Springboard–-DSC
Capstone Project 2

Optimizing Customer Segmentation and Marketing Strategies Through Advanced Clustering Techniques

By Ryan Mennemeier
July 2024

Introduction:

Understanding and anticipating customer purchasing behavior is critical for improving user experience and optimizing inventory management in e-commerce platforms. By leveraging customer segmentation we will provide actionable insights to enhance marketing strategies. The objective is to build predictive models capable of identifying distinct customer segments based on several key features of the data as well as user purchase frequency. These models will enable Instacart to tailor their marketing strategies, and be able to identify with their customer base better going forward.

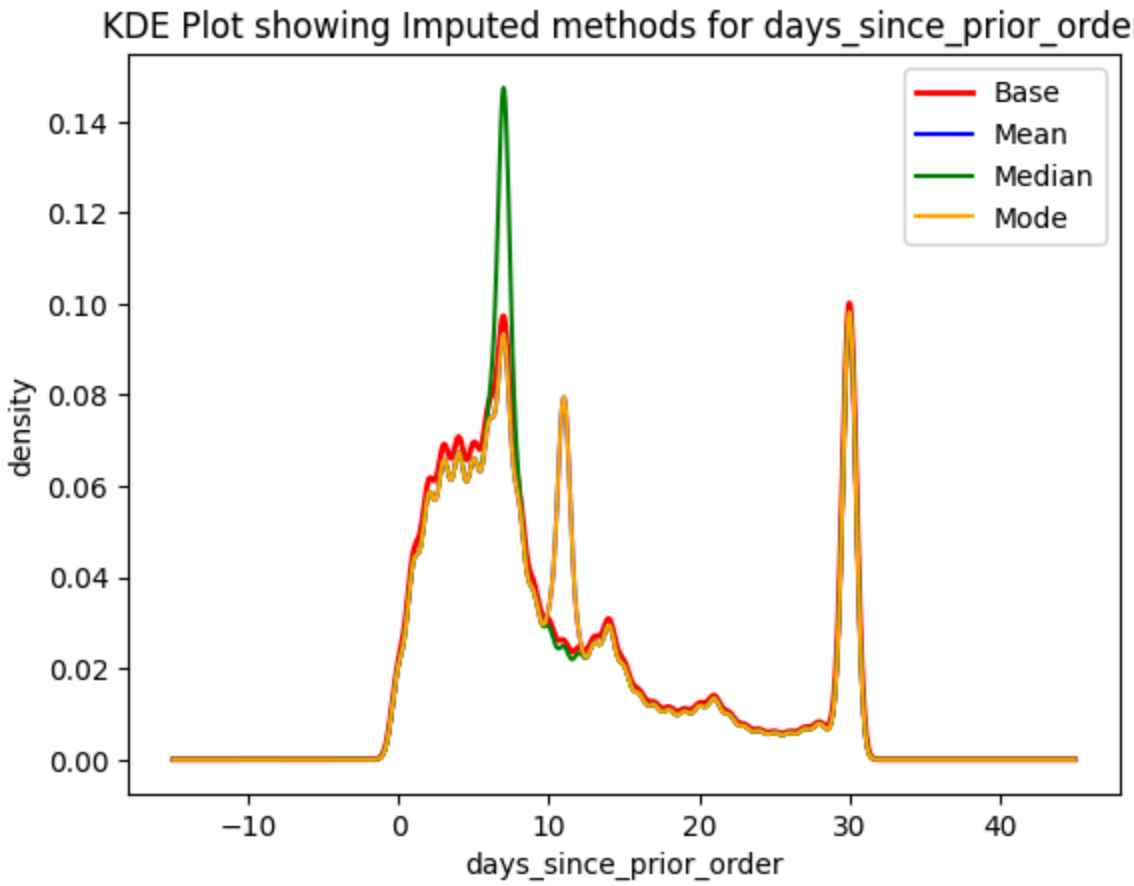
Implementation details can of course be found within the notebooks contained in the project itself:

Link to Project Repository

Approach:

[Data Acquisition and Wrangling]

For this section I began by importing each data link and making sure it came up in my notebook. I then used some basic pandas strategies to examine each dataframe, just doing some exploring into what this data was all about. I was looking for missing values, what the columns contained, and the shape and size of the datasets. I did discover some missing values in one of the datasets, and rather than delete them I used Imputation to fill in the mean value for all the missing data points.

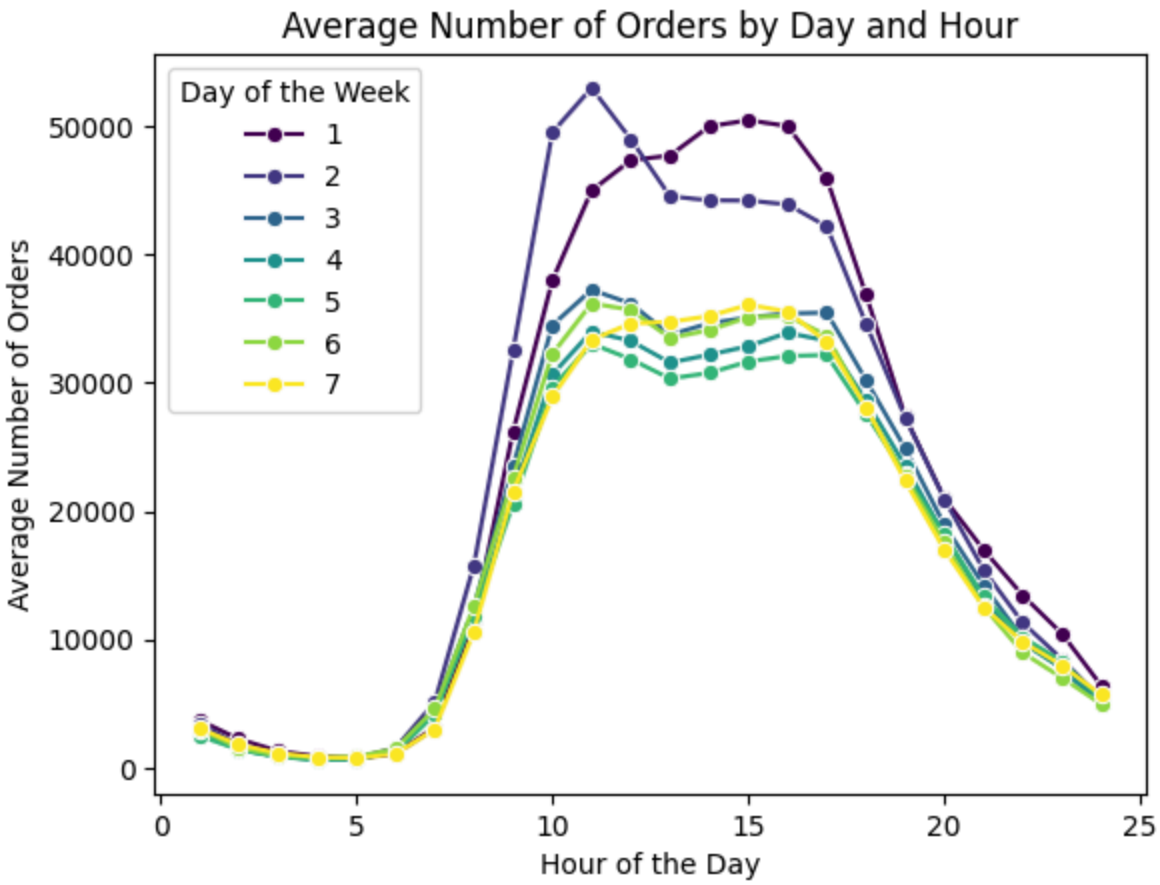


It appears that the Mean strategy (Blue) is the closest to the Base (and original) data, as you can't even see it.

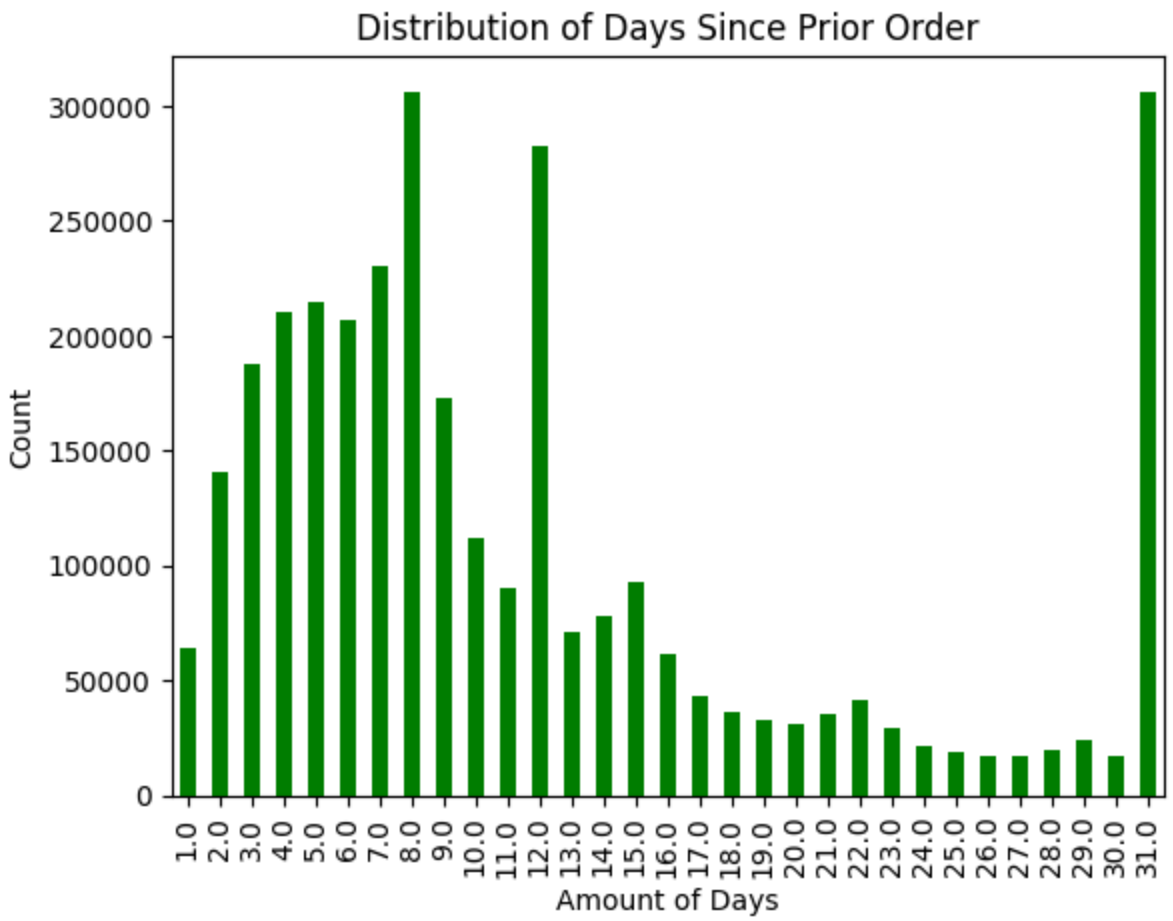
With that process done I merged all of the datasets together into one large one. While not completely necessary it made going from notebook to notebook a lot easier with saving progress as well as being able to recall the columns that I needed as they were in one place.

