Project Report

Table of Contents

This Report aims to identify important factors that might be influential in determining which employee might leave the firm and who may not. This report also includes steps that are used to find the results

Part1

DataSet

The dataset includes IBM HR Attrition data which includes information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. The dataset contains different features on the basis of attrition and is analyzed for different reasons.

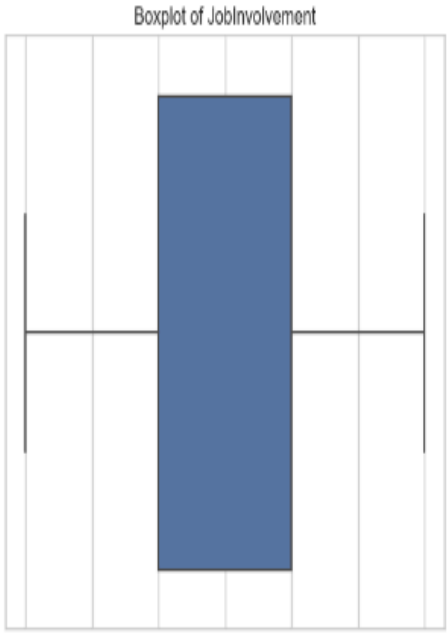
DataPreProcessing

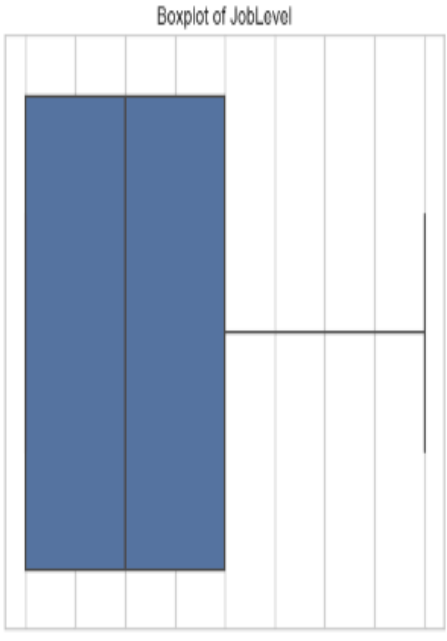
To improve models efficiency it is essential to perform data preprocessing on the data set which includes data cleaning, data transformation and data structuring.

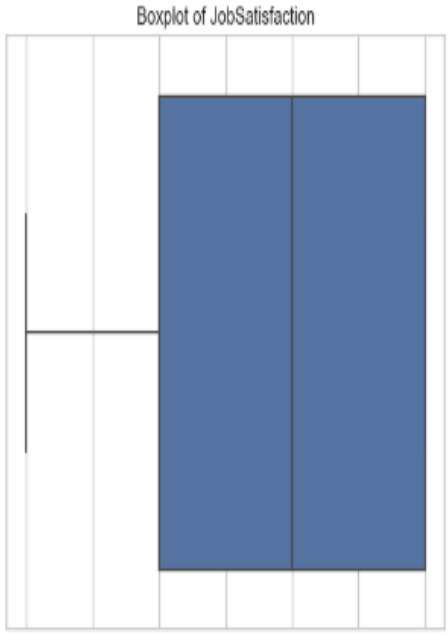
- Null Values: The dataset did not contain any null values and besides that appropriate data format is used to store values.

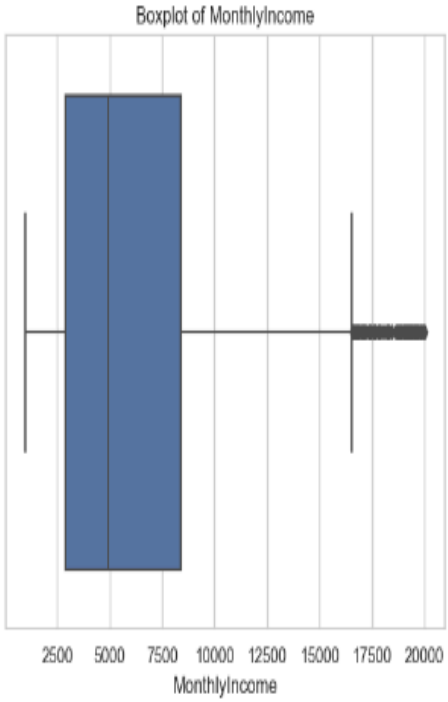
```
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
```

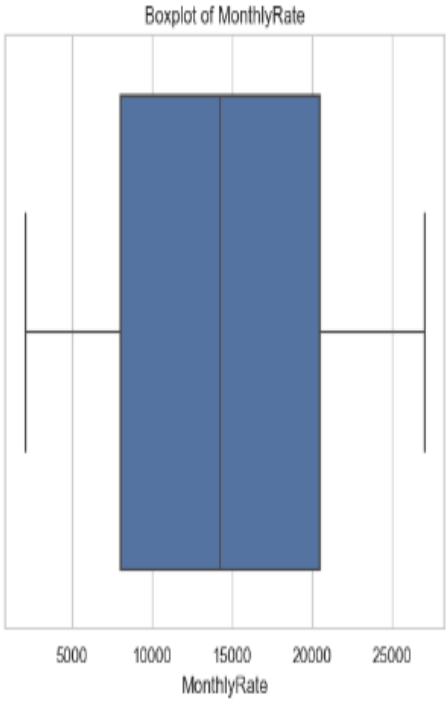
- Generated descriptive statistics to understand and check central tendency and distribution of the data. With the support of histograms and boxplots I detected outliers and removed them. The dataset contained very few outliers.

