Analysis of Dillard's Point of Sales (POS) Data

Table of Contents

Executive Summary.....1

Introduction.....1

Methodology.....1

Modeling Methods.....2

Clustering Results.....3

ROI Analysis.....4

Appendix.....6

Executive summary

Using a POS data set from the department store Dillard's between August 1st, 2004 and August 27th, 2005, the group performed a clustering machine-learning exercise for the stock-keeping-units (SKU) using K-Means, GMM, and Hierarchical Clustering algorithms. With limited features available and computing power, the team was able to cluster 687,795 SKUs into 7 clusters. Given the clustering results, the team proposes a naive product recommendation system for Dillard's eCommerce platform and, as well as at their physical retail stores as part of the salesforce intelligence system. Compared to the current recommendation system of Dillard's on their website (which is solely brand-based), we calculated an ROI of $2.2M or 362% based on the most pessimistic market-research assumptions.

Introduction

Dillard's point of sales (POS) data contains multiple tables about a certain item, including SKU, brand, color, etc. Since there are multiple characteristics of an item available, a logical business question was how similar items could be grouped together. By clustering similar items according to their characteristics, a recommendation system could be constructed. Thus, when a customer purchases something, the system will recommend similar products based on the characteristics of the item purchased. Studies have shown that more than 70 percent of consumers will purchase an item recommended to them through advertising or social media, so implementing a recommender system will likely increase the company's sales.

Methodology

Feature Engineering and Selection

Variables "Percent Discount" and "Percent Return" were created for the potential features. "Percent Discount" is calculated as the average percent discount of an item grouped by SKU; all the missing and negative values were removed because a negative percent discount does not make any sense. "Percent return" is simply the return rate of an item.

"Brand" is a significant categorical factor in the clustering, but since there are over two thousand distinct values, doing one-hot encoding would be a considerable computational burden. Therefore, "Department," a subdivision inside a department store such as a Nike boots shelf inside Dillard's, was selected to represent "Brand." With only 60 distinct values, "Department" makes a good candidate for one-hot encoding while capturing the major features of "Brand." The distribution of departments is shown in Fig. 1 in Appendix.

Different colors give people different impressions, so people tend to stick with only one or a few colors, which makes color a good clustering feature. The 200 most commonly used colors in SKU were classified and reduced into 17 color groups. Color groups were joined back with SKU, which assigned each SKU to a color group. For the SKUs with "color" not in the 200 most common colors, "Unknown" was assigned. The distribution of color is shown in Fig. 2 in Appendix.

Location is also an important component, as people in two regions might have entirely different preferences for items. Therefore, from the transaction data, the most common value for states for each SKU was selected, then joined with a states-division-region mapping table sourced from the most recent U.S. census. This found the regions where the items are most commonly sold for each SKU. The distribution of regions is shown in Fig. 3 in Appendix.

Data Manipulation

The average price of each SKU was calculated, but the bottom 1% and the top 1% were removed to deal with the outliers. Moreover, since the values (average price, percent discount, percentage return) are skewed as shown in the appendix in Fig. 4, Fig. 5, and Fig. 6, respectively, a log transformation was applied to the data and then standardized.

Modeling Methods

K-means and GMM algorithms require the number of centers to be given as a hyperparameter which determines the number of clusters to be made. This hyperparameter was determined using principal component analysis (PCA).

Principal Component Analysis (PCA): It breaks the features down into principal components such that each component is linearly independent of the other component. In doing so, some explainability is lost, but a nice set of linearly independent features can be achieved.

