IBM MACHINE LEARNING

**Fatima, Sayeda**

**8/17/2022**

A picture containing automaton

Description automatically generated­­­-­­

IBM EXPLORATORY DATA ANALYSIS FOR EMPLOYEE ATTRITION MODEL

**Table of Contents**

[**1) Project Overview** 2](#_Toc111643034)

[**2) About the Dataset** 3](#_Toc111643035)

[**2a) Brief description of the data set you chose:** 3](#_Toc111643036)

[**2b) Summary of Data Attributes** 3](#_Toc111643037)

[**2c) Main Objectives of Analysis** 3](#_Toc111643038)

[**3) Data Exploration, Data Cleansing and Features Engineering** 4](#_Toc111643039)

[**3a) Data Exploration:** 4](#_Toc111643040)

[**3b) Data Cleansing Actions:** 6](#_Toc111643041)

[**3c) Features Engineering** 8](#_Toc111643042)

[**4) Summary of Training Different Classifier Models** 12](#_Toc111643043)

[**4a) Machine Learning Algorithm Approaches** 12](#_Toc111643044)

[**I) Data Level Approaches:** 15](#_Toc111643045)

[**II) Algorithm Ensemble Approach:** 15](#_Toc111643046)

[**4b) Summarizing Employed Models** 16](#_Toc111643047)

[**1) Logistic Regression (LR) Models:** 16](#_Toc111643048)

[**2) Random Forest Models:** 18](#_Toc111643049)

[**3) XGB Model** 24](#_Toc111643050)

[**5) Recommended Model** 25](#_Toc111643051)

[**5a) Result Summary** 25](#_Toc111643052)

[**5b) Overall Visual Summary** 25](#_Toc111643053)

[**5c) Individual Model Visual Summary** 26](#_Toc111643054)

[**5d) Model Choice and Justification** 26](#_Toc111643055)

[**6) Summary Key Findings and Insights** 27](#_Toc111643056)

[**6a) Summarizing Model Drivers:** 27](#_Toc111643057)

[**6b) Enlisting Top Contributory Factors** 27](#_Toc111643058)

[**6c) Visualizing Top Contributory Factors to Employee Attrition** 28](#_Toc111643059)

[**7) Link to Other Useful Models** 29](#_Toc111643060)

[**8) Github Link to Assignment Notebook** 29](#_Toc111643061)

# **1) Project Overview**

A fundamental issue facing organisations is attraction and retention of best talent. Given the cost of retraining new employees, it is important for a business to prevent loss of good talent. Hence, identification of key factors driving employee churning or turnover is important for the organization's Human Resource (HR) Department.

It is here that machine Learning models can be very useful to gain deeper insight into underlying factors and their relationship in driving employee turnover.

Hence, the main aim of the following machine learning modelling and analysis is to enable the business to:

\* To identify different factors predict employee churn

\* To gain insight into factors contributing to employee churning

\* To enable the business maximize employee attrition

# **2) About the Dataset**

## **2a) Brief description of chosen data set:**

This project uses a hypothetical dataset 'IBM HR Analytics Employee Attrition & Performance' which was downloaded from the following link:

<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?resource=download>

## **2b) Summary of Data Attributes**

The dataset exhibits 1,470 data points (rows) and 35 features (columns) reflecting on employees' background and characteristics and can be downloaded from the following link:

The data also comes with ‘Attrition’ Column to show current employees and leavers which represents the Class we are trying to predict.

## **2c) Main Objectives of Analysis**

Organizational performance is largely dependent on its employees, their quality and experience. Hence, organizations are continuously faced with the challenge to reduce employee attrition and increase retention. Consequently, this analysis is targeted towards answering the following queries

* What are the various factors contributory to employee attrition?
* Which business units face higher employee attrition rate?

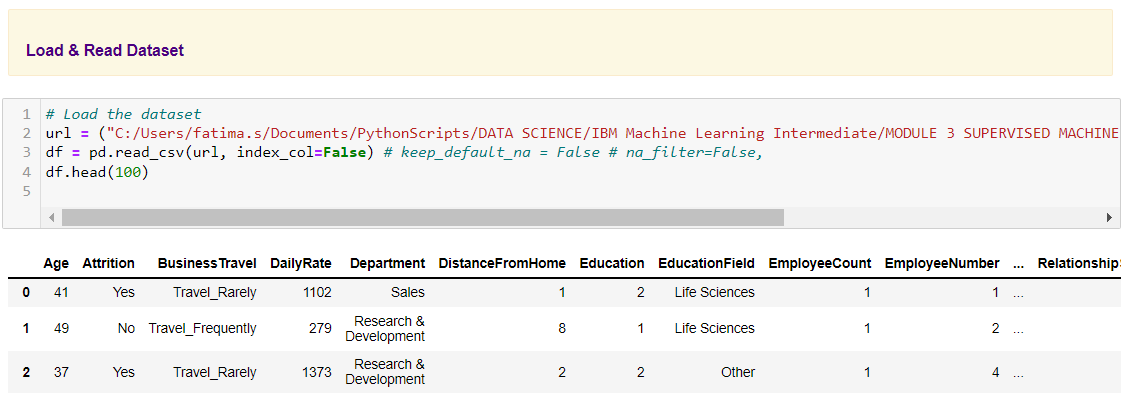
As a consequence, implementation of the model will enable the organization to:

* devise suitable measures to increase employee retention
* to save valuable resources in retraining new employees hired in place of leavers

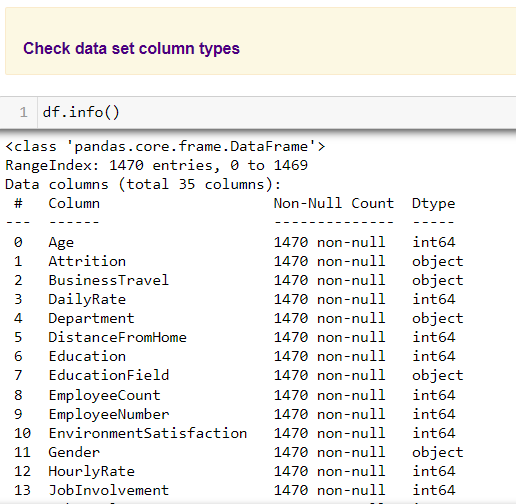
# **3) Initial Plan for Data Exploration**

## **3a) Data Exploration**

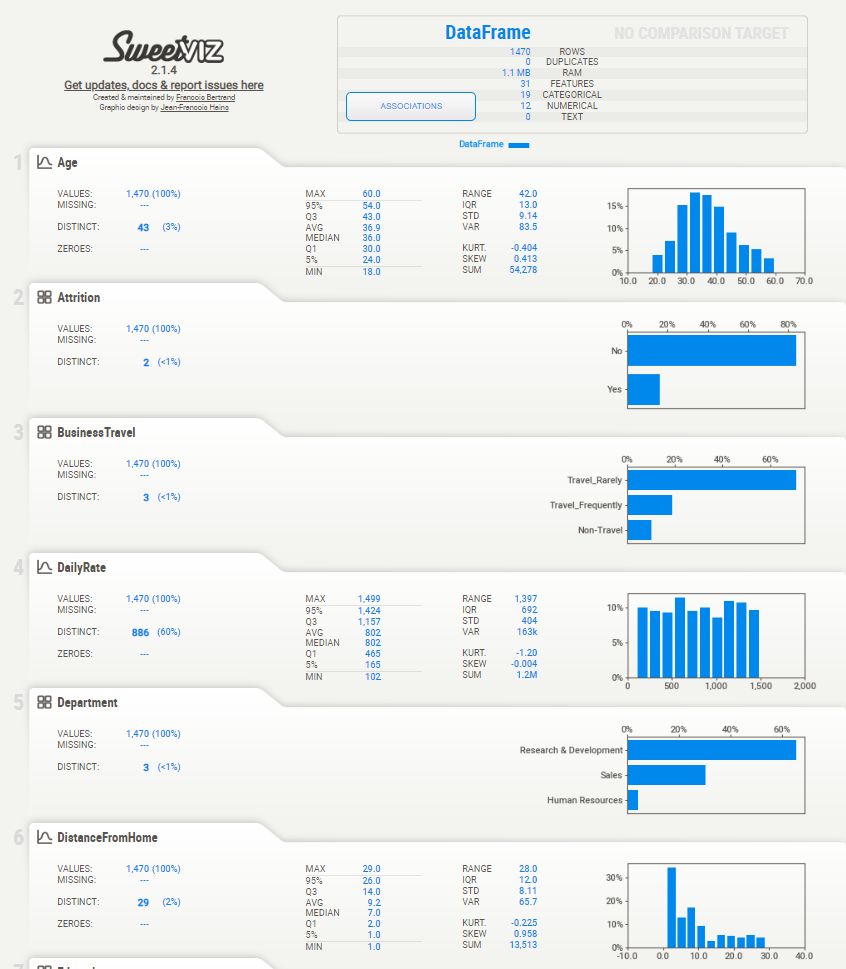
* Data was first loaded into pandas dataframe



