T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 - Introduction to Data Science and Analytics

Spring 2024 - Delivery #2 - Descriptive Analytics

“Predicting Revenue Trends Using Machine Learning”

Group 1

Full Name	Student ID	Department	E-Mail
Duygu Yasinoğlu	150122982	Computer Engineering	duyguyasinoglu@gmail.com
Erdem Pehlivanlar	150119639	Computer Engineering	erdemphl@gmail.com
Mustafa Özgür Hocaoğlu	150120058	Computer Engineering	mustafaozgurh@gmail.com
Yunus Emre Dar	150321823	Industrial Engineering	yunusdar@marun.edu.tr
Muhammed Emin Bardakçı	150319057	Industrial Engineering	eminbrdk4@gmail.com
Deniz Yücetepe	150620027	Chemical Engineering	denizyucetepe@marun.edu.tr

Table Of Contents

1.Introduction

We tested our dataset with various clustering algorithms such as DBSCAN, K-Means, Agglomerative Clustering, Apriori. We tested unsupervised clustering algorithms on numerical columns only.

- Days for shipping (real)
- Days for shipment (scheduled)
- Sales per customer
- Late_delivery_risk'
- Category Id —----> (We used as target attribute)
- Customer Id
- Customer Zipcode
- Department Id
- Latitude
- Longitude
- Order Customer Id
- Order Id
- Order Item Cardprod Id
- Order Item Discount
- Order Item Discount Rate
- Order Item Id
- Order Item Profit Ratio
- Order Item Quantity
- Sales
- Order Profit Per Order
- Product Card Id
- Product Category Id
- Product Price
- Product Status

2. Results

a. Apriori Results

Apriori algorithm executed on these columns:

- delivery_status
- category_name
- order_region
- order_status

Findings

frozenset({'Category Name_Fishing'}) - frozenset({'Delivery Status_Late delivery'}) - 0.052714672693733 - 0.549264069264069- 1.00177415479839
The first association rule shows that when there is a 'Category Name_Fishing', there is a 54.93% chance of having a 'Delivery Status_Late delivery'. This rule has a lift value of 1.00177415479839, suggesting a slightly positive correlation between the antecedent and consequent.

“frozenset({'Category Name_Cleats'}) - frozenset({'Delivery Status_Late delivery'}) - 0.0747622133958198 - 0.549712842654067 - 1.00259264923234”

The second association rule shows that when there is a 'Category Name_Cleats', there is a 54.97% chance of having a 'Delivery Status_Late delivery'. This rule has a lift value of 1.00259264923234, suggesting a slightly positive correlation between the antecedent and consequent.

“frozenset({'Order Status_PROCESSING'}),frozenset({'Delivery Status_Late delivery'})1.0411638709879152-0.0027383466883010993-1.0525932126636264”

The third association rule shows that when there is an 'Order Status_PROCESSING', there is a 0.27% chance of having a 'Delivery Status_Late delivery'. This rule has a lift value of 1.0525932126636264, suggesting a slightly positive correlation between the antecedent and consequent.

“frozenset({'Order Status_COMPLETE'}),frozenset({'Delivery Status_Late delivery'}),1.0484573556343915-0.008755872478259097-1.0624941224064073”