[Storytelling and Inferential Statistics]

-Upon initially plotting the data, patterns developed that showed customer behavior even before they were segmented.



The above shows that on the first couple days of the week there is a significantly higher rate of purchasing than the rest of the week, as well as the time of the day being during general working hours (between 9AM - 5PM).



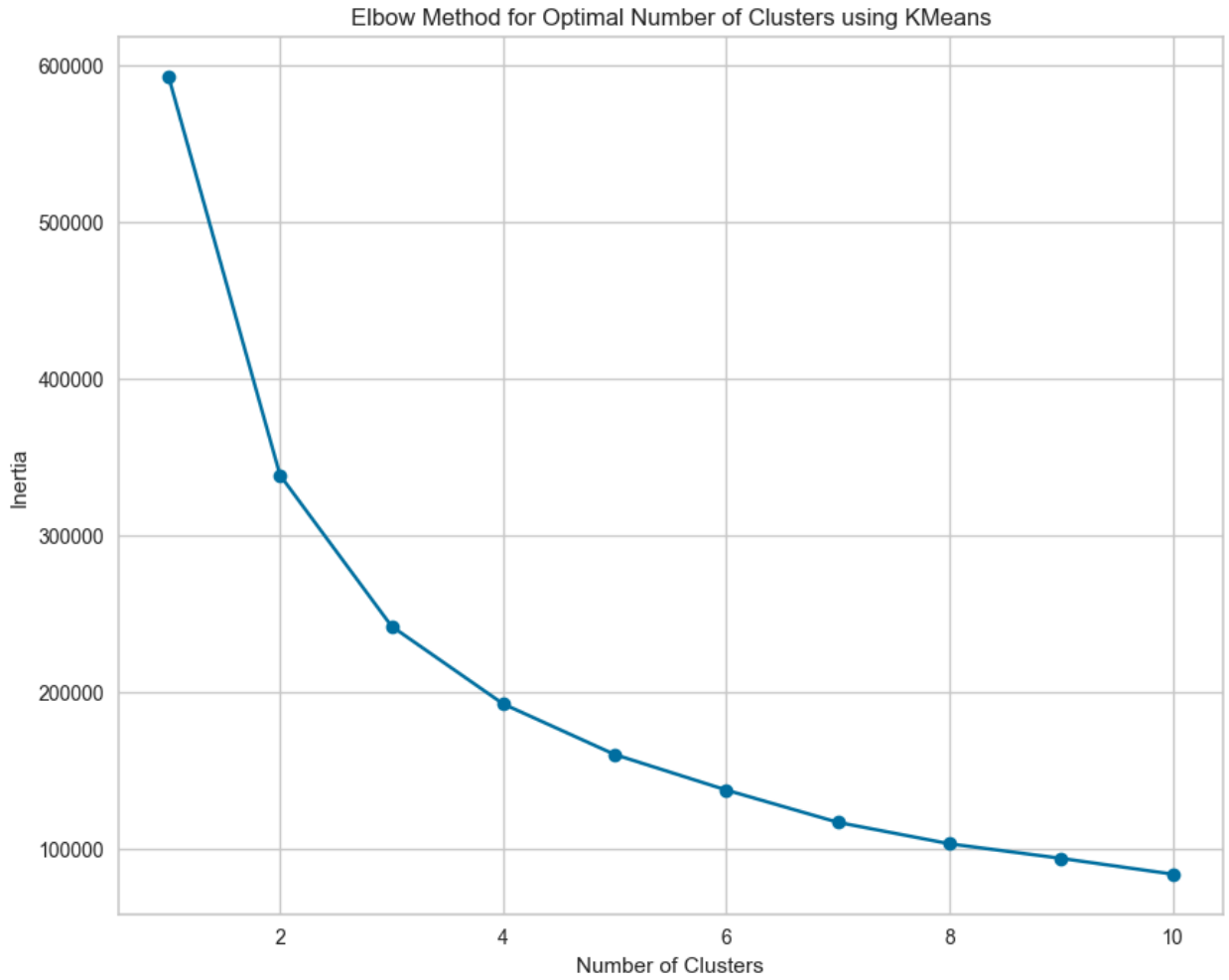
I also viewed the Days Since Prior Order, seeing the majority of the customer base ordering again right away up until the first 10 days.

Between these visuals (and others) I had a basis for some of the key features that would be included in the clustering models.

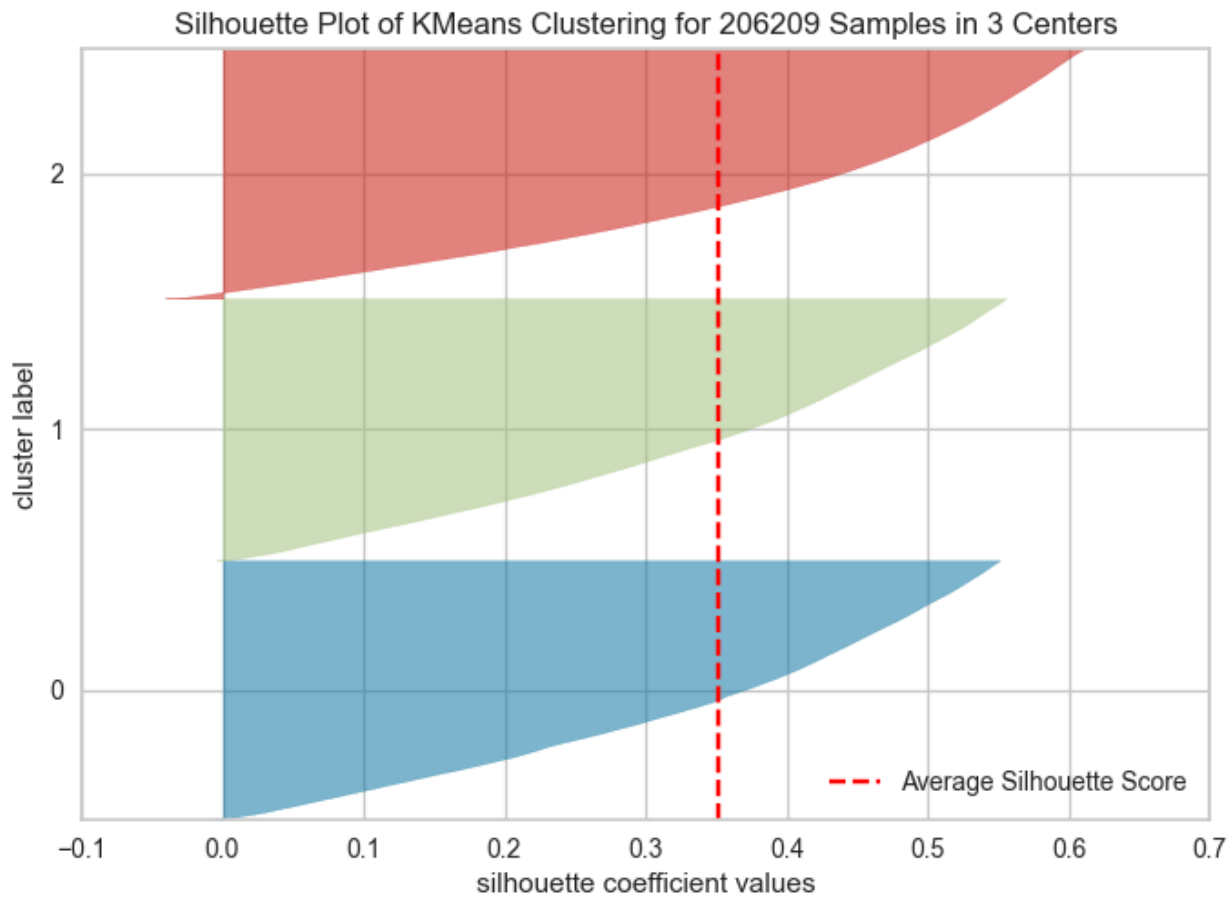
[Baseline Modeling]

As a baseline model I chose KMeans Clustering first, as it is straightforward and I wanted to get results from that model first before going on to some of the more nuanced versions of clustering.

