

Abstract geometric lines in black on a white background, forming various overlapping polygons and shapes, primarily concentrated on the left side of the image.

UNSUPERVISED LEARNING CAPSTONE PROJECT

Marcus Hendricks

CONTENTS

Executive Summary	3
Problem Summary	4
Solution Summary	5
Recommendations for Implementation	8
Appendix	11

EXECUTIVE SUMMARY

- Companies want to get to know their customers better and to increase their profits
- **Customer Segmentation** is the process of dividing a dataset of customers into groups of similar customers based on certain common characteristics
- **Benefits**
 - Optimize Return on Investment
 - Makes marketing more efficient
 - More Customer retention and engagement
 - Brand awareness



CUSTOMER DATASET

A company has given a dataset with 27 Columns and 2240 data entries, with numerical and categorical variables ranging from income to education to various spending habits, and would like the best possible customer segments created using Unsupervised Learning ideas such as Dimensionality Reduction and Clustering

THINGS TO CONSIDER: VARIABLES

- What features/variables have the most correlation?
- What variables have the most variance in the customers?
- What are the most important variables to consider when segmenting customers?

THINGS TO CONSIDER: CLUSTERS

- What is an appropriate number of clusters for segmenting customers?
- What are the cluster profiles, and how are they different?
- How can the cluster profiles be used to target customers for growth and increased revenue?

THINGS TO CONSIDER: CAMPAIGNS

- How can campaigns be customized based on cluster profiles for a higher acceptance rate?

PROBLEM SUMMARY

SOLUTION SUMMARY

Explore/Clean Up the Data

Explore the data, clean up the data to make more sense of variables, feature engineer new variables, and univariate and bivariate analysis (See [Correlation Plot](#))

Use different Unsupervised Learning Methods and Algorithms

Use K-Means, K-Medoids, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models (GMM) to cluster customers into different segments

T-SNE and PCA

Reduce the dimensionality of the data, visualize it in 2 dimensions, and help reduce the multicollinearity between the variables (See [T-SNE and PCA Plots](#))

Compare and decide on best Clustering method

Compare segment customer profiles and Silhouette scores and decide on which method produced the best segmentation

SOLUTION SUMMARY CONT.

K-MEANS

- Best K Value: K=3
- Silhouette Score: 0.2691
- Produced 3 fairly evenly distributed segments
- Created High, Mid, and low-income groups
- *However, there are still outliers in the data and K-Means is affected by outliers, so these insights may not be totally accurate and should not be used*
- See [K-Means PCA Scatter Plot](#)

K-MEDIODS

- Best K Value: K=3
- Silhouette Score: 0.2889
- Produced 3 evenly distributed segments
- Created High, Mid, and low-income groups with more distinct separations and insights in other variables than K-Means
- See [K-Mediods PCA Scatter Plot](#)

HIERARCHICAL CLUSTERING (WARD LINKAGE AND EUCLIDEAN DISTANCE)

- Best K Value: K=3
- Silhouette Score: 0.3149
- Produced 3 less evenly distributed segments
- Created very similar customer groups and K-Mediods, but there were some variables that had the groups not quite as distinct as K-Mediods
- See [HC Clustering PCA Scatter Plot](#)

SOLUTION SUMMARY CONT.

DBSCAN

- Best K Value: K=3
- Silhouette Score: 0.3441
- Highest Silhouette Score
- *But resulted in only 2 clusters and they were not evenly distributed, so this method should not be considered*
- See [DBSCAN PCA Plot](#)

