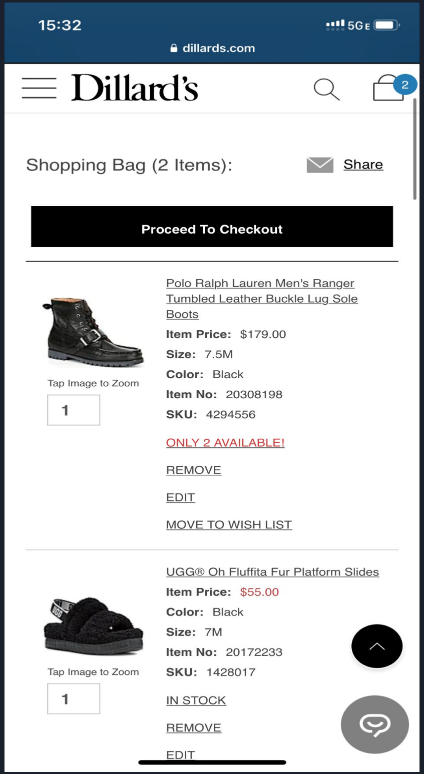
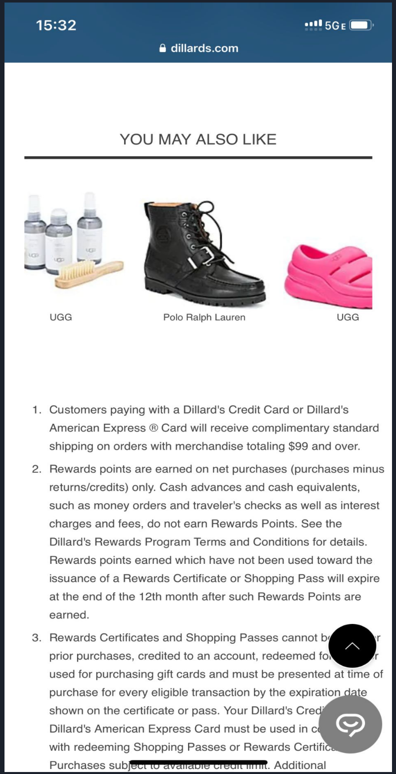
Product (SKU) Clustering for Dillard’s

MSiA 400 Group 1

Motivations





Dillard's App Today:

- Simple and outdated recommendation system

  - Only products from same brands are recommended.

  - Very similar type of products.

  - Barely a personalized experience.

Challenges

- **Large data set**

  - Too many categories for some desired variables, hard for one hot encoding

  - Limited computing sources to handle

- **Primarily offline stores**

  - Not much online information

- **Limited data on customers**

  - No information about customer themselves to make better recommendations

  - Dependent upon proxy information from Dillard's sale data

Solutions

- **Chose similar variables with less categories as substitute**

  - Still capture most features of the original ones.

  - Easier to compute and do one hot encoding.

- **Sale data as proxy for customer likes**

  - Sale data can serve to model the likes of customers as products belongs to the same cluster can serve us recommendations

Methods

- **PCA**:

  - It breaks the features down into principal components such that each component is linearly independent of the other component

  - We stop when the inertia by clusters elbows. We used 7 clusters in our analysis

- **K-Means**:

  - Select the number of 7 groups with corresponding randomly initialized center points.

  - Classify each data point by computing its distance and repeat the steps till it converges or we reach the maximum number of iterations.

Methods

- **<u>Hierarchical Clustering</u>**:

  - Computationally expensive. Cannot handle 680,000 rows. We ran it on 10000 rows and it gave 2 clusters

  - Can explore it if more computational resources were available

- **<u>Gaussian Mixture Modeling</u>**:

  - Assumes each data point from a gaussian distribution. There is a latent variable γ for each data point that determines which type of gaussian distribution was used.

  - Similar to k-means, we assumed 7 clusters.