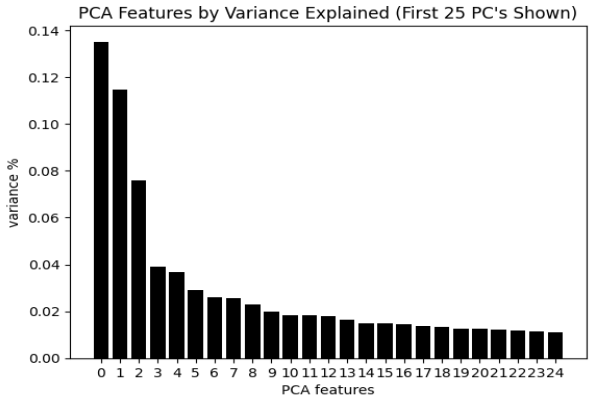

Fig. 7: Variance explained per component

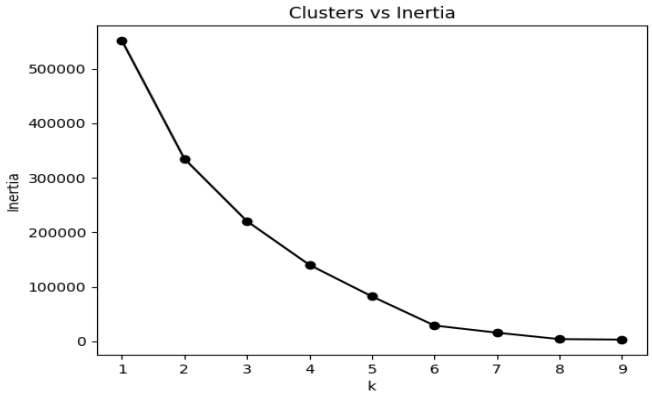

Fig. 8: Elbow point at cluster 7

To do PCA, the dataset was first standardized since PCA is sensitive to such data. Then PCA was implemented to find the number of components and the percent variance explained by each component. Note that the percent variance explained drops off sharply after the third component (Fig. 7), and the elbow point occurs at 7 clusters (Fig. 8). Therefore, seven was the number of clusters used for the rest of the models.

K-Means: The number of classes/groups to use was selected, and their respective center points (seven in this case) were randomly initialized. Then, each data point was classified by computing the distance between that

point and each group center and classifying the point as the group whose center is closest to it. Based on these classified points, the group center is recomputed by taking the mean of all the vectors in the group. These steps were repeated until convergence, or the maximum number of iterations was reached.

Gaussian Mixture Modeling (GMM): It was assumed that each data point is generated from any type of gaussian function with a corresponding probability rather than hard assignment into clusters like k-means. Each distribution has some "responsibility" for generating a particular data point. There is a latent variable γ for each data point that determines which type of gaussian distribution was used.

The data used to create the GMM model has been scaled. The number of components (i.e. number of clusters) was chosen using PCA and finding the "elbow point" of the inertia vs K's graph. K is chosen to be the point where inertia no longer decreases significantly. In our case, K was seven.

Hierarchical clustering: Hierarchical clustering is a very computationally intensive approach. The whole dataset has 687,795 rows, and this clustering technique will not run within a reasonable timeframe using that dataset. After sampling 10,000 rows and running hierarchical clustering, this method gives two clusters. Because of this, hierarchical clustering was not chosen to create the final recommendation system; however, if more computational resources were available, it might be a method that can be attempted in the future.

Clustering Results

Out of the three clustering methods used, only K-Means and GMM could give concrete results. Thus, these two methods were chosen to cluster the original data and group new data into the clusters determined by each method. This is further illustrated in Table 3 in Appendix.

K-Means

The number of clusters chosen for this method is seven (chosen using PCA), and the number of SKUs in each cluster is shown below in Table 1. For example, the SKUs 119323 and 128002 are both the most popular in the South Atlantic region. Also, the brand for 119323 is Z-Cavari, and the brand for 128002 is Phoenix; these are both men's clothing brands. The K-Means clustering algorithm can use the association between departments of SKUs to capture the features explained by the brand and cluster them into cluster 2 (the label is arbitrary).

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
97254	97472	97404	99711	97588	98980	99386

Table 1 of value counts from K-means clustering

GMM

The number of clusters chosen for this method is seven (chosen using PCA and same as K-means), and the number of SKUs in each cluster is shown below in Table 2. For example, using the same SKUs 119323 and 128002, both are the most popular in the South Atlantic region. Also, the brand for 119323 is Z-Cavari, and the brand for 128002 is Phoenix; these are both men's clothing brands. The GMM algorithm can use the association between departments of SKUs to capture the features explained by the brand. Thus, GMM is also able to cluster them into the same group similar to K-Means. This shows that both methods are able to take advantage of the variables created and associations between apartments to cluster products in a similar fashion.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
133623	132599	167897	67623	16404	74236	95413

Table 2 of value counts from GMM clustering

ROI Analysis

Use Cases

Since the business goal is to improve the recommendation system of Dillard's, one must connect the business value of the clustering methods with the recommendation system product itself.

