IBM MACHINE LEARNING

**Fatima, Sayeda**

**10/23/2022**

­­­­­­­­

UNSUPERVISED MACHINE LEARNING MODELS FOR MALL CUSTOMER SEGMENTATION

**Table of Contents**

[**1) Project Overview and Objective** 2](#_Toc117458255)

[**1a) Problem Overview** 2](#_Toc117458256)

[**1b) Objectives** 2](#_Toc117458257)

[**1c) Implications for Business:** 2](#_Toc117458258)

[**2) About the Dataset** 3](#_Toc117458259)

[**2a) Brief Description of Chosen Data Set** 3](#_Toc117458260)

[**2b) Summary of Data Attributes** 3](#_Toc117458261)

[**2c) Main Aim of Analysis** 3](#_Toc117458262)

[**3) Data Exploration, Data Cleansing and Features Engineering** 4](#_Toc117458263)

[**3a) Data Exploration:** 4](#_Toc117458264)

[**3b) Data Cleansing Actions & Features Engineering:** 7](#_Toc117458265)

[**4) Summary of Training Different Clustering Models** 13](#_Toc117458266)

[**4a) Machine Learning Algorithm Approaches** 13](#_Toc117458267)

[**I) Data Level Approaches:** 16](#_Toc117458268)

[**II) Algorithm Ensemble Approach:** 16](#_Toc117458269)

[**4b) Summarizing Employed Models** 17](#_Toc117458270)

[**1) K-Means Clustering Algorithm:** 17](#_Toc117458271)

[**2) Mini-Batch K-Means Clustering Algorithm:** 19](#_Toc117458272)

[**3) Hierarchical Agglomerative Clustering Algorithm:** 21](#_Toc117458273)

[**4) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm:** 23](#_Toc117458274)

[**5) Ordering Points to Identify the Clustering Structure (OPTICS) Algorithm:** 25](#_Toc117458275)

[**5) Result Summary and Recommended Model** 27](#_Toc117458276)

[**5a) Result Summary** 27](#_Toc117458277)

[**5b) Model Choice and Justification** 27](#_Toc117458278)

[**6) Summary Key Findings and Insights** 28](#_Toc117458279)

[**6a) Cluster Description** 28](#_Toc117458280)

[**6b) Cluster Information Assignment** 28](#_Toc117458281)

[**6c) Cluster Assignment Result** 29](#_Toc117458282)

[**7) Future Recommendations** 30](#_Toc117458283)

[**8 Useful Links** 31](#_Toc117458284)

[**8a) Link to Other Useful Models** 31](#_Toc117458285)

[**8b) Model Evaluation and Scoring** 31](#_Toc117458286)

[**8c) SKLEARN Unsupervised Model Types and their Parameters** 31](#_Toc117458287)

[**9) Github Link to Assignment Notebook** 31](#_Toc117458288)

# **1) Project Overview and Objective**

## **1a) Problem Overview**

One of the main post-Covid challenges faced by shopping malls is sustaining revenue and maintaining profits. This challenge has been even more compounded by a falling pounds and global onset of another recession. Consequently, both big and small shopping malls have to devote more efforts to ensure attracting customers through discounts and offers specific to their interests to ensure repeat purchase and customer loyalty. It is here that unsupervised machine learning models can be very useful to gain deeper insight into underlying factors and their relationship in driving customer segmentation.

## **1b) Objectives**

Hence, the main objective of the following unsupervised machine learning modelling and analysis is targeted towards answering the following queries:

* Into how many clusters can the customers be segmented?
* What are the main characteristic driving this segmentation?

## **1c) Implications for Business:**

As a consequence, implementation of the model will enable the business:

* To segment customers into separate groups.
* To gain insight into customer characteristics contributing to identified grouping for sales strategy formulation.
* Target specific offers based on customers segmentation and cluster contributory characteristicescan which will likely lead to increased purchase, customer loyalty and sustainable profits.

# **2) About the Dataset**

## **2a) Brief Description of Chosen Data Set**

This project uses a hypothetical dataset ‘Customer Mall’ which seems to have been acquired for regular customers visiting a shopping a mall and was downloaded from the following link:

<https://www.kaggle.com/code/vjchoudhary7/kmeans-clustering-in-customer-segmentation/data>

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

The dataset exhibits 200 data points (rows) and 5 features (columns) reflecting on customers’ characteristics where, based on their spending, each customer has been assigned a Spending Score. Of these, the main four features are 'Age', 'Annual Income', 'Spending Score' and 'Gender'.

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

Hence, the aim of this project is segmentation of mall customers to aid formulation of target marketing strategy and its implementation. By segmenting customers into clusters, specific offers can be targeted to each cluster which will likely lead to increased purchase, customer loyalty and sustainable profits.

Therefore, the aim of this analysis is to:

* Segment customers into different groups and clusters
* Reflect on the main characteristic driving this segmentation

A2T63J6JM74V7uafcf

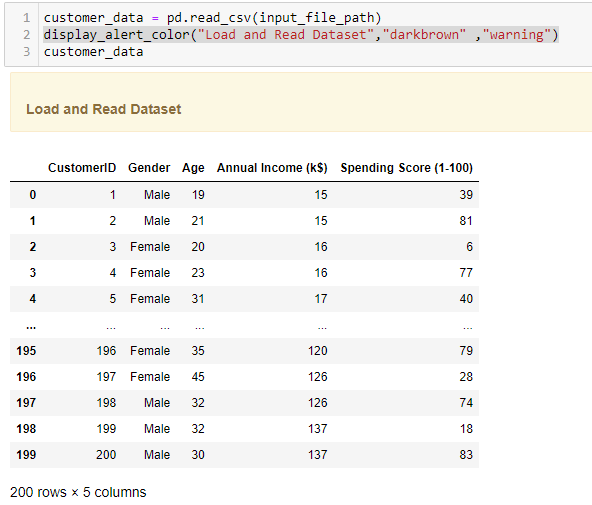
https://scentscientists.com/products/rose-geraniol-10ml?variant=37899469029552&currency=GBP&utm\_medium=product\_sync&utm\_source=google&utm\_content=sag\_organic&utm\_campaign=sag\_organic&utm\_campaign=gs-2021-01-08&utm\_source=google&utm\_medium=smart\_campaign&gclid=Cj0KCQjw48OaBhDWARIsAMd966BNXOfbyPl62YecvCqLoRog1oo7ok0zvYvxMtrKOpnWpW-QBPhjbgQaApKHEALw\_wcB

# **3) Data Exploration, Data Cleansing and Features Engineering**

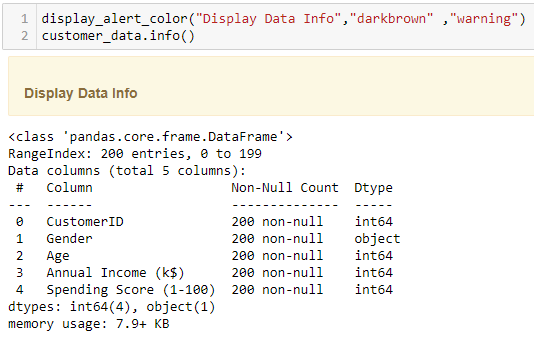
Since the quality of any machine learning model highly depends on quality of data, hence, this stage is not only most important but is also time consuming. Hence, it was conducted in a step-by-step process.

