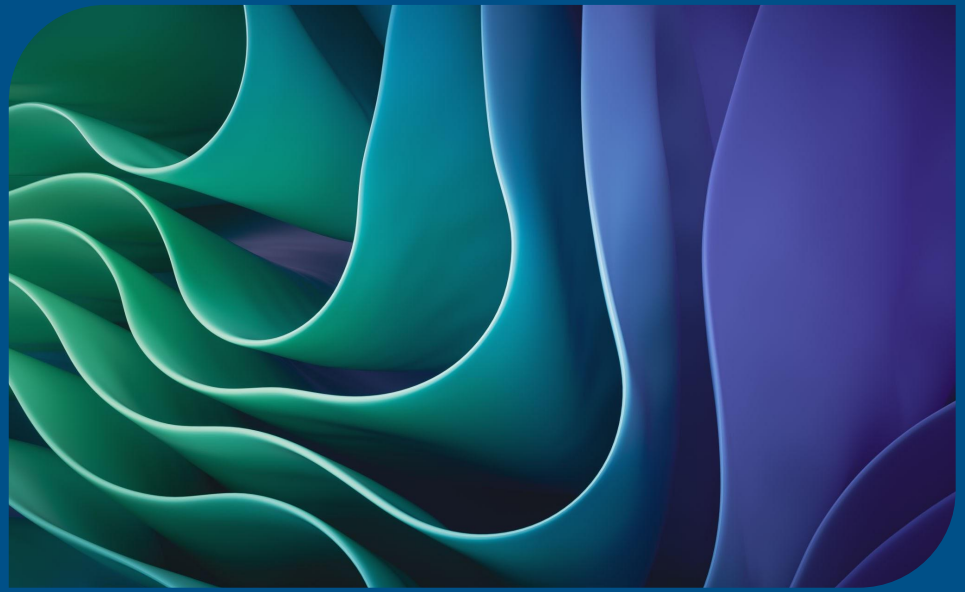
Unsupervised Algorithms in Machine Learning

CSCA 5632



Online Retail

Project Description

This dataset contains transactional data which contains all the transactions between 01/12/2010 to 09/12/2011 for a UK based and registered non-store online retail.

Main Goal of the project:

- Understand Customer purchasing behaviour using RFM strategy (Recency/Frequency/Monetary)
- Identify distinct customer groups using k-means clustering , HDBSCAN/DBSCAN and hierarchical clustering algorithms.
- Detect Outliers (Customers with unusual purchasing patterns)
- Providing actionable insights for the marketing teams.

Citation

Citation:

Seo, J. S. (2015). Online Retail Data Set from UCI ML repo. Kaggle.com. https://doi.org/10.1057/dbm.2012.17)

Chen, D. (2015). Online Retail [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.

Source:

Dr Daqing Chen, Director: Public Analytics group. chend ′@′ lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK

Project Steps

Topics we'll Be Covering in this project

Data Loading and Inspection

Data Cleaning and Preprocessing

EDA

Feature Engineering

Feature Scaling

Unsupervised Learning

Dimensionality Reduction

Clustering Algorithms

Market Basket Analysis

Conclusion

Things to Improve

References

Data Loading and Inspection

```
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

```
df.shape
```

```
(541909, 8)
```

+ Code + Text

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

Data Cleaning and Preprocessing

Preprocessing Stage 1

- **Handle Missing values**
- **Check for data types**
- **Cancelled invoices should be removed**
- **Check for outliers**

```
#Print and display the invoice number starting with 'C'