Assumptions

The group envisioned the products to be implemented in 2 use cases, broken down by Dillard's online and offline presence.

The first use case takes place in the physical Dillard's department stores across the nation, where the clustered products are integrated into sales representatives' smart devices (tablet, smartphone, etc), which would allow them to scan the barcode of a customer's items picked for SKU and recommend products from the same cluster, in addition to their knowledge as a sales representative.

The second use case is applied to Dillards.com, which is Dillard's online shopping platform. Upon observation, the group noticed the current recommendation system on Dillards.com is a simple rule-based model entirely based on brand. To improve the model, the group proposes to replace the current system with an algorithm that recommends products in the same cluster as the items in shopping cars, having half of the recommendations with brand filter on, achieving at least the same result as the original brand-based recommendations.

Risk Scenarios

Since the clustering (and all unsupervised learning) does not involve Ground Truth, it is difficult to quantify the exact amount the clustering method improves the recommendation systems by (Lift Rate), if at all. To perform

the ROI analysis, the group leveraged market research data to retrieve baseline conversion rates for each use case and assumed 3 risk scenarios to calculate our ROIs:

a. Bull (optimistic): assumes lift rate = 20%, or that our model outperforms the baseline model by 20% in achieving conversion.
b. Neutral: assumes lift rate = 10%, or our model outperforms the baseline model by 10% in achieving conversion.
c. Bear (pessimistic): lift rate = 1%, or our model outperforms the baseline model by 1% in achieving conversion.

Aware that the proposed recommendation system product requires further software development, the group included 4 front-end developers and 4 full-stack developers in the personnel, with their market wage entering the investment equation.

Results & Impact

The final output of the ROI Analysis (see Figure 9) suggests the proposed new recommendation system based on the group's clustering result can generate approximately $60.4M or 7235% in ROI under the most optimistic assumption, which is extremely large. Under the neutral assumption, the new recommendation system can generate approximately $29.8M or 3618% ROI. Under the most pessimistic assumption (where the clustering model only improves the existing baseline conversion rate by 1%), the new recommendation system can generate approximately $2.2M or 362% ROI. Even under the most pessimistic assumptions, the return on investing in the project is sizable.