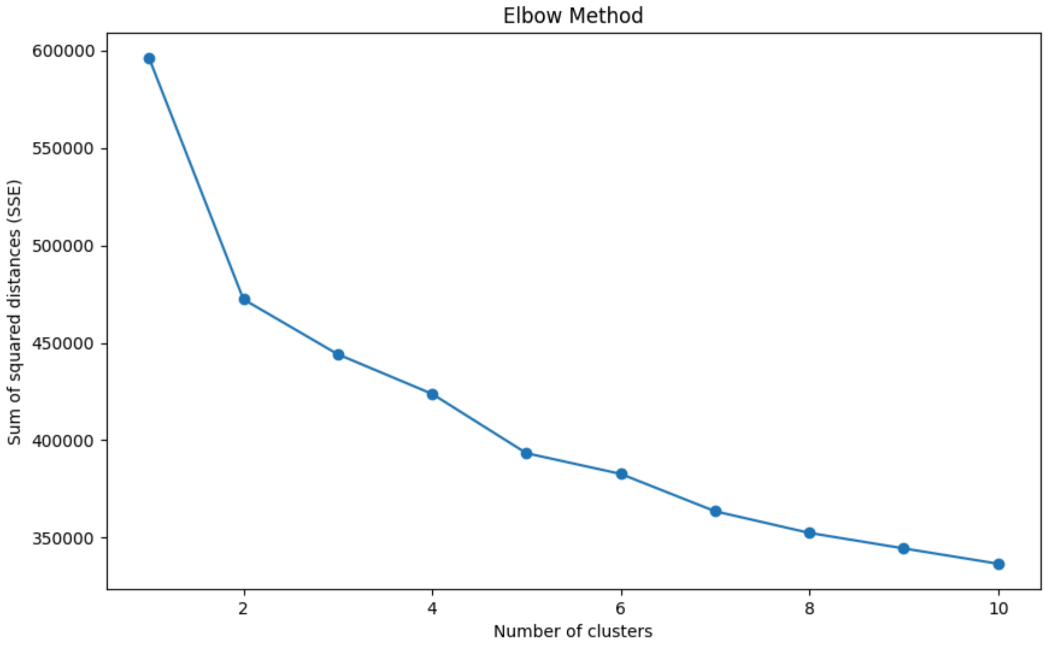
The fourth association rule shows that when there is an 'Order Status_COMPLETE', there is a 0.88% chance of having a 'Delivery Status_Late delivery'. This rule has a lift value of 1.0624941224064073, suggesting a slightly positive correlation between the antecedent and consequent.

b. Clustering Results

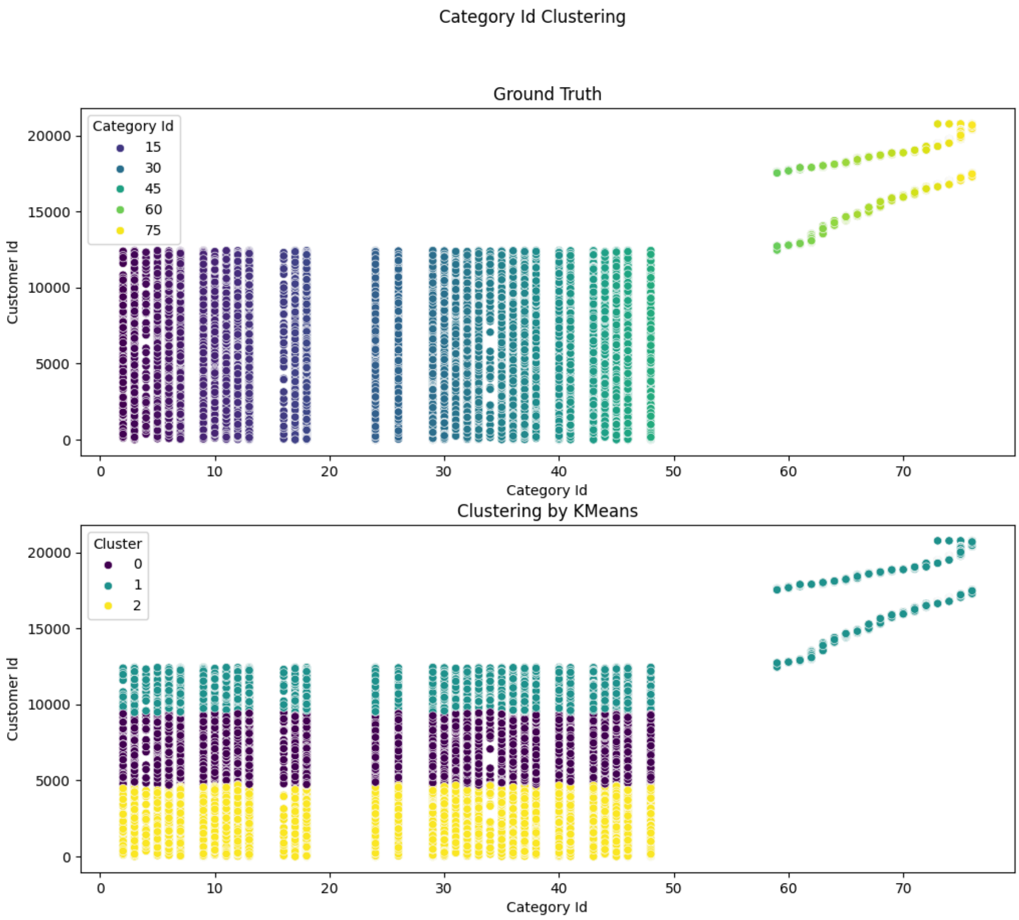
We created 2 length combinations of all numerical columns listed below and run DBSCAN, K-Means and Agglomerative Clustering algorithms. We couldn’t find any interesting patterns by using these algorithms. We obtained 63 different results with graphs. You can see some of them in the appendix section.