cancelled_invoices = df[df['InvoiceNo'].astype(str).str.startswith('C')]['InvoiceNo'].unique()
cancelled_count = df['InvoiceNo'].astype(str).str.startswith('C').sum()
print(cancelled_invoices)
print('Total Canceleld Invoices:',cancelled_count)
```

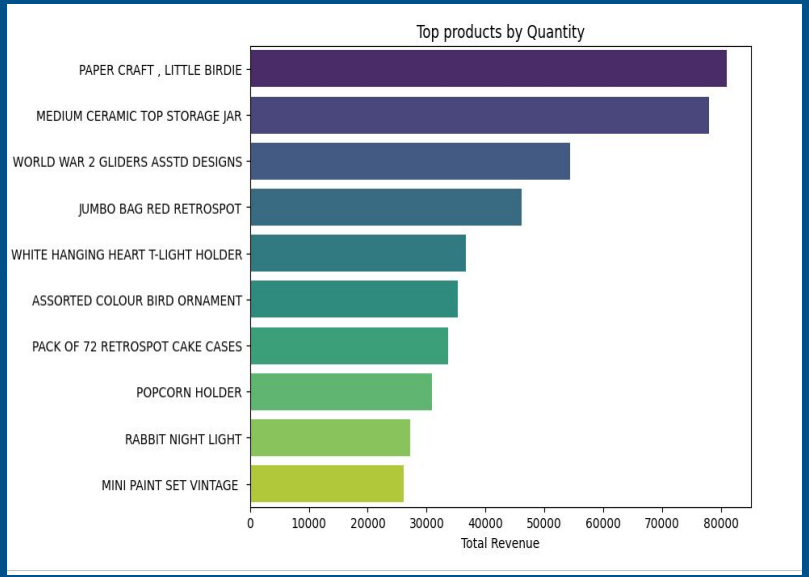
```
['C536379' 'C536383' 'C536391' ... 'C581499' 'C581568' 'C581569']
Total Canceleld Invoices: 8905
```

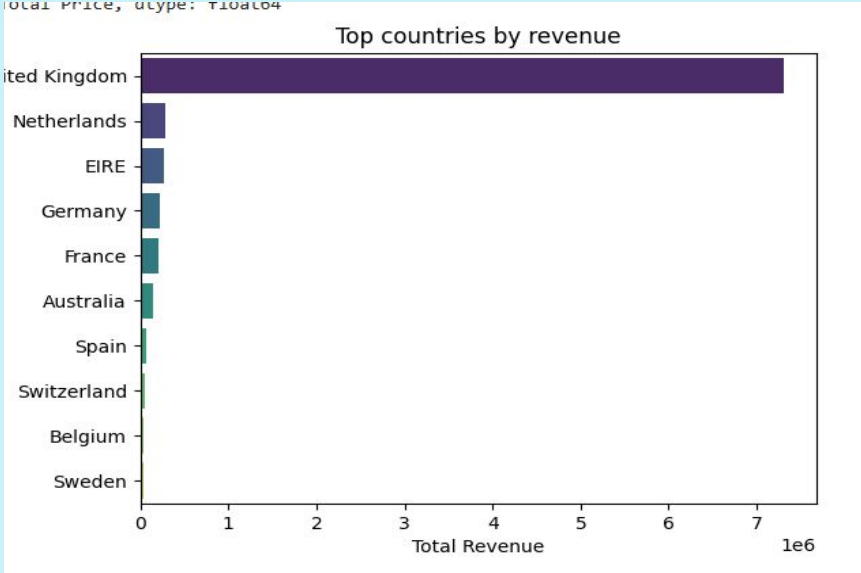
```
#Remove Cancelled Invoices
#Transactions with InvoiceNumber starting with 'C'are cancellations .They can be dropped.

df_cleaned = df[~df['InvoiceNo'].astype(str).str.startswith('C')].copy()
df_cleaned = df_cleaned[(df_cleaned['Quantity']>0) & (df_cleaned['UnitPrice']>0)]
df_cleaned.shape
```

```
(397884, 8)
```

Exploratory Data Analysis

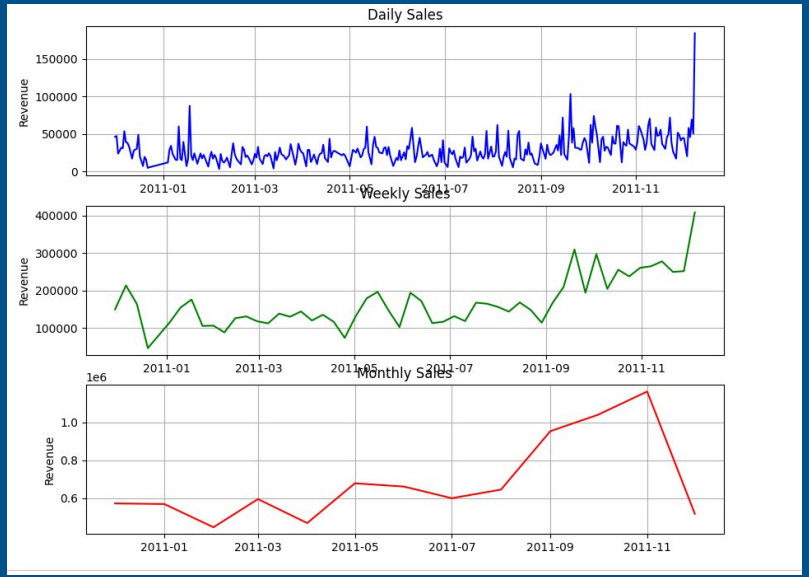
Top 10

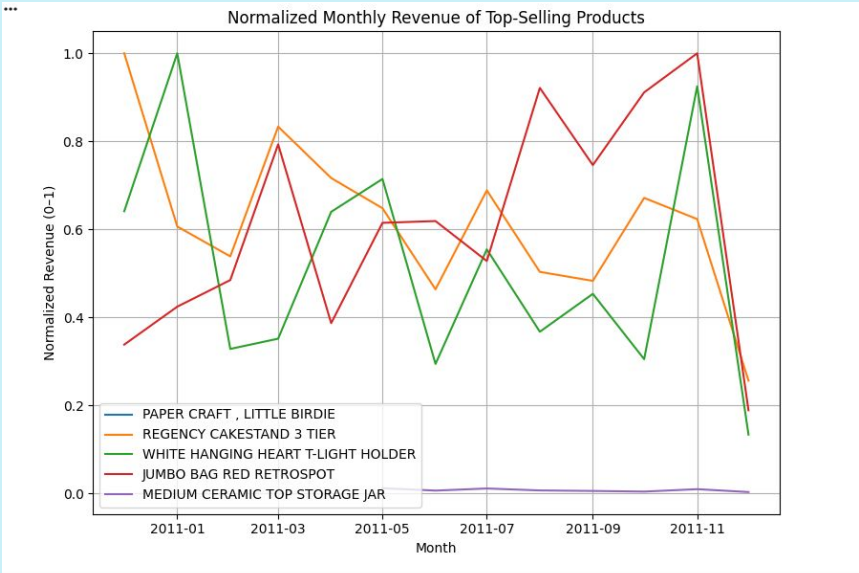




Exploratory Data Analysis

Time Series Analysis

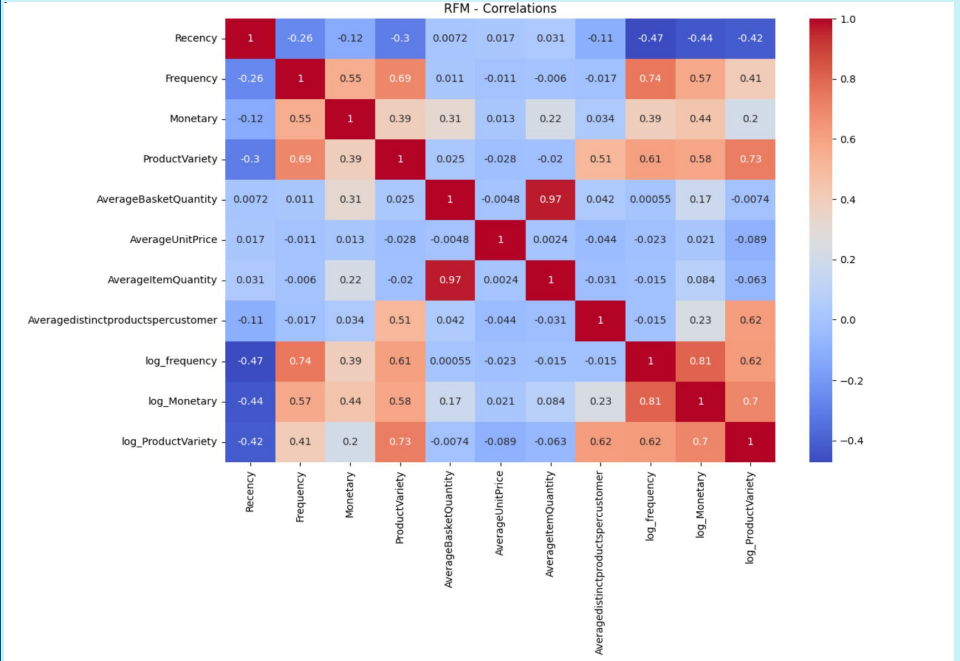




Feature Engineering

RFM

- Recency Frequency and Monetary dataframe creation to prioritize marketing strategies.
- Adding columns to RFM dataframe to improve clustering



Feature Scaling

Preprocessing Stage 2

