BUSI4370 - Analytics Specialization and Applications

Coursework 1
Student ID: 20492770

EXECUTIVE SUMMARY

Customer segmentation is a process of dividing customers into groups based on their characteristics, behaviours, and preferences. The purpose of customer segmentation is to tailor marketing strategies and offers to each group and increase customer satisfaction and loyalty. This report presents the results of a customer segmentation analysis for a national convenience store. The report uses point-of-sale transactional data from the store's customer database, to identify five main segments of customers: (loyalists, planners, need-based shoppers, extravagant shoppers, and convenience seekers). The report describes the profile, needs, and expectations of each segment and provides recommendations on how to target them effectively. The report also discusses the process of data cleaning, feature engineering, methods used to reduce dimensions as to visualise segments and algorithms used for clustering. As a final note, the report suggests areas for further research.

DATA AND FEATURES SUMMARY

The data used for the analysis is transactional data of 3000 customers who use loyalty cards at a national convenience store over 6 months, using which behavioural trends of different customer types can be extracted.

DATA AVAILABLE

The file 'baskets_sample.csv' contains the details of customer ID, and each row provides details of each visit, their purchase time, quantity of items bought, and amount spent during that visit. The file 'lineitems_sample.csv' gives much more detailed transactional details of each item purchased by a customer, the category the item belongs to, quantity of the item bought and its cost. The data does not provide any demographic or geographical data, hence the analysis will be performed on behavioural inferences made using these transactions.

There also is a file names 'customer_sample.csv' which aggregates lineitems and baskets data to create one row for one customer details of total spend and quantity bought by the customer. And also average spend and quantities. The file 'category_spends_sample.csv' uses the lineitems table to aggregate customer's spends on different category of items. The python notebook attached, provides details of shape and overall statistics of the data.

DATA CLEANING

The first step before choosing required features is to clean the data, i.e. checking for missing values, data types and outliers, if any. The following table provides details of issues observed in the data and how they are handled. (# All the cleaning methods are done using pandas library including importing the data frames).

FEATURE AND TABLE	CONCERN	RESOLUTION
'spend' in all tables	It is entered as a string with pound sign.	Convert it to float by removing the pound sign so that the calculations are easier.
'purchase_time' in baskets_sample.csv	It is stored as a string value.	Convert to date time format so that differences between days can be easily calculated.
'spends' in category_spends_sample.csv	Not correctly calculated for product category 'bakery' and negative spends are not handled.	Recalculate categorical spends using lineitems_sample.csv and converting spends to absolute values. (It is assumed that customer has returned the item when spend is negative and we will consider it as positive for initial purchase made).

FEATURE SELECTION

Using all the four data frames of data collected, many features that help assign customers to different segments can be engineered. The below table summarizes all features calculated for the analysis and also the features that were excluded, with justification.

FEATURE	DESCRIPTION	JUSTIFICATION
customer_number	Customer ID number.	To identify unique customer.
total_quantity	The total quantity of products the customer has purchased in the store.	To identify total products bought of a customer in the store.
total_spend	The total amount spent by the customer in the store.	To identify total expenditure of a customer in the store.
average_quantity_per_visit	The average quantity of products bought by the customer per visit.	To distinguish customers who buy more/less products in one visit.
average_spend_per_visit	The average amount spent by the customer per visit.	To distinguish customers who spend more/less in one visit.
average_spend_per_item	The average amount spent by the customer per item.	To distinguish customers who buy expensive products.
frequency	The number of visits made by a customer.	To identify frequent visitors and casual shoppers.
recency	Number of days since last visited by a customer as calculated by offset from max purchase_time in data + 1 day.	To identify if a customer is still visiting the store or not.
Spend for each category of product (20 categories)	Total spend of a customer on a particular category, calculated for all product categories.	To identify customers who spend on particular categories of products.

The following features from original data are dropped since they do not provide much information.

FEATURE/TABLE	DESCRIPTION	JUSTIFICATION
purchase_time/ baskets_sample	Time of purchase in a visit by a customer.	The recency feature is extracted from this and dropped, since no other information such as time of day had good variance.
basket_quantity, basket_spend, basket_categories /baskets_sample	The basket quantity and spend are customers spends and products bought per visit. Basket categories is number of categories purchased per visit.	Included as average_quantity_per_visit and average_spend_per_visit calculated from line items.
baskets/ customer_sample	The total different categories bought by a customer (sum of distinct categories in each visit)	Frequency gives a better information of number of visits by customer than this feature.