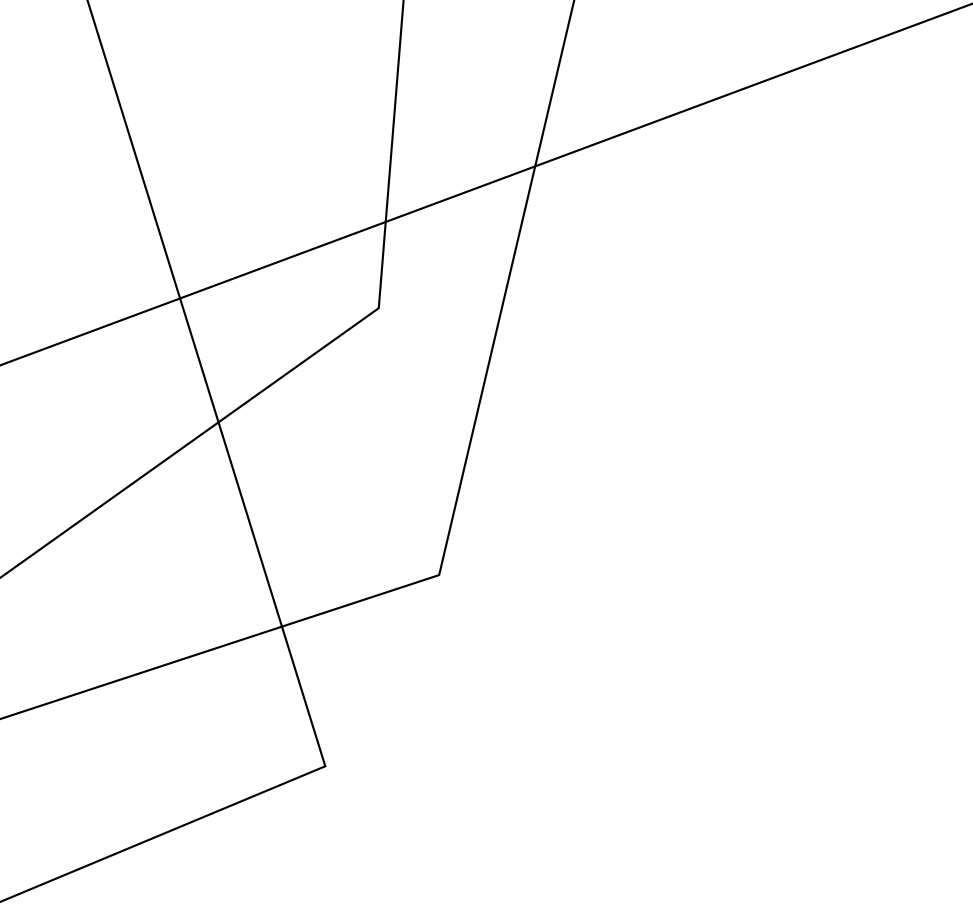
GAUSSIAN MIXTURE MODEL (GMM)

- Best K Value: K=3
- Silhouette Score: 0.1667 (Lowest score)
- Explored K=5 segments (S score = 0.1462)
- Created 5 segments that had some smaller groupings
- Created 2x High, 1x Mid-High, 1x Mid, and 1x Low-income groups
- This gave more insight into different income groups, but was not as clear and separated on many of the other variables
- See [GMM PCA Plot](#)

SUMMARY & DECISION

Algorithms	K Value	Silhouette Score
K-Means	3	0.2691
K-Medoids	3	0.2889
Heirarchial Clustering (Ward Linkage and Euclidean distance)	3	0.3149
DBSCAN	3	0.3441
Gaussian Mixture Model	3	0.1667

- **Selected K-Medoids clustering**
 - Most insightful profiles
 - Kept it simple
 - Higher Silhouette Score
 - Evenly Distributed Groups



Targeted Marketing using the 3 Clusters created by K-Medoids Clustering

- Associate customer IDs and Cluster groups and create the 3 customer segments
- **Cluster 0 (Mid-income group)** consists of mid to high income customers, who are the oldest age group, and typically have family size between 2-3, they have high web purchases and store purchases, spend a lot on wine similarly to cluster 2, they take advantage of the most deal purchases, have the most total purchases, and most have accepted 1 campaign.
 - ❖ This segment could be targeted with web and store deals, most of them are willing to accept campaigns, tend to make lots of smaller purchases, and could be very interested in wine deals.
- **Cluster 1 (Low-income group)** consists of the lowest income group, they are the youngest age group and typically have the largest family size near 3, they have the least number of purchases across the board and typically accept 0 campaigns.
 - ❖ This segment may be harder to target due to their limited discretionary spending. However, they may be more willing to purchase deals or maybe products for their family since they typically have the largest family.
- **Cluster 2 (High-Income group)** consists of the high-income group, who are middle age group in late 40s, who have a smaller family typically of 2 with 0 kids or teens at home, spend the most on wine and other products, they make lots of purchases via the catalog and store, they have the most expenses, highest amount per purchase, and accept the most campaigns on average around 1.
 - ❖ This group can be easily targeted by sending them catalogs for any products, especially wine, or have promotions in the store. They are also the most willing group to accept campaigns, so make sure to include them with those.