## **3a) Data Exploration:**

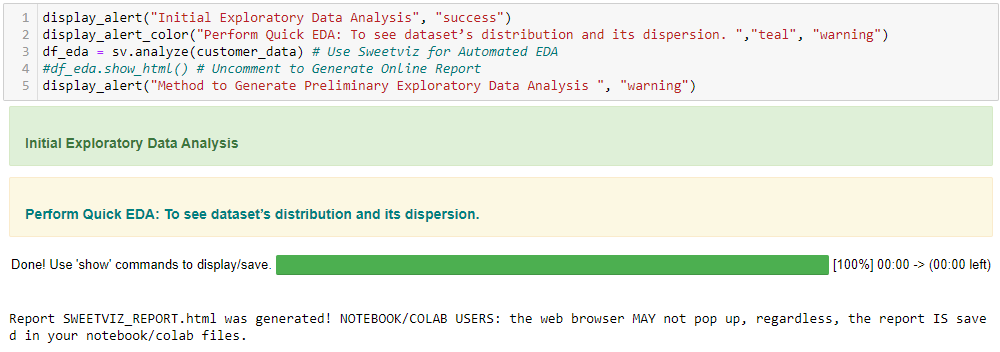
* Data was first loaded into pandas dataframe

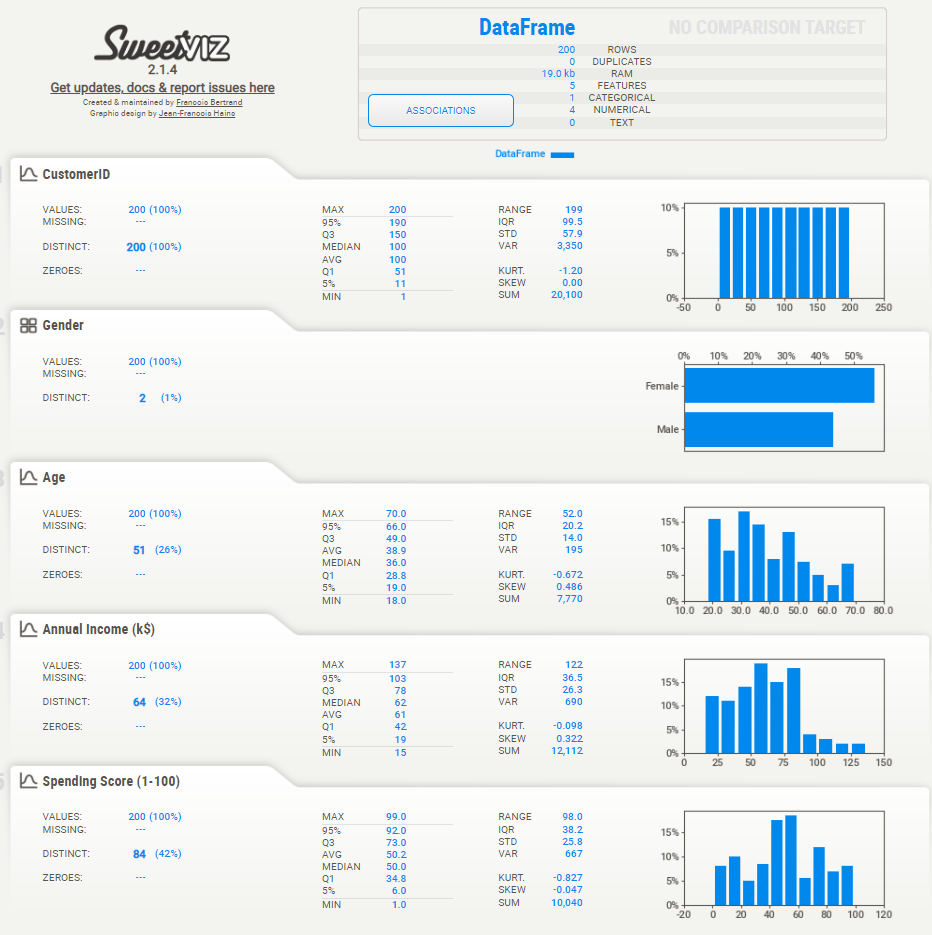


* Column types were then explored

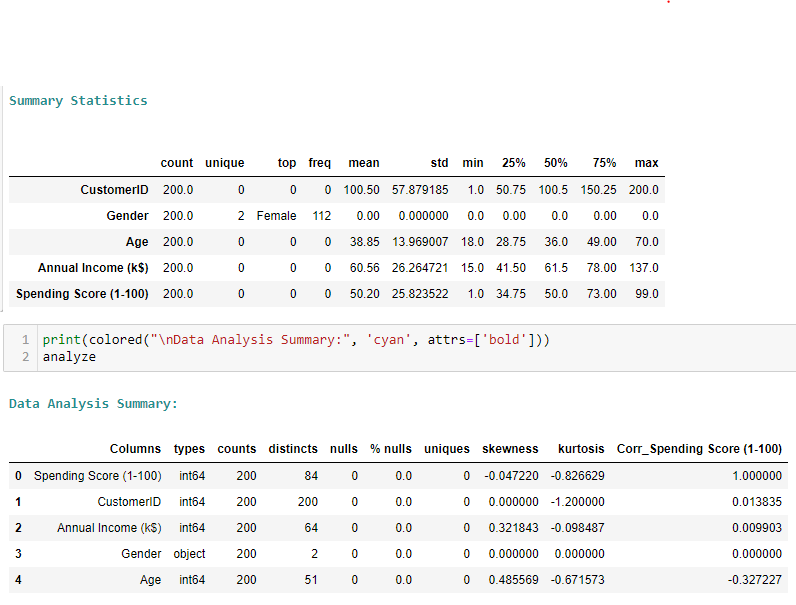


* Automated Exploratory Data Analysis was performed using Sweetviz





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



## **3b) Data Cleansing Actions & 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.

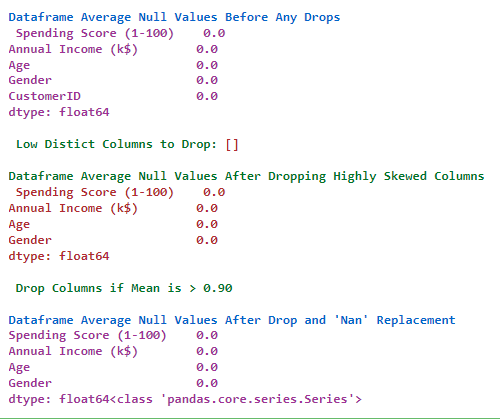
Consequently, adopting filter-based feature selection methods, the project approached filter engineering in the following steps.

An automated data cleansing method was created to do the following:

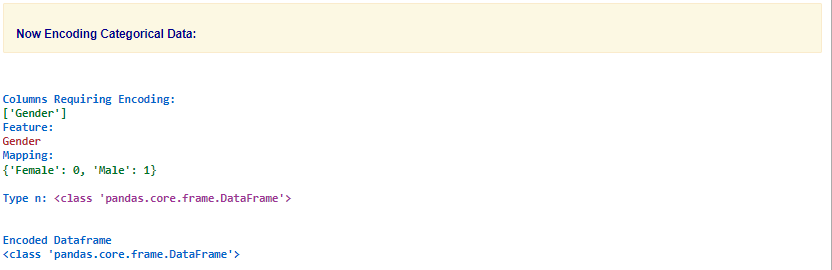
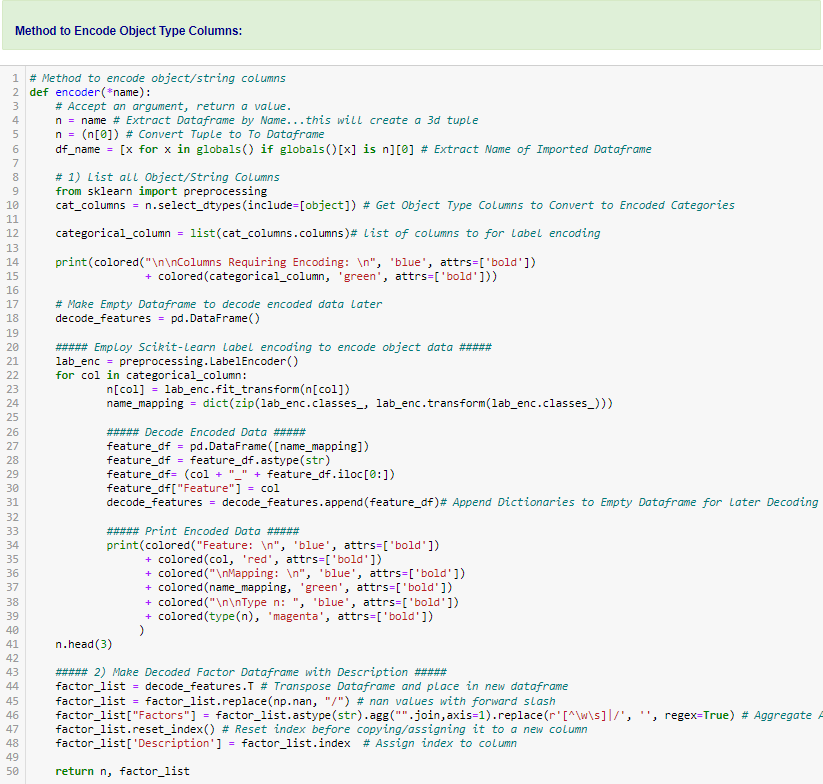
* Drop Columns with Unique Values Less than threshold of 2
* Drop Highly Skewed & Low Correlation Columns with target
* Drop Columns with High Nan Values