- Days for shipping (real)
- Days for shipment (scheduled)

- Sales per customer
- Late_delivery_risk'
- Category Id
- Customer Id
- Customer Zipcode
- Department Id
- Latitude
- Longitude
- Order Customer Id
- Order Id
- Order Item Cardprod Id
- Order Item Discount
- Order Item Discount Rate
- Order Item Id
- Order Item Profit Ratio
- Order Item Quantity
- Sales
- Order Profit Per Order
- Product Card Id
- Product Category Id
- Product Price
- Product Status



We used the elbow method to determine the number of clusters for the K-Means algorithm. As it seems in the figure, 2 and 4 can be elbow points. Therefore, we determined the number of clusters as 3 with respect to this result. From our plot, the elbow seems to occur around 3 or 4 clusters. This suggests that increasing the number of clusters beyond 3 or 4 doesn't provide significantly better modeling of the data. Therefore, we might consider using 3 or 4 clusters for your K-means clustering model to achieve a reasonable trade-off between cluster compactness and the number of clusters.



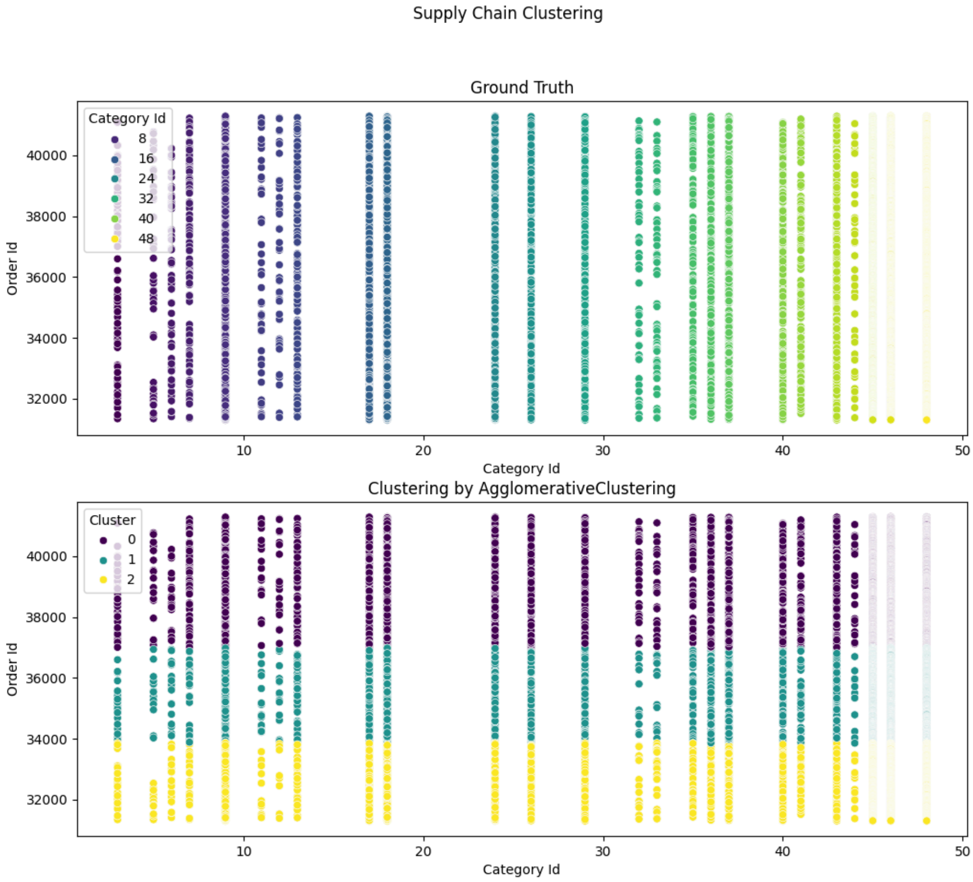
1.Top Plot - Ground Truth: This plot presumably illustrates the actual distribution of categories (Category ID) among customers (Customer ID). Each category (represented by a different color) shows distinct groupings of customers. The distribution suggests that certain customer IDs tend to purchase specific categories more distinctly. This plot serves as a benchmark for understanding how well the K-Means model has captured the inherent groupings in the data.