RECOMMENDATIONS FOR IMPLEMENTATION



RECOMMENDATIONS FOR IMPLEMENTATION

Benefits

- Increased Return on Investment for advertising/marketing
- More tailored experience for the different customer groups
 - More willing to spend more
 - Improve customer retention
- More brand awareness

Risks

- Customers misclassified or belong to multiple segments
 - Causing customers to feel misaligned or alienated with the company
- Customer segments could be too narrow or too wide

Further Analysis

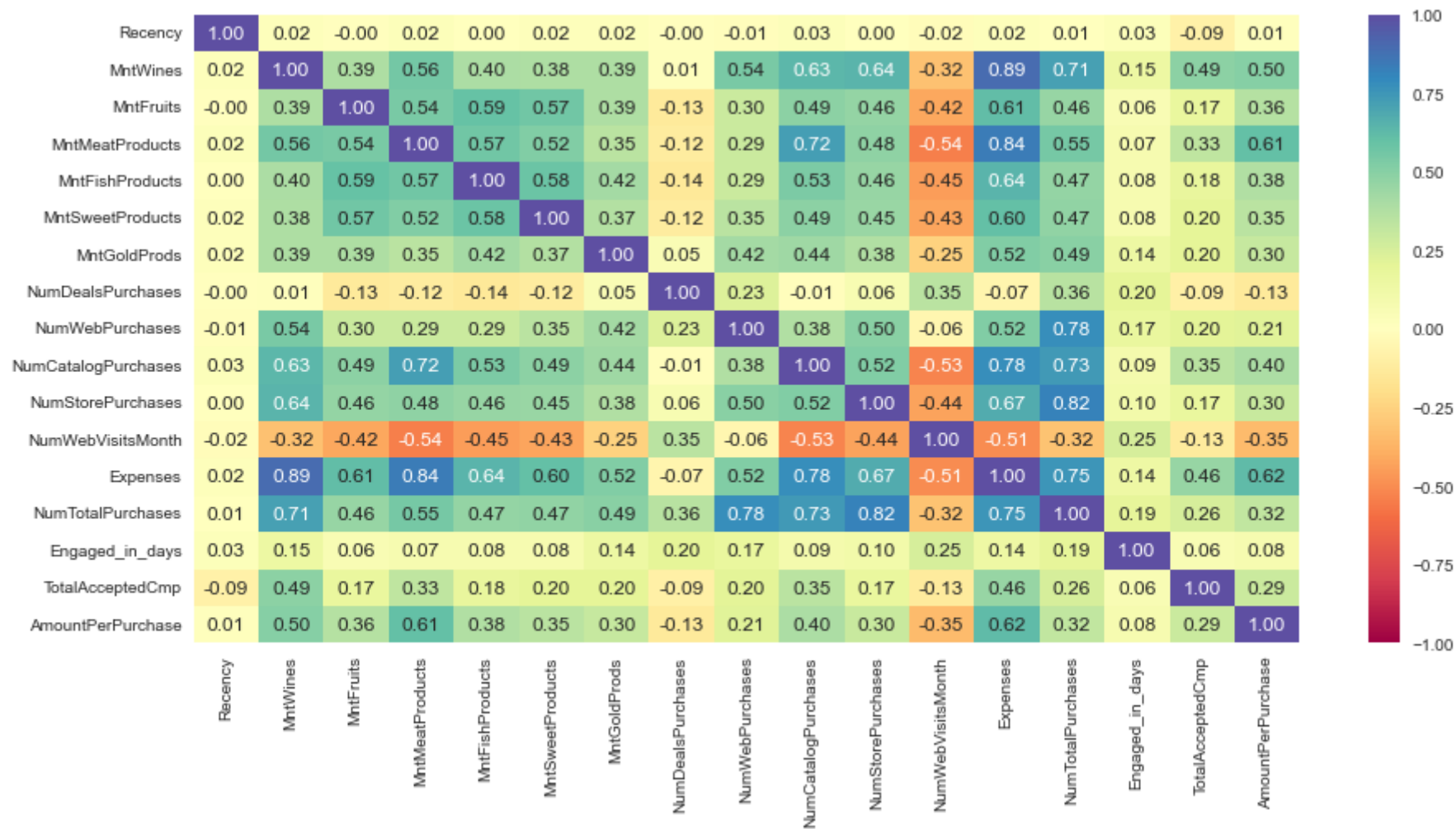
- Get more data to make segments more clearly defined
- Clearly Define segment parameters so that you add new customers to their respective segment quickly

A series of white, thin, overlapping geometric lines and polygons on a black background, located on the left side of the slide.