DATA MANIPULATION

After data for the 3000 customers is cleaned and the required features are engineered, the distribution of all features can be observed using scatter plots. From the plot as shown in Figure 1 in the appendix, it can be seen the data is right skewed for all the features. An appropriate transformation is required to transform the data into Gaussian/Normal distributions.

Also since the data has a lot of dimensions (27 excluding customer_number), the dimensions have to be reduced using techniques such as Principal component analysis (PCA) or Non-negative matrix factorization (NMF). These techniques require the data to be scaled so that it is not affected by features with huge values.

POWER TRANSFORMER

Clustering algorithms such as KMeans clustering, assume that the data is normally distributed, since the clusters generated are globular. Hence, it is highly important to make sure the data is appropriately transformed. sci-kit library provides a pre-processing module called power transformer which, as the name implies, transforms the data to be more Gaussian-like. By taking the log or square root of the variable, we can perform a power transform directly, but this may not be the optimal power transform for that variable. Instead, power transformer is a generalized version that can find a parameter that optimally transforms a variable to a Gaussian probability distribution.

Using Yeo-Johnson method (which handles both positive and negative values) the data is transformed. After transformation is applied the distributions are visualised again to check if the result is desirable. The transformed distributions can be observed in Figure 2 of the appendix.

MINMAX SCALER

As discussed, both dimensionality reduction techniques and KMeans clustering algorithms rely on Euclidean distances which will be affected by higher value features and outliers. Hence scaling of the data is a must.

Sci-kit learn library provides many scaling techniques. Among them, minmax scaler is a popular algorithm that scales each feature to a given range, usually [0,1] or [-1,1], by using the minimum and maximum values of each feature. It preserves the shape of the original distribution.

So MinMaxScaler is used to scale the data before dimensionality reduction techniques are applied.

DIMENSIONALITY REDUCTION

If the data has a lot of dimensions, machine learning algorithms used to fit the data will be computationally expensive which reduces performance. Also, data visualisation is quite hard which is important, especially when it comes to presenting clusters for business reports. There are a lot of dimensionality reduction techniques available, one of which is Non-negative matrix factorisation (NMF).

NMF decomposes a matrix into two matrices such that their products approximate the original matrix. One advantage of NMF over principal component analysis (PCA) is that it can preserve more information than PCA by imposing non-negativity constraint on the data. We would like reduce the 27 components into two or three dimensions so that it easier for us visualise the clusters as well as make KMeans clustering faster and efficient.

After performing multiple iterations it can be learned that 3 components of NMF reduced data is a good approximation of original data. The weights of different features for the 3 components are as follows. (These weights are as observed from the bar plots generated as shown in Figure 3 of appendix).

DIMENSION	WEIGHTS OF FEATURES	DESCRIPTIVE NAME
Dimension 1	The first dimension gives weightage to product categories that are groceries and food and also total spend and quantities of purchases.	FOOD_AND_TOTALEXPENSE
Dimension 2	Dimension 2 gives weightage to non-food products such as lottery, newspaper and magazines, and practical items	NON_FOOD_ITEMS
Dimension 3	This dimension mainly varies based on recency of customers and their average expenditures	RECENCY_AND_AVGEXPENSE

SEGMENTATION METHODOLOGY

Clustering is an unsupervised learning method that groups different data point, customers in this case, into clusters based on similarities of data points.

KMeans clustering requires us to choose a value of k, and then the algorithm creates k clusters based on similarity of data. K centroids are first initialised, and iteratively reached to centres of its cluster so that the data points in the cluster is least distant to that centroid. These centroids represent the mean of the data points in that cluster.

KMeans is chosen over other clustering algorithms because of its simplicity and scalability. It adapts to new examples easily which will enable the business to update their segments without reengineering all the steps to create segments. And also the algorithm guarantees convergence.

METHODOLOGY

Before fitting the reduced data, i.e., data reduced by NMF, we should find the best number of clusters to define k. This can be achieved by calculating silhouette scores for different values of k and choosing the one with the highest silhouette score. Since the requirement is to create 5 to 7 segments of customer profiles, the algorithm will be tested with k=5,6, and 7.