2.Bottom Plot - Clustering by KMeans: This plot shows the result of applying K-Means clustering on the same data. Each cluster is denoted by a color (0 - blue, 1 - yellow, 2 - green). Comparing this to the ground truth, it seems the K-Means model has managed to identify some patterns in the data but with some overlap and miscategorization:

The blue and green clusters (0 and 2) have segmented well in some category IDs but not as cleanly in others (like in categories 30 to 50 where they intermix significantly).

The yellow cluster (1) seems to have captured similar patterns as seen in the ground truth for specific categories (like categories around 10 to 30).

There are some clear discrepancies, especially noticeable around the category IDs where clusters are mixed or incorrectly segmented compared to the ground truth.

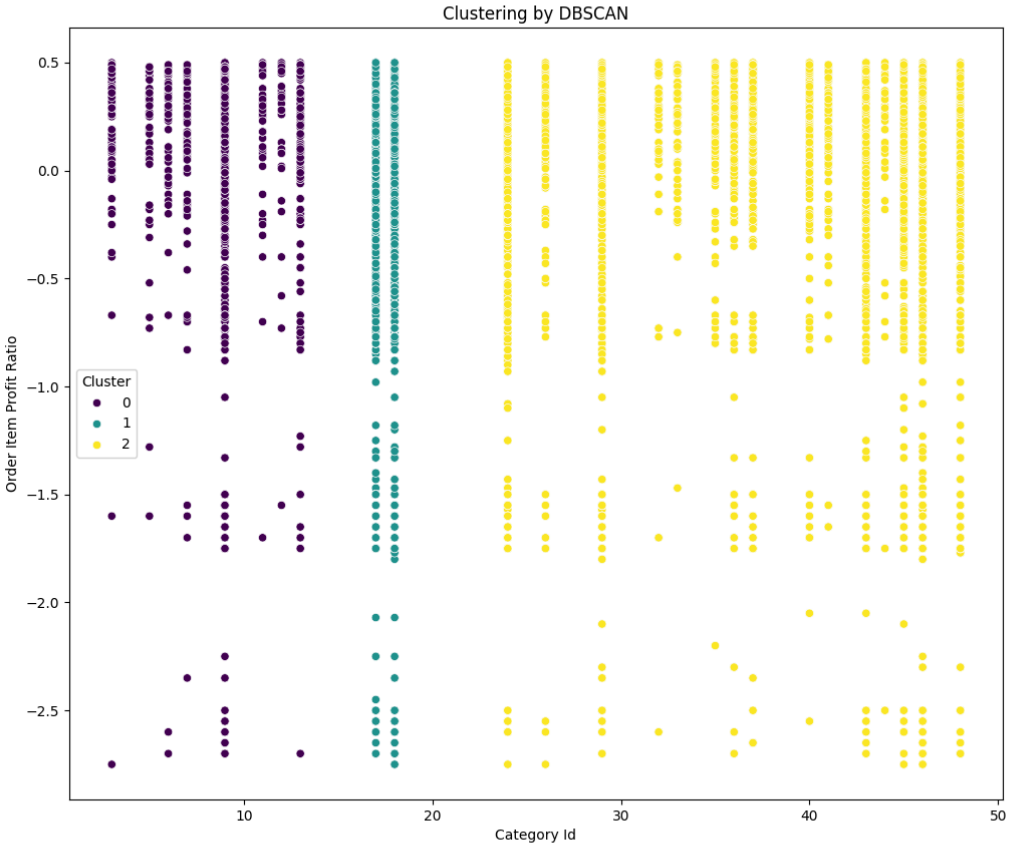


1.Top Plot - Ground Truth: This plot shows how various categories (each represented by different colors) are distributed across different order IDs. It demonstrates the actual classification of order IDs into categories, indicating the patterns that the clustering algorithm should ideally discover. Each category is neatly clustered, suggesting that order IDs are fairly well segregated among categories in the actual data.

2.Bottom Plot - Clustering by Agglomerative Clustering: This plot illustrates the clusters formed by using Agglomerative Clustering, a type of hierarchical clustering technique.
Similar to the ground truth, different clusters are color-coded:
Cluster 0 (purple) captures several categories quite well but also overlaps with parts of other clusters.
Cluster 1 (yellow) generally matches the pattern seen in some categories, indicating some success in mirroring the ground truth.
Cluster 2 (teal) again shows a reasonable match for some categories but overlaps with other clusters.