The KMeans clustering model produced a k = 3 result, as you can see below:

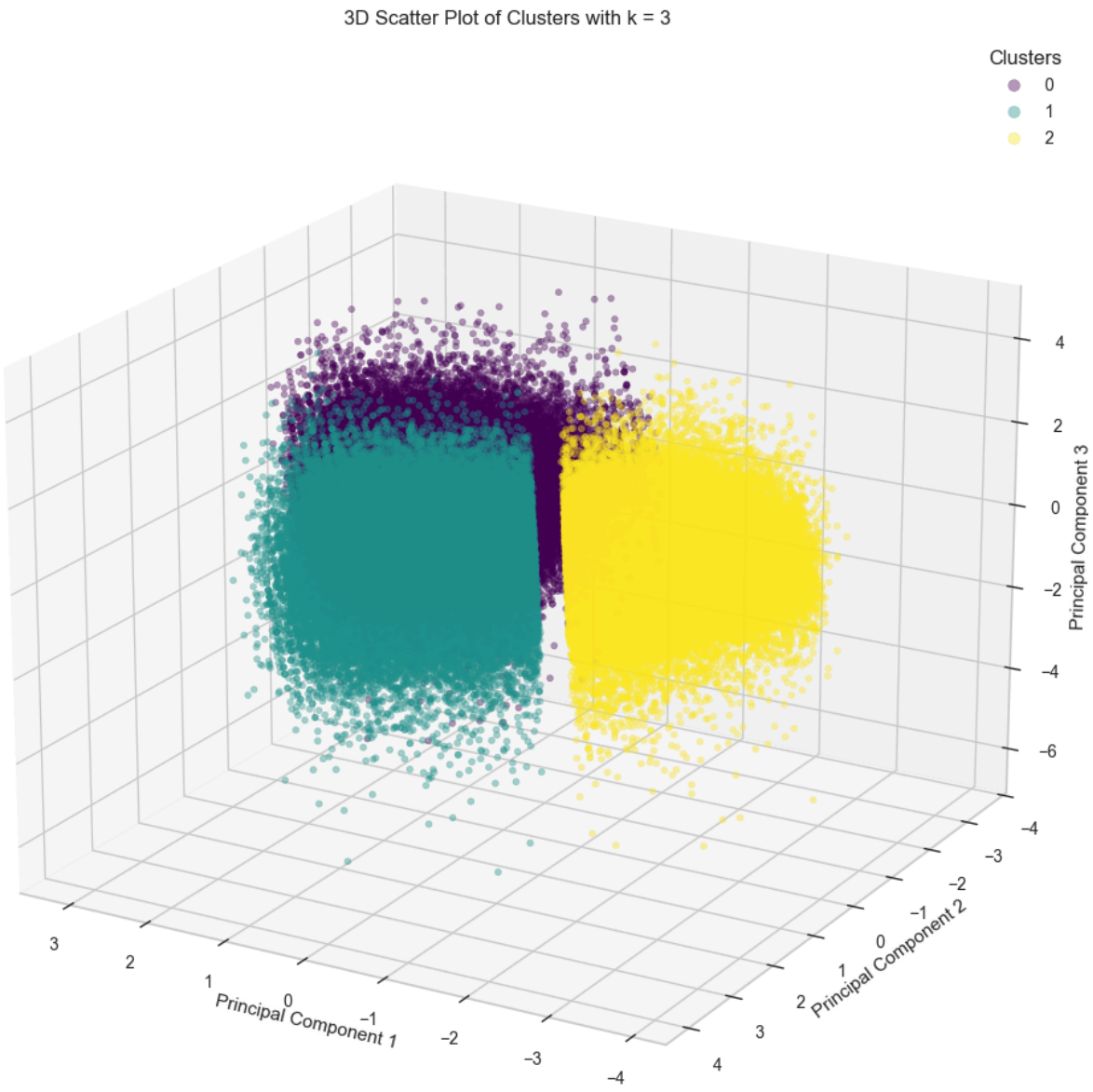


Looks like at least from this elbow plot the best k value is k = 3 due to how the plot starts to 'flatten' or more specifically when the Inertia decreases more slowly at that point. I still observed a Silhouette plot for this data just to confirm this k value so we can revise our Clustering models and more efficiently extract our customer clusters.