Steps to choose silhouette score:

- Fit KMeans algorithm from sci-kit learn to the reduced data for k=3.
- Predict the reduced data points using this cluster model.
- Use sci-kit learn's silhouette score module to calculate the score using these predictions and reduced data.
- Print the score and repeat for k=4,5,6,7, and 8.

After running through the above loop, the best silhouette score was found for k=5, which was 0.34. So the number of clusters is chosen to be 5.

With the value of k as 5, the clustering model is created by fitting the reduced data using the KMeans algorithm. Now the clusters can be visualised using plots to analyse the final results of clusters and centroids and make inferences of the behaviours of customers of each segment.

VISUALISATION

Since the number of dimensions of reduced data was 3, the library plotly can be used to visualise these clusters.

Steps:

- Extract centres of clusters from the model that was fit.
- Predict the clusters of reduced data using the model. These will be the cluster assignments of data points.
- Merge the predictions with reduced data and plot the data points using 3d scatter plot provided by plotly.

The resulting visualisation is attached as Figure 4 in the appendix.

The file 'final_assignments.csv' contains details of which customer belongs to which cluster.

INSIGHTS

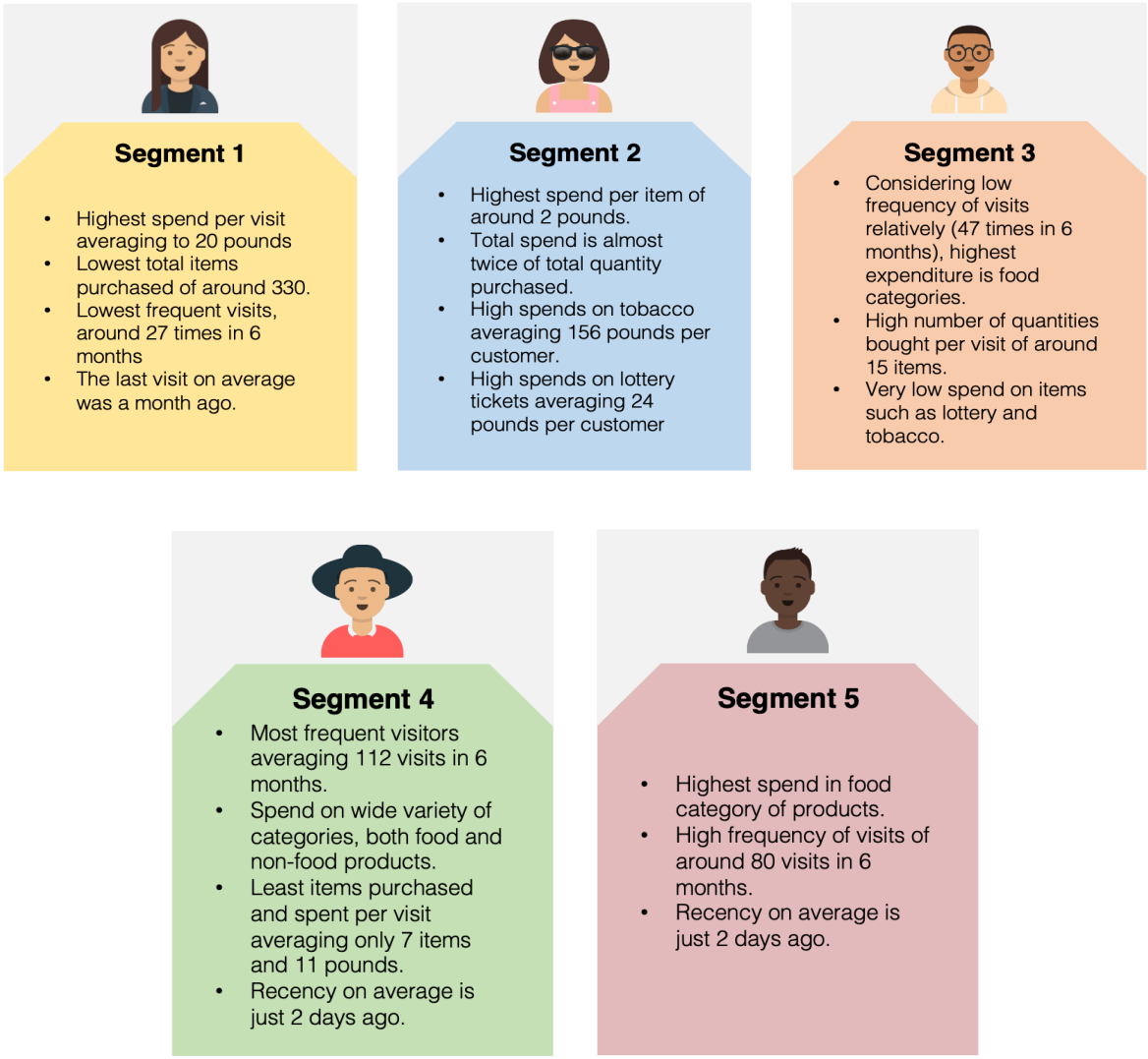
Every retail store encounters different types of customers who have different shopping preferences, shop at different frequencies, and have varying levels of loyalty. The above clusters of customers give insights on the major 5 different types of such customer bases. We will first explore the overall statistics of the customer base of the store and then move on describing each segment.