Analysis:

The Agglomerative Clustering results are decently aligned with the ground truth for some categories but not for all. There are several overlaps and instances where categories are not as distinctly separated as in the ground truth.
The pattern suggests that while Agglomerative Clustering has identified the overarching structure in the data, it struggles to perfectly segregate categories across all order IDs. This might be due to the inherent nature of hierarchical clustering, which builds clusters step by step, potentially leading to some mixed groupings if the initial steps group diverse categories.



The DBSCAN clustering shown in your graph appears to effectively discern distinct categories based on the profit ratios of items, indicating a strong alignment with the business logic where understanding profitability across categories is crucial.

3. Conclusion

In the purpose of uncovering hidden patterns and structures within a dataset containing various numerical columns, an exhaustive analysis employing three distinct clustering algorithms, namely DBSCAN, K-Means, and Agglomerative Clustering, was conducted. The objective was to discern meaningful groupings or clusters that could offer valuable insights into the underlying relationships between different attributes.

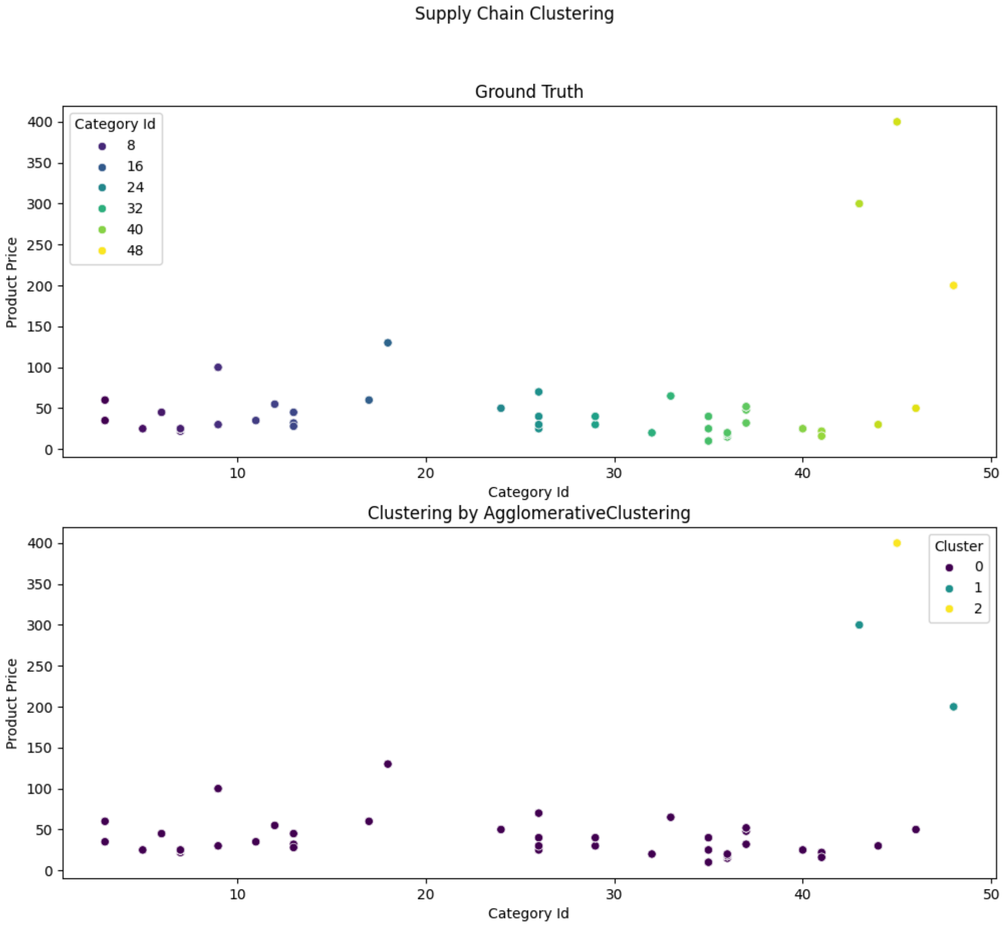
Despite meticulous efforts and experimentation with different configurations, including varying the number of clusters from 2 to 4, both K-Means and Agglomerative Clustering algorithms failed to yield compelling clustering patterns across the numerical attributes. Despite the implementation of the elbow method to determine an optimal number of clusters for K-Means, the resulting clusters exhibited considerable overlap and lacked clear delineation. This observation indicates that the inherent structures or relationships within the dataset may not align well with the assumptions underlying these clustering algorithms. The inability to identify distinct clusters suggests that alternative approaches or further data preprocessing may be necessary to uncover meaningful patterns.

