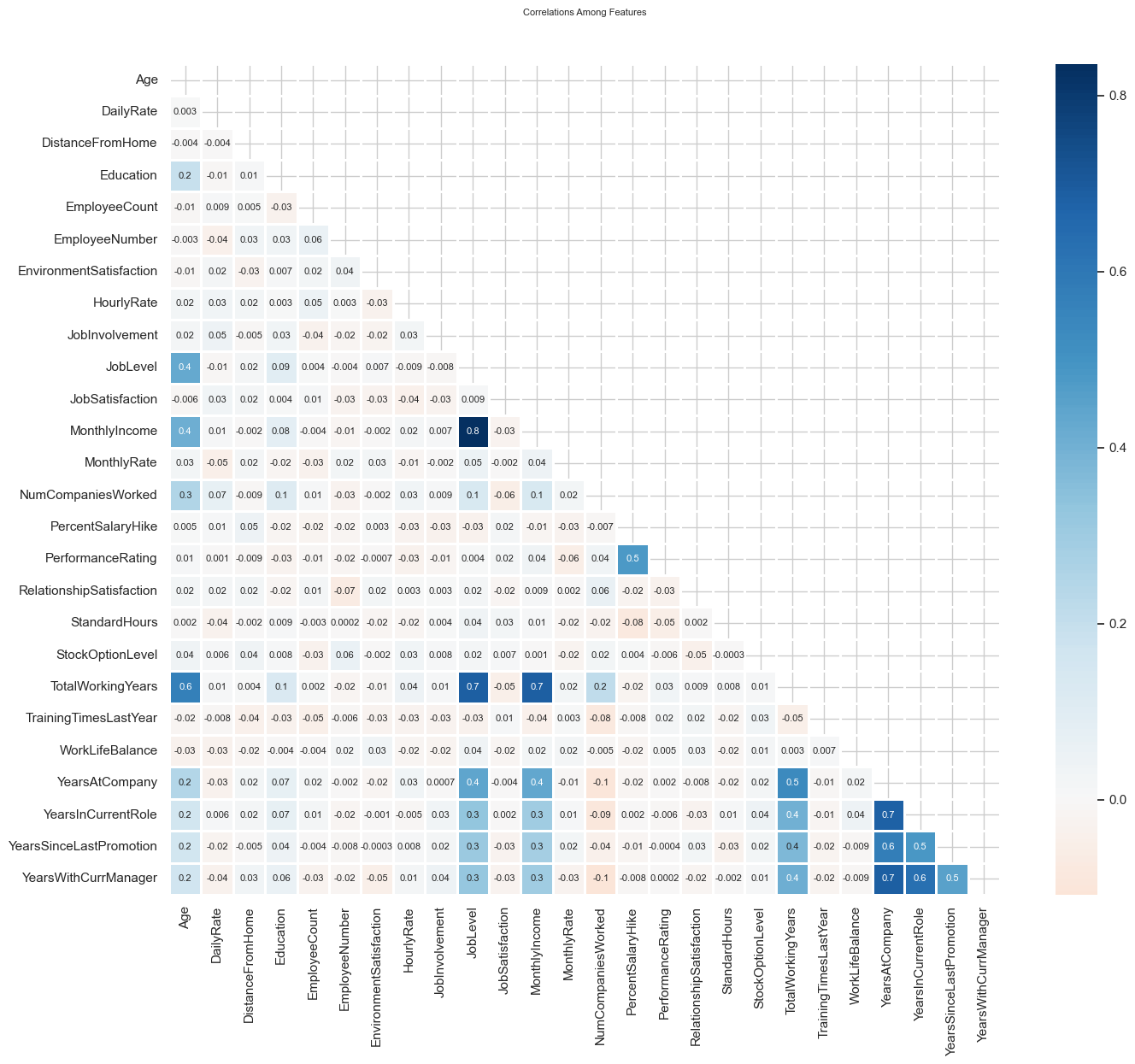


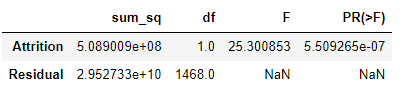
Figure Correlation Matrix

**ANOVA**

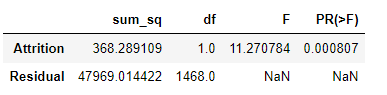
  Analysis of variance (ANOVA) is a statistical model that us used to analysed the differences amongst means. It is commonly used to test the difference between two or more means.

The higher the F-value in an ANOVA, the higher the variation between sample means relative to the variation within the samples. In another words, the higher the F-value, the lower the corresponding p-value.

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



Whereas for the Years in the company, the results suggest it does not affect attrition.



ANOVA table (fig. 17) were obtained from the good p-values, which are higher than 0.05, the significance level us typically set at 0.05 or 5%. This really means that results only have a 5% chance occurring or less.