OVERALL CUSTOMER BASE

All the customers in the data collected are loyalty card holders. On average 50% of the customers visit the store 9 times a month or twice a week and spend 100 pounds a month. Their average spend on the products is around 1.25 pounds per item. 25% of shoppers visit once in a fortnight. Most of the categories of products purchased are groceries and food items with only 25% of customers purchasing commodities such as tobacco, lottery tickets, newspaper and magazines and practical items. Around 20% of the customers buy expensive products with average spend per item close to 2 pounds. Looking at statistics of each segment of the clusters will provide more details to profile the segments for which specific marketing techniques can be designed. A summary statistic of overall customers can be found in the python notebook.

SUMMARY STATISTICS OF DIFFERENT CUSTOMER SEGMENTS

The statistics of each segment considering highly distinguishable features of those segments can be summarized as below:



The complete list of statistics can be found in these files attached: 'segment_description.csv' and 'true_centres.csv'. True centres provides details of central customer of each segment.

PEN PROFILES

SUMMAY AND RECOMMENDATIONS

The segmentation of customers based on their shopping preferences has provided behavioural insights of those customer segments. It can be used to design and develop marketing strategies that increase profitability of the business, with a level of confidence.

It is advisable to mainly choose the following two customer segments for initial marketing and observe the results.

- The planner: This customer base are a major chunk of the customer base and they plan their shopping before visiting the store. So they are customers who might not have time to visit the shop frequently. Offering them discounts when bulk buying products and providing quick check out options are attractive marketing strategies.
- The loyalist: These customers visit very often and buy almost all categories of products. Suggesting they are a loyal customer base. Offering them coupons and points system which add up every time they buy products and can redeem the points as discounts will make sure the customers remain loyal.

FURTHER RECOMMENDATIONS

To understand a wider customer base, collecting and analysing transactional data of customers without a loyalty card will be beneficial. It can provide information of customers who choose not to use a loyalty card and the reasons for it, which ultimately generate techniques that can used to make the loyalty card itself more attractive.

Also it is advisable to update the clusters with demographical data and geographical data which informs whether the customer shops for a single person or a family, how far they travel. This kind of data opens up to lot of possible analysis and profitable business solutions.