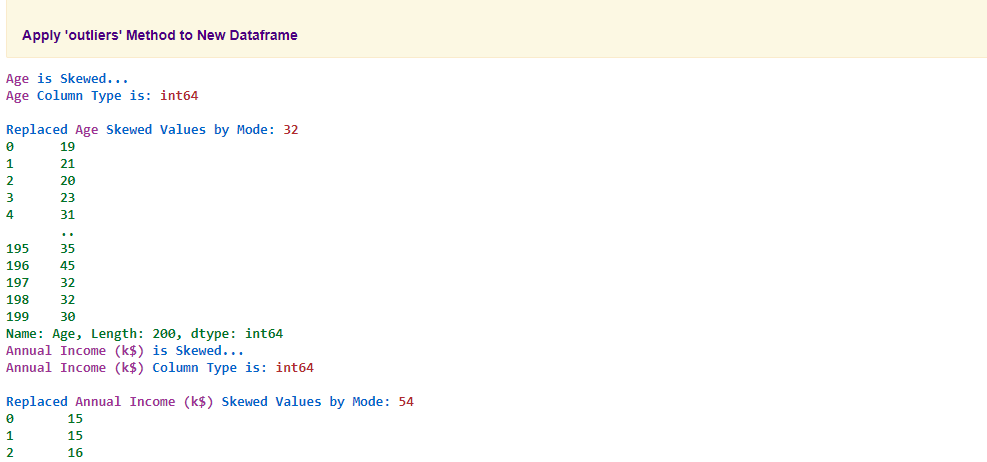
Null values were summed, and data was found to exhibit zero null values. Thus, no filling of null values was required.



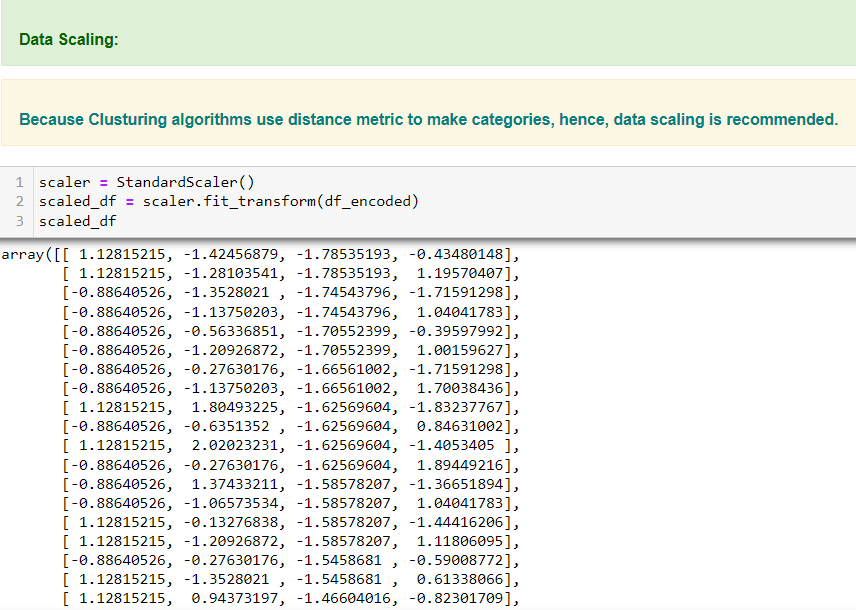
**Data Encoding:** Additionally, Data Encoding of Object or String Columns was carried out to facilitate any statistical computation during features selection process. Hence, after deep copying of original dataset, a function was created and employed to encode object data using Scikit-learn label encoder.



**Outlier Treatment:** Supervised learning models like K\_Means are sensitive to outliers. Hence, an automated method was created to replace outliers with “Mode”, that is the most common value.

­­­­ 

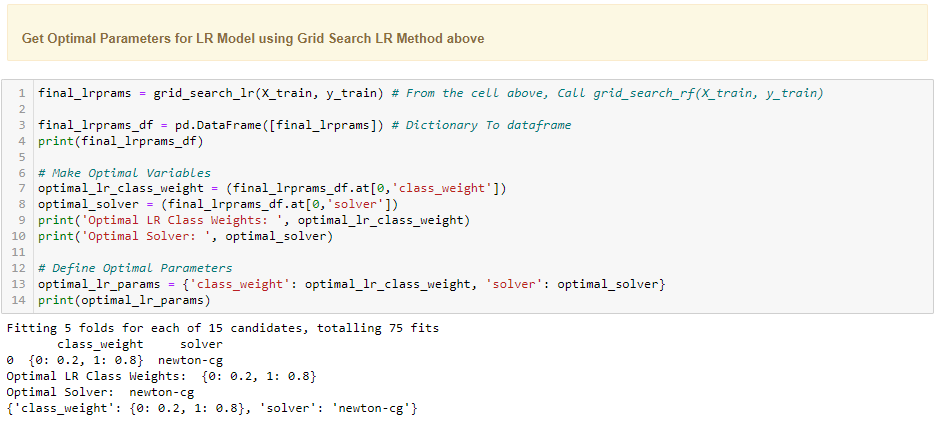
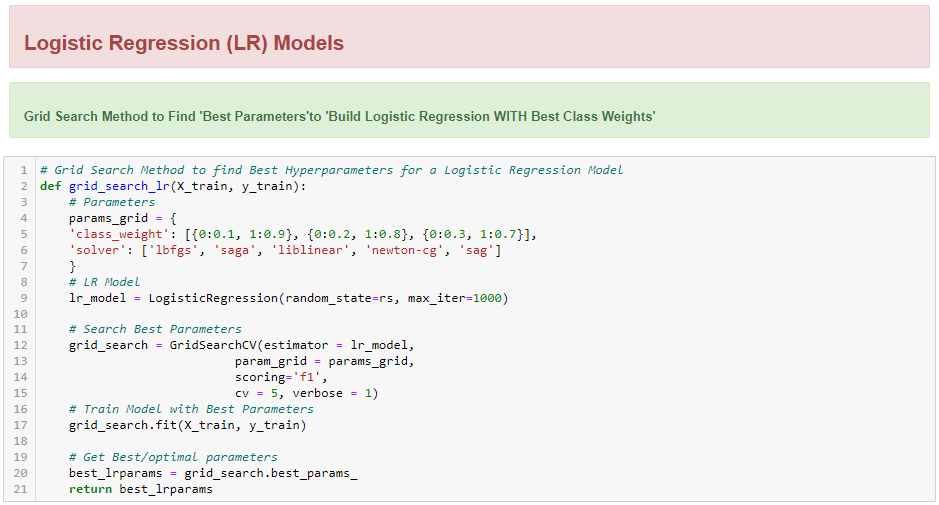
**Data Scaling:** Due to dependency of clustering algorithms on distance matrix, data scaling was carried out.