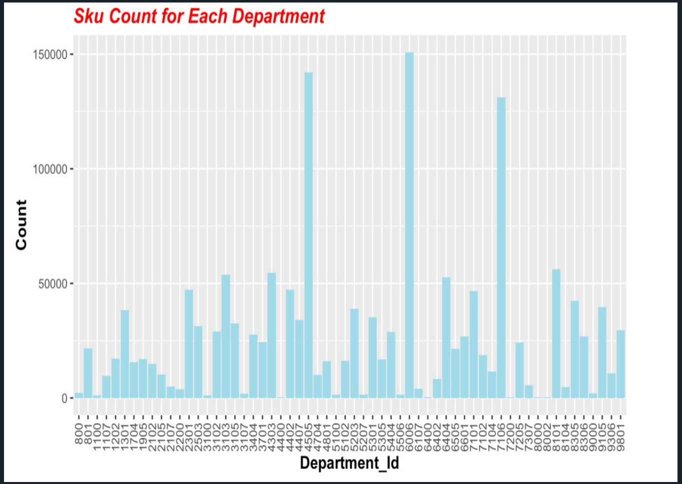
Features of Choice

- **Brand**:

  - Good predictor as consumers tend to stick with certain brands.

  - But over 2000 factors, tough for one hot encoding.

  - "**Department**" (60 columns) used instead to capture the features of "Brand".



Top Three:

	id [PK] bigint	dept_name character varying	count bigint
1	4505	POLOMEN	142108
2	6006	INVEST	150815
3	7106	BRIOSO	131106

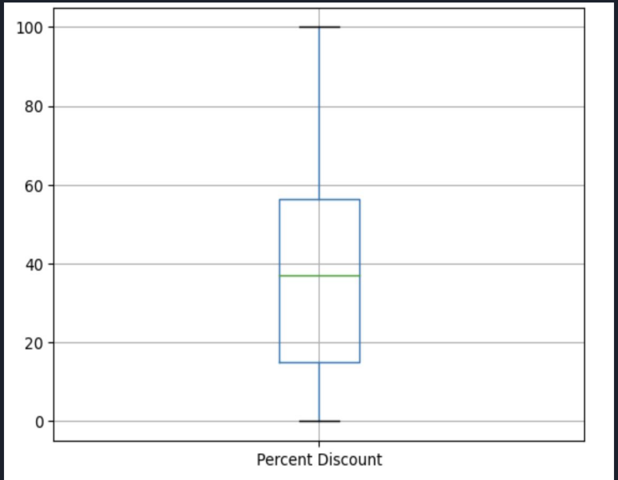
Feature of Choice

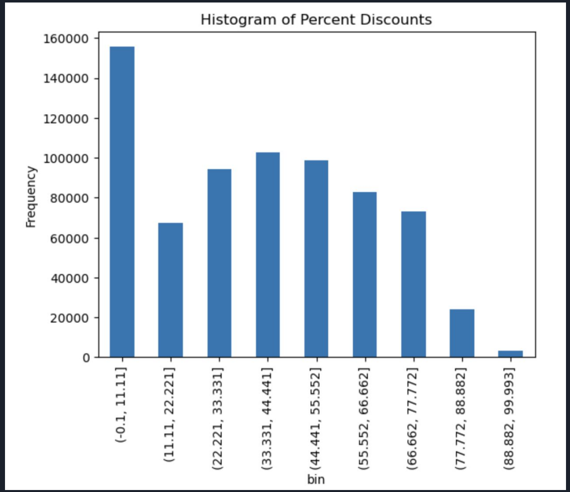
- When it comes to any form of business, money related topic never goes away:

  - **<u>Avg Price</u>** : average price of each sku.

  - **<u>Percent Discount</u>**: the average discount percent of an item.

  - **<u>Percent Return</u>**: the average rate of return of an item.