```
num_features = ['Recency','log_frequency','log_Monetary','log_ProductVariety','AverageBasketQuantity','AverageUnitPrice','AverageItemQuantity','Averagedistinctproductspercustomer']
cat_features = ['Country']

#Preprocessing pipeline

pre = ColumnTransformer(transformers=[("num",StandardScaler(),num_features),
                                      ("cat",OneHotEncoder(handle_unknown="ignore",sparse_output=False),cat_features)],
                                      remainder="drop"  )

#Fit+Transform

X = pre.fit_transform(rfm)

#Build Feature names

num_out = [f"sc_{c}" for c in num_features]
cat_out = list(pre.named_transformers_['cat'].get_feature_names_out(cat_features))
feature_names = num_out + cat_out

#Final dataframe

X_scaled = pd.DataFrame(X,columns=feature_names)
print("X_scaled shape:",X_scaled.shape)
```

```
X_scaled shape: (4338, 45)
```

Unsupervised Learning Techniques

Dimensionality Reduction - PCA and UMAP

```
#Fit PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

#Explained Variance
explained_var =  pca.explained_variance_ratio_.sum()
print('PCA1 and PCA2 total variance:', explained_var)

plt.figure(figsize=(10,5))
plt.plot(np.arange(1, len(pca.explained_variance_ratio_)+1),
         pca.explained_variance_ratio_, marker='o')
plt.title("PCA Explained Variance per Component")
plt.xlabel("Principal Component")
plt.ylabel("Explained Variance Ratio")
plt.grid(True)
plt.show()
```

```
PCA1 and PCA2 total variance: 0.5956586805548374
```

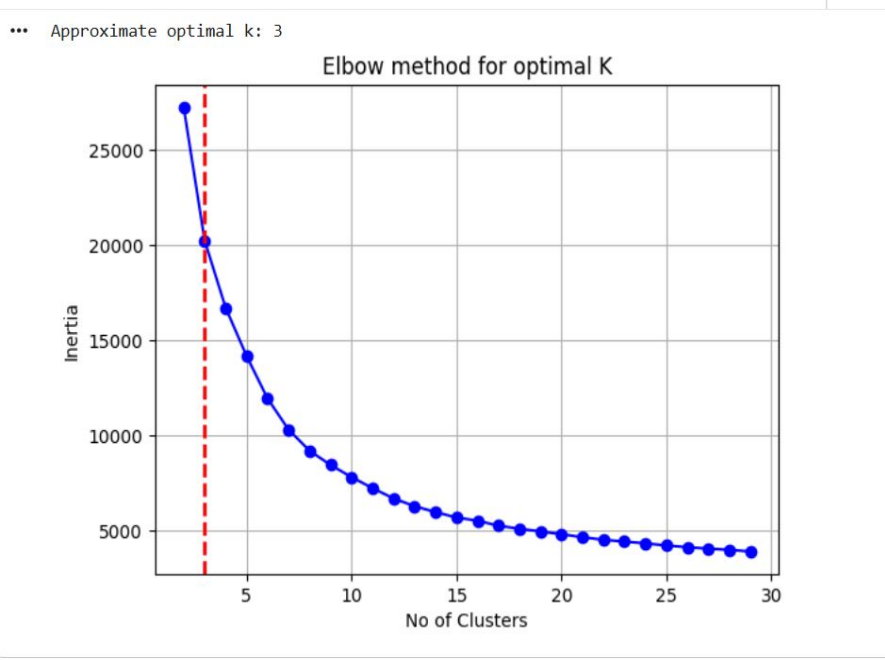
```
#UMAP
import umap