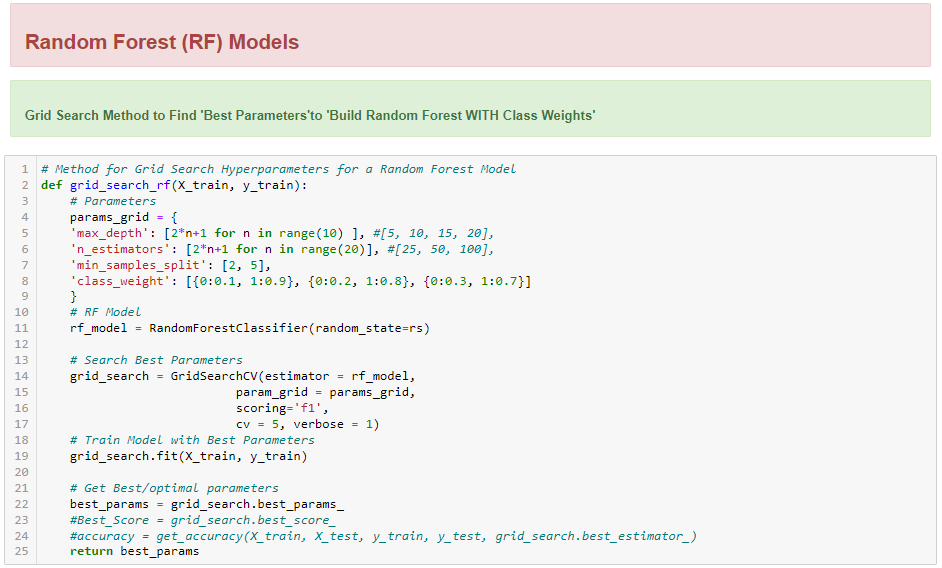


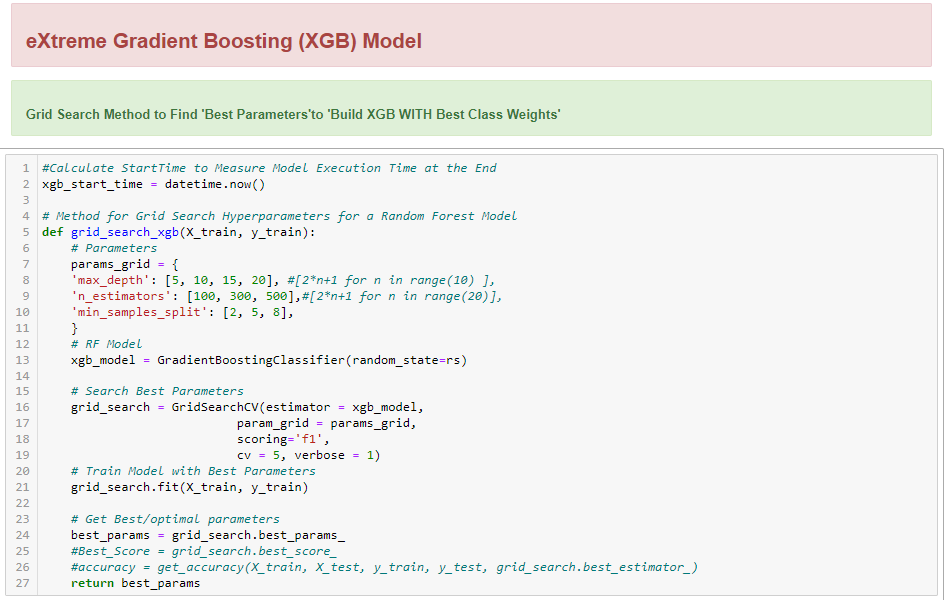
# **4) Summary of Training Different Clustering Models**

## **4a) Machine Learning Algorithm Approaches**

Although, While data level and algorithm ensemble approaches do exist for dealing with imbalanced datasets, nevertheless, an automated optimal parameter search method was created to achieve best class reweighting along with isolating other optimal model parameters. This approach was employed because best hyper-parameters are not automatically learnt within estimators and its manual search not only slows down model development but may also lead to ineffective model construction. Hence, exhaustive cv grid search approach was used to pass parameter arguments to the constructor in order to find optimal parameters for each model.







Additionally, the following approaches were also combined with optimal parameters to find a model with best scores:

### **I) Data Level Approaches:**

**i) Synthetic Minority Over-sampling Technique (SMOTE):**

Due to highly imbalance class distribution, employee data contains very few instances of minority class for any classification model to explicitly learn decision boundary. A popular approach to tackle this problem is oversampling minority class examples which are close in the feature space using SMOTE. This approach allowed us to achieve a balanced class distribution.

**ii) Random under-sampling:**

Using random under-sampling examples from majority class were deleted to achieve a balanced class distribution.

**iii) Random over-sampling:**

Random oversampling was employed to duplicate examples from minority class to achieve a balanced class distribution.

### **II) Algorithm Ensemble Approach:**

1. **Boosting:**

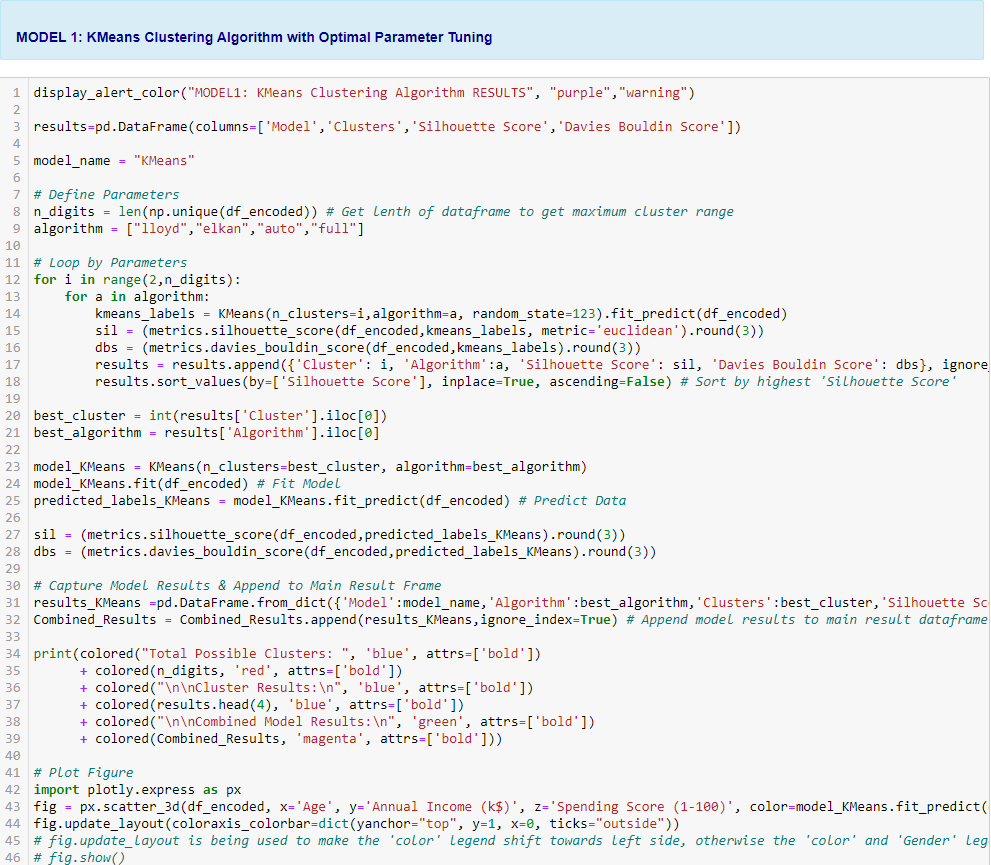
Using boosting, a sequential aggregate of base classifier was created on weighted versions of training data set which focused on misclassified samples at every stage of creating new classifiers based on sample weights that were altered as per classifier’s performance. Boosting was achieved using XGB Classifier.