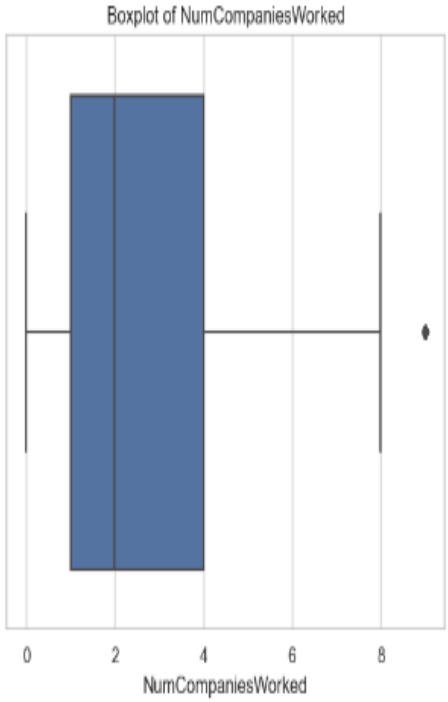

Boxplot of JobInvolvement
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobInvolvement

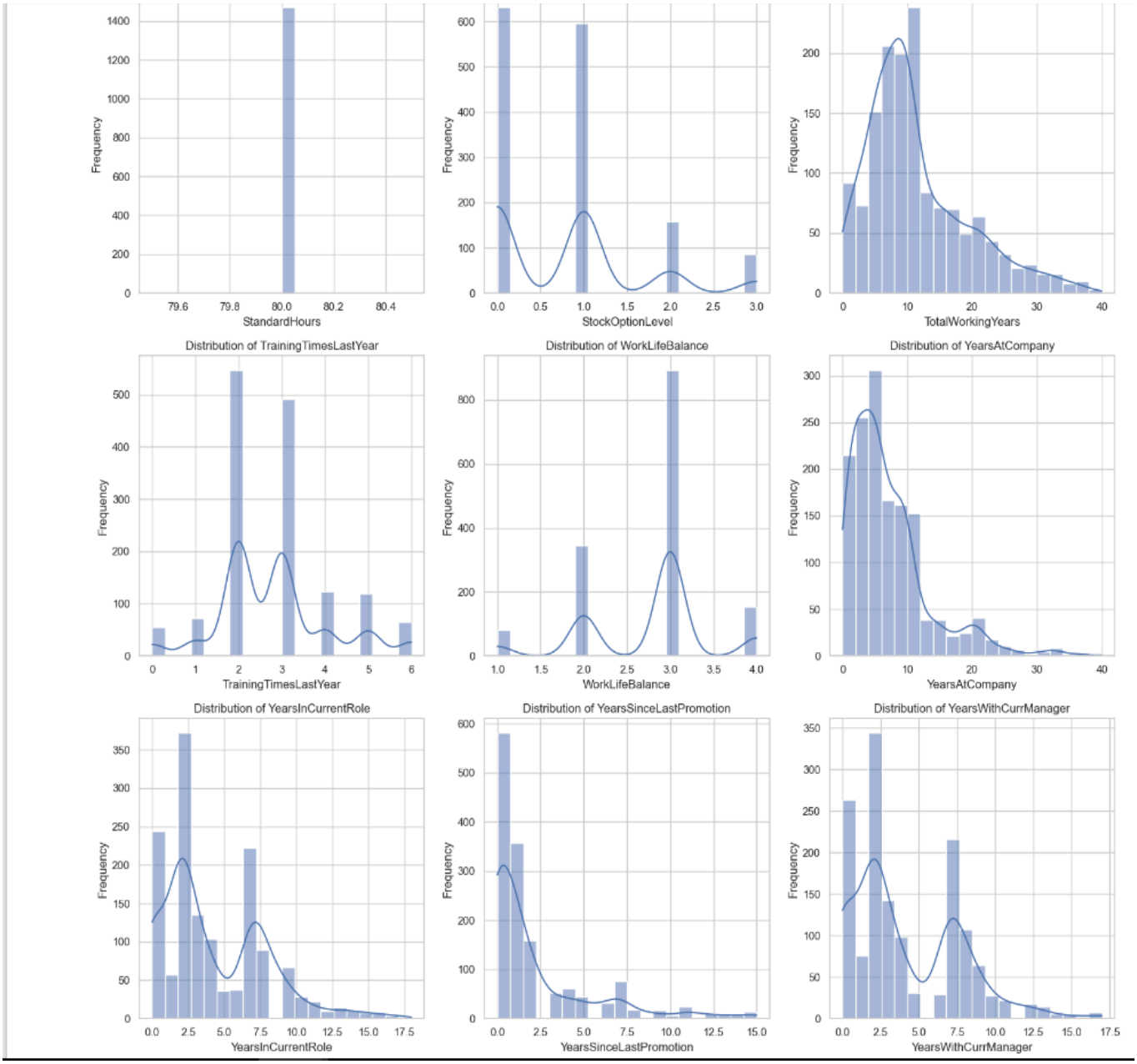

Boxplot of JobLevel
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
JobLevel