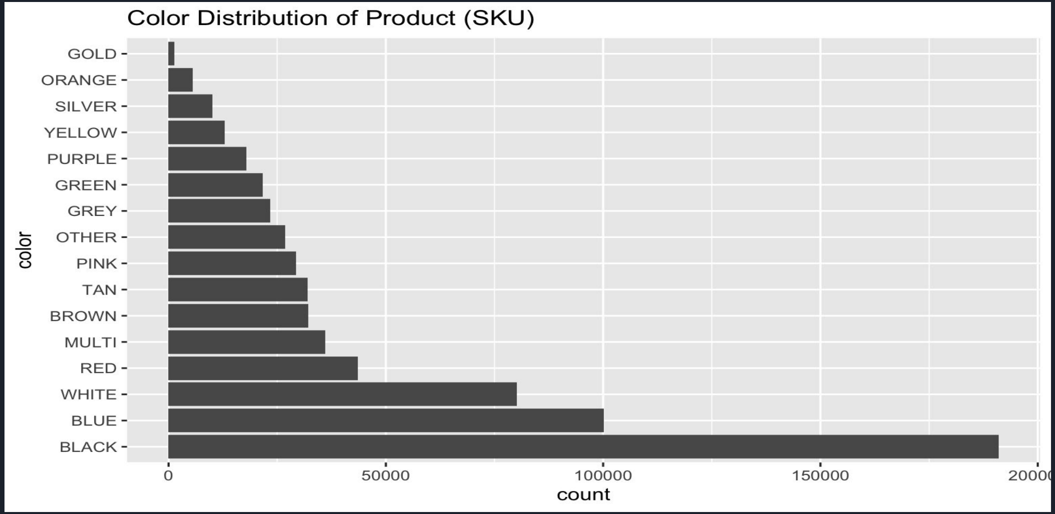
Feature of Choice

- **<u>Color</u>**:

  - Over 200 used colors in SKU, too many for one hot encoding.

  - Reduced to 17 most-common color groups.

  - Each sku assigned to a color.

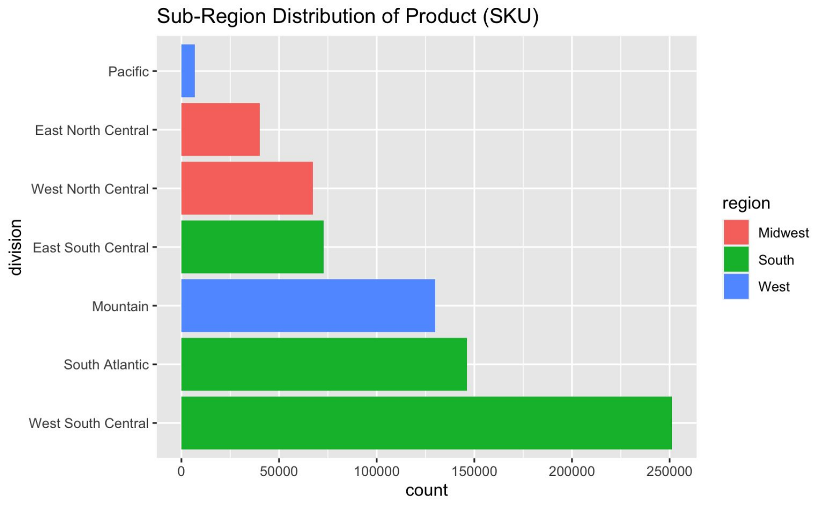


Feature of Choice

- **<u>Location</u>**:

  - People in different regions might have entirely different preferences.

  - Used a mapping table from US census to get the "Region"



Results

- Randomly select 25 skus and perform two clustering methods (K-means and Gmm).

- The colored ones are the pairs that both algorithms put into the same cluster.

sku	K-means Cluster	Gmm Labels
5157585	3	6
4786297	3	6
3558696	7	2
3829286	7	2
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
50316	2	0
2474743	5	6
6988824	1	0

Use Case 1

Physical Shopping Site

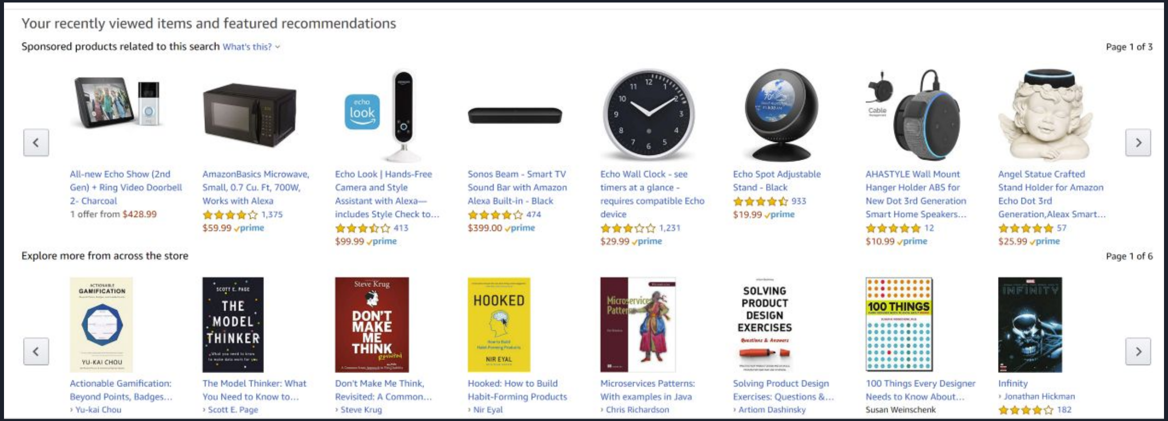
- Scan QR code to get sku when customer takes an item to fitting room
- System automatically recommends items from corresponding cluster
- Record customer's decision for further data collection.