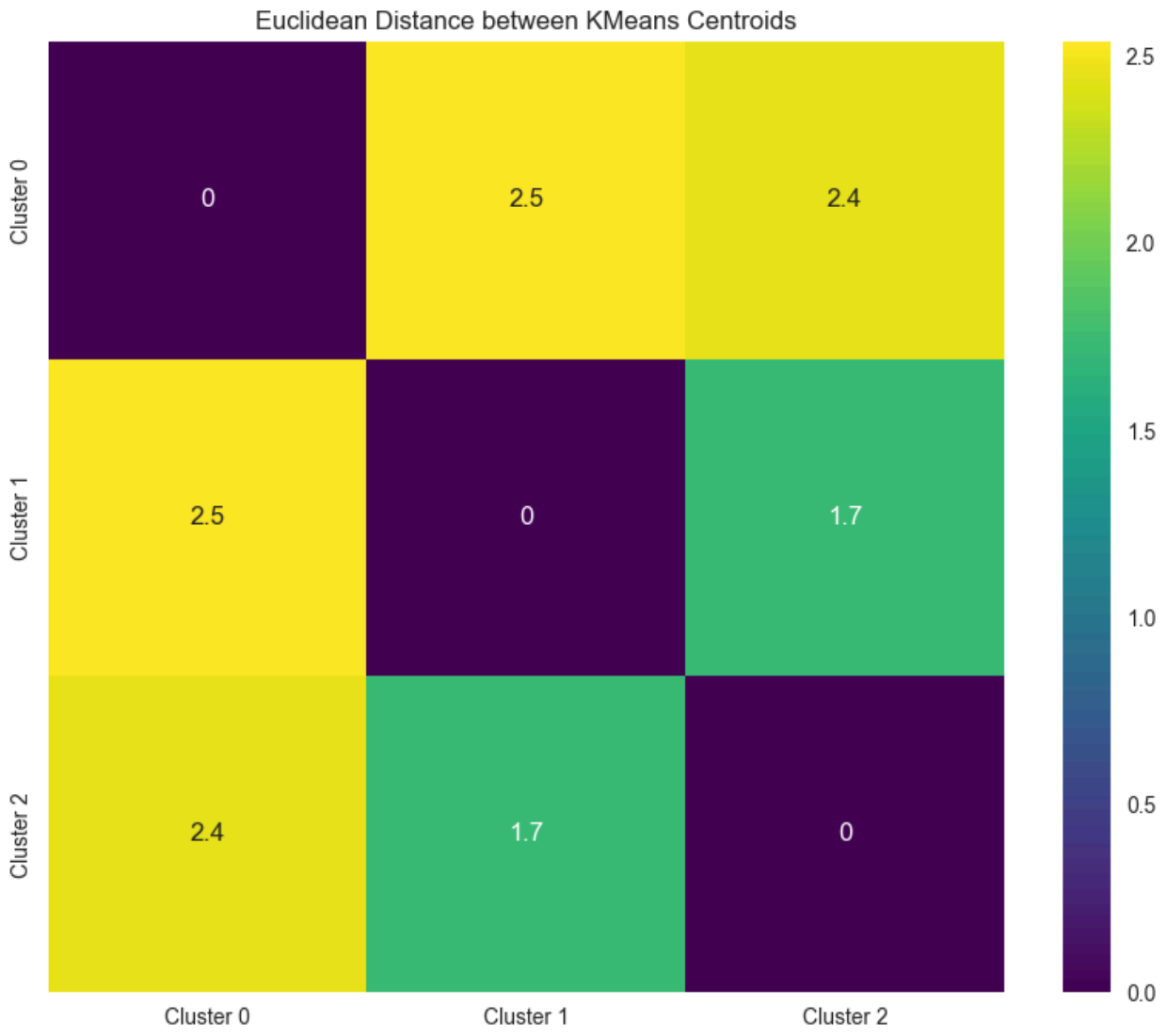


Silhouette score for n_clusters = 3 : 0.35

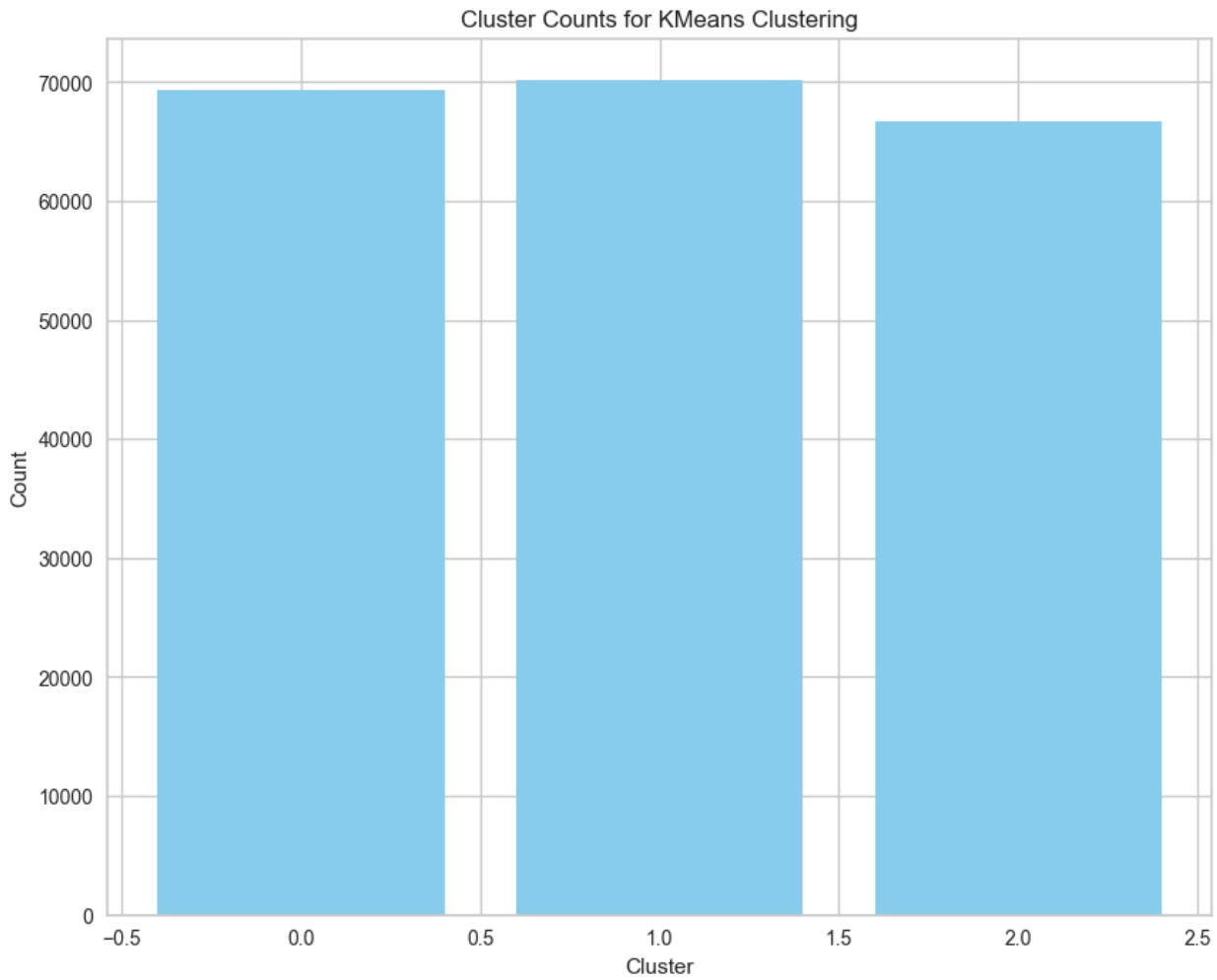
In observing our silhouette plots the results align mostly with our previous plots. The silhouette analysis shows our best k value to be k = 3. There are other values that we can make an argument for here, however with the elbow plot showing k = 3 and now the silhouette plot to back it up we can feel confident going forward at this time that k = 3 is a good value.



As you can see above, the clusters are very nicely separated and contained together really well. There is some degree of variance but that will be discussed later.



In the above heatmap for showing Euclidean distances you can see that Cluster 0 and Cluster 1 are well separated, as are Cluster 0 and Cluster 2. Cluster 1 and Cluster 2 still have a decent degree of separation, however further diagnosis may need to be done.

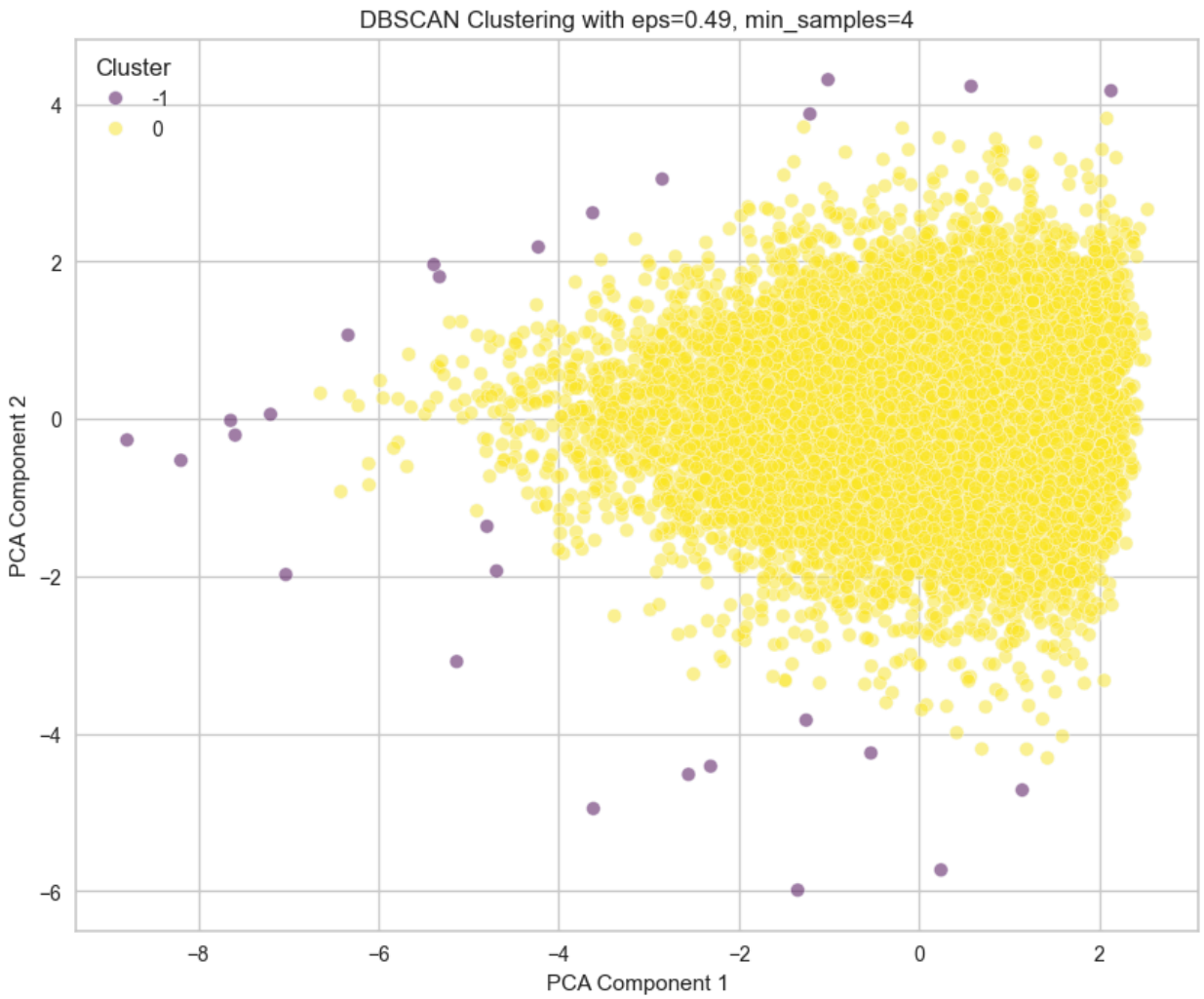


This plot shows the population distribution for the KMeans Clustering at k = 3. As you can see it is a very nice distribution; very even which is what you are looking for.

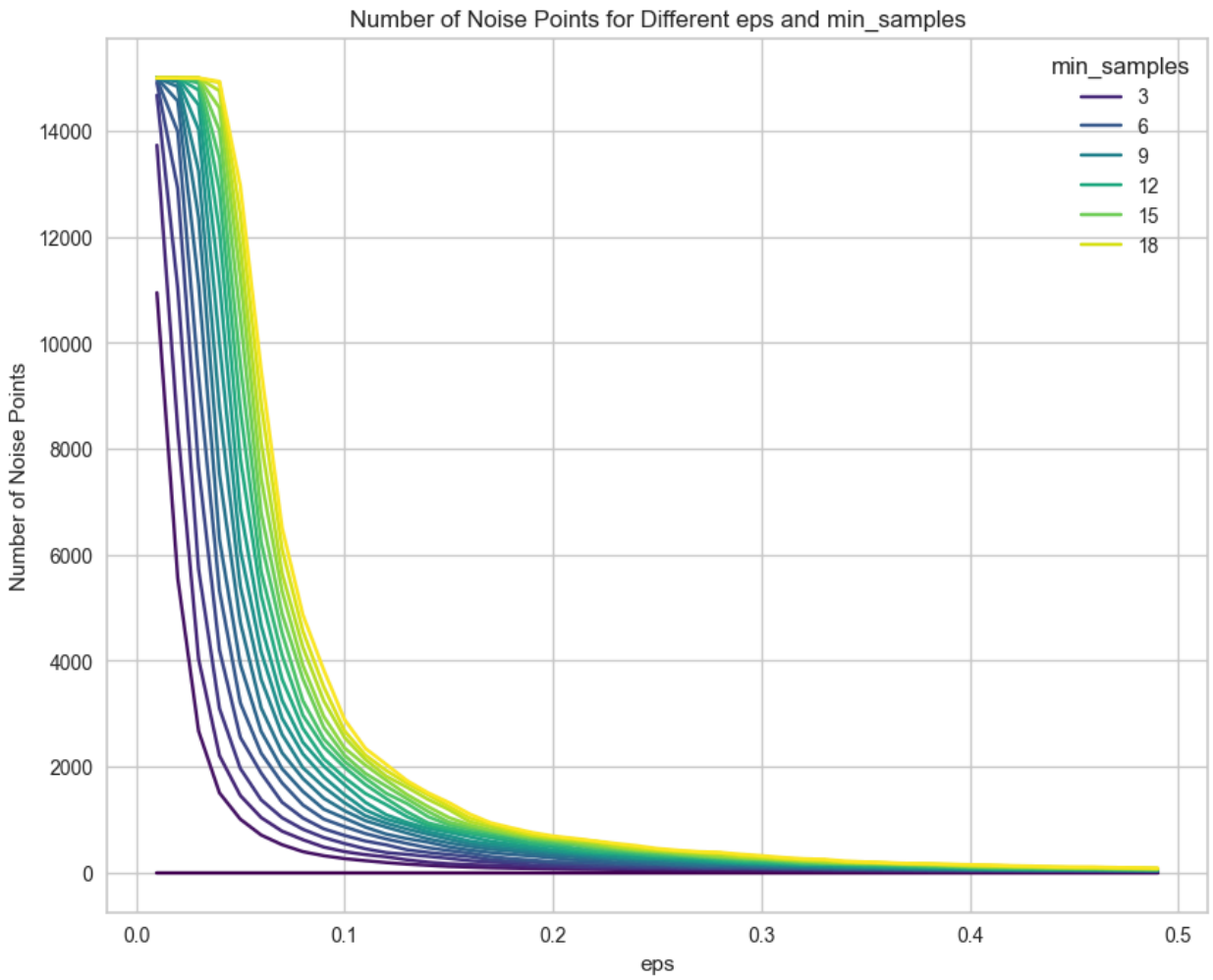
[Extended Modeling]

The motivation behind other models was because there could potentially be something 'better'. I didn't want to just try one clustering method and be satisfied with the results given, so upon setting out in doing this project there was always going to be at least 2 clustering methods performed as a comparison. For this project as the additional methods I chose Agglomerative Clustering and DBSCAN.