Boxplot of JobSatisfaction
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobSatisfaction

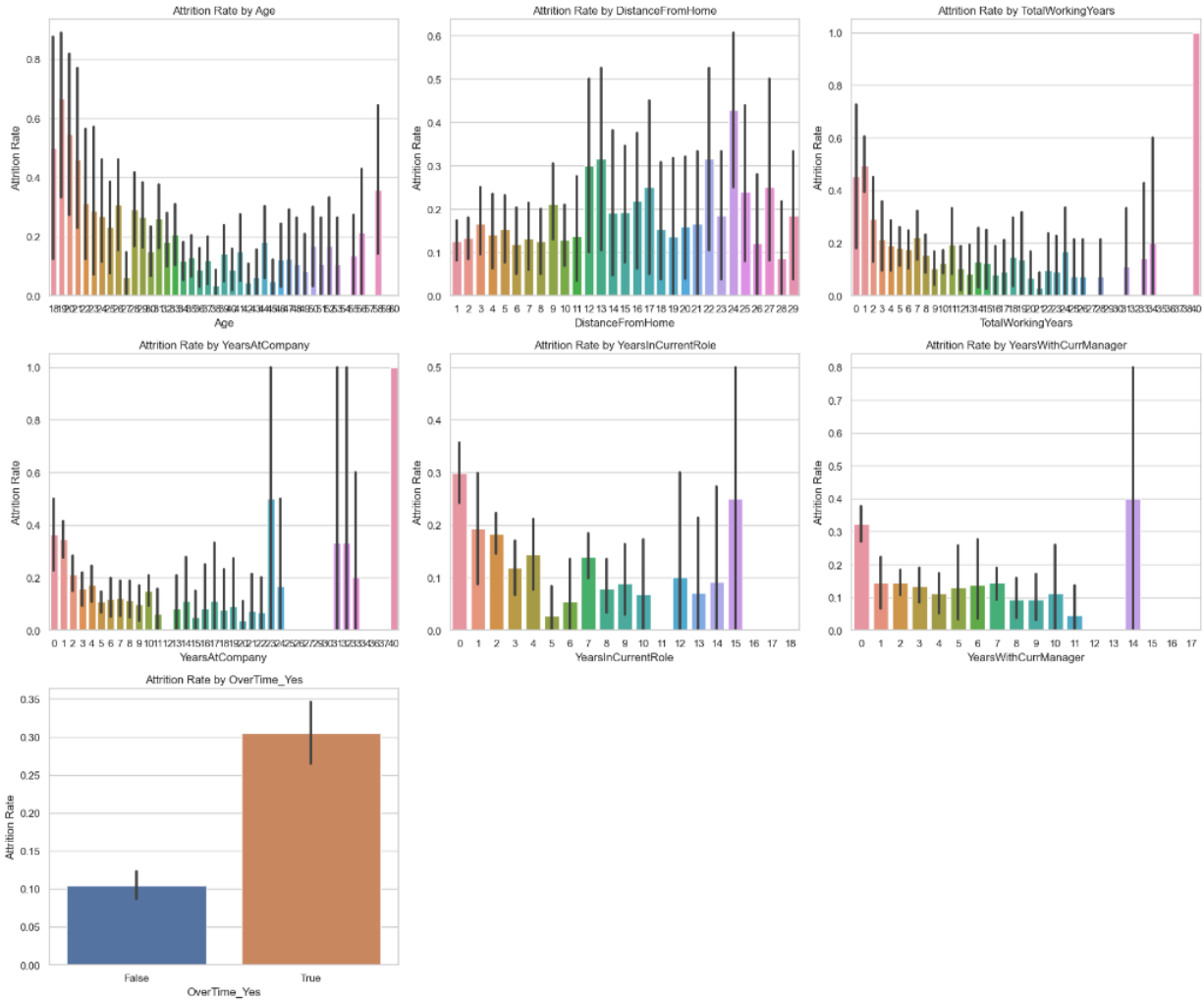

Boxplot of MonthlyIncome
2500
5000
7500
10000
12500
15000
17500
20000
MonthlyIncome


Boxplot of MonthlyRate
5000
10000
15000
20000
25000
MonthlyRate


Boxplot of NumCompaniesWorked
0
2
4
6
8
NumCompaniesWorked



- Check the balance of the target value as an unbalance targets can negatively reflect model results