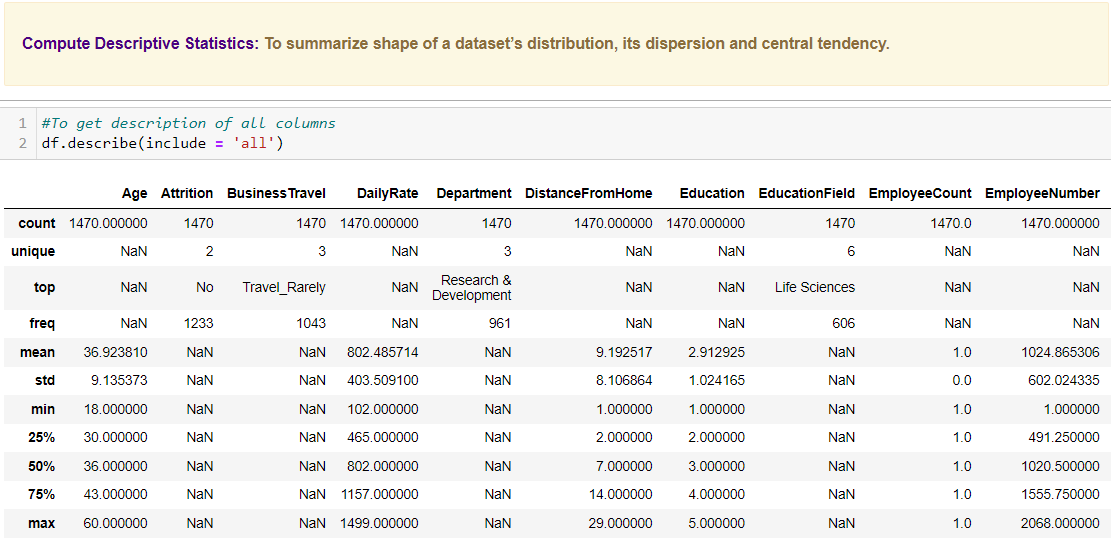
* Column types were explored



* Automated Exploratory Data Analysis was performed using Sweetviz to check



* Descriptive statistics were computed to summarize shape of a dataset’s distribution, its dispersion and central tendency

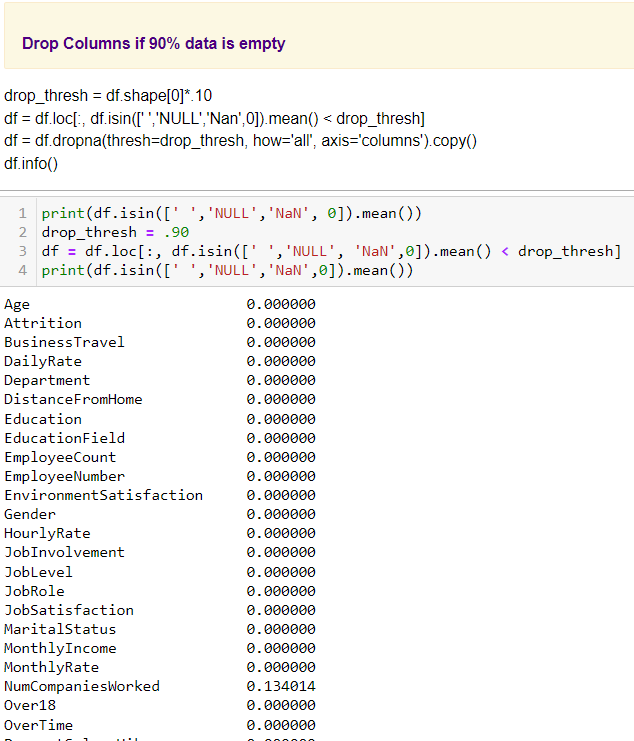


# **4) Actions taken for Data Cleansing and Features Engineering**

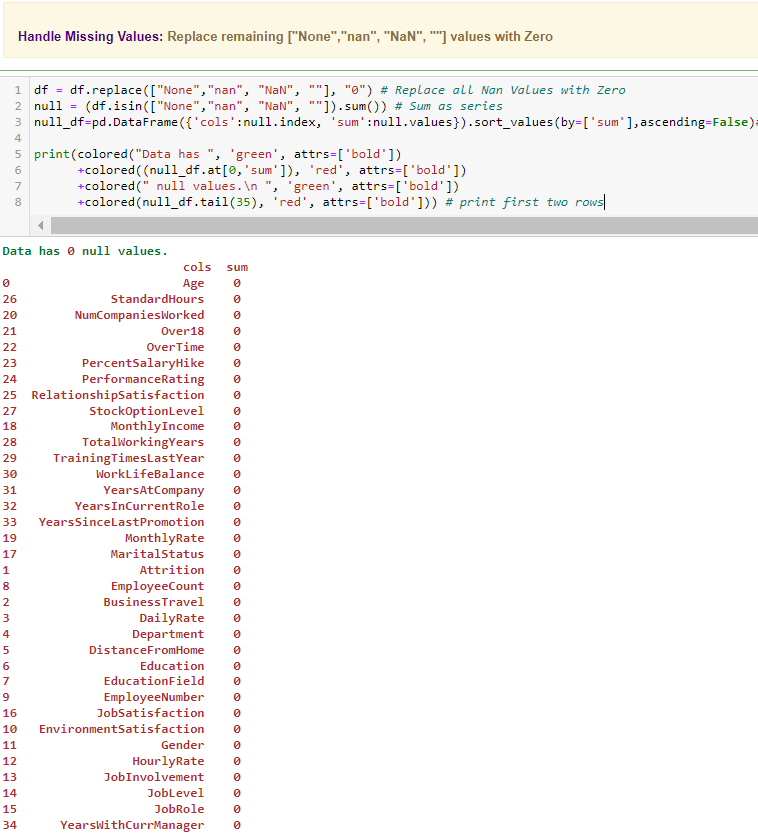
Since the quality of any machine learning model highly depends on quality of data, hence, this stage is not only most important but is also time consuming. Hence, it was conducted in a step-by-step process.