X_umap = umap.UMAP(n_neighbors = 15,min_dist = 0.1,n_components=2,random_state=42)
X_umap = X_umap.fit_transform(X_scaled)

#umap_df = pd.DataFrame(X_umap,columns=["UMAP1","UMAP2"])
```

Unsupervised Learning Techniques

K-Means Clustering Algorithm

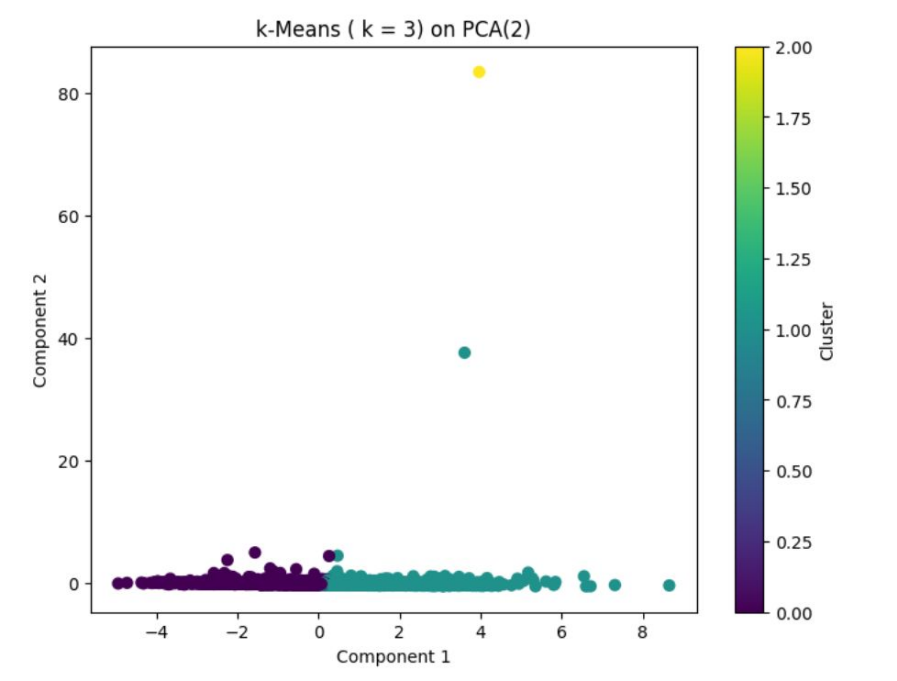
Analysis:

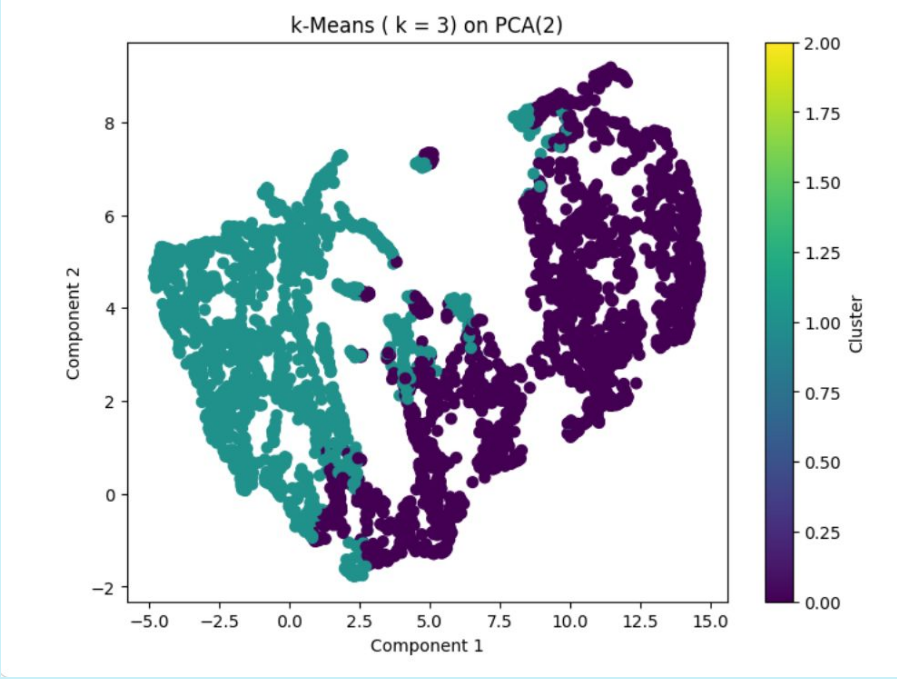
- k value found using:
- Elbow Method (k=3)
- Silhouette Score (for k=3 , score = 0.32142239788983323)
- DB Score (for k=4 , according to the plot above)



Unsupervised Learning Techniques

K-Means Visualization with PCA and UMAP

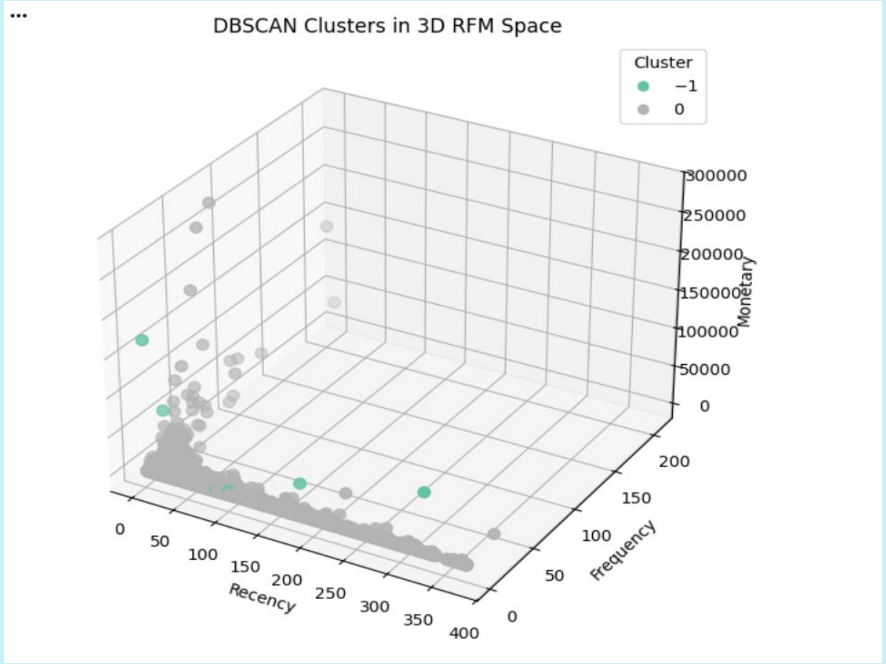




Unsupervised Learning Techniques

DBSCAN

- In DBSCAN Clustering results,outliers are the data points labelled as -1.
- These are the customers whose RFM behavior doesn't fit into any of the dense clusters DBSCAN identified.
- We can see that only outliers and 1 cluster is formed and that is the reason why Silhouette score is very high.
- This means that DBSCAN didn't predict useful customer segments and only a dense group is formed.



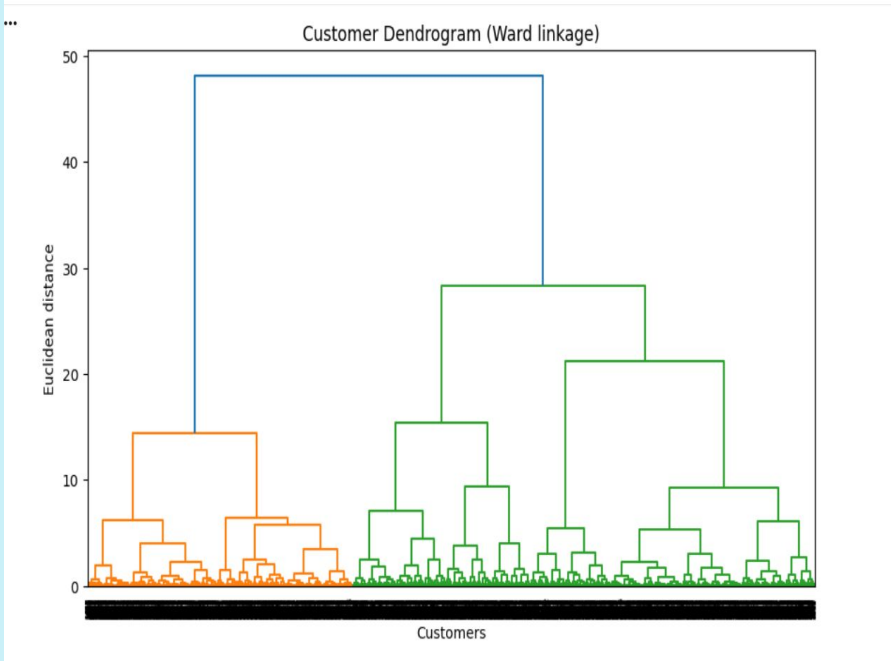