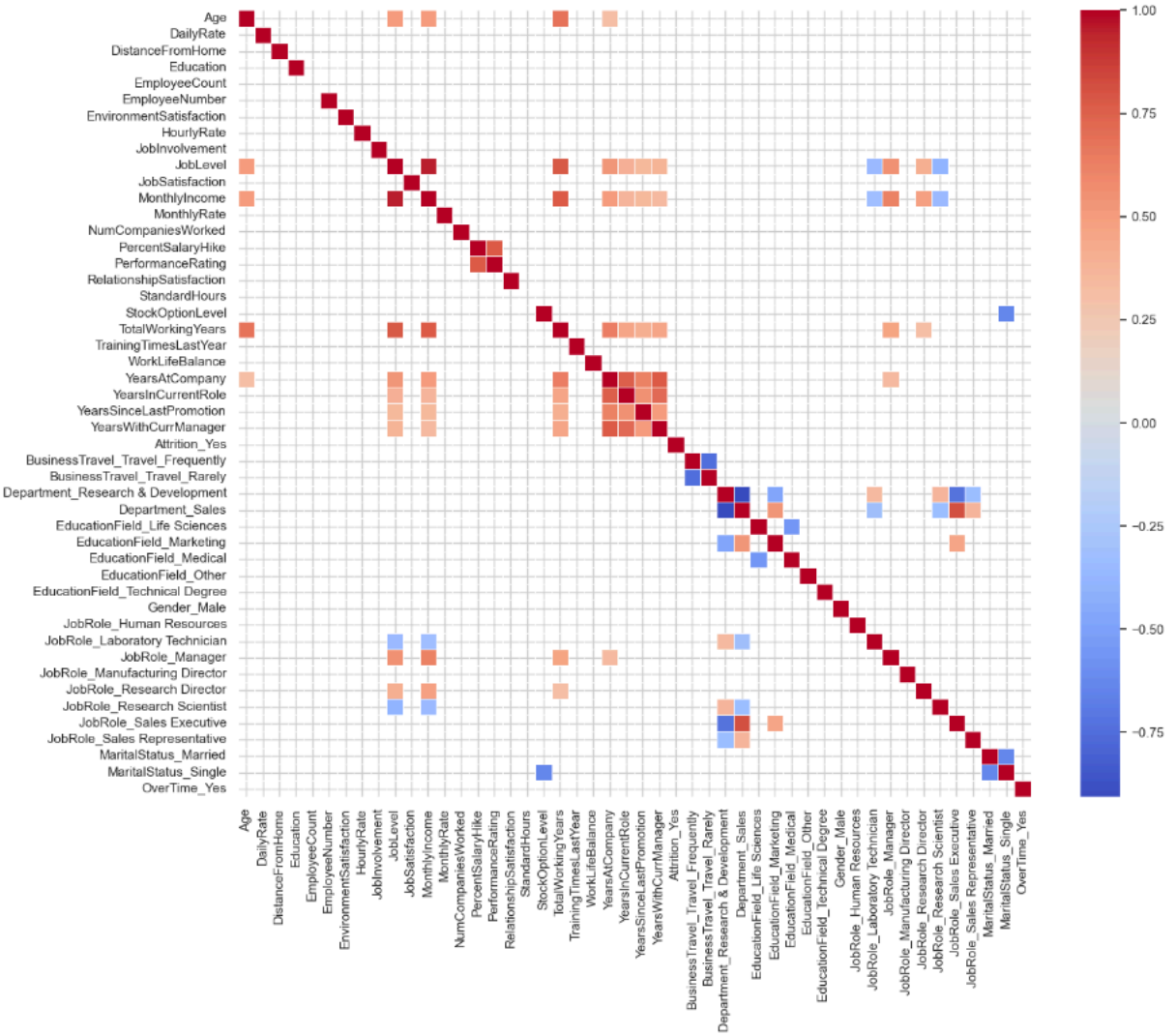


- Calculated correlation coefficient between each feature to extract important features

  Selected important features based on chi-square test.

  Important Features:

	Age	DailyRate	DistanceFromHome	MonthlyIncome	MonthlyRate	TotalWorkingYears	YearsAtCompany	YearsInCurrentRole	YearsWithCurrManager	OverTime_Yes
0	41	1102	1	5993	19479	8	6	4	5	True
1	49	279	8	5130	24907	10	10	7	7	False
2	37	1373	2	2090	2396	7	0	0	0	True
3	33	1392	3	2909	23159	8	8	7	0	True



Classification

Models

Models used for Classification:

1. Logistic regression
2. Decision Tree
3. Random Forest

1. Logistic Regression

Advantages:

- Simplicity and Interpretability: Logistic regression is easy to understand and interpret. The coefficients can be directly interpreted as the change in the log-odds of the outcome for a one-unit change in the predictor.
- Efficiency: It is computationally efficient and works well with smaller datasets.
- Probabilistic Output: It provides probabilities for class membership, which can be useful for decision-making.