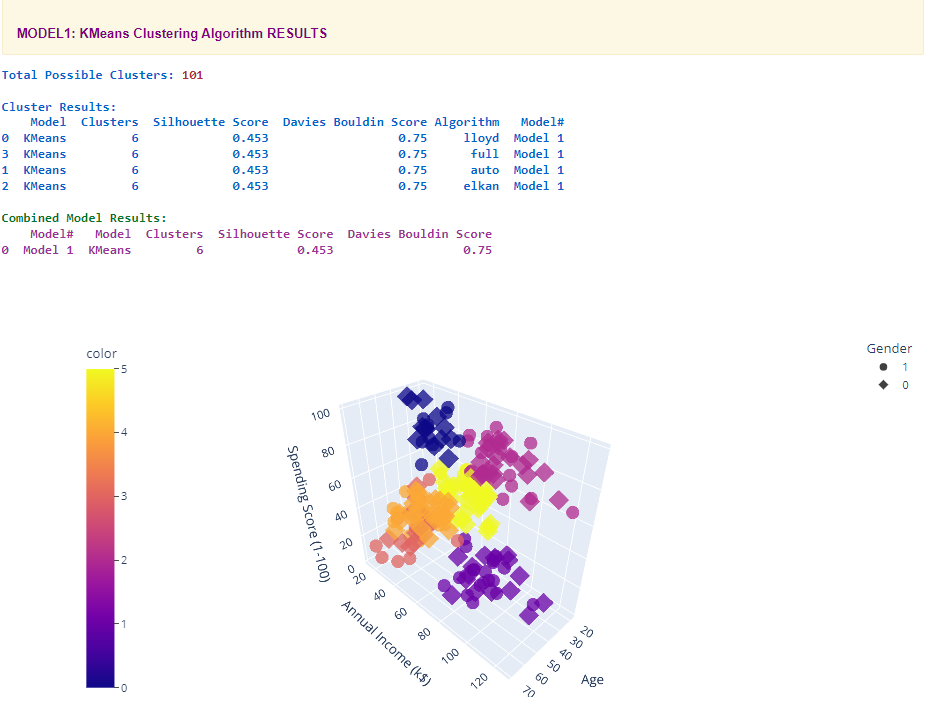
## **4b) Summarizing Employed Models**

Following three main classifier models have been used to predict employee attrition.

### **1) K-Means Clustering Algorithm:**

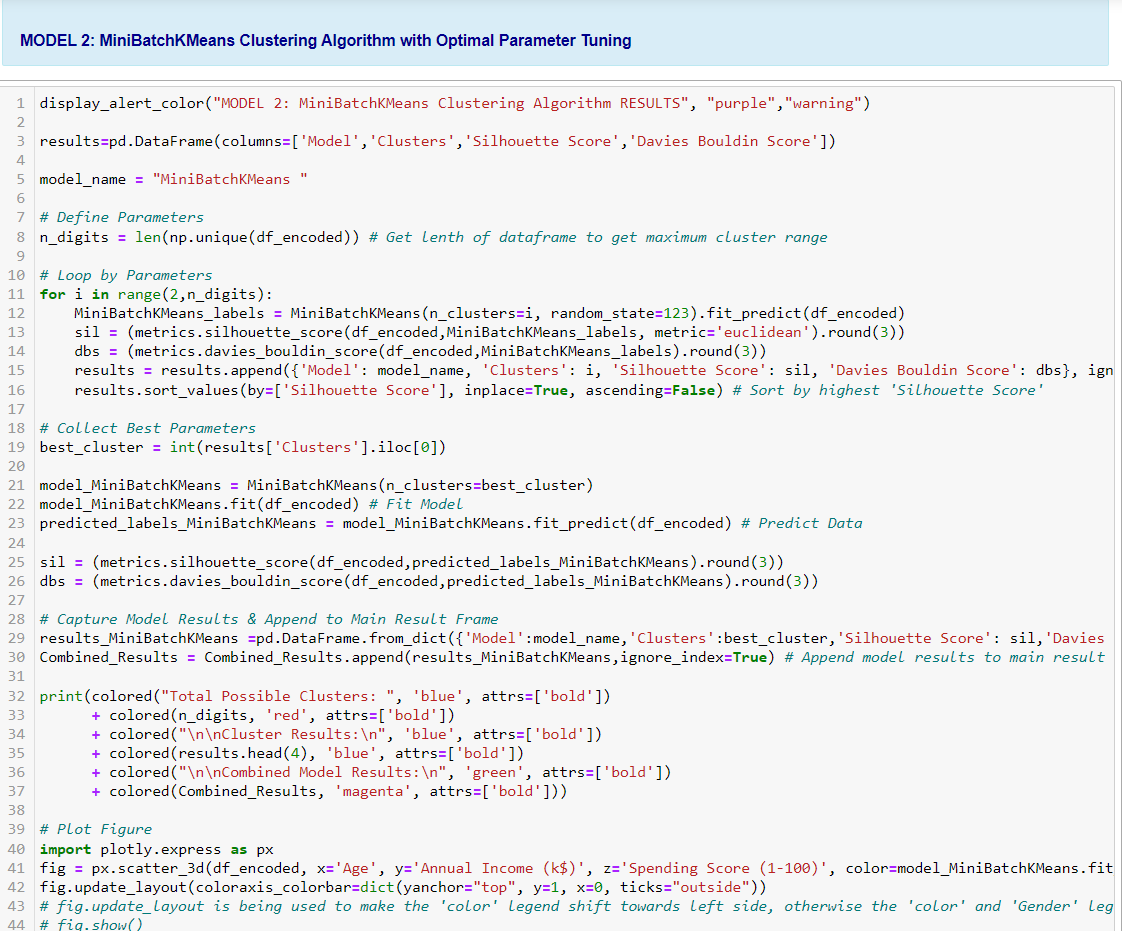
Due to the focus on segmentation, the popular K-Means Clustering Algorithm with optimal parameters was employed.