## **4a) Data Cleansing Actions**

* Empty or nearly empty columns were removed using "drop\_thresh" to drop columns if 90% of data was empty



* Duplicates were dropped using pandas "df.drop\_duplicates()" method



* Null values were summed and Data was found to exhibit zero null values. Thus, no filling of null values was required

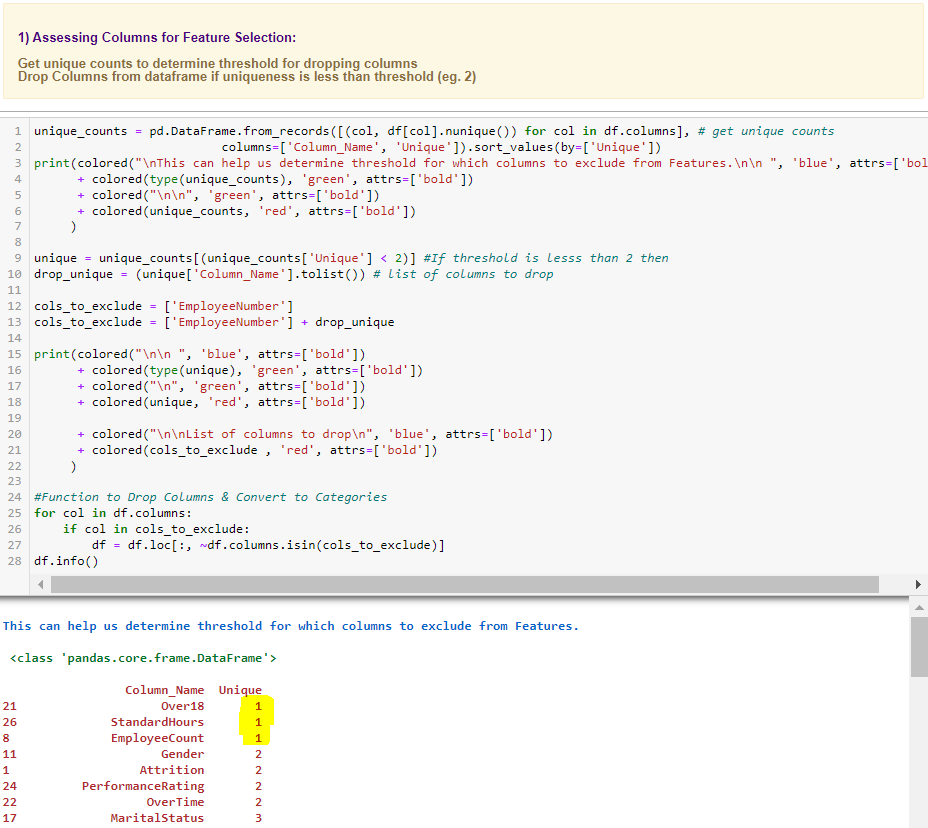
## **4b) Features Engineering**

In machine learning, feature selection is the method to reduce the number of input variables during developing predictive modelling. This reduction in input variables is necessary not only to minimize computational cost of modeling but also to achieve performance improvement of the model.

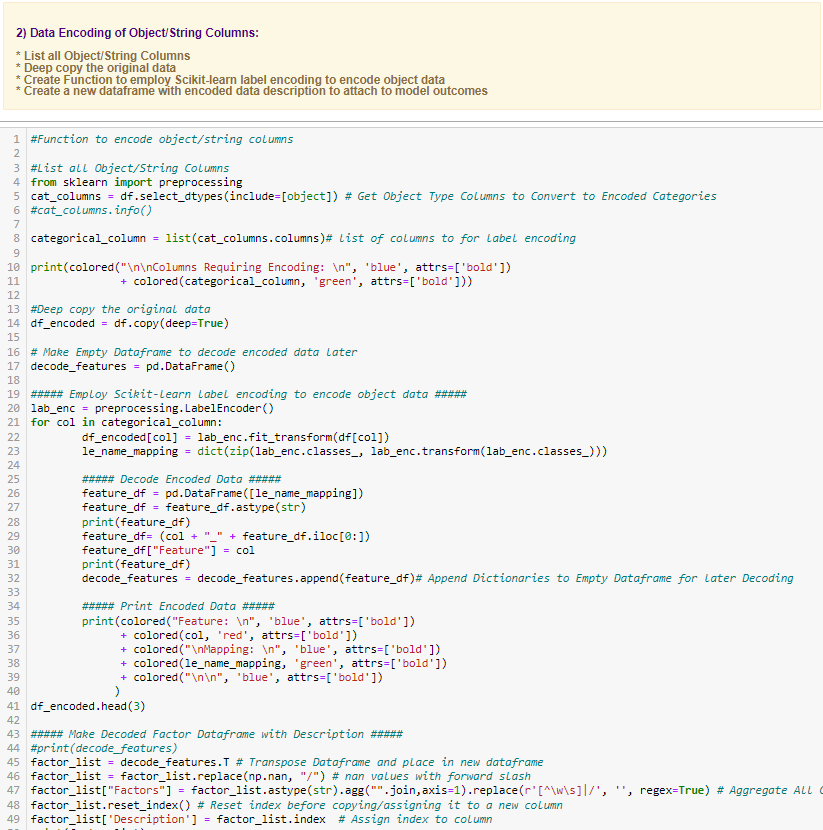
Among widely practices feature selection approaches include statistical-based feature selection methods which use statistical measures to evaluate relationship between each input variable and the target variable and then select those exhibiting strongest relationship with the latter. While these methods can be both speedy and effective, however, the ultimate choice of statistical measure is largely dependant on data types of both of these variables.