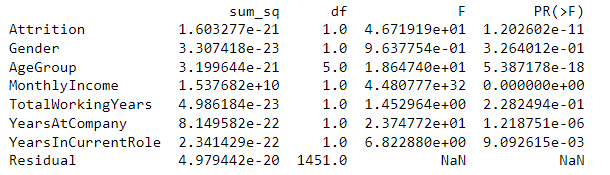


Figure ANOVA Table

# Data Modelling & Hyperparameter Testing

After training and evaluating models with various hyperparameter combinations, the optimal set of hyperparameters was discovered using performance measures. The model arrangement utilized is as follows:

**Selection of Hyperparameters**

The n estimators used:

'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**

Accuracy of 80.27% - the model accurately predicted the target variable ('Attrition') for about 80.27% of cases in the test data. In other words, the model made a high percentage of right predictions. Accuracy is a popular indicator, particularly in balanced datasets, since it provides a broad picture of how well the model performs across all classes.

## Machine Learning using Best Hyperparameters

Individual Model Evaluation:

* KNN (K-Nearest Neighbors): 78%
* Random Forest (PCA): 80%
* Random Forest (LDA): 80%
* Logistic Regression: 84%

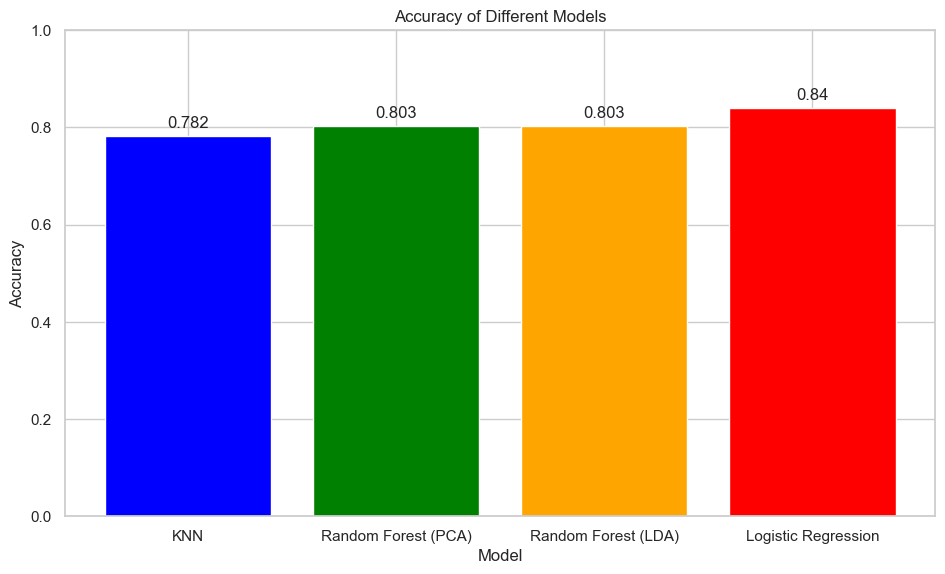
Cross-Validation Results:

* Method: K-Fold Cross-Validation with 10 folds.
  + Fold 1: 0.7891156462585034
  + Fold 2: 0.7891156462585034
  + Fold 3: 0.7619047619047619
  + Fold 4: 0.8503401360544217
  + Fold 5: 0.8095238095238095
  + Fold 6: 0.8231292517006803
  + Fold 7: 0.8163265306122449
  + Fold 8: 0.8095238095238095
  + Fold 9: 0.8367346938775511
  + Fold 10: 0.8367346938775511
  + Mean Accuracy: 0.8122448979591835

## Analysis of Results

**Model Performance Overview:**

* Logistic Regression has the highest accuracy amongst the individual models, at 84%, indicating that it performs well in predicting 'Attrition' based on the given data.
* Random Forest models (PCA and LDA) closely follow with 80% accuracy, suggesting strong predictive potential.
* KNN, while still good at 78%, has slightly poorer accuracy when compared to the other models.



**Cross-Validation Folds:**

* The K-Fold Cross-Validation is done 10,20 and 30 and our findings indicate very constant accuracies across the folds, showing that the models generalize well to diverse subsets of the dataset.
* The mean accuracy of 81.22% provides a comprehensive estimate of how well the models are likely to perform on unknown data

In conclusion, while Logistic Regression looks to be a solid choice, it is important to analyse the problem's specific needs, the model's interpretability, and the potential benefits of ensemble approaches such as Random Forest. The trade-off between precision, interpretability, and computing efficiency.

**Milestones**

Researching while working on this project helped me to gain better understanding of visualisation techniques, statistics and how to interpret the various graphs and charts, gradually gaining some confidence in this subject. I enjoyed learning and attempting to understand such a complex subject of study as Data Analytics or Machine learning are.

While challenging at first, I've additionally developed a better understanding of GitHub, including version control and how to create repositories. It was overall a fairly complex task, however, it gave me a good foundation for future learning.

## 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:**

[**https://github.com/Miroslava888/CA2\_ML**](https://github.com/Miroslava888/CA2_ML)

**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://www.kaggle.com/)

[**https://pypi.org/project/fasteda/**](https://pypi.org/project/fasteda/)

[**https://towardsdatascience.com/how-to-clean-your-data-in-python-8f178638b98d**](https://towardsdatascience.com/how-to-clean-your-data-in-python-8f178638b98d)

[**https://realpython.com/**](https://realpython.com/)

[**https://www.geeksforgeeks.org/**](https://www.geeksforgeeks.org/)

[**https://pandas.pydata.org/**](https://pandas.pydata.org/)