Disadvantages:

- Linearity Assumption: Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable, which may not hold true in all cases.
- Sensitivity to Outliers: It can be sensitive to outliers, which can skew the results.
- Imbalanced Classes: As seen in the results, it struggles with imbalanced datasets, leading to poor performance in predicting the minority class (1).

2. Decision Tree

Advantages:

- Non-Linear Relationships: Decision trees can capture non-linear relationships between features and the target variable.
- Interpretability: The model is easy to visualize and interpret, making it accessible for stakeholders.
- No Need for Feature Scaling: Decision trees do not require normalization or standardization of features.

Disadvantages:

- Overfitting: Decision trees are prone to overfitting, especially with complex trees. This can lead to poor generalization on unseen data.
- Instability: Small changes in the data can lead to different tree structures, making them less robust.

- Bias towards Dominant Classes: Like logistic regression, decision trees can also struggle with imbalanced datasets, as seen in the results.

3. Random Forest

Advantages:

- Robustness: Random forests are less prone to overfitting compared to individual decision trees due to the averaging of multiple trees.
- Handling Imbalanced Data: They can handle imbalanced datasets better than logistic regression and decision trees, as they aggregate predictions from multiple trees.
- Feature Importance: Random forests provide insights into feature importance, which can be useful for understanding the model.

Disadvantages:

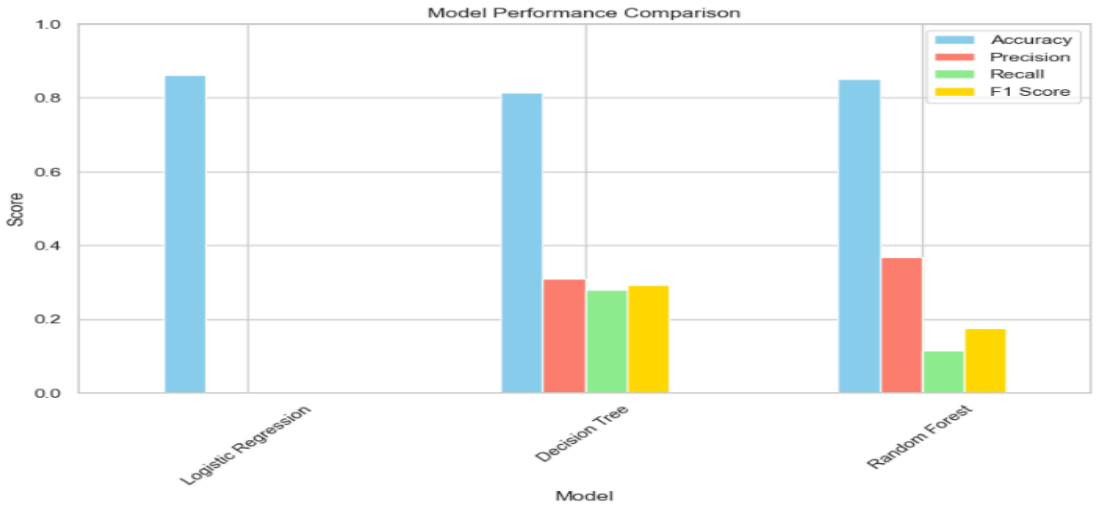
- Complexity: The model is more complex and less interpretable than logistic regression and decision trees.
- Computationally Intensive: Training multiple trees can be computationally expensive and time-consuming, especially with large datasets.
- Memory Usage: Random forests can consume a lot of memory due to the storage of multiple trees.

Analysis of Results and Performance Comparison

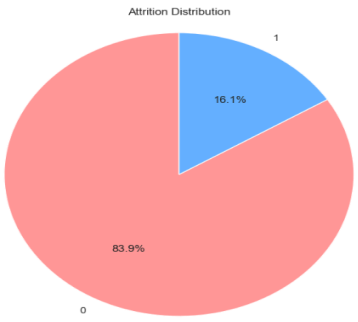
```
Logistic Regression Accuracy: 0.8616780045351474
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.93       380
           1       0.00      0.00      0.00        61

    accuracy                           0.86       441
   macro avg       0.43      0.50      0.46       441
weighted avg       0.74      0.86      0.80       441
```

```
Decision Tree Accuracy: 0.7959183673469388
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.88      0.88       380
           1       0.28      0.30      0.29        61

    accuracy                           0.80       441
   macro avg       0.58      0.59      0.58       441
weighted avg       0.80      0.80      0.80       441
```

```
Random Forest Accuracy: 0.8503401360544217
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.97      0.92       380
           1       0.37      0.11      0.17        61

    accuracy                           0.85       441
   macro avg       0.62      0.54      0.55       441
weighted avg       0.80      0.85      0.81       441
```