Unsupervised Learning Techniques

Hierarchical Clustering - Agglomerative

Based on the visual selection of this denodram,the most significant groupings are

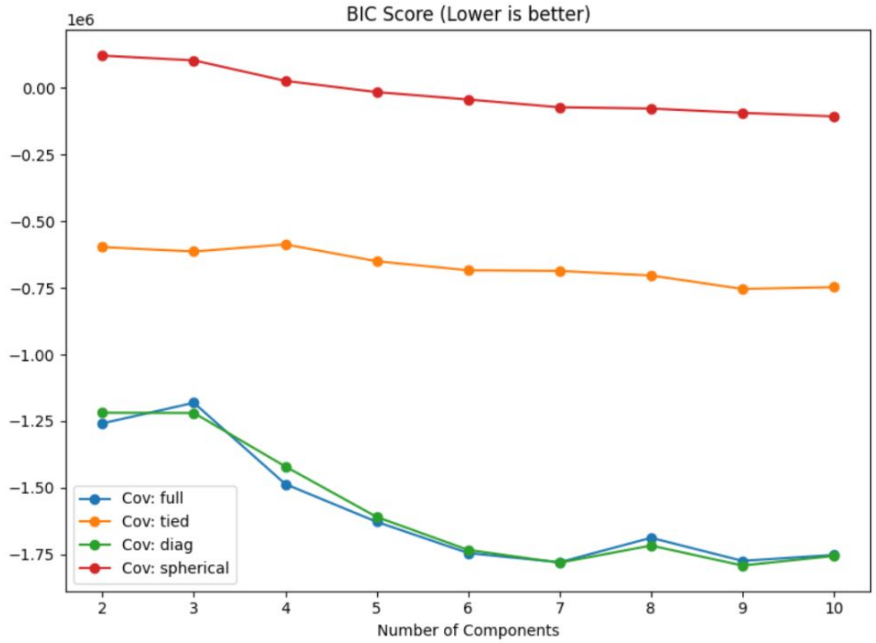
- 3-Cluster Selection
- 4-Cluster Selection
- 2 -Cluster selection

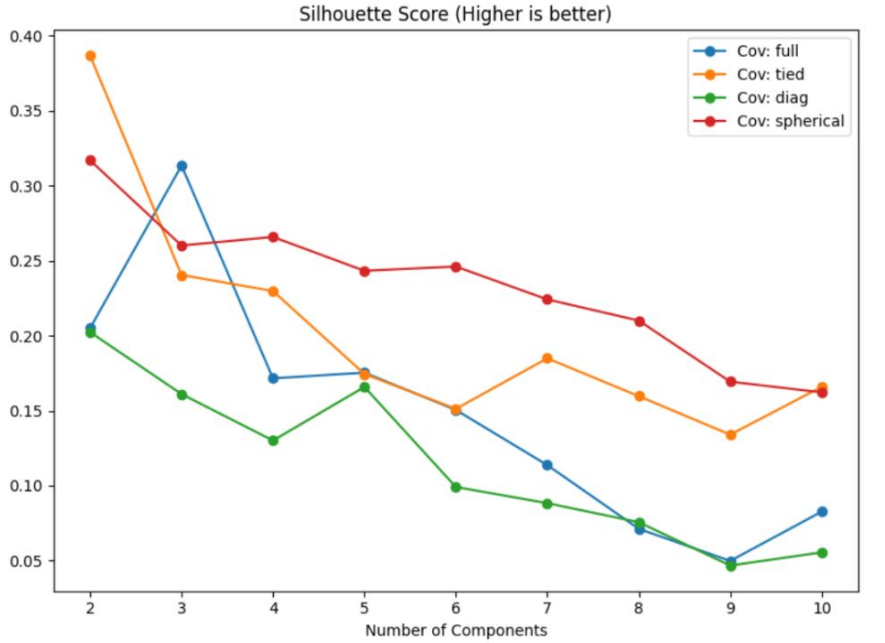
According to Silhouette score, k=2/3 can be chosen.



Unsupervised Learning Techniques

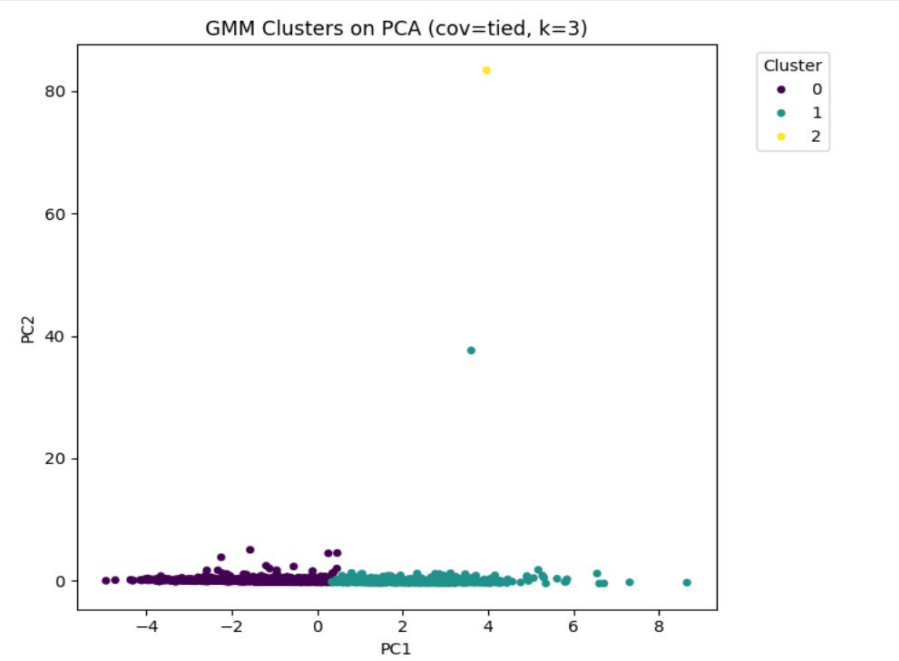
GMM Clustering

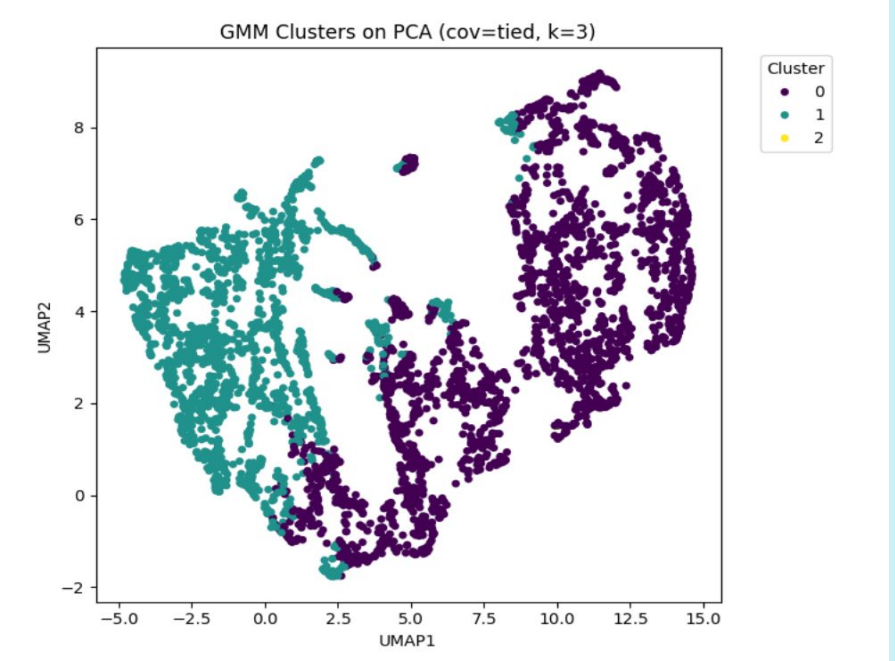




Unsupervised Learning Techniques

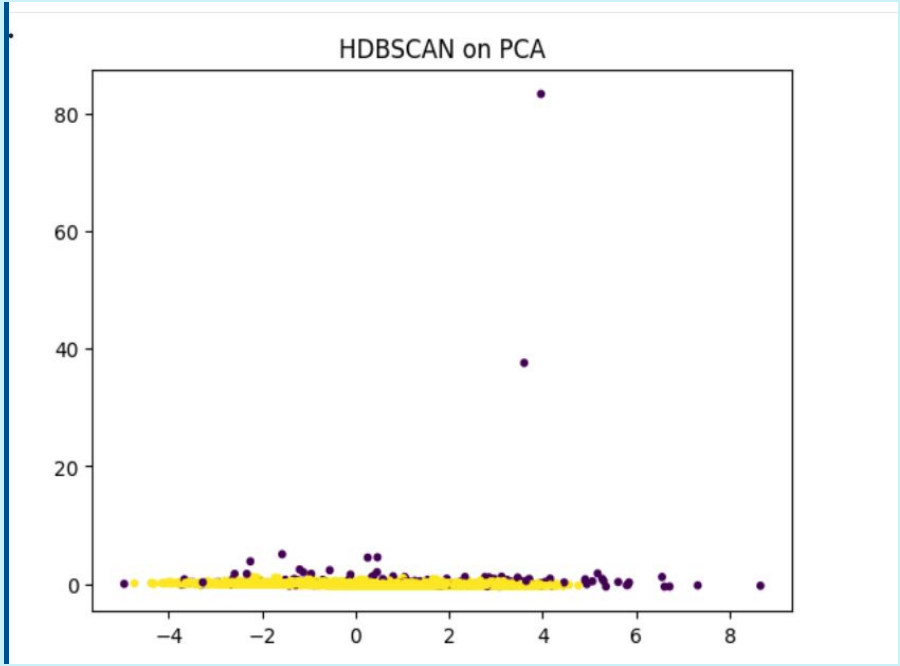
GMM Visualization on PCA and UMAP

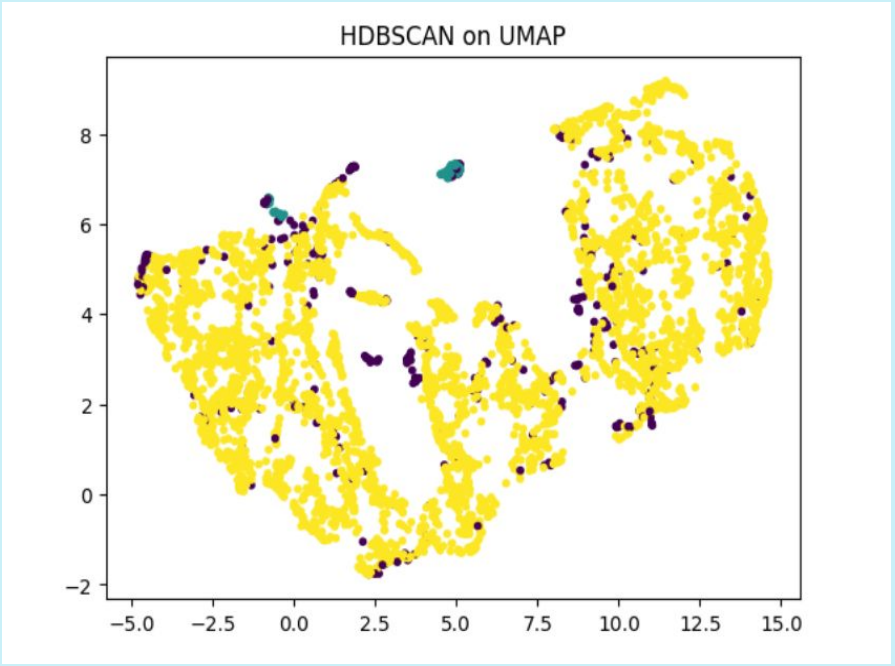




Unsupervised Learning Techniques

HDBSCAN





Unsupervised Learning Techniques

Market Basket Analysis

	antecedent	consequent	support	confidence	lift	leverage
2762	PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.022394	0.821782	24.173440	0.021467
2763	GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.022394	0.658730	24.173440	0.021467
3493	GARDENERS KNEELING PAD KEEP CALM	GARDENERS KNEELING PAD CUP OF TEA	0.020721	0.606635	21.454506	0.019755