Appendix

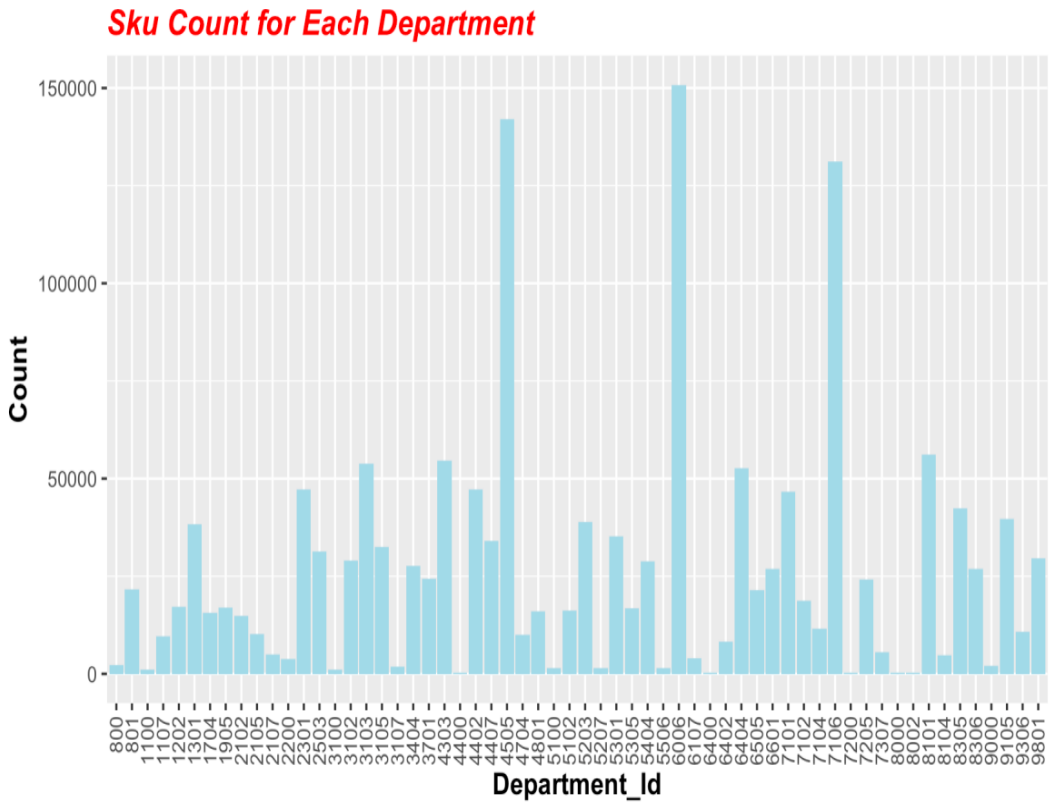

Fig. 1: Distribution for Departments

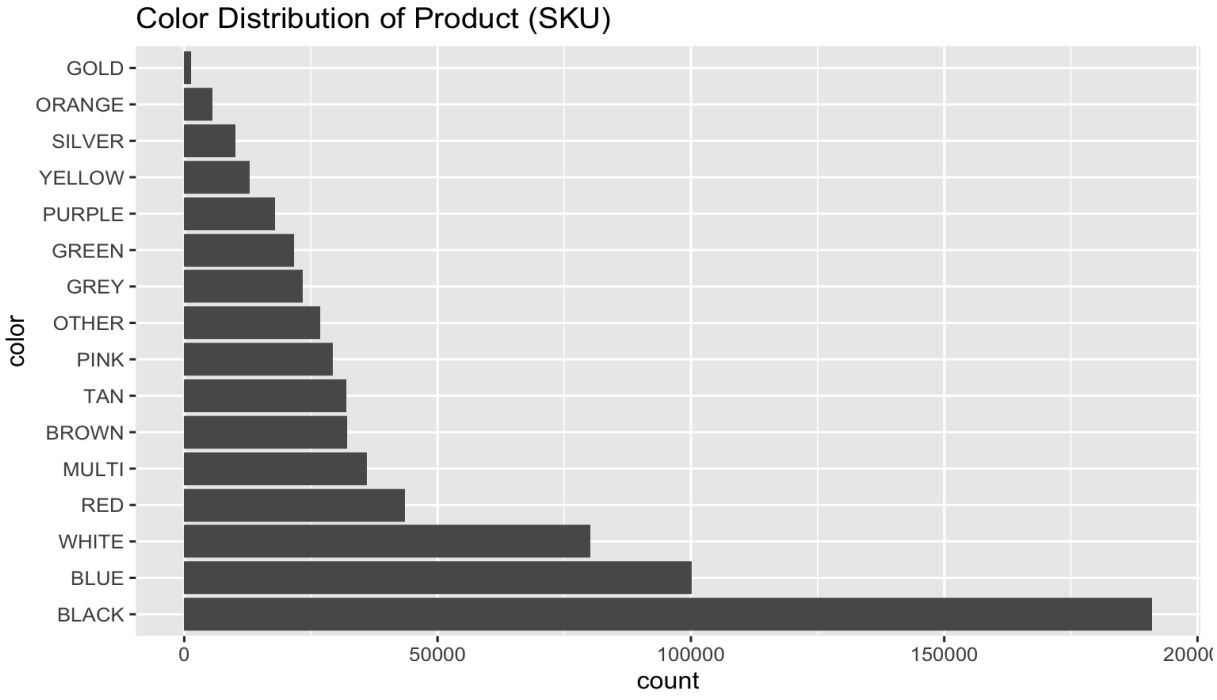

Fig. 2: Distribution of Colors of SKUs

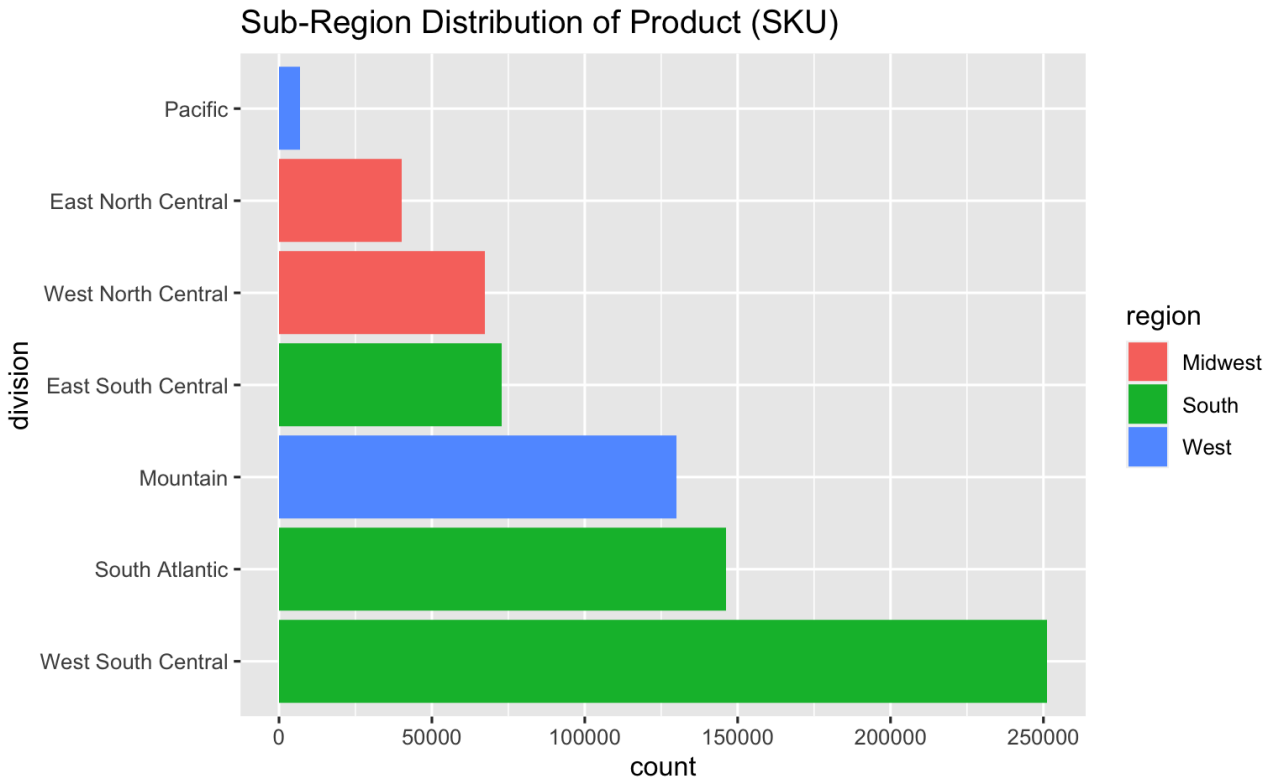