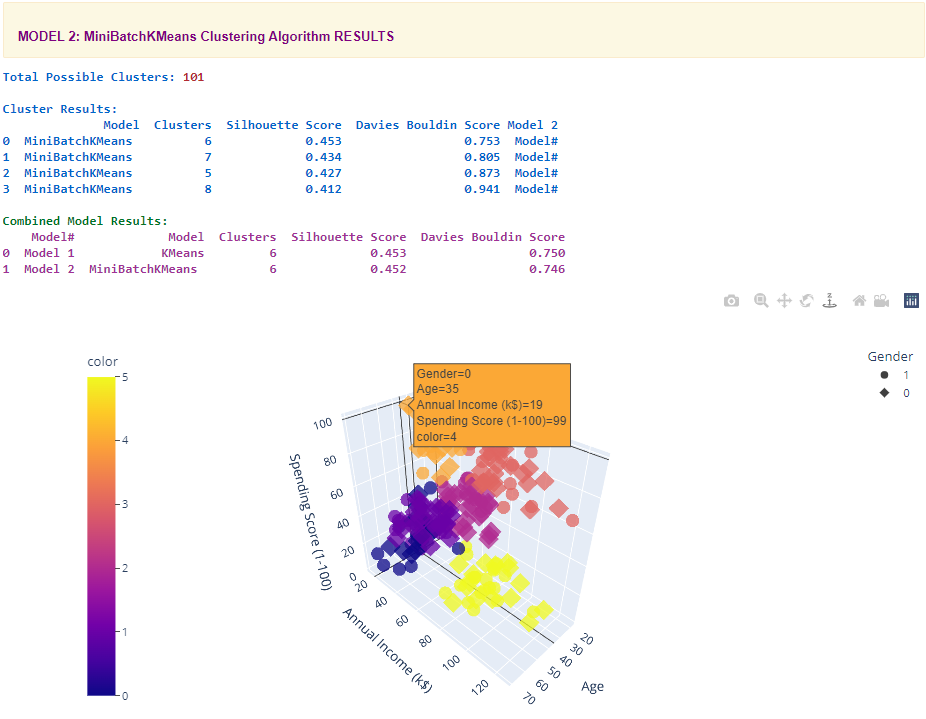




### **2) Mini-Batch K-Means Clustering Algorithm:**

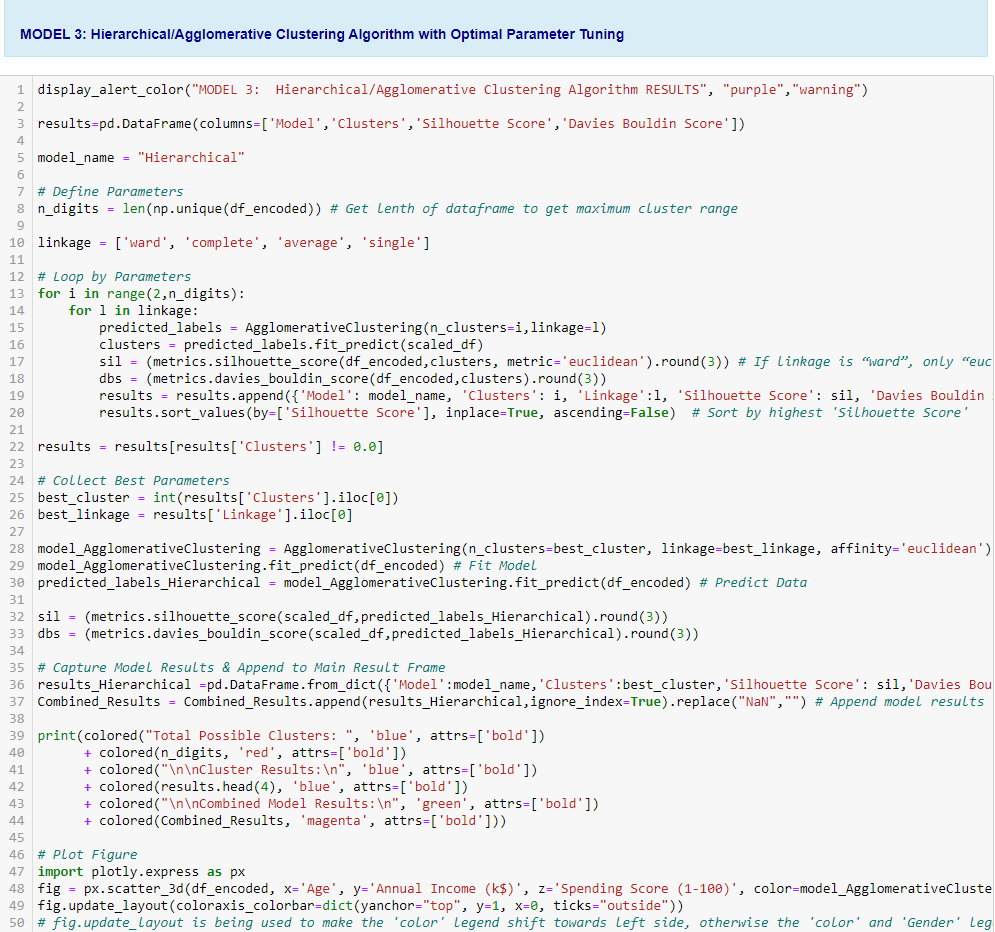
Mini-Batch K-Means Clustering Algorithm with optimal parameters, which is faster than K-Means due to utilizing random fixed size data batches, was employed.

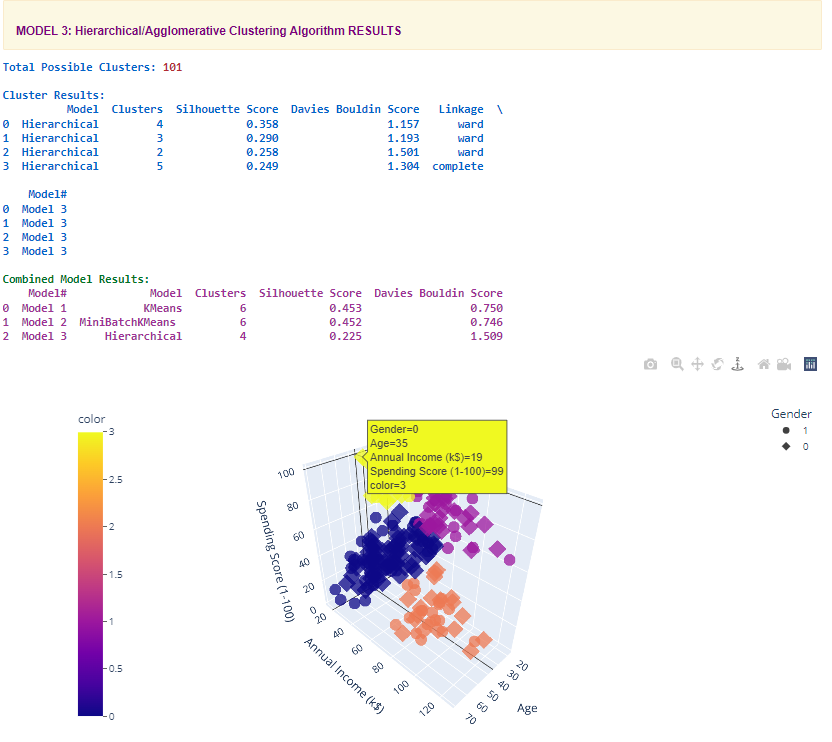




### **3) Hierarchical Agglomerative Clustering Algorithm:**

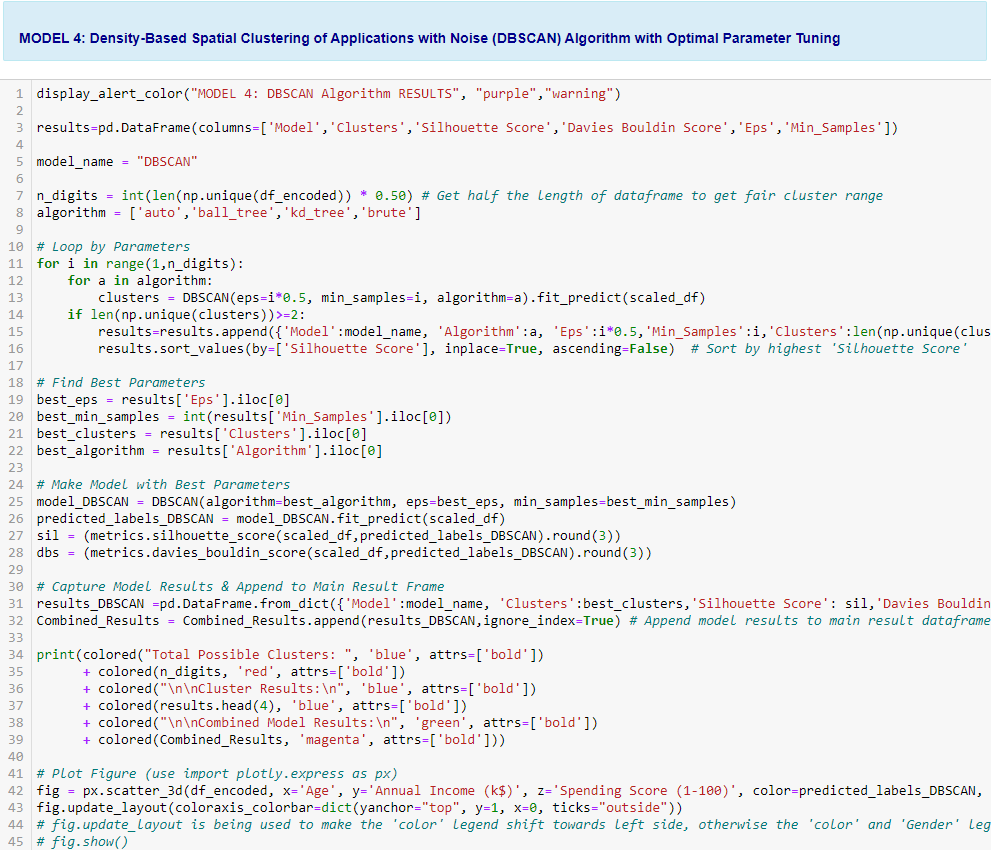
Hierarchical Agglomerative Clustering Algorithm with optimal parameters, which starts with smaller clusters to merge them into bigger ones, was employed.