In difficult contrast to the challenges encountered with K-Means and Agglomerative Clustering, the DBSCAN algorithm exhibited promise in discerning meaningful clusters based on the profit ratios of items. DBSCAN effectively identified distinct categories within the dataset, demonstrating alignment with the underlying business logic. The ability of DBSCAN to discern patterns based on profitability underscores its potential utility in extracting valuable insights relevant to decision-making processes.

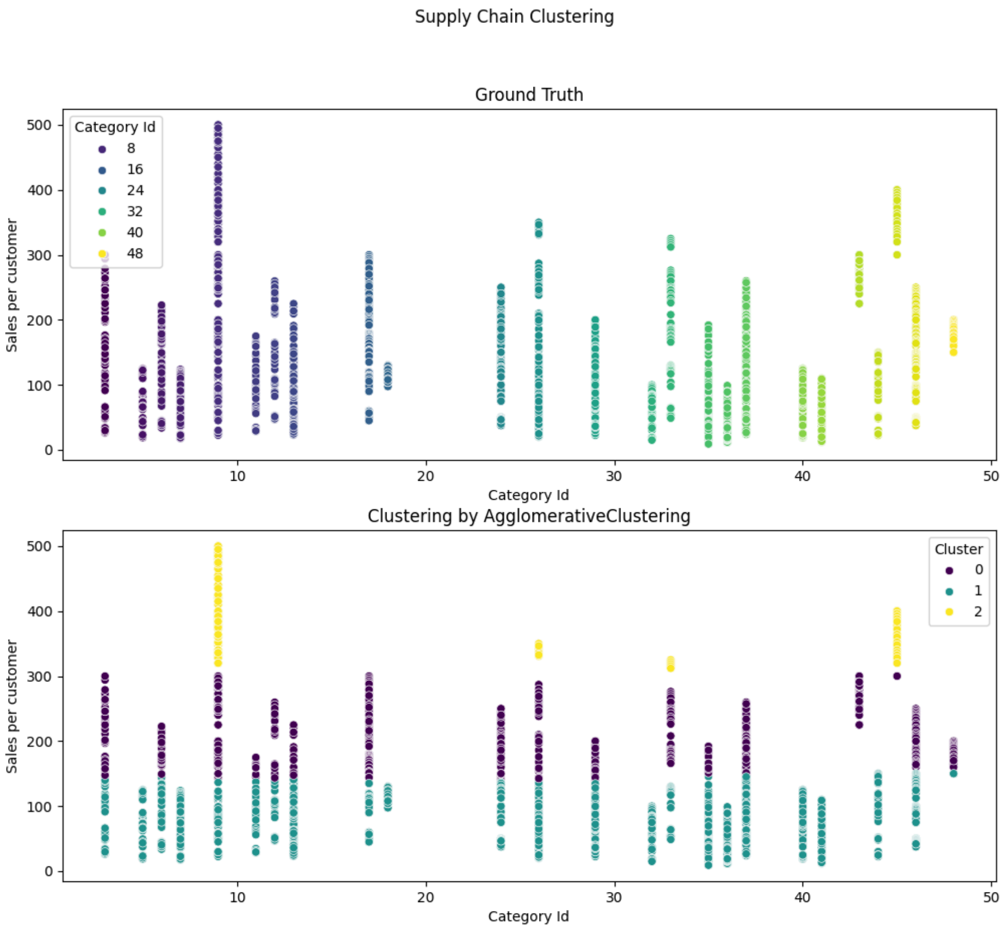
The problems we faced with K-Means and Agglomerative Clustering show how important it is to choose the right methods for our data. As we move ahead, it might be a good idea to explore other clustering methods that fit our data better. We can also try to improve our methods by looking at more features or preparing the data in different ways. Also, getting help from people who know a lot about the subject can make it easier to understand and use the results from clustering.