Fig. 3: Distribution of Regions by SKU

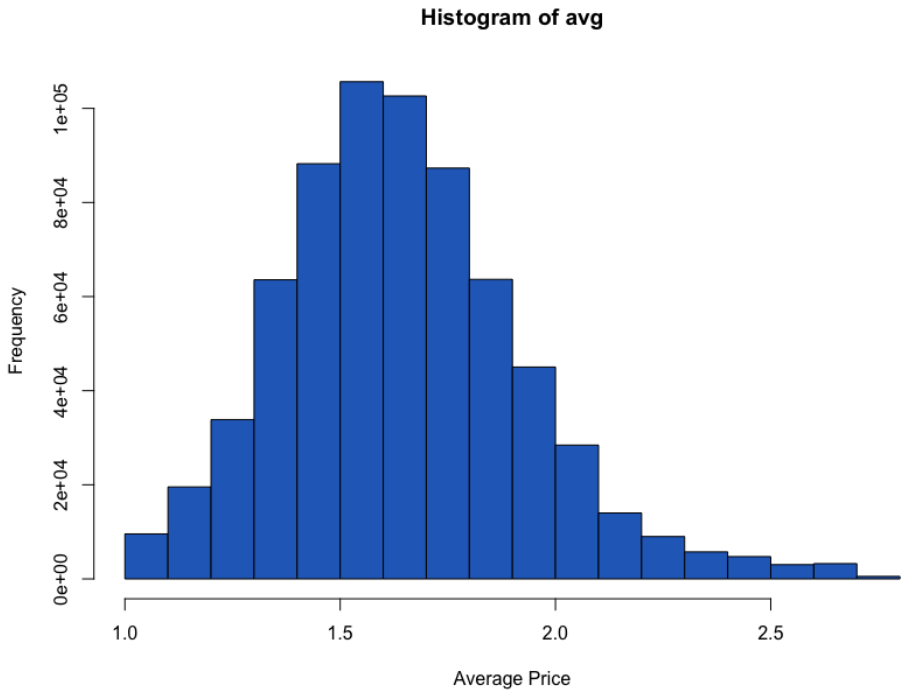

Fig. 4: Histogram of Average Price

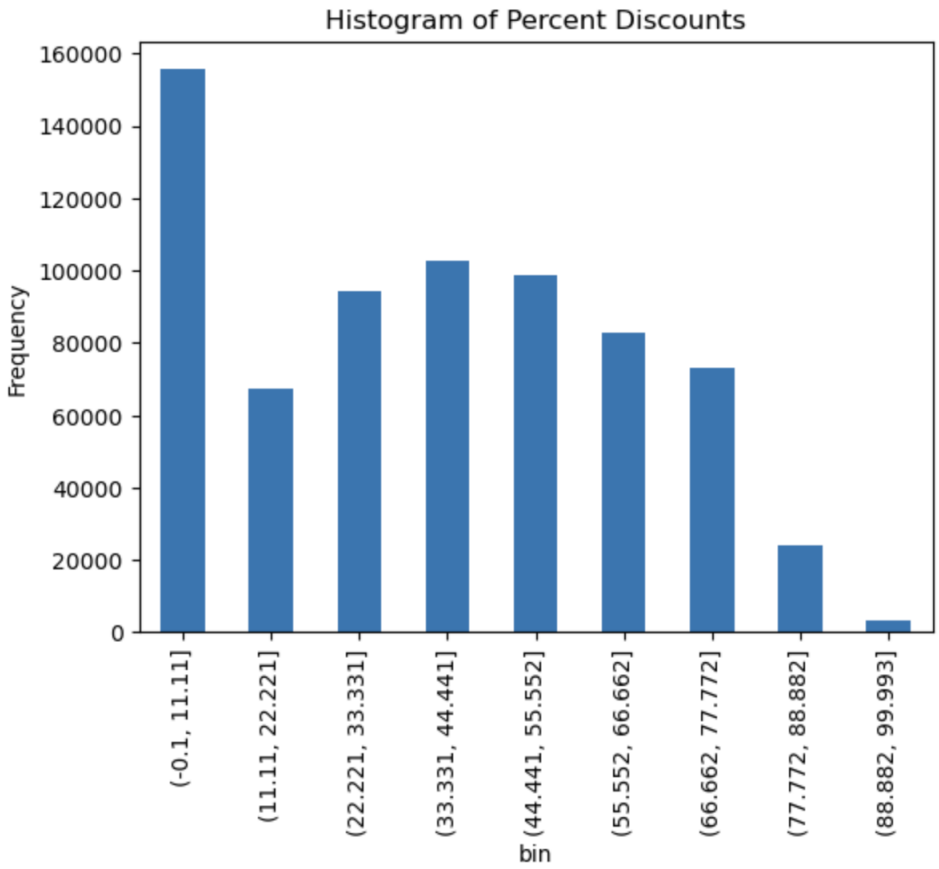

Fig. 5: Histogram of Percent Discounts

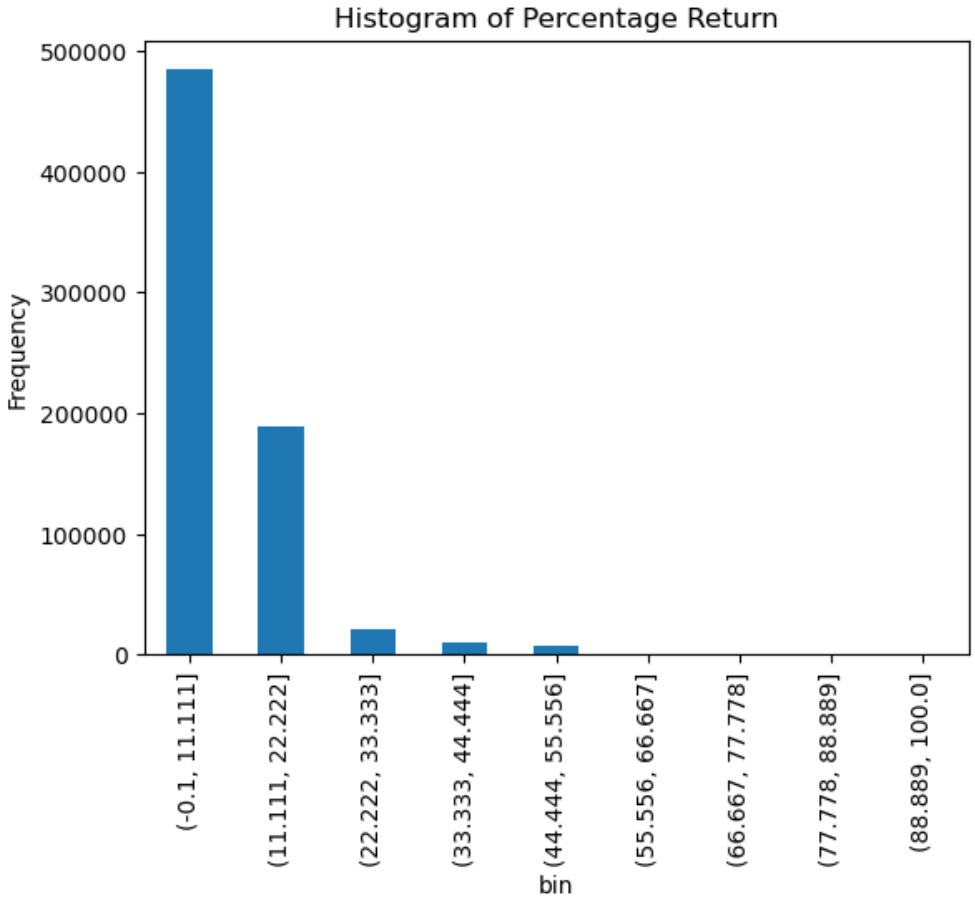