Irrespective of the statistical measure being employed, two dominant feature selection techniques, that is supervised and unsupervised, exist where the former can be further categorized into wrapper, filter and intrinsic techniques. Filter-based feature selection methods employs statistical measures to evaluate correlation between input and output variables so that those exhibiting highest correlations are selected. Statistical measures employed in filter-based feature selection are normally univariate in nature since they evaluate relationship of single input variables one by one with target variable, disregarding their interaction with each other.

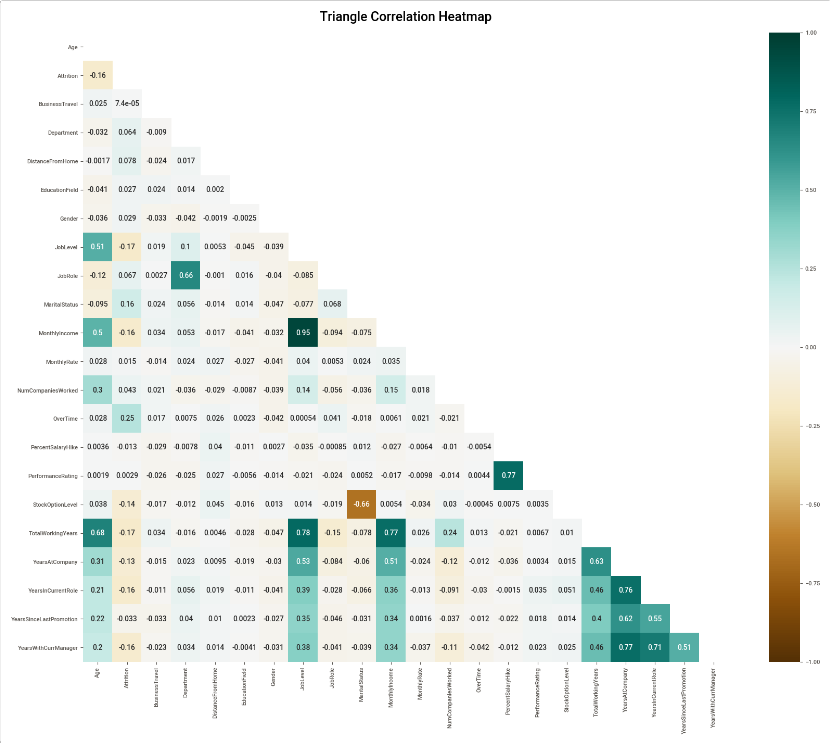
Consequently, adopting filter-based feature selection methods, the employee attrition model approached filter engineering in three steps. Firstly, unique values for all columns were computed after which columns with unique values less than 2 were dropped.



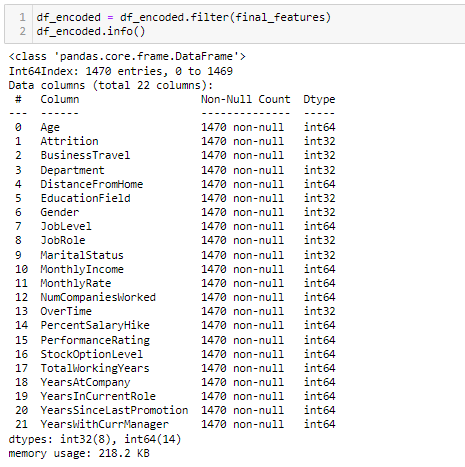
Prior to final features selection Data Encoding of Object or String Columns was carried out to facilitate any statistical computation during features selection process. Hence, after deep copying of original dataset, a function was created and employed to encode object data using Scikit-learn label encoder.



Statistical measures were then employed with supervised filter-based feature selection technique. Using Pearson's Correlation, the first set of features are selected based on the strength of positive correlation with taget variable 'Attrition'. Additionaly, Pearson's Correlation Matrix was also computed to select feature pairs exhibiting positive correlations with each other.