Clustering: The dataset contains data about employees who perform attriti

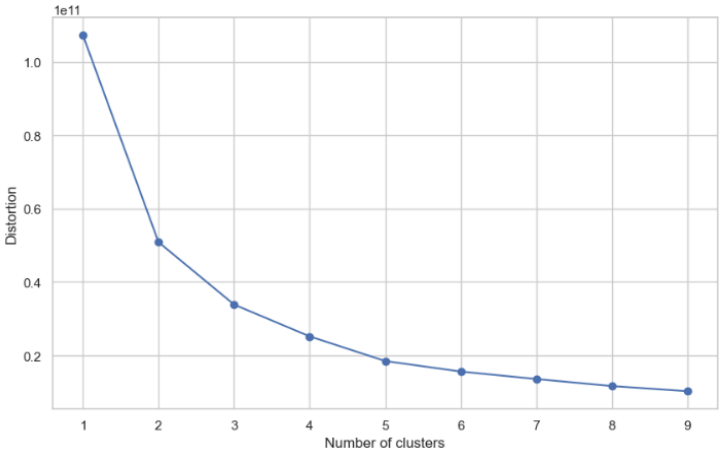


Different Clustering for each feature

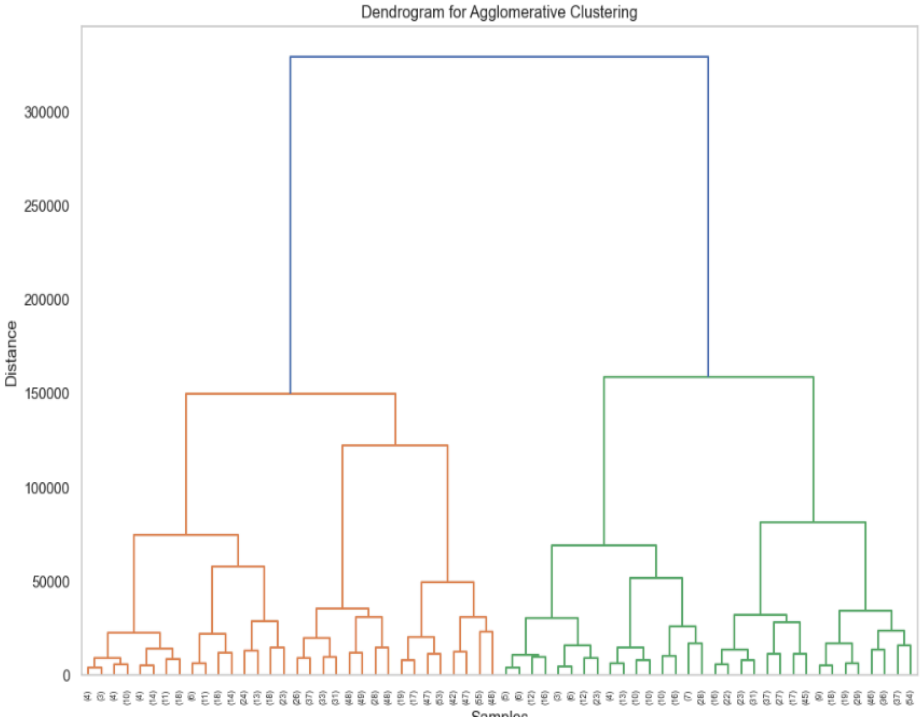
Choosing Appropriate cluster number

There were factors taken into account to find appropriate number of cluster

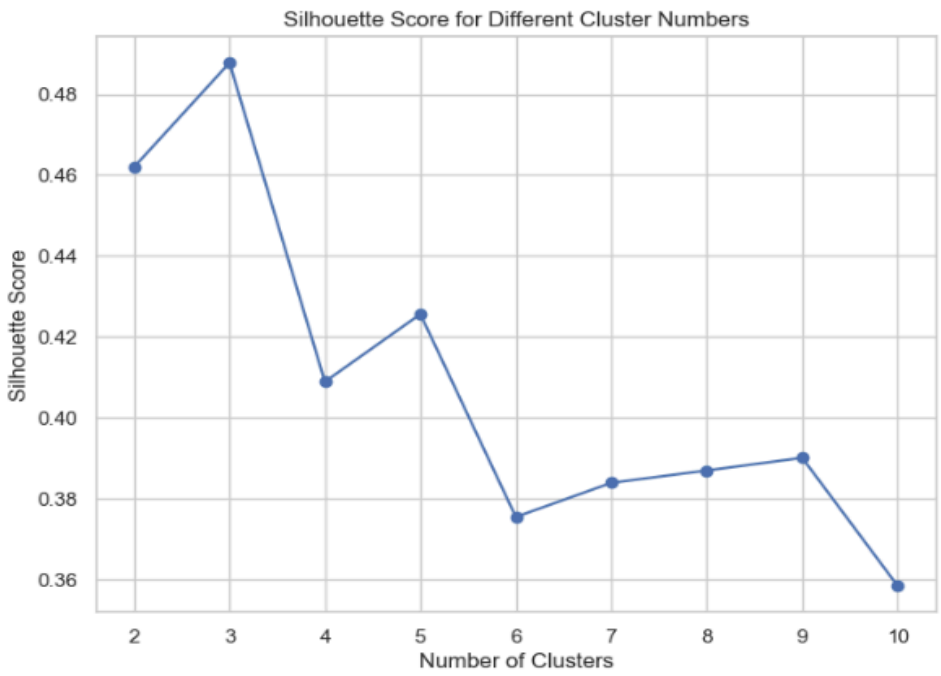
- Distortion Rate: As we can see distortion is decreasing with increase in the number of clusters.