QUESTIONS?

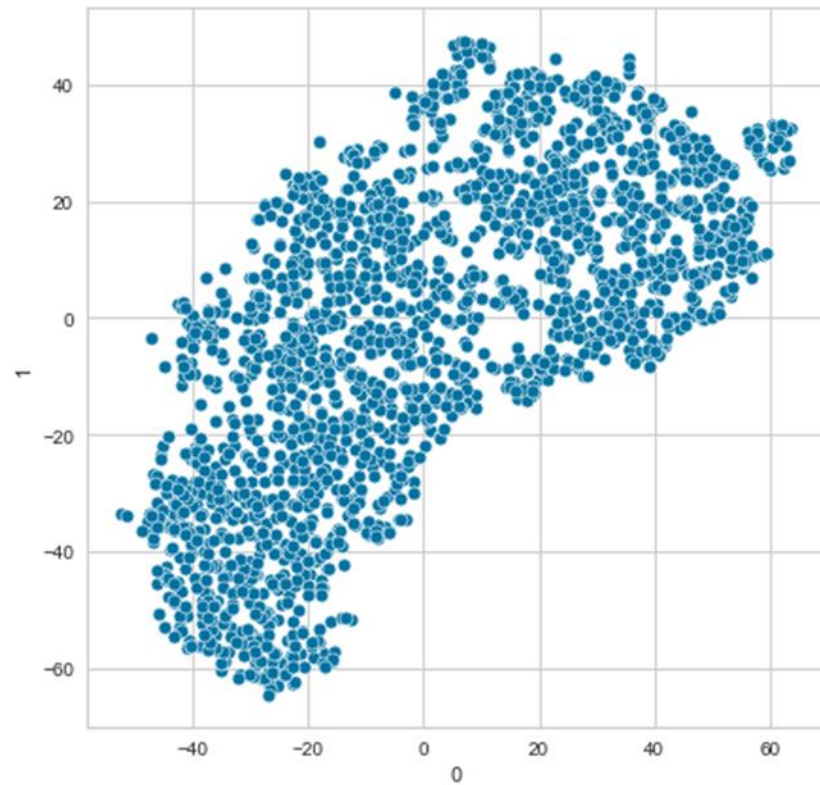
APPENDIX

CORRELATION PLOT

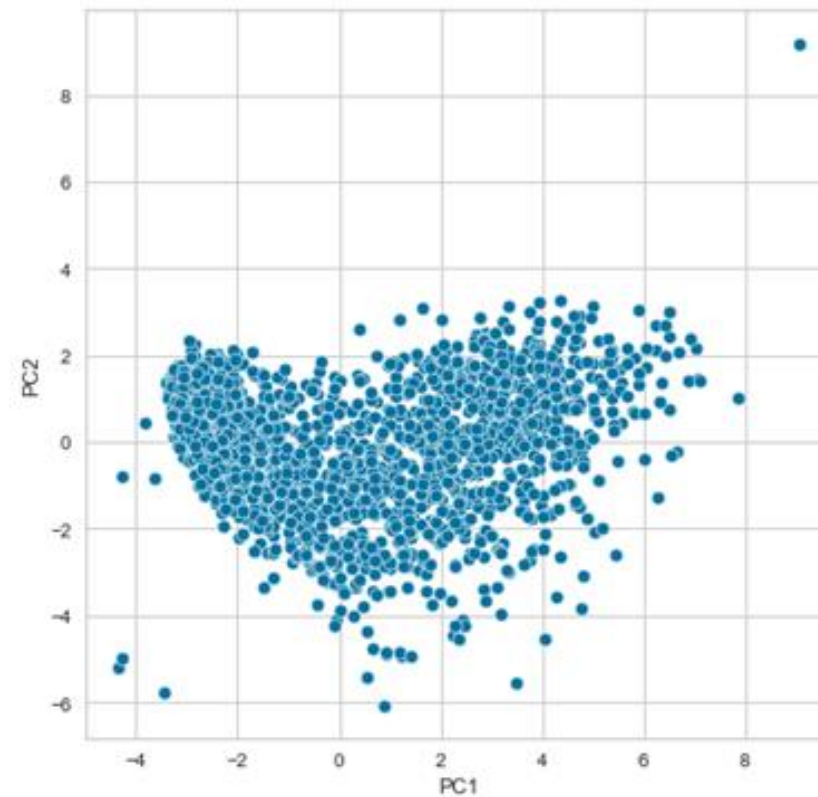


APPENDIX

T-SNE PLOT

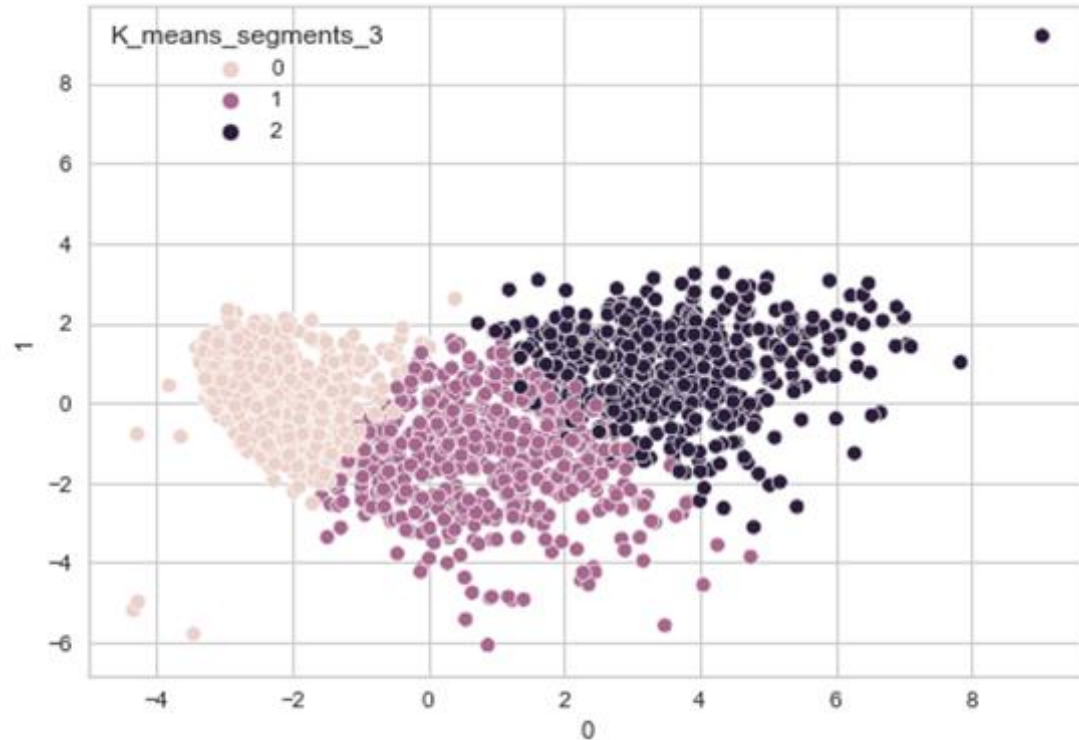


PCA PLOT



APPENDIX – K-MEANS

K-MEANS K=3 PCA PLOT