Now to talk about DBSCAN first, I was not able to render a reliable clustering for it. The data was too noisy for it (see below) and although at a later date if desired I could go back to it and keep trying to squeeze out a potential parameter combination between the eps and min_sample at this time there isn't a reliable clustering found for DBSCAN.

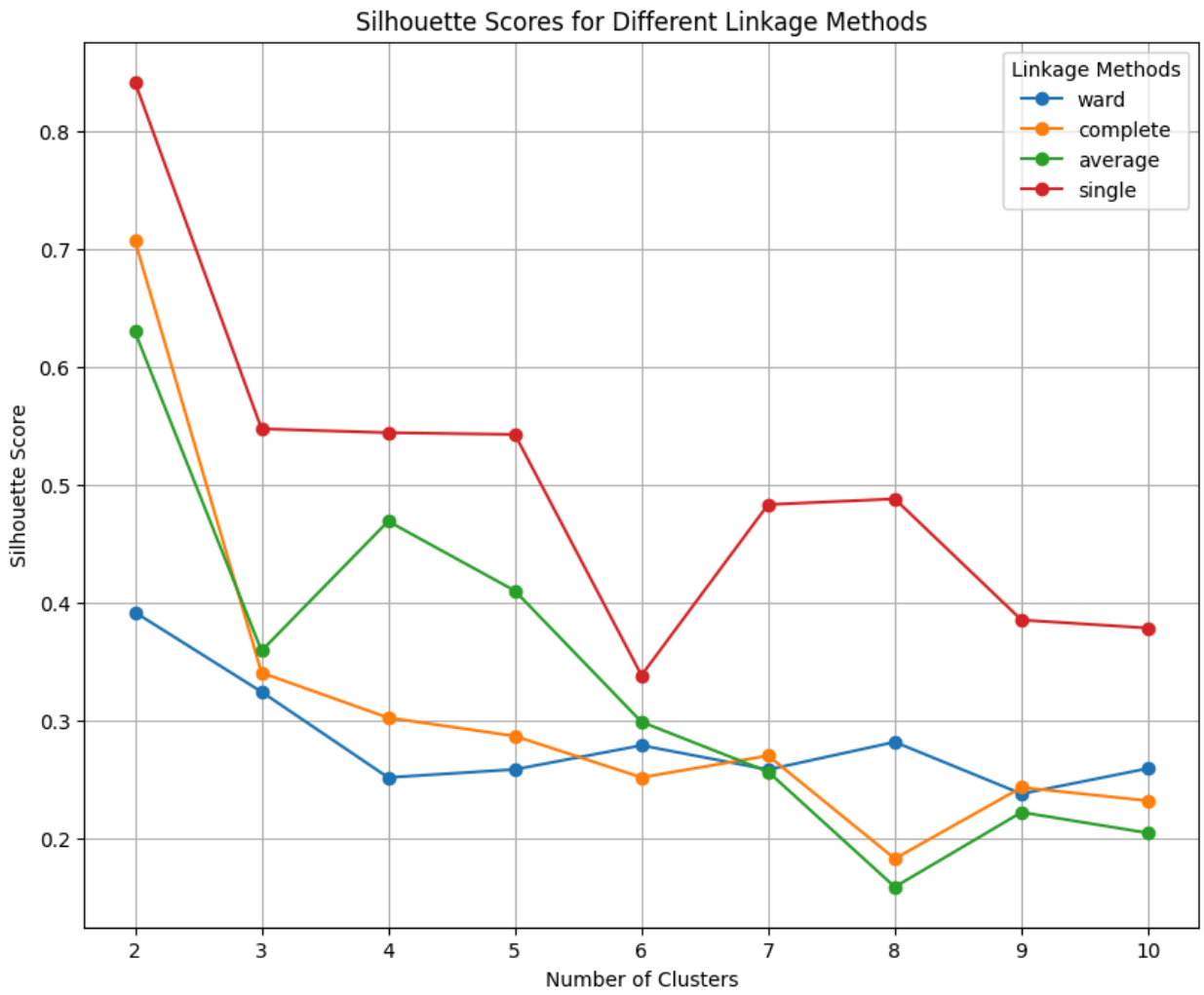


In the above you can see almost all of the data points get relegated to Cluster 0 (yellow), and the remaining are surrounding it and shown as Cluster -1. In DBSCAN the denotations of -1 are the noise points. These points don't belong to any cluster and are considered outliers or anomalies.

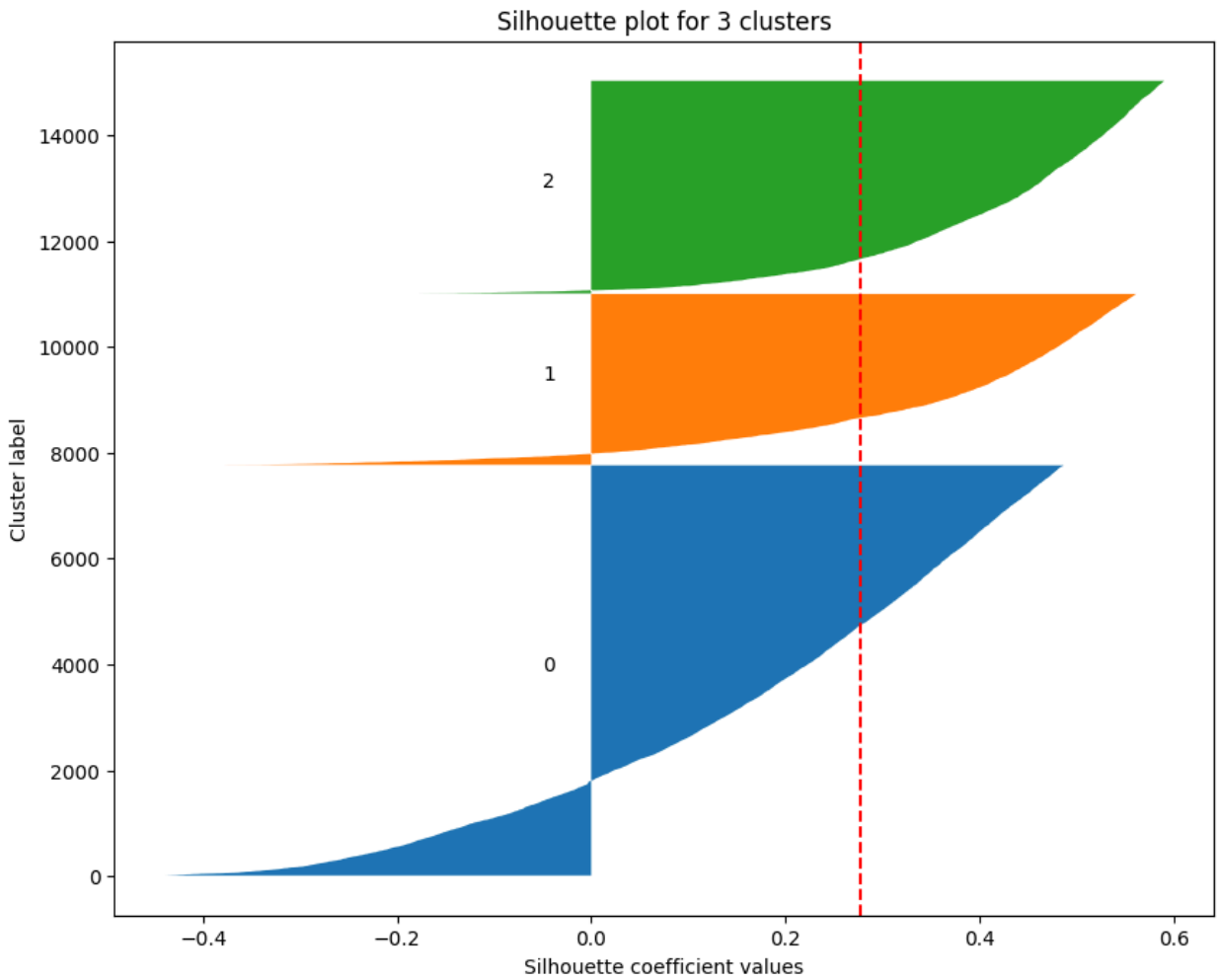


In the above plots you can see the abundance of noise points produced by various parameter combinations of the DBSCAN model, as well as a DBSCAN clustering that shows the vast majority in one single cluster with noise points surrounding it (first plot).

Now Agglomerative Clustering was successful in yielding results from the data. It was indeed more nuanced than KMeans as we had to decide which among 4 linkage methods was best. After going through them Ward linkage was best, with a k = 3 value for that clustering method as well. Even though the k value ended up being the same the results were a bit different. Note that for Agglomerative Clustering we had to utilize sampling (at n = 15000) in order to be able to process it, as this clustering method is quite a bit more computationally expensive to use.