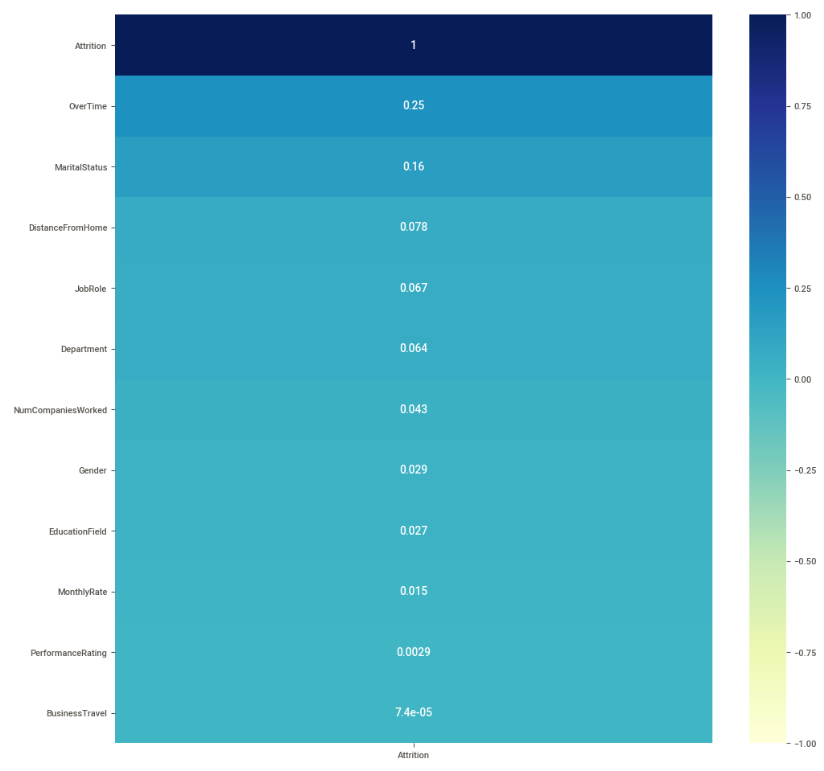
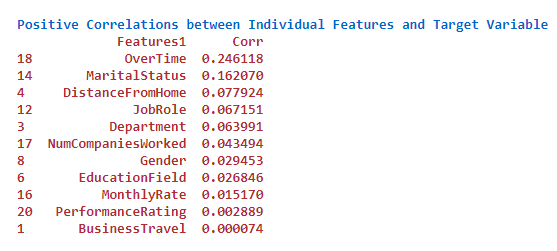
All feature lists were then combined to filter out dataframe columns not included in the ‘final\_features’ list.



# **5) Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner**

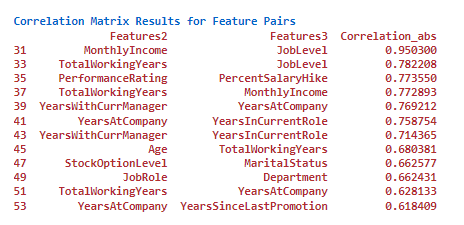
## **5a) Summary of Features Exhibiting Positive Correlation with Target**

The following features exhibited positive correlations with target variable:

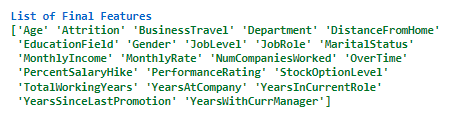


## **5b) Summary of Positive Correlations for Feature Pairs**

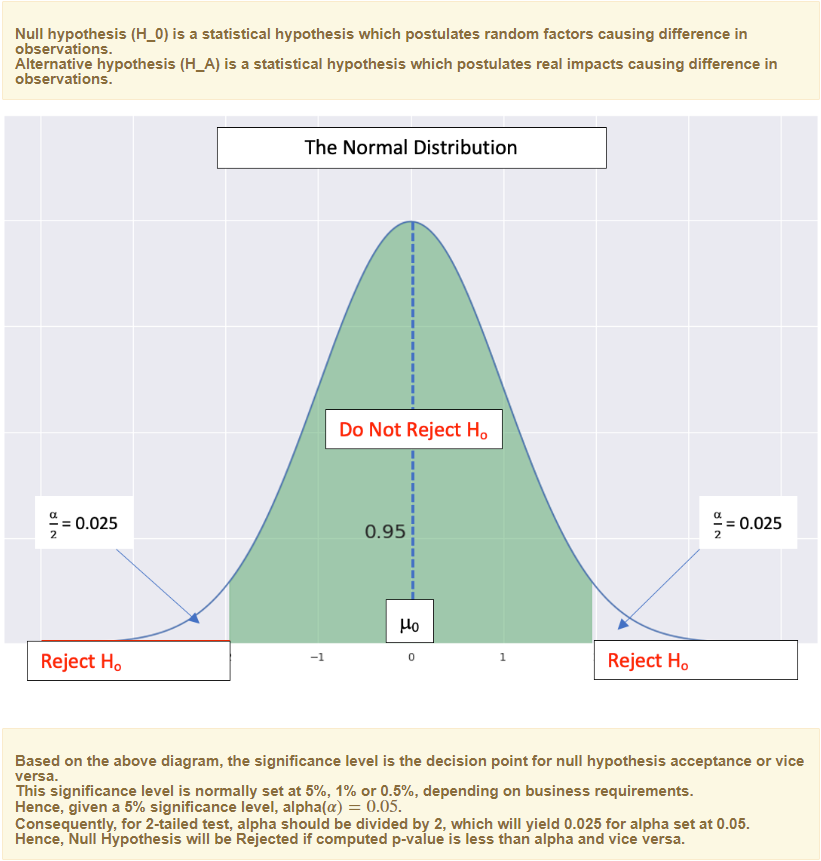
In addition, the following feature pairs displayed high correlation