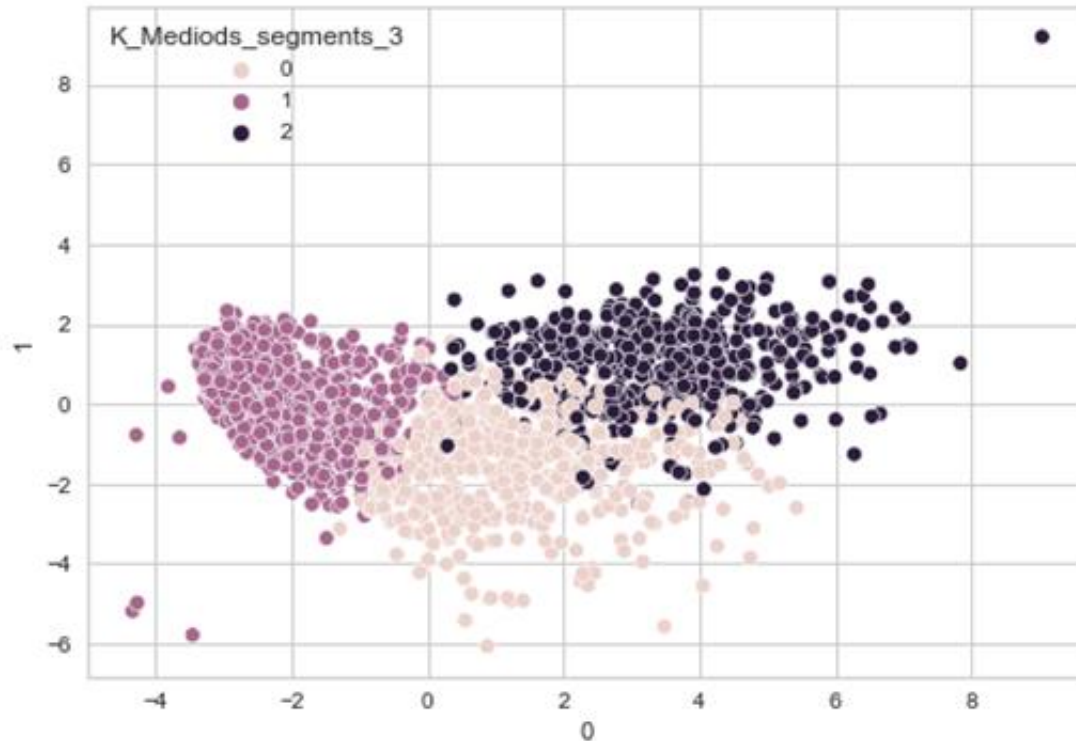


```
1 data.K_means_segments_3.value_counts()

0    1058
1     606
2     566
Name: K_means_segments_3, dtype: int64
```

APPENDIX – K-MEDIODS

K-MEDIODS K=3 PCA PLOT



```
1 data.K_Mediods_segments_3.value_counts()
```

```
1    1152
```