Use Case 2

- Current systems recommendation system is a rule based one which recommends same brand to the customer

- We recommend to replace it with clustering based system so that products in the same cluster as the items in the cart are recommended to the user

- Half of the results will use same brand so as to achieve the similar result as the current baseline

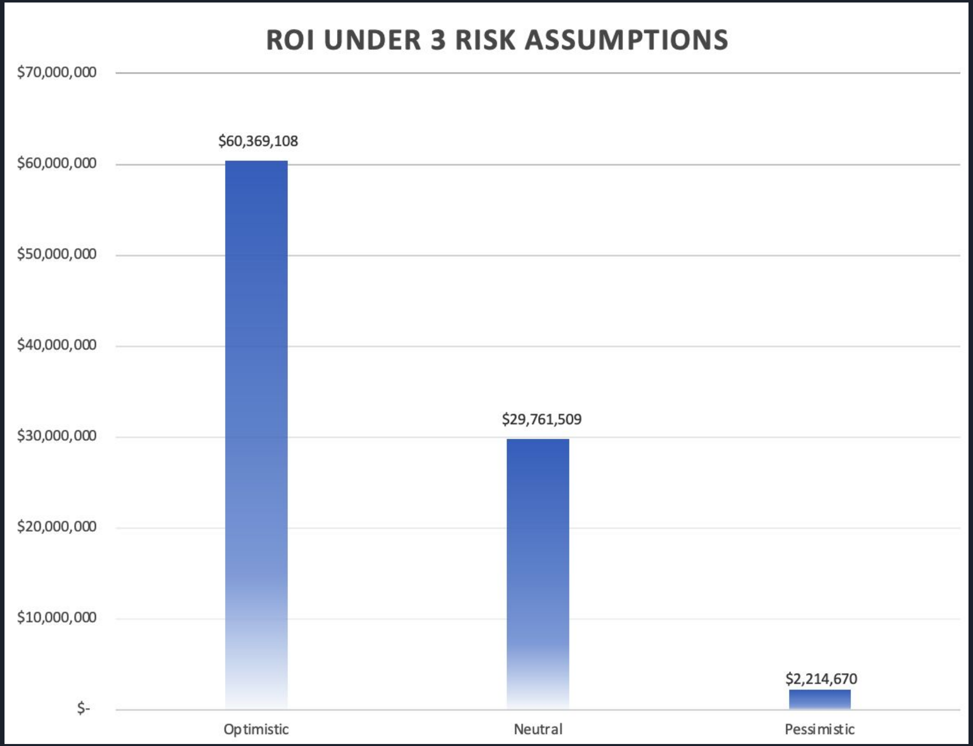


ROI

- Without A/B Testing results, we can only assume the lift percentage of implementing our model

- ROI Analysis constructed under 3 sets of assumptions:
    - Bull case: lift = 20%
    - Neutral case: lift = 10%
    - Bear case: lift = 1%

- Even under the bear case, the ROI amount is $2.2million dollars; ROI rate is 362%.



Given More Time/Resources/Data....

If more time & resources are given:

- Use brand as feature.

  - More accurate cluster based on over 2000 brands.

- Try hierarchical clustering on entire dataset

  - Obtain the optimal number of clusters, human intervention not required.

  - Clear visualization from dendrograms, practical and easy to understand.

- Get more and more recent data for training

  - Never a bad thing to have more data.