3492	GARDENERS KNEELING PAD CUP OF TEA	GARDENERS KNEELING PAD KEEP CALM	0.020721	0.732824	21.454506	0.019755
2089	SPACEBOY LUNCH BOX	DOLLY GIRL LUNCH BOX	0.019858	0.602291	21.020081	0.018913
2088	DOLLY GIRL LUNCH BOX	SPACEBOY LUNCH BOX	0.019858	0.693032	21.020081	0.018913
2203	GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.026765	0.787302	19.986675	0.025425
2202	ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.026765	0.679452	19.986675	0.025425
2764	ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.021368	0.542466	19.906882	0.020295
2765	PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.021368	0.784158	19.906882	0.020295

Conclusion

Analysis

- Cluster 0: At Risk/Inactive Customers
- Cluster 1: Champion/Loyal Customers
- Cluster 2: Outliers

	k_Cluster	Recency	Frequency	Monetary	ProductVariety	\
0	0	138.389474	1.611789	448.742418	20.973895	
1	1	36.912334	7.493884	3959.460276	110.590214	
2	2	326.000000	1.000000	77183.600000	1.000000	

	AverageBasketQuantity	AverageUnitPrice	AverageItemQuantity	Num_Customers
0	178.155635	5.494693	25.696051	2375
1	306.959300	3.226434	30.709090	1962
2	74215.000000	1.040000	74215.000000	1

Conclusion

Analysis

- Cluster 0: Decor Items
- Cluster 1: Seasonal Stuff
- Cluster 2: Outliers

	cluster	Description	Revenue
0	1	PAPER CRAFT , LITTLE BIRDIE	168469.60
1	1	REGENCY CAKESTAND 3 TIER	128440.15
2	1	WHITE HANGING HEART T-LIGHT HOLDER	89810.95
3	2	MEDIUM CERAMIC TOP STORAGE JAR	77183.60
4	1	JUMBO BAG RED RETROSPOT	76474.09
5	1	POSTAGE	67606.61
6	1	PARTY BUNTING	57352.13
7	1	ASSORTED COLOUR BIRD ORNAMENT	49832.73
8	1	RABBIT NIGHT LIGHT	49274.60
9	1	Manual	48829.39
10	0	PICNIC BASKET WICKER 60 PIECES	39619.50
11	1	BLACK RECORD COVER FRAME	36681.86
12	0	REGENCY CAKESTAND 3 TIER	14152.80
13	0	PARTY BUNTING	11492.20
14	0	WHITE HANGING HEART T-LIGHT HOLDER	10637.20
15	0	CHILLI LIGHTS	10200.30
16	0	POSTAGE	10197.35
17	0	JUMBO BAG RED RETROSPOT	8746.69
18	0	PAPER CHAIN KIT 50'S CHRISTMAS	7262.25
19	0	ASSORTED COLOUR BIRD ORNAMENT	6747.61
20	0	Manual	4950.54

Things to Improve

- Feature engineering enhancements can be done by adding time based features.
- Create Pairwise ratios like (Monetary/Frequency) etc.
- Dimensionality reduction techniques like t-SNE can be used.They are good for visualizing small datasets with complex structure.
- NMF can be done.
- Evaluation scores like Davies-Bouldin and Calinski-Harabasz score can be calculated.

References

- https://medium.com/@hasan.unlu/online-retail-clustering-3bbf860e249b
- https://medium.com/@hasan.unlu/online-retail-clustering-3bbf860e249b
- https://towardsdatascience.com/market-basket-analysis-with-pandas-246fb8ee10a5/
- https://www.geeksforgeeks.org/machine-learning/hdbscan/
- https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/
- https://www.geeksforgeeks.org/machine-learning/hierarchical-clustering/

Links

The deliverables for Unsupervised Learning project can be found below:

Video Link :

https://youtu.be/w5QH4oOOW_U

GitHub Repository:

https://github.com/kmadhu181090/MS-AI/tree/main/Pathway%20Machine%20Learning%3A%20Theory%20%26%20Hands-On%20Practice%20with%20Python%20Specialization/CSCA%205632%3A%20Unsupervised%20Algorithms%20in%20Machine%20Learning

Thank You