Fig. 6: Histogram of Percentage Return

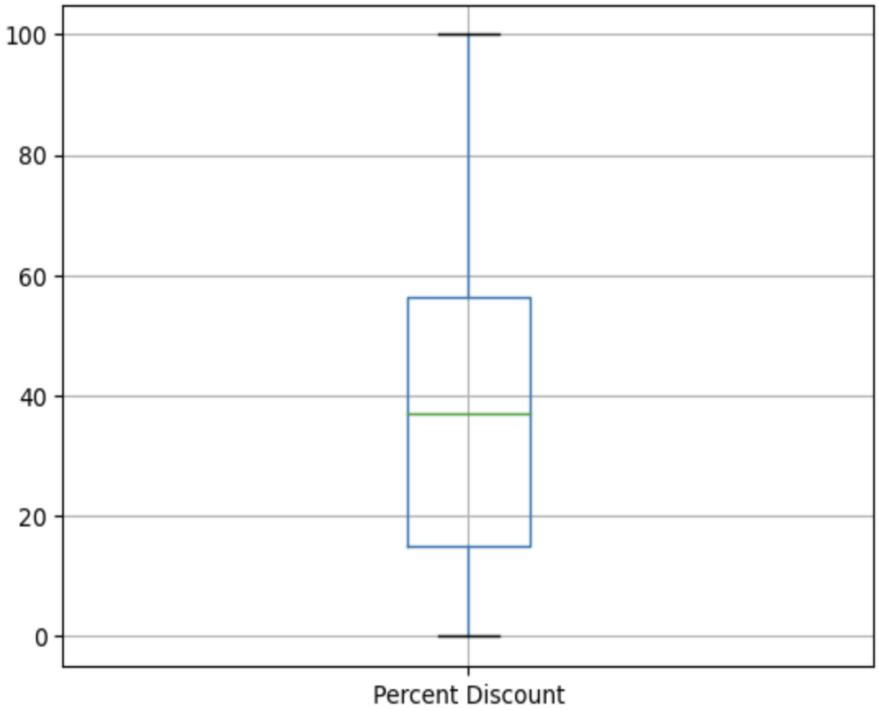

Fig. 9: Boxplot of Percent Discount

SKU	K-means Cluster	GMM Labels
5157585	3	6
4786297	3	6
3558696	7	2
3829286	7	2
870767	2	2