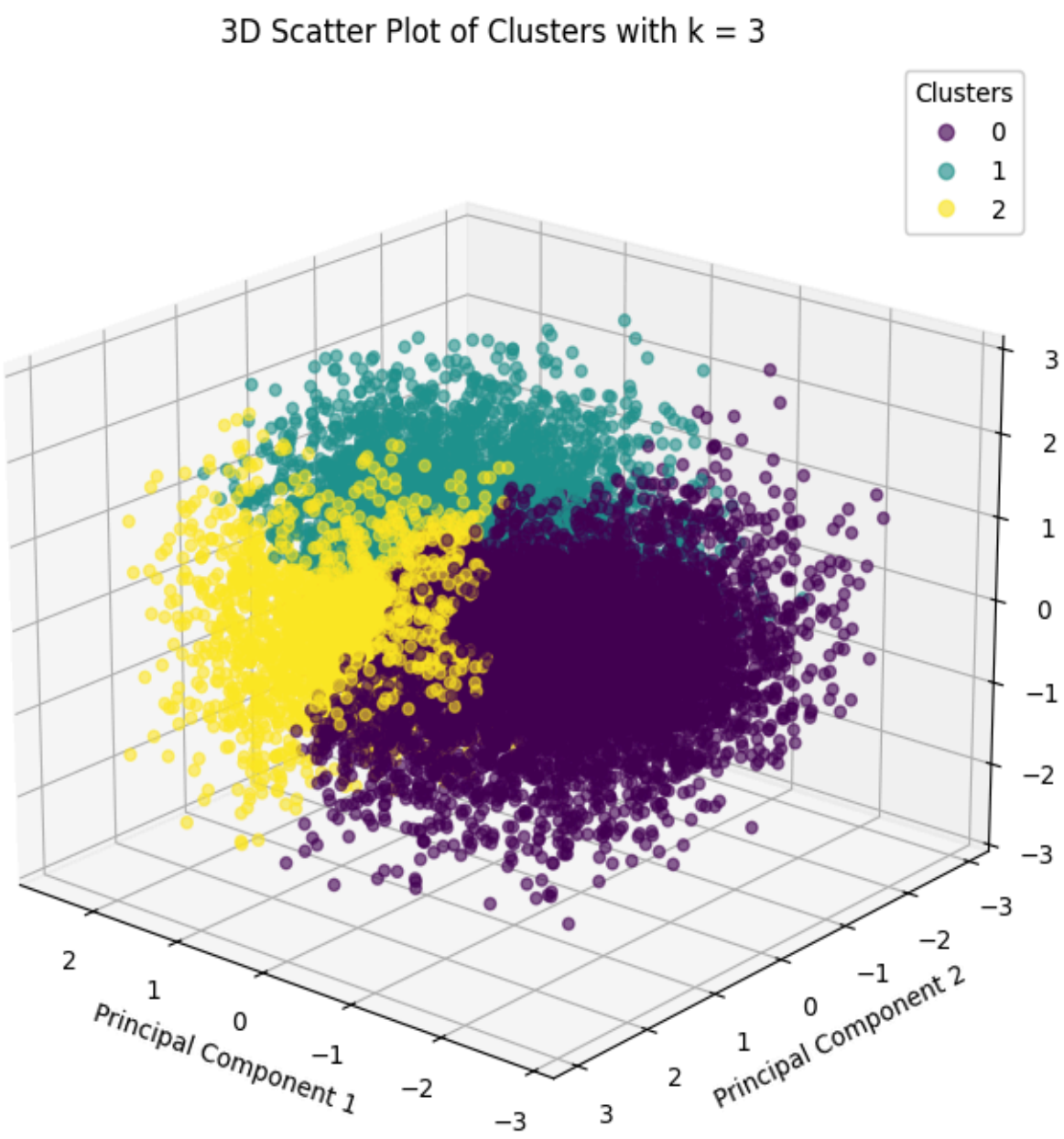


The above plot tells us quite a bit. Overall we can see that the silhouette scores decrease as the number of clusters increase for each linkage method, so we should be looking for a smaller number of clusters independent of which linkage method we choose. SIngle linkage in particular shows the highest silhouette scores, however for this project it is not ideal in that single linkage can create chain-like clusters that aren't necessarily compact which is what we are trying to establish with our customer groups. The Average linkage method also has strong silhouette scores in the first few represented clusters defined, and produces clusters that are moderately compact. Complete linkage only shows to have good scores at its 2nd cluster, and beyond that its scores drop considerably. Finally Ward linkage, while having some of the lower scores shows consistency across all plotted clusters; its scores aren't impressive but the key point about Ward linkage is that it tends to produce compact and relatively even-sized clusters which is a key focus for our project. Before we choose a linkage method however we still need to look at other factors in our consideration such as cluster size (read population) and their own relationships with each other as the silhouette scores themselves aren't enough to determine which method is the one to choose yet. Further analysis is needed.