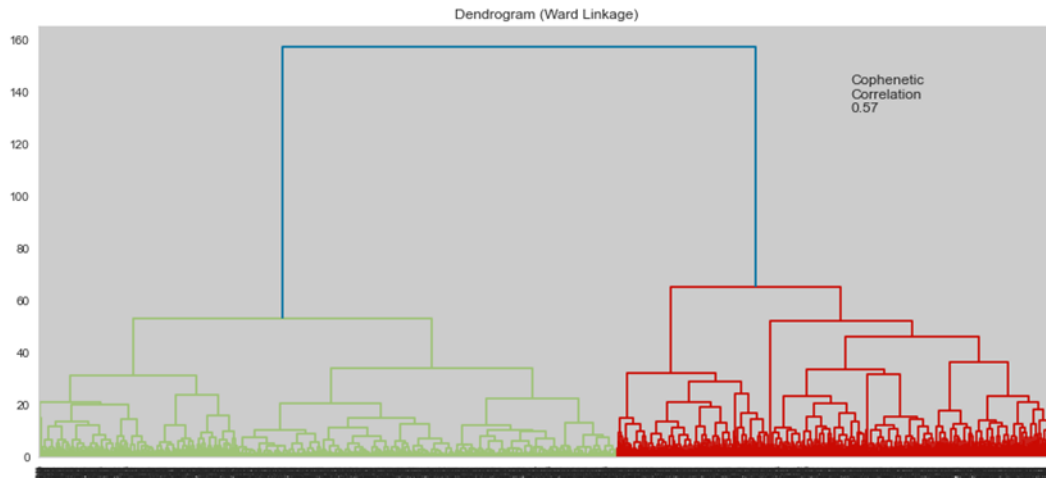
```
0     540
```

```
2     538
```

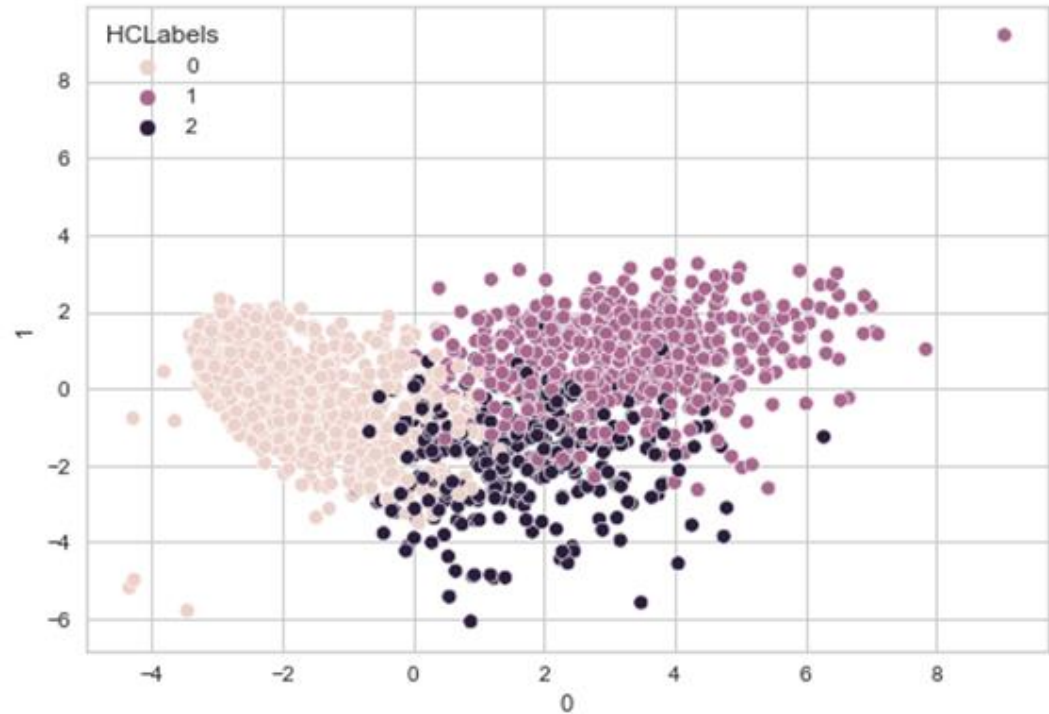
```
Name: K_Mediods_segments_3, dtype: int64
```

APPENDIX – HC CLUSTERING