7689462	4	3
2528788	5	0
4192124	7	1
754438	2	6
6876630	1	1
5711256	1	1
9007336	6	2
7664037	4	6
8411572	4	6
2444420	5	1
5044109	3	1
4198166	7	6
2948120	7	6
3181271	7	5
3613505	7	5
5243871	3	2
9890968	6	4
5736285	1	2
9626390	6	0
7629750	4	2
4836932	3	6
4258943	3	3
50316	2	0
2474743	5	6
6988824	1	0

Table 3: Thirty SKUs and Their Respective Clusters by Modeling Method

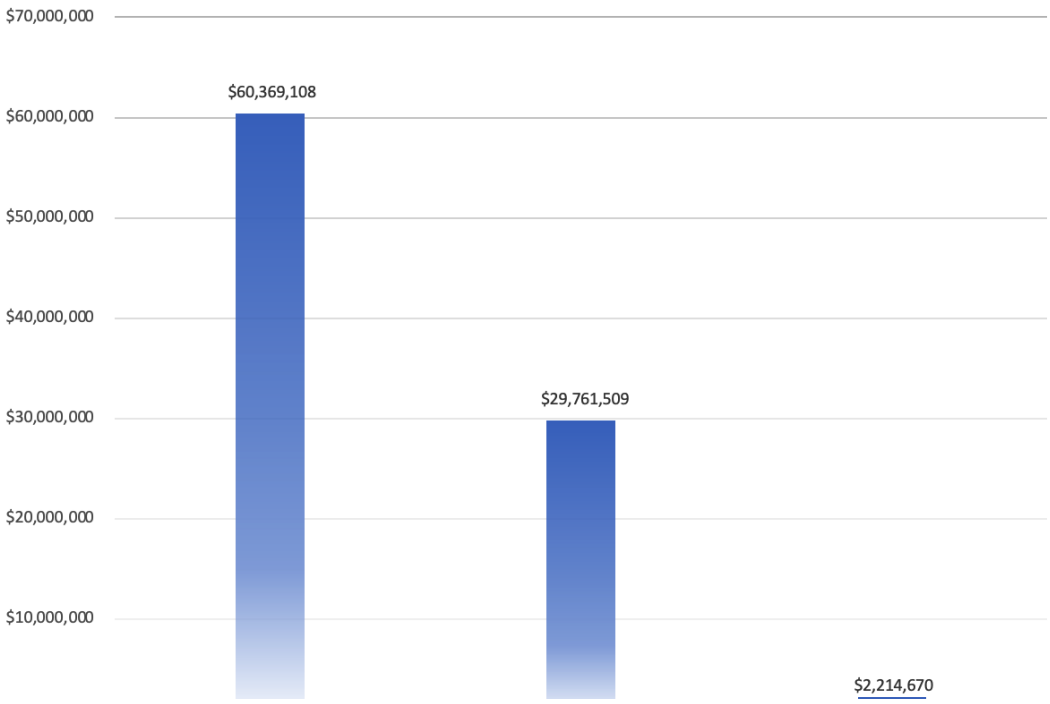

Fig. 10: ROI Analysis Results