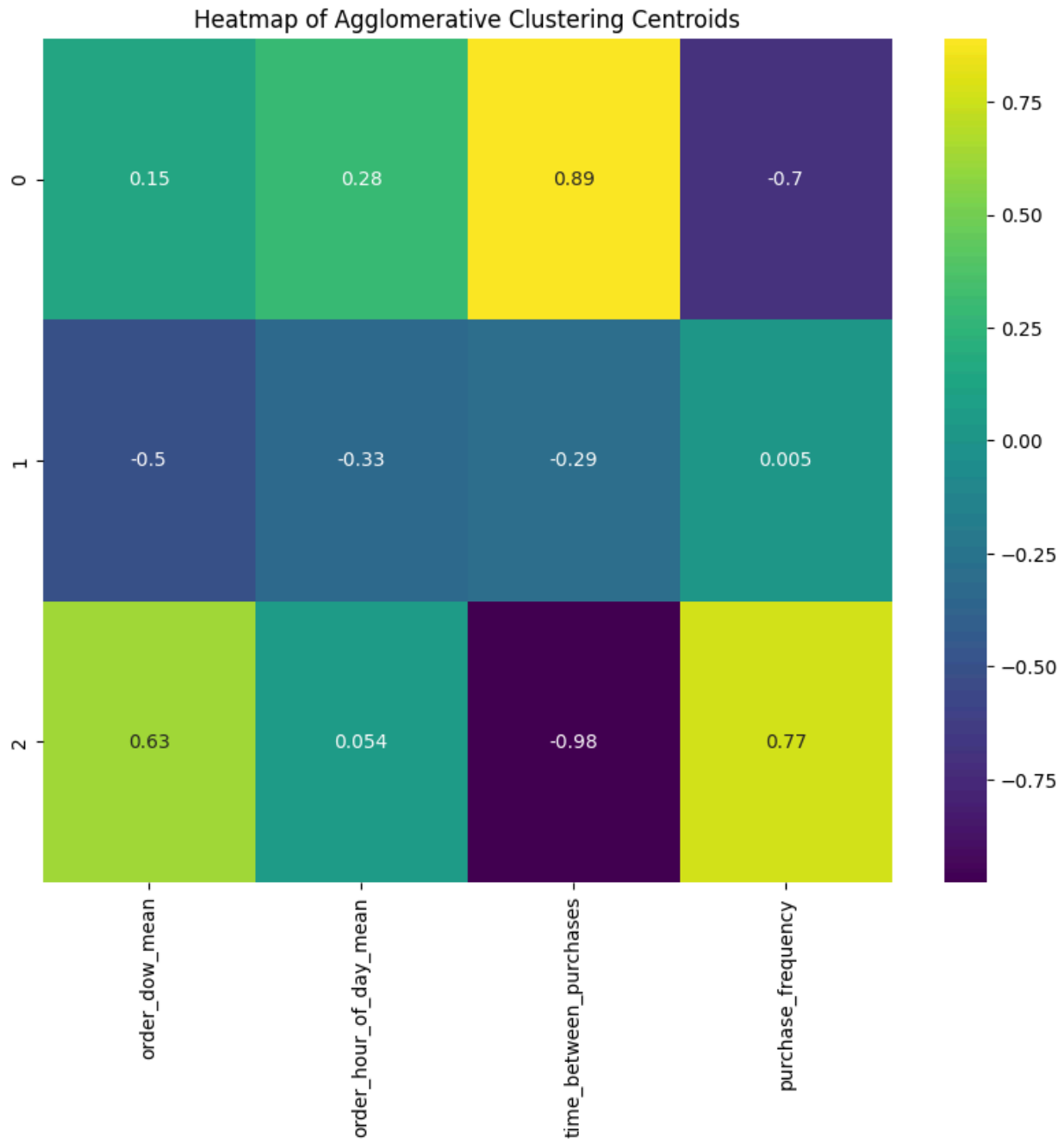


Silhouette score for n_clusters = 3: 0.27

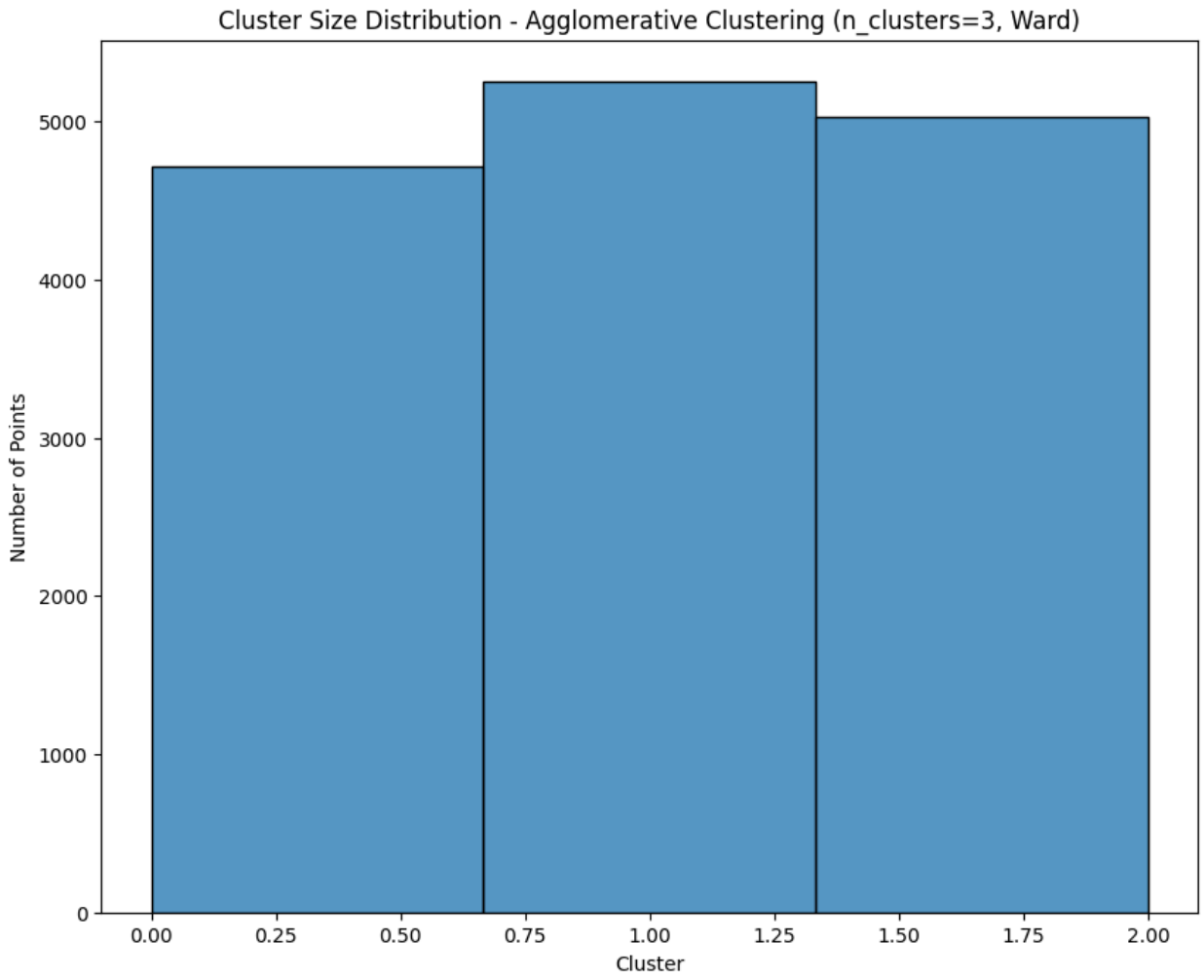
Looking at the plot above for Ward linkage you can see that the 3 clusters are close to each other's scores and the sizes are pretty similar as well. This representation could be better no doubt but when compared to some of the others this selection was one of the better ones for our Agglomerative Clustering method.



Above you can see the 3D render for the Agglomerative Clustering model. In this case while there are 3 different groups with separation, that separation isn't as definitive as was shown in KMeans. Granted again we are only looking at a sampled version however there are enough data points represented to give an idea of the model.



In the heatmap above you can see that the Clusters are starting to identify with their chosen characteristics. Cluster 0 has the longest time between purchases as well as most infrequent purchasing rates. Cluster 1 has average rates across the board, and Cluster 2 is much more active with quicker purchase activity.



Viewing all the plots above for Agglomerative Clustering, there are certainly similarities at least from an observational perspective. The k value is the same, the population distribution looks good as well for Agglomerative Clustering, and the 3D render, while not nearly as clean as KMeans, still shows good grouping.

Findings:

In this project we set out to look into Customer Segmentation of the Instacart Dataset from Kaggle using several Clustering methods, doing individual analysis and modeling on each then comparing all of the methods at the end (which we will do below) to see which one is the best fit for the project.

As a summary of how each model performed:

KMeans Clustering -

Identified three distinct customer segments (Frequent, After Hours, and Weekend Shoppers) with clear separation between clusters.

Agglomerative Clustering -

Also identified three similar segments (Infrequent, Average/Moderate, and Frequent Shoppers) with a slightly different distribution, showing consistent patterns in customer behavior.

DBSCAN Clustering -

Revealed the presence of noise in the data and identified core points, but with fewer clear segments compared to KMeans and Agglomerative Clustering. No clustering identified outside of the core cluster was able to be reliably identified due to the noisy data points.

#	Method	Number of Clusters	WCSS	BCSS	Silhouette Score
1	KMeans	3	47325.830	351379.47	0.3521
2	Agglomerative	3	28520.669	15327.63	0.2776
3	DBSCAN	1	N/A	N/A	N/A

Metrics Overview -

WCSS (Within Cluster Sum of Squares) shows and determines compactness of each cluster or group of customers; we are looking for lower scores here. Higher scores are going to indicate variance, which is a measure of how far the data points are spread out from each other.

BCSS (Between Cluster Sum of Squares) - shows and determines separation of each cluster or group of customers. In this metric we are looking for higher scores.

Silhouette Scores - are a measure of how similar an object (read data point) is to its own cluster compared to other clusters. The higher score the better here. Higher scores indicate that the clusters are well-defined and distinct from each other, are matched well to their own cluster and poorly matched to other clusters.

Clustering Method Selection:

There are really just the two methods to choose from as DBSCAN was unable to render a reliable cluster. For this project I will be selecting the **KMeans Clustering** results as the best method for our customer segmentation. While KMeans has a noticeably higher variance in its WCSS score, this method has a much better separation in its clusters with the BCSS score and the silhouette score, while not great, is comparably better than that of the Agglomerative Clustering model. The 3D render comparison shows us as well how well the separation and clustering performed for a visual aid. If desired the variance measure can be investigated through the use of an alternative clustering method (for example, GMM).

In conclusion KMeans Clustering is the choice at this time to go forward with for this project based on the clustering methods and the work performed.

[Future Work]

Listed below are several items that could be pursued further if desired to achieve further results from the existing modeling results.

-As noted above can look into the variance of the chosen KMeans model to optimize even further, opting to utilize something along the lines of GMM that is similar but focuses on reducing the variance of the clustering.

-Could look deeper into the DBSCAN model for potential clusters. May not change the results however could be worth a future look.

-The big value add for this project from the beginning was being able to create custom marketing promotions for each customer cluster once they were properly identified and segmented. If desired and continue on and create a list of custom promotions, then run their predicted success using a model like Linear Regression and plot the results for each cluster group. Report the results based on how successful each promotion is for each unique cluster group so the stakeholders can use going forward.

[Recommendations for the Clients]

Before going into the bulleted list it is important to discuss a key finding from the selected clustering method (as it is appropriate here). Also the recommendations are being issued without the use (read testing) of the custom promotions, which were to be a core feature of this endeavor.

It has been mentioned that KMeans Clustering identified three clusters with the following general characteristics:

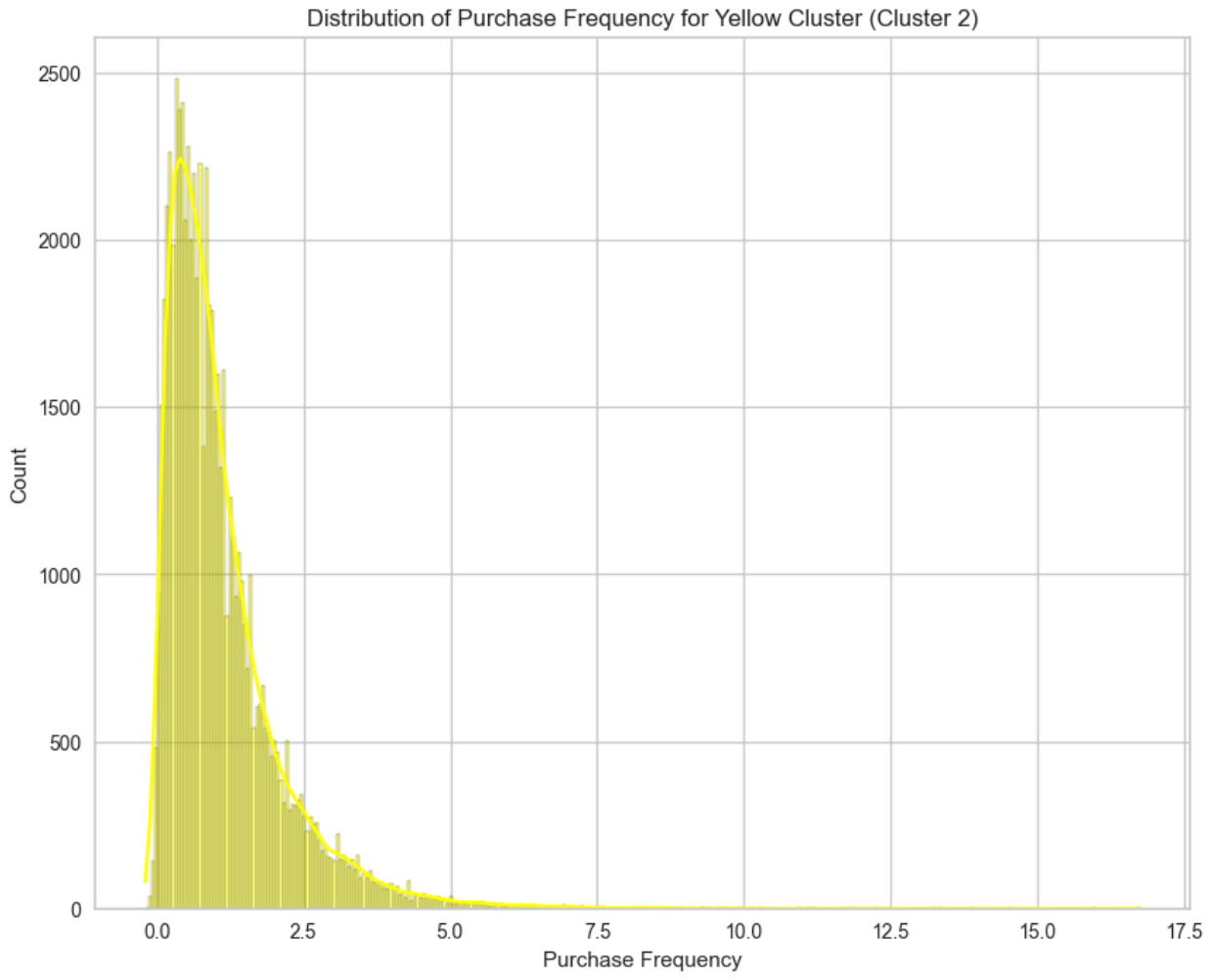
Cluster 2 (Frequent Shoppers): This cluster group has been shown to have a high purchase frequency, short time between purchases, orders slightly earlier in the day, and average on popular days.