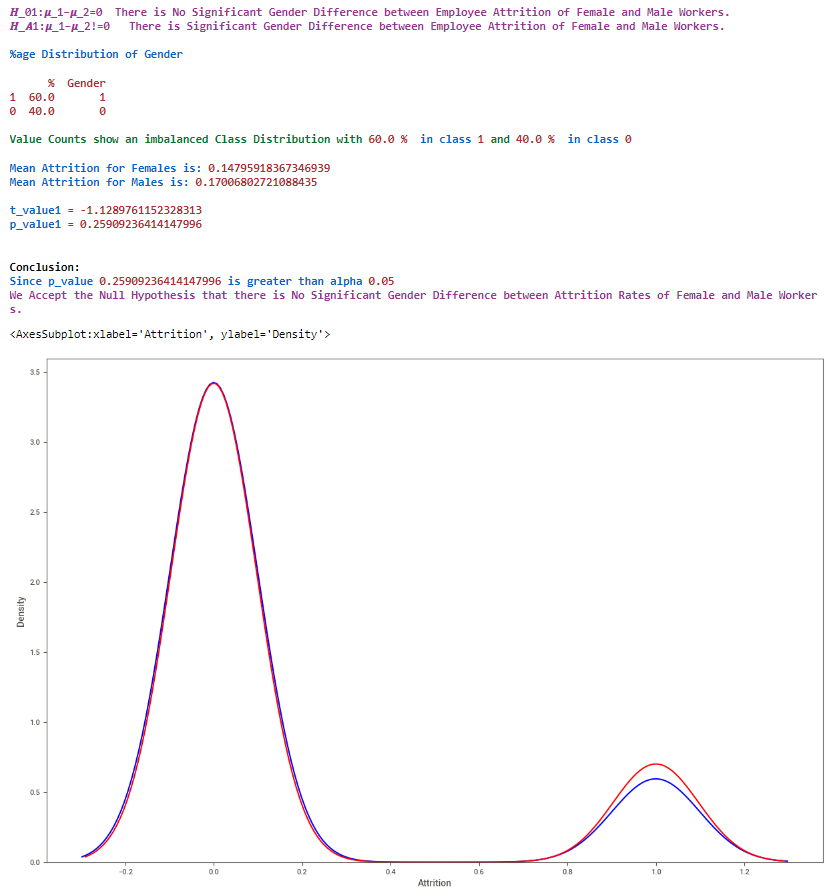
Based on the above two correlation findings, the following final features were selected while others were dropped as they were considered to bear no or less impacts towards employee attrition.



