AIML | MODULE PROJECT

Unsupervised Learning

TOTAL SCORE 60

General Instructions:

1. Submission of all the parts is expected in 1 notebook only

2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only

3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.

4. If output for any code cell is missing, 50% marks will be deducted.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and

2. '.html' (Jupyter Notebook > File > Download as > HTML)

5 Marks will be deducted if submission in any of the

Part A - 40 Marks

- **DOMAIN :** Business
- **CASE STUDY** : Product Segmentation Case Study
- **CONTEXT :** When you think of sneakers for a trip, the importance of good footwear cannot be discarded, and the obvious brands that come to mind are Adidas and Nike. Adidas vs Nike is a constant debate as the two giants in the apparel market, with a large market cap and market share, battle it out to come on top. As a newly hired Data Scientist in a market research company, you have been given the task of extracting insights from the data of men's and women's shoes, and grouping products together to identify similarities and differences between the product range of these renowned brands
- **DATA DESCRIPTION :**

The dataset consists of 3268 products from Nike and Adidas with features of information including their ratings, discount, sales price, listed price, product name, and the number of reviews.

- **Product Name:** Name of the product
- **Product ID**: ID of the product
- **Listing Price**: Listed price of the product
- **Sale Price**: Sale price of the product
- **Discount**: Percentage of discount on the product
- **Brand:** Brand of the product
- **Rating**: Rating of the product
- **Reviews**: Number of reviews for the product

- **PROJECT OBJECTIVE:**

To perform an exploratory data analysis and cluster the products based on various factors

- **STEPS AND TASKS [40 Marks]:**

1. Data Understanding:

a. Read the Data set "data_add_nik.csv" and print the shape of the dataset. **[2 marks]**

b. Check if there is any redundant column in the dataset and drop them. **[1 mark]**

c. Check if there are any duplicate records in the dataset. If yes, drop them. **[1 mark]**

d. Check the 5 points summary of the whole data and share your observations. **[1 mark]**

2. Exploratory Data Analysis:

a. Check how many products have Listing_price '0'. Check it's 5 points summary and share your insights. **[2 marks]**

b. Records having Listing_price '0', replace them with Sale_Price of that record. **[2 marks]**

c. Check and print feature-wise percentage of missing values present in the data and impute with the best suitable approach. **[2 Marks]**

d. Perform univariate analysis on the data and share your insights. **[2 marks]**

e. Perform bivariate and multivariate analysis and share your insights. **[3 marks]**

3. Data Preprocessing :

a. Scale all the numeric variables using suitable technique. **[2 marks]**

4. K-Means Clustering :

a. Apply K-Means clustering for 2 to 10 clusters. **[3 Marks]**

b. Plot a visual and find elbow point. **[2 Marks]**

c. Considering the above visual, mention which are the optimal clusters. **[1 Mark]**

d. Calculate silhouette scores of all K clusters 2 to 10. **[2 Marks]**

e. Considering 4.c and 4.d, what is the optimal value of K? **[1 Mark]**

f. Train a K-means clustering model once again on the optimal number of clusters. **[2 Marks]**

g. Add K-means cluster labels to the original data. **[3 Marks]**

h. Do cluster profiling and display. **[3 Marks]**

i. Share your insights on the clusters which are built. Here try to explain what each cluster is saying. **[3 Marks]**

j. Considering the cluster profiles, give any 2-business recommendation which will help the business. **[2 Marks]**

- **DOMAIN :** Business
- **CASE STUDY :**Tourism Case Study
- **CONTEXT :** Tourism is now recognised as a directly measurable activity, enabling more accurate analysis and more effective policies can be made for tourism. Whereas previously the sector relied mostly on approximations from related areas of measurement (e.g. Balance of Payments statistics), tourism nowadays is a productive activity that can be analysed using factors like economic indicators, social indicators, environmental & infrastructure indicators, etc. As a Data Scientist in a leading tours and travels company, you have been assigned the task of analysing several of these factors and group countries based on them to help understand the key locations where the company can invest in tourism services
- **DATA DESCRIPTION:**

This dataset contains key statistical indicators of the countries. It covers sections like general information, economic indicators, social indicators, environmental & infrastructural indicators

Data Dictionary :

- **country**: country
- **Region**: region of the country
- **Surface area**: Surface area in sq. km
- **Population in thousands**: Population of the country, in thousands, as in the year 2017
- **Population density**: Population density per km2, as in the year 2017
- **GDP**: Gross domestic product: GDP of the country in million USD
- **Economy**: **Agriculture:** Contribution of agriculture to the economy as a percentage of Gross Value Added
- **Economy: Industry:** Contribution of the industry to the economy as a percentage of Gross Value Added
- **Economy: Services and other activity:** Contribution of services and other activities to the economy as a percentage of Gross Value Added
- **International trade:** Balance: Amount, in million USD, of balance between international exports and imports
- **Health: Total expenditure:** Total expenditure on healthcare facilities as a percentage of GDP
- **Education:** Government expenditure: Total expenditure on education as a percentage of GDP
- **Mobile-cellular subscriptions:** no. of mobile/cellular subscriptions per 100 people
- **Individuals using the Internet:** no. of individuals using the Internet per 100 people
- **CO2 emission estimates**: CO2 emission estimates in million tons

- **PROJECT OBJECTIVE:** To explore the data and identify different groups of countries based on important factors to find key locations where investments can be made to promote tourism services

- **STEPS AND TASKS [20 Marks]:**

1. **Data Understanding:**
    a. Read the Data set "country_stats.csv" and print the shape of the dataset. **[1 Mark]**
    b. Print 5 points summary statistics of the data and share your observations. **[1 Mark]**
2. **Exploratory Data Analysis:**
    a. Perform Univariate Analysis on the data and find out which country has more population as per 2017. **[1 Mark]**

    Hint: plot population in thousand and country to know which country has more population

    b. Perform bivariate and multivariate analysis and share your insights. Any 2 plots which explains the relationship better. **[2 Marks]**
3. **Data Pre-processing**
    a. Check if the data has any missing values, if any, impute those with suitable approach **[1 Mark]**
    b. Scale all the numeric variables in the data. **[1 Mark]**

4. **Hierarchical Clustering**
    a. Apply Hierarchical clustering to the scaled data. **[2 Marks]**
    b. Identify the number of optimum clusters using Dendrogram and briefly describe them **[2 Marks]**
    c. Do Cluster profiling and display. **[2 Marks]**
    d. Share your insights on cluster profiles and also give any 2 business recommendations. **[2 Marks]**
5. **Dimensionality Reduction using PCA**
    a. Apply PCA on the scaled data with 2 components. **[2 Marks]**
    b. How much Cumulative Variance is Explained by 2 PCA's? **[1 Marks]**
    c. Plot a scatter plot on PCA's formed i.e. PCA 1 and PCA 2 with hue as cluster profiles from
    d. Mention which are the major clusters. **[2 Marks]**

    Hint: sns.scatterplot(data=reduced_df_pca, x="Component 1", y="Component 2", hue=df["HC_Clusters"], palette="rainbow")

    Above code is a hint, please use the names in which you have defined your functions.