- To establish the number of clusters I have used the Dendrogram.

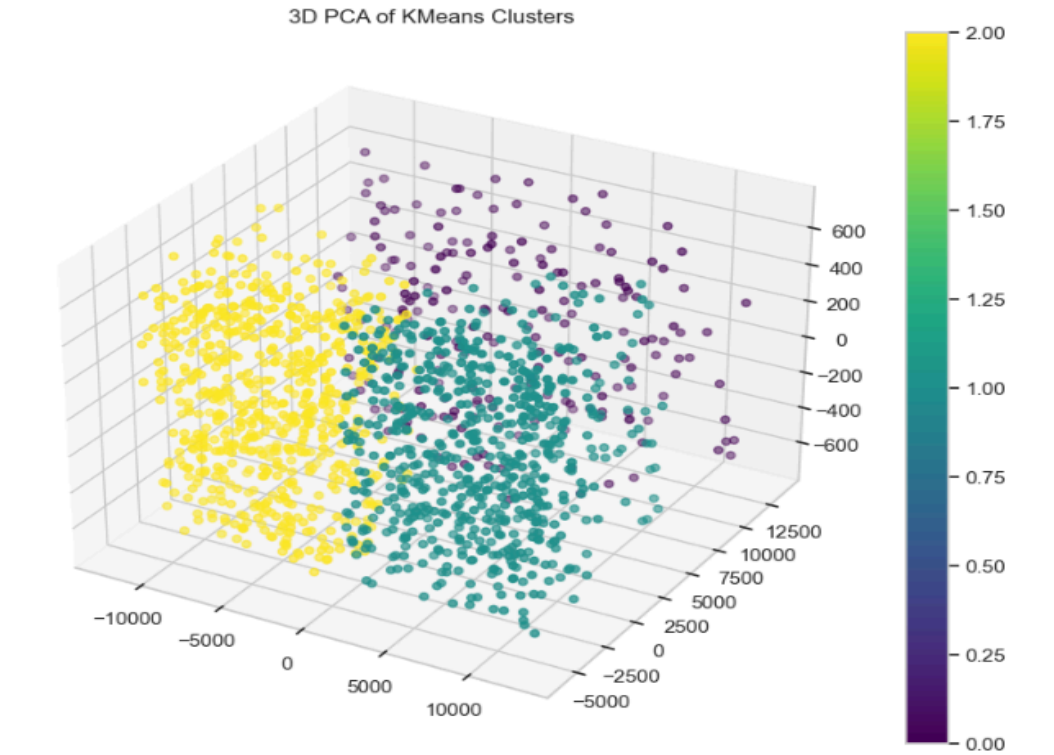


- Used Silhouette Average Score

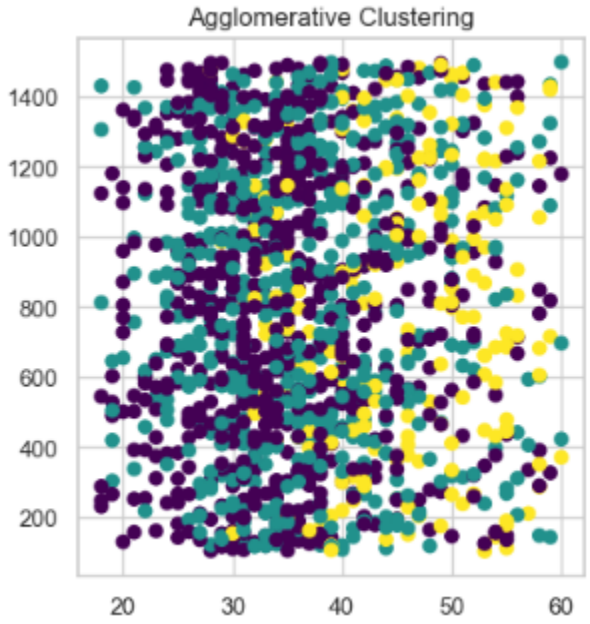


Models

1. K-Mean Clustering



2. Agglomerative Clustering



Clustering Results Summary

Clustering Method	Silhouette Score	Davies-Bouldin Score	Cluster Counts
KMeans	0.493	0.797	Cluster 0: 203 Cluster 1: 622 Cluster 2: 645
Agglomerative Clustering	0.411	0.911	Cluster 0: 683 Cluster 1: 586 Cluster 2: 201

Key Finding:

Category	Aspect	Details
Key Findings	Cluster Distribution	KMeans produced three balanced clusters, while Agglomerative Clustering had one small cluster (201 members).
	Cluster Characteristics	KMeans results suggest three distinct employee groups. Agglomerative Clustering indicates a unique group with specific attributes.
	Silhouette Scores	KMeans and Agglomerative Clustering have higher silhouette scores, indicating better-defined clusters.
Next Steps	Analyze Cluster Features	Examine mean values of features within each cluster to identify key characteristics associated with attrition.
	Create Visualizations	Develop visualizations to illustrate differences between clusters and their attributes.
	Develop- Attrition-Strategies	Based on findings, create strategies to address employee attrition effectively.