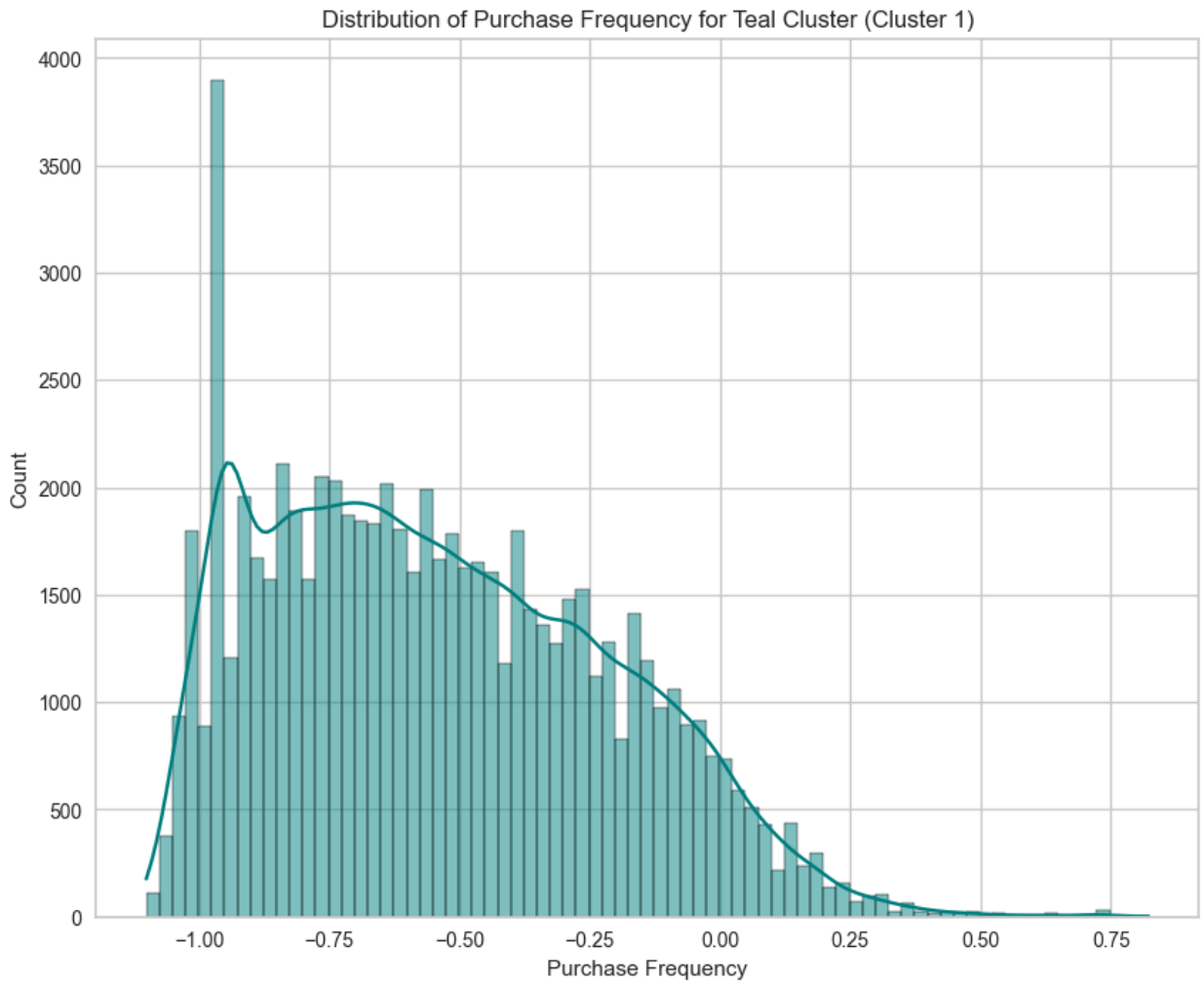
Cluster 1 (After Hours Shoppers): The cluster group has been shown to have a lower purchase frequency, longer time between purchases, orders later in the day, and less on popular days.

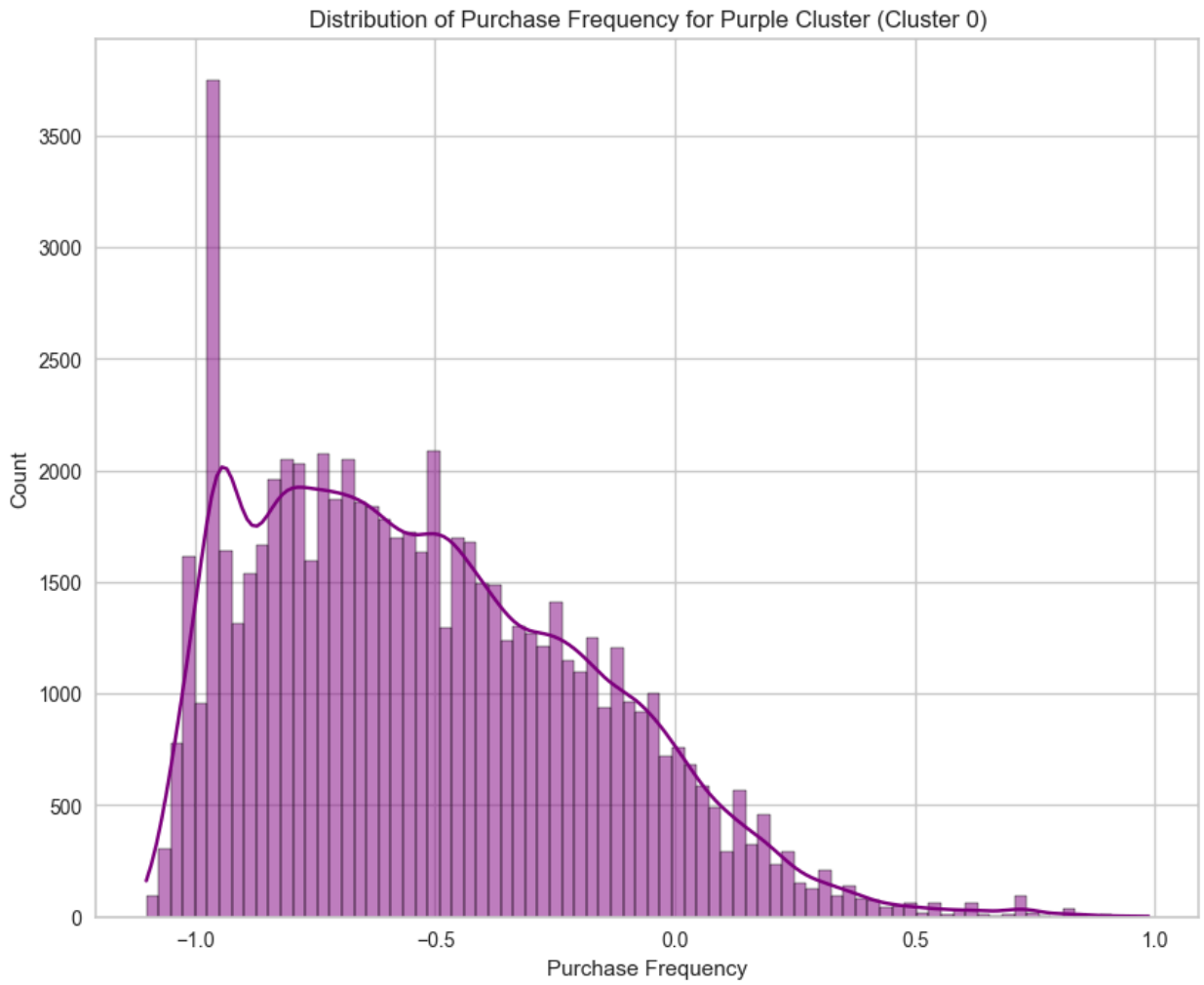
Cluster 0 (Weekend Shoppers): Lower purchase frequency, longer time between purchases, orders slightly earlier in the day, and on popular days.

Using distribution plots we can get more specific:

Purchase Frequency was the most telling and utilized metric out of our features.




Distribution of Purchase Frequency for Teal Cluster (Cluster 1)
Count
Purchase Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75



As you can see above from the individual cluster plots for Purchase Frequency, Cluster 2's behavior shows the increased spike in Purchase Frequency that was mentioned in the summary compared to Cluster 1 and Cluster 2.

My recommendations for each cluster are as follows, though without the custom promotions these are solely based on what behaviors have been extracted from the data. Should further analysis be conducted and the promotions be tested then the following is subject to change.

Cluster 2 (Frequent Shoppers - denoted by yellow) -> For this cluster group I recommend implementing loyalty programs to reward their shown high level of engagement.

Cluster 1 (After Hours Shoppers - denoted by teal) -> For this cluster group I recommend implementing time-specific promotions to encourage purchases during off-peak hours.

Cluster 0 (Weekend Shoppers - denoted by purple) -> For this cluster group I recommend personalized reminders to reduce the time between purchases.

[Consulted Resources]

Link Here for Data!