EUCLIDEAN DISTANCE AND WARD LINKAGE DENDOGRAM



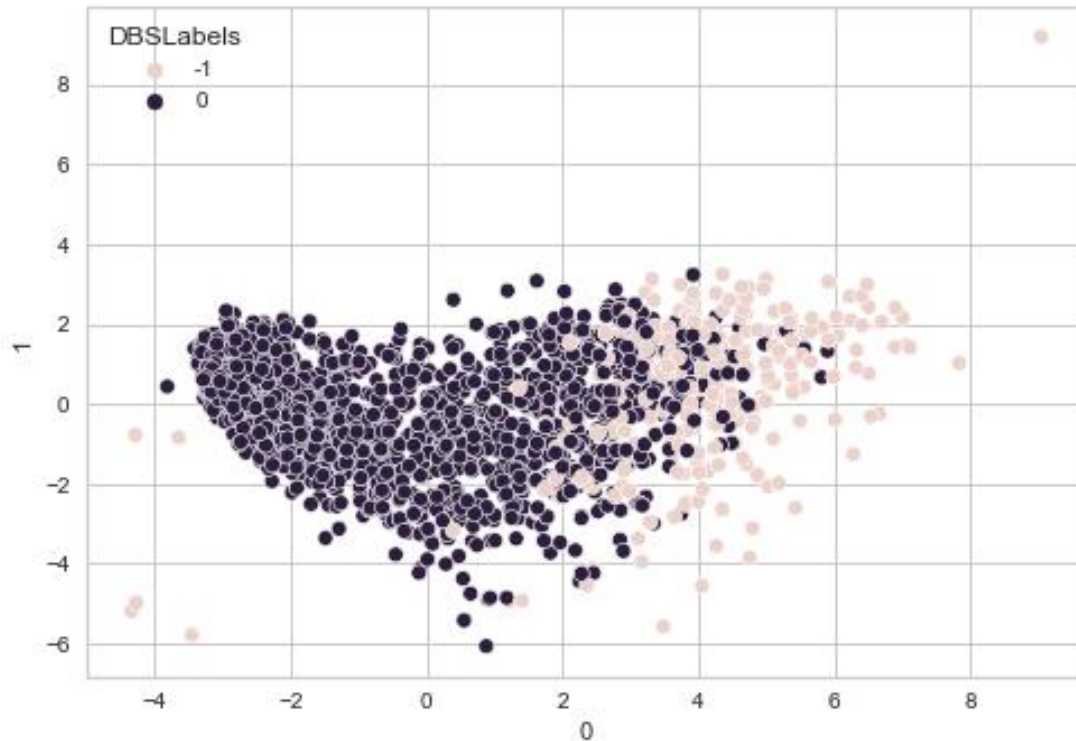
HIERARCHICAL CLUSTERING K=3 PCA PLOT



```
1 data.HCLabels.value_counts()
0    1271
1     624
2     335
Name: HCLabels, dtype: int64
```