Part 2

Choosing the Best Model for Medical Insurance Data

- **Data Characteristics:**
    - The dataset contains both numerical and categorical variables.
    - There are likely non-linear relationships, particularly between features like BMI, smoking status, and charges.

    - Categorical variables (e.g., region, smoker) may interact in ways that linear regression cannot easily capture.
- **Model Selection Based on Data:**
    - **Linear Regression:** Suitable only if a strong linear relationship exists between predictors and charges (unlikely given the complexity of the dataset).
    - **Decision Tree:** Good for understanding feature splits but may overfit on small datasets.
    - **Random Forest:** Suitable for handling complex relationships and reducing overfitting but may require more computational resources.
    - **Gradient Boosting:** Best for capturing non-linear interactions and achieving high accuracy, especially with proper hyperparameter tuning.

In conclusion, for datasets with non-linear relationships and complex feature interactions, such as insurance data, **Gradient Boosting** is theoretically the best choice. However, if interpretability is crucial, **Decision Trees** or **Linear Regression** might be preferred despite potentially lower accuracy.

Model Performance:

Strengths:

- The R^2 score of 0.88 is quite impressive, suggesting that the model is able to capture the majority of the variance in the insurance charges. This means that the features selected (age, bmi, children, sex, smoker, and region) are highly predictive of the target variable.
- The MAE value of 2446.20 is reasonable given the scale of the target variable, suggesting the model is generally accurate in its predictions on average.

Areas for Improvement:

- The MSE value being relatively high suggests that the model might be overfitting or struggling with large deviations in some cases (possibly outliers). These large errors might need to be addressed by further data cleaning (e.g., removing or transforming outliers) or tuning the model.
- The presence of high MSE could also imply that the model might benefit from additional regularization or tuning of the Gradient Boosting parameters, such as adjusting the learning rate, max depth, or the number of estimators.

Overall, the Gradient Boosting Regressor performs quite well with an **R^2 score of 0.88**, indicating that it explains a large portion of the variability in the charges. However, the **MAE** and **MSE** values suggest that there are still opportunities to improve, especially in reducing the impact of large prediction errors or outliers. Adjusting hyperparameters, applying more data transformations, or using ensemble methods might help improve the model further.

Part3

Ways to use the rules:

1. **Personalized Recommendations**:
    - Netflix can leverage these association rules to recommend shows to users based on what they have already watched. For example, if a user has watched "Family Guy" and "Atypical," the service might recommend "Sex Education" as a next watch, since the rule shows a strong likelihood that viewers who watch these shows also enjoy "Sex Education."
2. **Cross-Promotion**:
    - By identifying which shows tend to be watched together (e.g., "Mr. Robot" and "Ozark"), Netflix can suggest a **watchlist or themed collection** of related shows. This can increase user engagement and time spent on the platform by introducing them to shows they might not have discovered otherwise.
3. **Content Acquisition Strategy**:
    - If certain combinations of genres or shows show a high degree of association (e.g., shows like "Ozark" and "Mr. Robot"), Netflix could use this information to guide its content acquisition strategy. It could look for or create new content that aligns with these patterns to better cater to audience preferences.
4. **Targeted Marketing**:
    - Netflix could use these rules to create more **targeted advertising** or promotional campaigns. For instance, if a user watches "The Blacklist" and "Mr. Robot," Netflix could push notifications or ads for other similar shows that share high association scores (like "Sex Education").
5. **Dynamic Content Placement**:
    - For users who have watched a particular set of shows, Netflix could dynamically adjust the **homepage layout** or **content grid** to prioritize recommendations that align with frequent associations, thereby increasing the

chances of a user discovering relevant content.

6. **Improving Content Recommendations Algorithms**:
    - The association rules can be fed into **collaborative filtering** models to enhance recommendations by supplementing user history with broader patterns in viewing behavior. This approach could help identify hidden preferences or interests that the user might not have explicitly expressed.
7. **Improved Search Functionality**:
    - Netflix could improve its search functionality by incorporating these rules. For example, if a user searches for a specific show, Netflix could suggest a list of other shows that are often watched with it, based on the antecedent-consequent rules.

The rules derived from association rule mining provide valuable insights into how users engage with content. By utilizing these insights, services like Netflix can enhance **personalization**, **cross-promotion**, and **targeted marketing**, improving the overall user experience and increasing engagement. The metrics such as **lift**, **confidence**, and **support** indicate strong relationships between shows, allowing Netflix to make recommendations that are more likely to appeal to its audience.