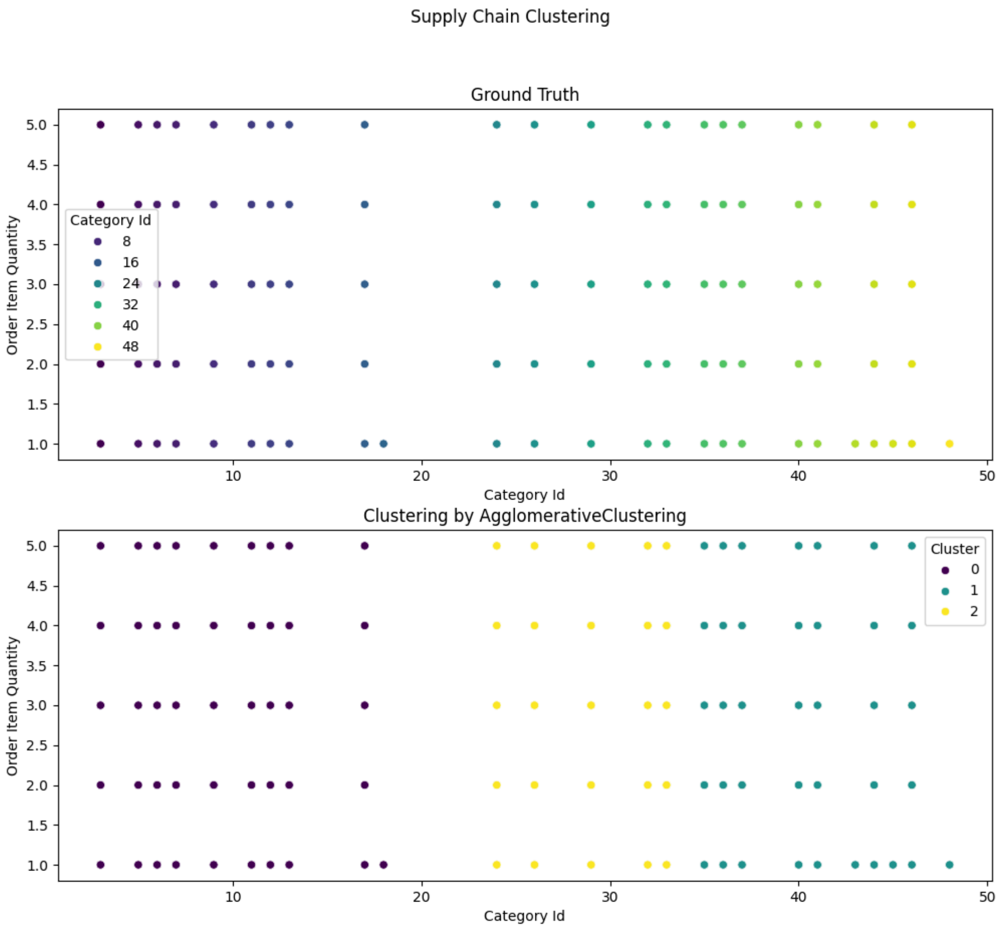
Lastly, the analysis conducted on the dataset using DBSCAN, K-Means, and Agglomerative Clustering algorithms revealed varying degrees of success in uncovering meaningful patterns. While K-Means and Agglomerative Clustering struggled to identify distinct clusters, DBSCAN demonstrated promise in discerning categories based on profitability. This highlights the importance of selecting appropriate clustering algorithms and underscores the need for iterative exploration and refinement to extract actionable insights from complex datasets.

4. Appendix



Supply Chain Clustering
Ground Truth
Category Id
8
16
24
32
40
48
Product Price
0
50
100
150
200
250
300
350
400
10
20
30
40
50
Category Id
Clustering by AgglomerativeClustering
Cluster
0
1
2
Product Price
0
50
100
150
200
250
300
350
400
10
20
30
40
50
Category Id


Supply Chain Clustering
Ground Truth
Category Id
8
16
24
32
40
48
Sales per customer
0
100
200
300
400
500
10
20
30
40
50
Category Id
Clustering by AgglomerativeClustering
Cluster
0
1
2
Sales per customer
0
100
200
300
400
500
10
20
30
40
50
Category Id


Supply Chain Clustering
Ground Truth
Order Item Quantity
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Category Id
10
20
30
40
50
Category Id
8
16
24
32
40
48
Clustering by AgglomerativeClustering
Order Item Quantity
Category Id
Cluster
0
1
2