APPENDIX – DBSCAN

DBSCAN K=3 PCA PLOT

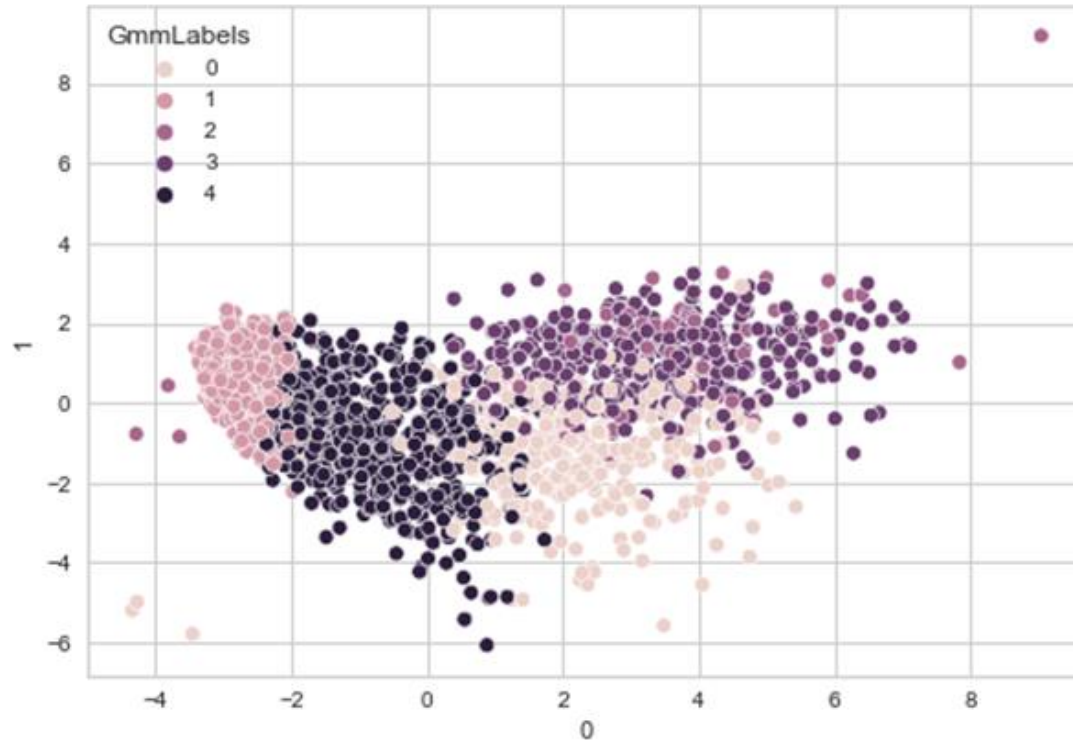


```
data["DBSLabels"].value_counts()
```

```
0      1911  
-1      319  
Name: DBSLabels, dtype: int64
```


APPENDIX – GMM

GAUSSIAN MIXTURE MODEL (GMM) K=3 PCA PLOT



	data.GmmLabels.value_counts()
1	756
4	574
3	420
5	363
2	59
0	58

Name: GmmLabels, dtype: int64