APPENDIX

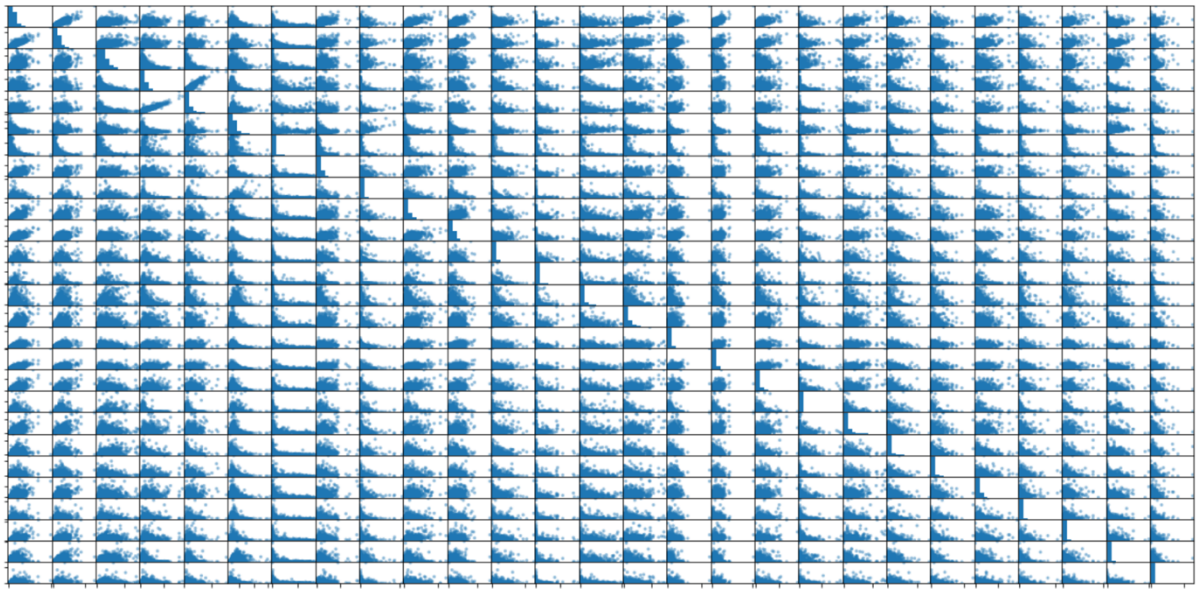


Figure 1: Checking skewness. The diagonal scatter plots are all right skewed which means all the features have skewed distributions

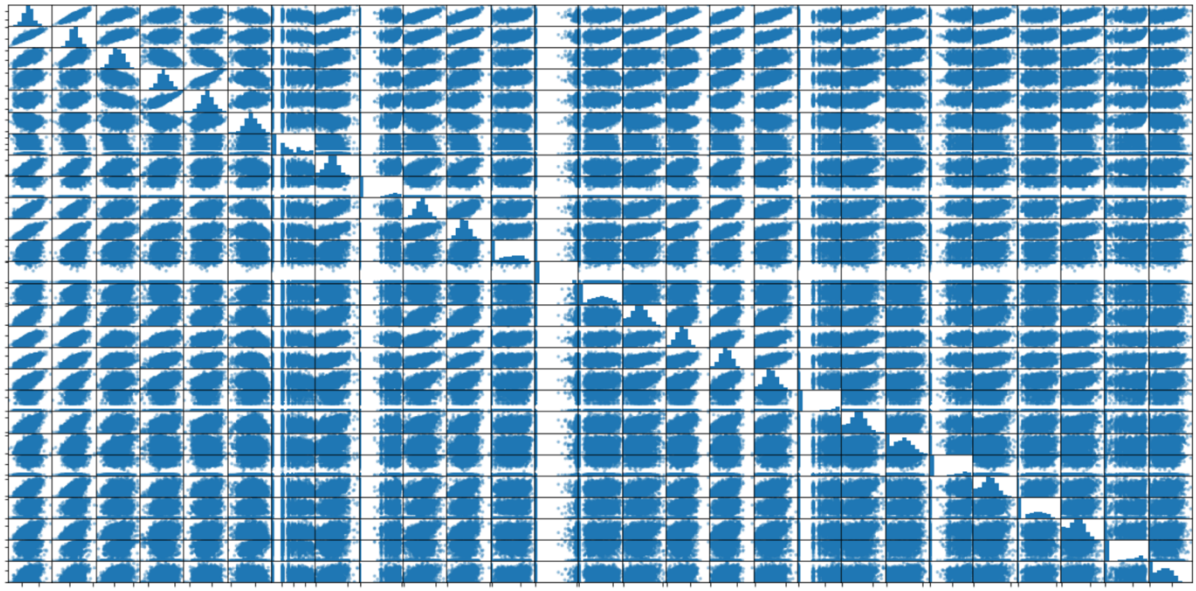


Figure 2: Scatter plots to show all the features are normally distributed after transformation