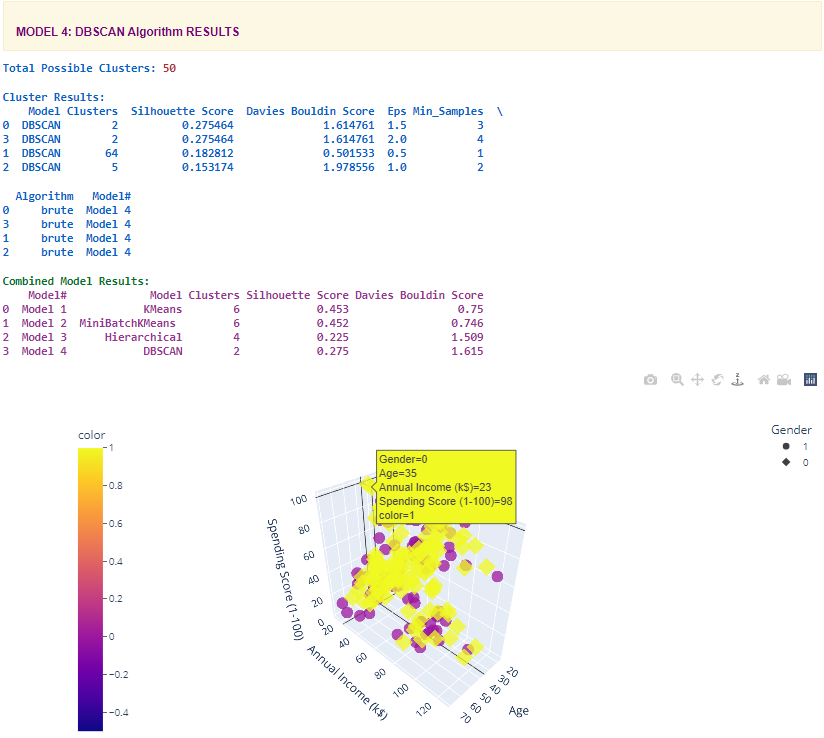




### **4) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm:**

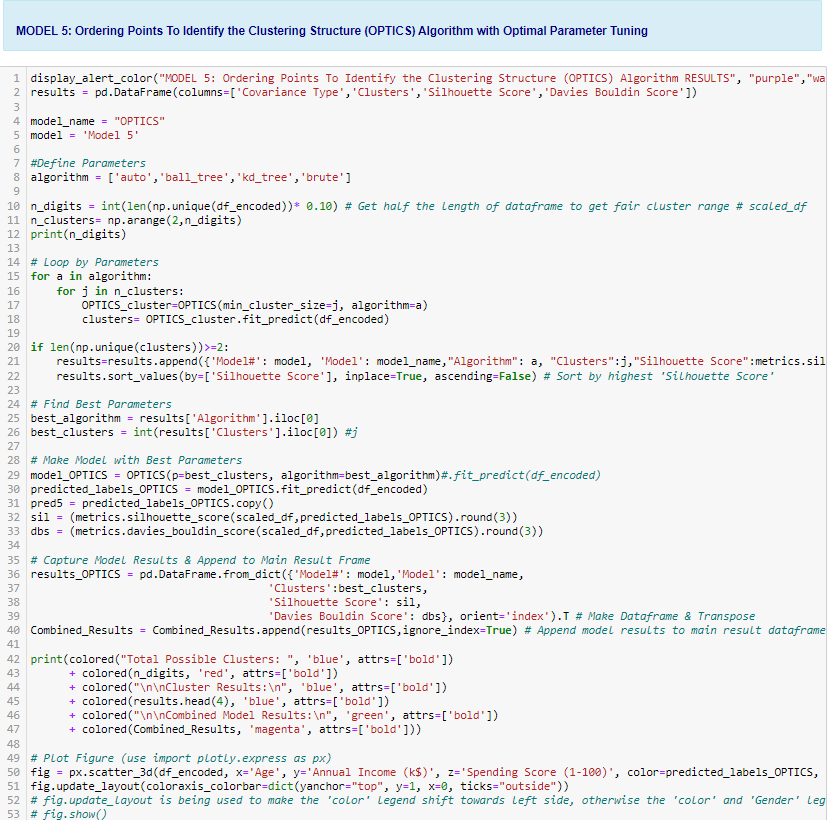
DBSCAN Algorithm with optimal parameters, which groups closely packed points together, was employed.

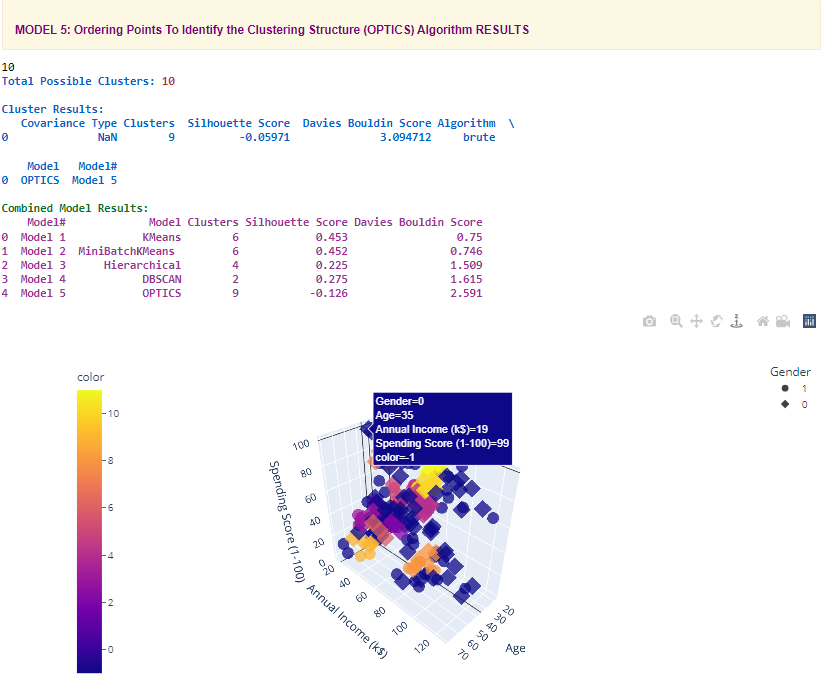




### **5) Ordering Points to Identify the Clustering Structure (OPTICS) Algorithm:**

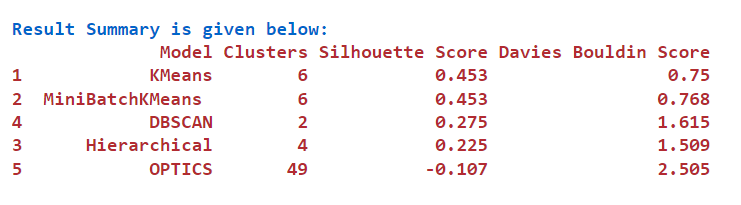
Ordering Points to Identify the Clustering Structure (OPTICS) Algorithm with optimal parameters, which orders data points so that spatially closest points become neighbours in the ordering, was employed.