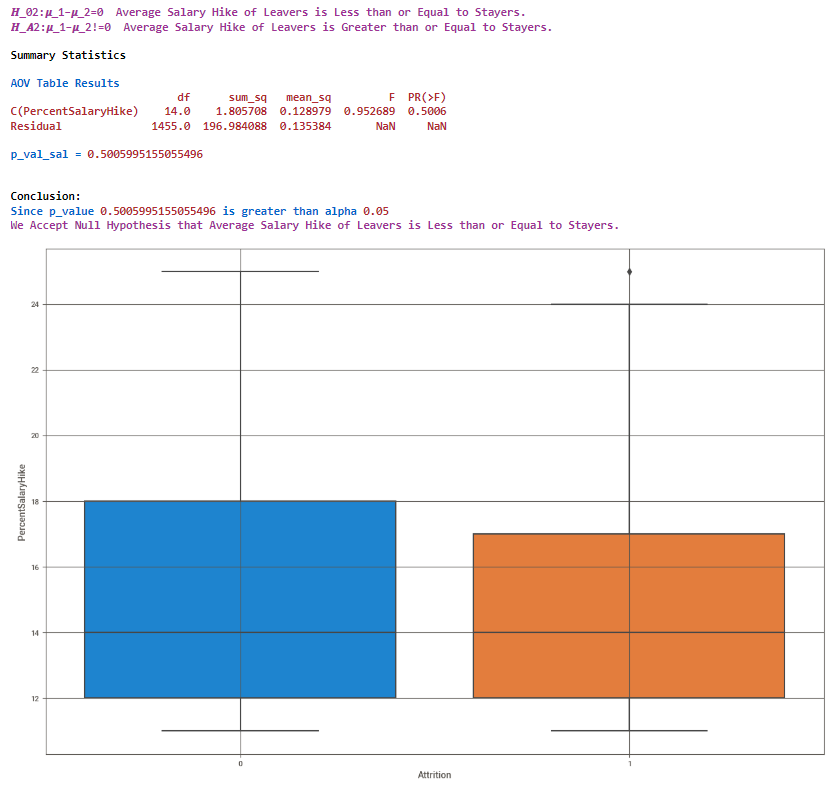
# **6) Three Major Hypothesis**



## **6a) Hypothesizing Gender Differences in Attrition Rates**



## **6b) Hypothesizing Differences in Salary Hike between Leavers and Stayers**



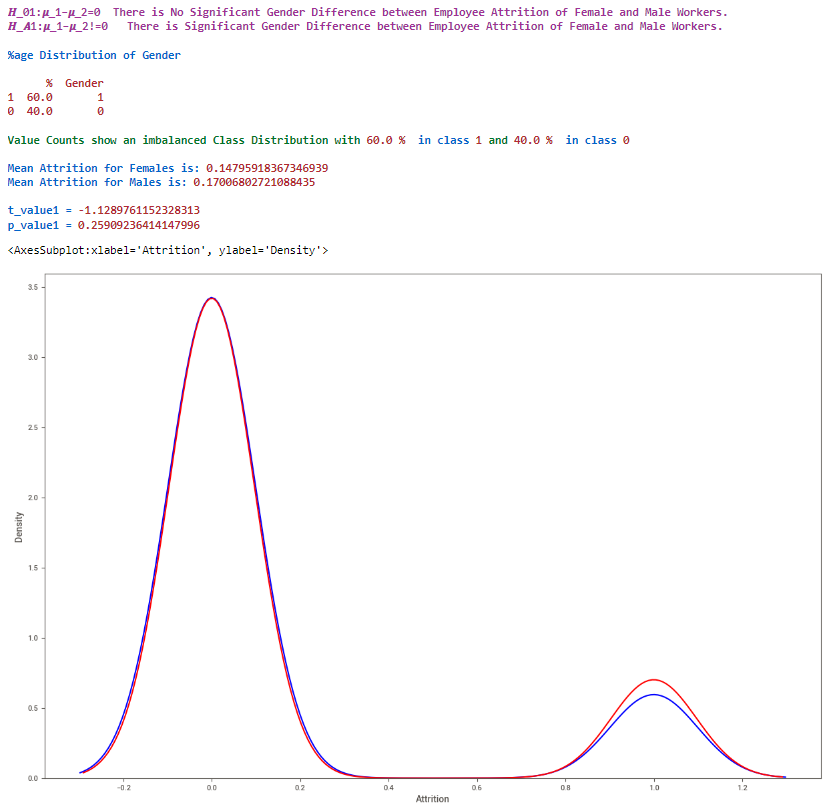
## **6c) Hypothesizing Differences in Salary Hike between Leavers and Stayers**



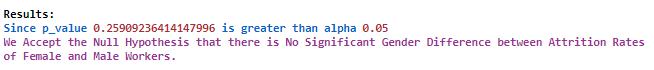
# **7) Conducting a formal significance test for one of the hypotheses and discuss the results**

## **7a) Significance Test for Hypothesis 1**





## **7b) Results and Discussion**



# **8) Suggestions for next steps in analysing this data**

* EDA now should proceed towards analysing problematic data outliners for all final features to increase statistical significance of the model.
* Using Gower Distancing, further cluster analysis can be done to gain in-depth understanding of employee clusters at risk of turnover rather than utilizing one size fit all approach (see, Section 10(e).

# **9) A paragraph that summarizes the quality of this data set and request for additional data if needed**

Overall data quality was quite good since EDA exhibited zero null values and contained extensive variables to work on. However, while the data did contain extensive parameters, nevertheless, including other pull factors like community fit, workload, etc. may prove to be useful. For example, based on social relationship theory, pull factors like community fit, industry, etc has been found to exhibit a negative correlation with intentions to leave, especially with increased perceptions of needs fulfilment and social networking (see, Ramesh and Gelfand, 2010). Hence, including this information can shed further light on employee turnover. Then, including other mooring factors, such as, personal life involvement may also improve data quality as well as model’s predictive power since these may also serve as underlying factors in employee turnover. For instance, numerous studies found higher turnover rates among employees exhibiting high family centrality and work interference with family life (see, Bagger et al., 2008; Haldorai, et al., 2019). Hence, provision of these additional parameters may render enhanced insight into employee attrition and may even change predictive outcomes.

# **10) Link to Other Useful Models**

1. <https://github.com/IBM/employee-attrition-aif360/blob/master/notebooks/employee-attrition.ipynb>
2. <https://github.com/JNYH/employee_attrition/blob/master/employee_attrition.ipynb>
3. <https://github.com/elastic/examples/tree/master/Machine%20Learning/Analytics%20Jupyter%20Notebooks>
4. <https://github.com/ganesh10-india/HR_Analytics-Employee_Attrition-Classification-Models/blob/main/HR_Analytics_Employee_Attrition_Classification_Models.ipynb>
5. <https://www.adam-d-mckinnon.com/posts/2020-08-04-clusteranalysis/>

# **11) Github Link to Assignment Notebook**

https://github.com/FATIMASP/IBM-MACHINE-LEARNING-CERTIFICATION/blob/main/Exploratory%20Data%20Analysis%20for%20Machine%20Learning/EDA%20SUPERVISED%20CLASSIFICATION%20EMPLOYEE%20ATTRITION.ipynb