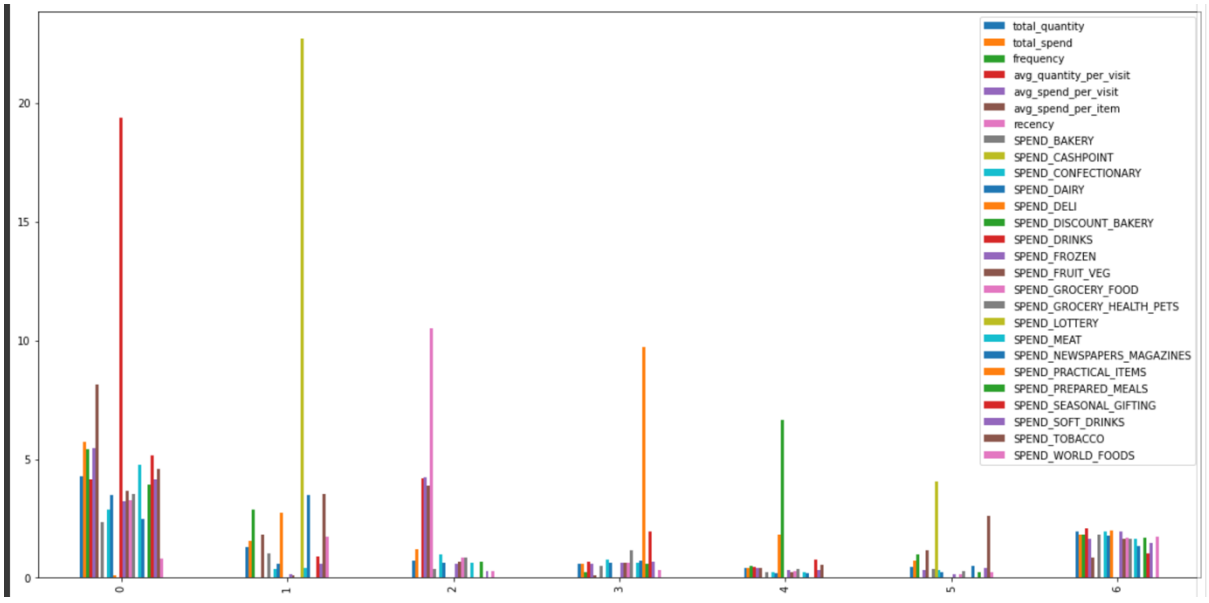


Figure 3: Results after features are reduced using NMF to 7 dimensions

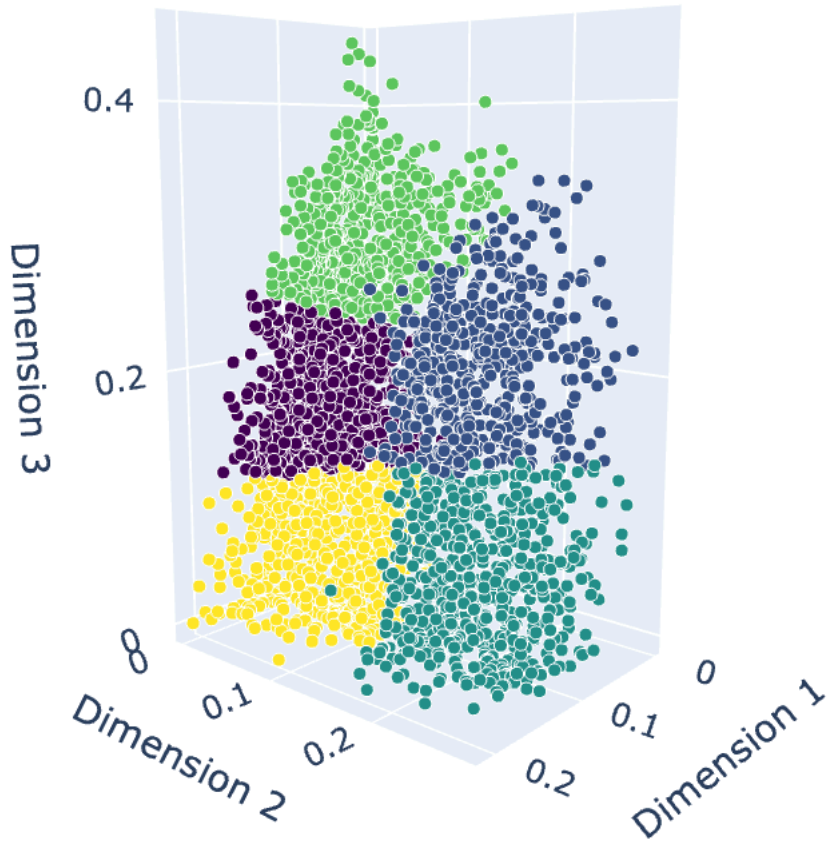


Figure 4: Scatter plot of final cluster assignments of data points