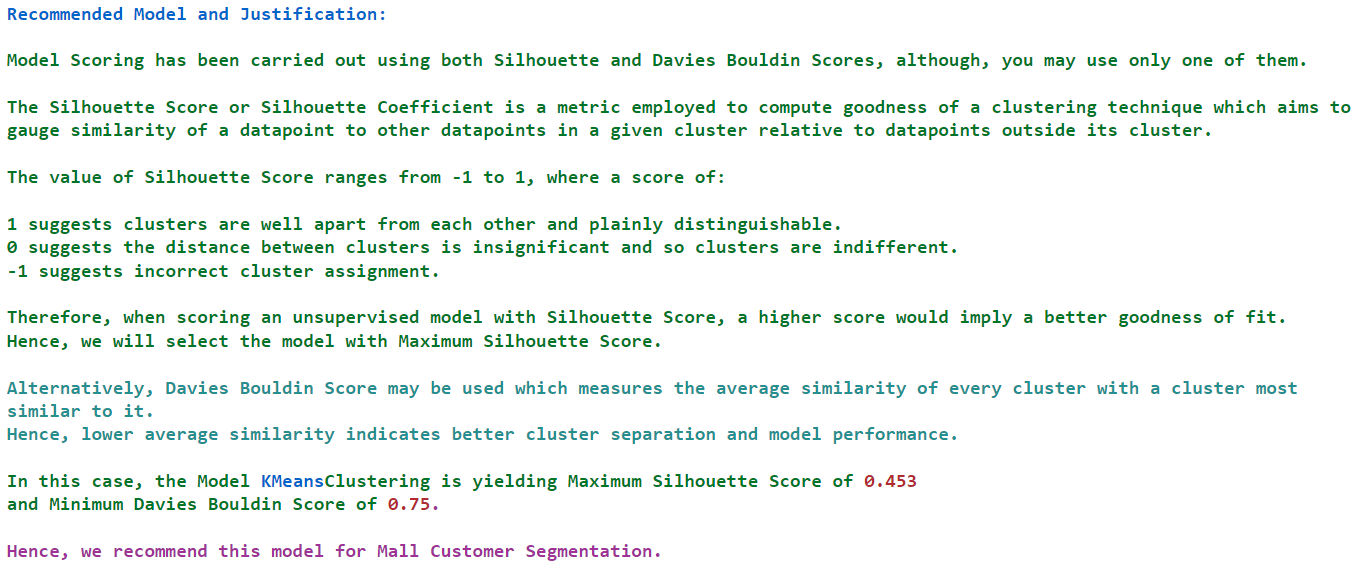


# **5) Result Summary and Recommended Model**

## **5a) Result Summary**



## **5b) Model Choice and Justification**



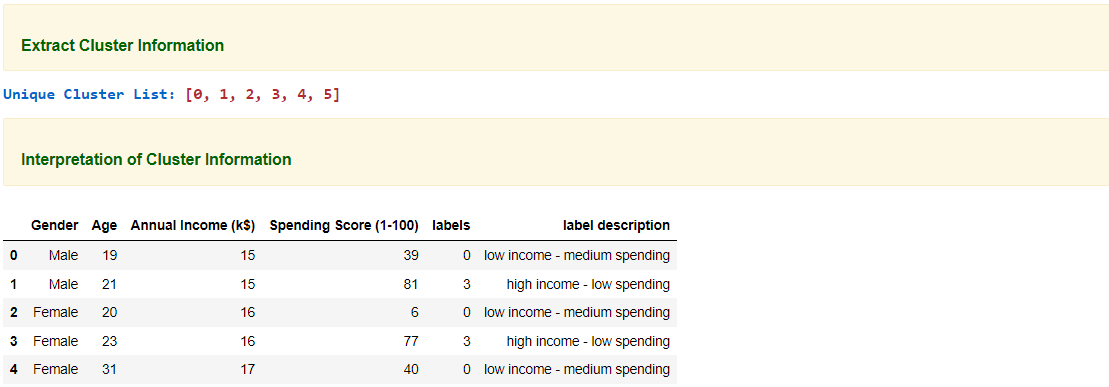
# **6) Summary Key Findings and Insights**

Model figures clearly show age, annual income and spending score values for each cluster which aided in gaining deeper insight into cluster characteristics. Since age is scattered across all clusters, it does not seem to be a distinctive feature. Consequently, based on further descriptive analysis, cluster descriptions were assigned as follows:

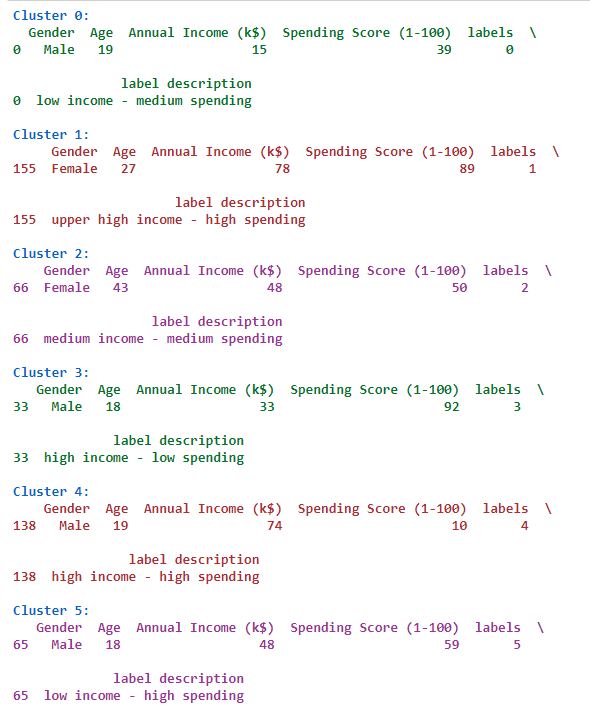
## **6a) Cluster Description**



## **6b) Cluster Information Assignment**



## **6c) Cluster Assignment Result**



# **7) Future Recommendations**

Despite its novel approach to automatically pick best model with best parameters, the project is not without its shortcomings.

To begin with, it only utilized a very small dataset with merely five features.

Hence, it may have discounted other important factors that may impact customer segmentation.

Furthermore, the project could have employed other unsupervised models to ensure even more improved results.

Future recommendations are, hence, as follows:

1) Use a larger dataset with more features.

2) Employ more unsupervised models like MeanShift, Birch etc.

3) Try other evaluation metrics like Calinski-Harabasz Index or Adjusted Rand Index and see if there is a difference in results.

# **8 Useful Links**

## **8a) Link to Other Useful Models**

* [**https://medium.com/@mbektas/customer-segmentation-with-clustering-algorithms-in-python-be2e021035a**](https://medium.com/@mbektas/customer-segmentation-with-clustering-algorithms-in-python-be2e021035a)
* [**https://www.kaggle.com/code/aayush7kumar/clustering-using-k-means-hierarchical-and-dbscan/notebook**](https://www.kaggle.com/code/aayush7kumar/clustering-using-k-means-hierarchical-and-dbscan/notebook)
* [**https://github.com/eklavyasaxena/IBM-Unsupervised-Machine-Learning-Capstone-k-Means-Clustering/blob/master/Exploring-the-Taste-of-NYC-Neighborhoods/Project\_Notebook.ipynb**](https://github.com/eklavyasaxena/IBM-Unsupervised-Machine-Learning-Capstone-k-Means-Clustering/blob/master/Exploring-the-Taste-of-NYC-Neighborhoods/Project_Notebook.ipynb)
* [**https://goodboychan.github.io/python/machine\_learning/natural\_language\_processing/vision/2020/10/26/01-K-Means-Clustering-for-Imagery-Analysis.html**](https://goodboychan.github.io/python/machine_learning/natural_language_processing/vision/2020/10/26/01-K-Means-Clustering-for-Imagery-Analysis.html)

## **8b) Model Evaluation and Scoring**

* [**https://towardsdatascience.com/how-to-evaluate-unsupervised-learning-models-3aa85bd98aa2**](https://towardsdatascience.com/how-to-evaluate-unsupervised-learning-models-3aa85bd98aa2)
* [**https://medium.com/@mbektas/customer-segmentation-with-clustering-algorithms-in-python-be2e021035a**](https://medium.com/@mbektas/customer-segmentation-with-clustering-algorithms-in-python-be2e021035a)
* [**https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad**](https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad)

## **8c) SKLEARN Unsupervised Model Types and their Parameters**

* [**https://scikit-learn.org/stable/unsupervised\_learning.html**](https://scikit-learn.org/stable/unsupervised_learning.html)
* [**https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html**](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)
* [**https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html**](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html)

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

[**https://github.com/FATIMASP/IBM-MACHINE-LEARNING-CERTIFICATION/blob/main/Unsupervised%20Machine%20Learning/UNSUPERVISED%20LEARNING%20MALL%20CUSTOMERS%20FINAL%20MODEL.ipynb**](https://github.com/FATIMASP/IBM-MACHINE-LEARNING-CERTIFICATION/blob/main/Unsupervised%20Machine%20Learning/UNSUPERVISED%20LEARNING%20MALL%20CUSTOMERS%20FINAL%20MODEL.ipynb)