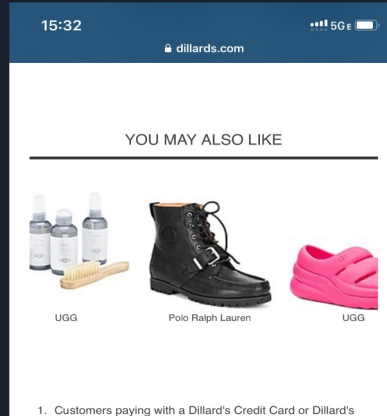
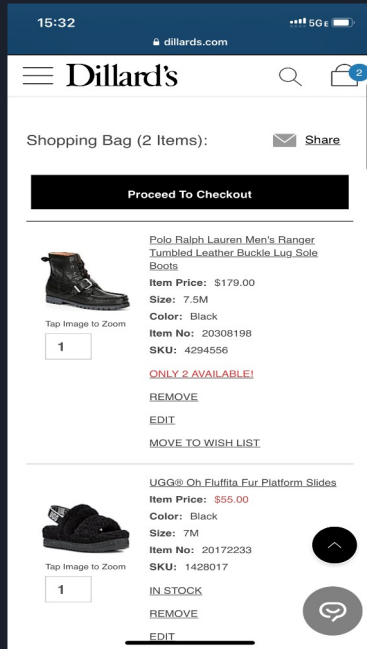





# Product (SKU) Clustering for Dillard's

# Motivations: Dillards.com Today



- Completely **rule-based** recommendation system
- Only products from **same brands** as items in cart are recommended
- Very **similar** type of products
- Barely a personalized experience



# Challenges: Why is making recommendations difficult for Dillard's?



➤ Large data sets

- Too much transactional data from various system databases



➤ Primarily offline stores

- Limited access to customer data comparing to other ecommerce platforms
- Difficult to track revealed preferences at customer level

# Solutions

## ➤ Utilize POS data and master data on product (SKU)

- Leverage relational database to join and perform analysis
- Use SKU features to place products into clusters
- Without customer-level purchase history data, recommend similar items to current transaction (at checkout or fitting stage)

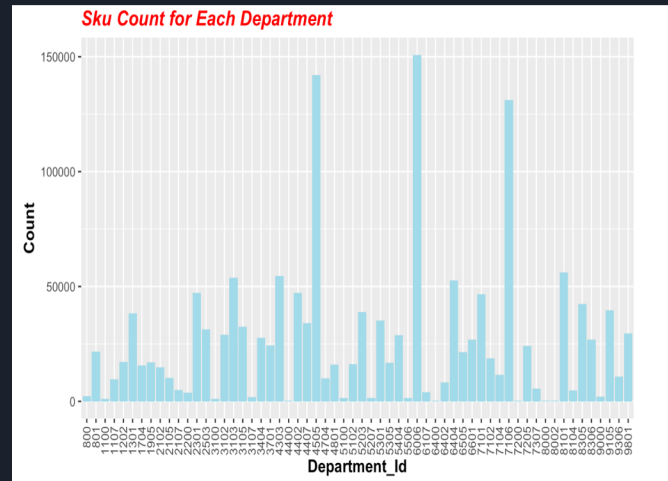


# Features of Choice

- 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"

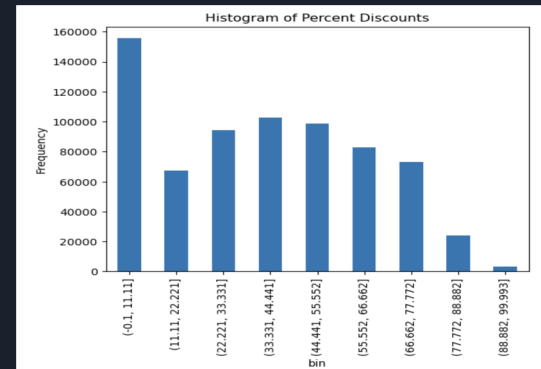
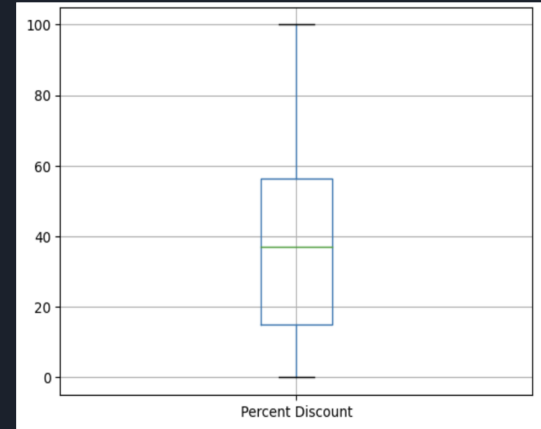
## Top Three:

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



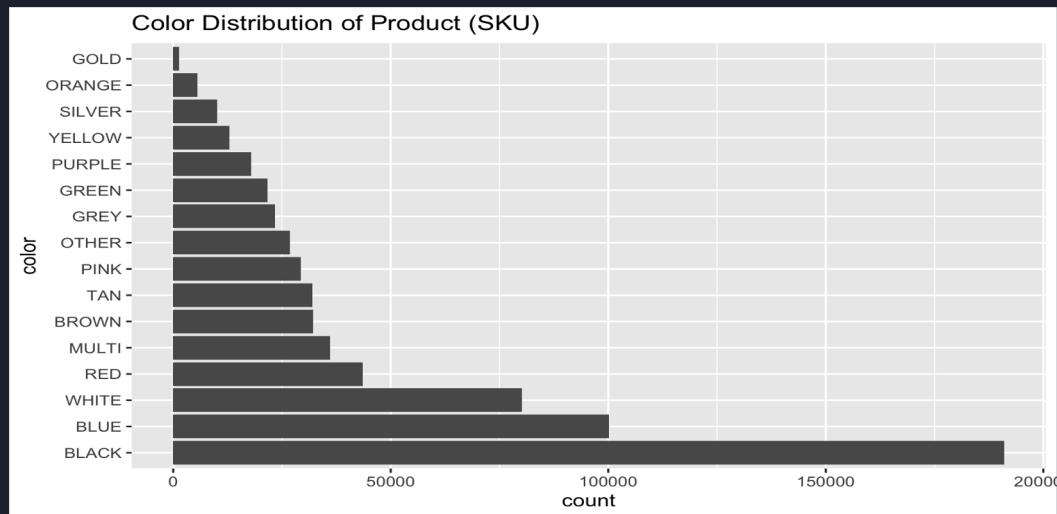
# Features of Choice

- When it comes to any form of business, money related topic never goes away:
  - Avg Price : average price of each sku
  - Percent Discount: the average discount percent of an item
  - Percent Return: the average rate of return of an item



# Features of Choice

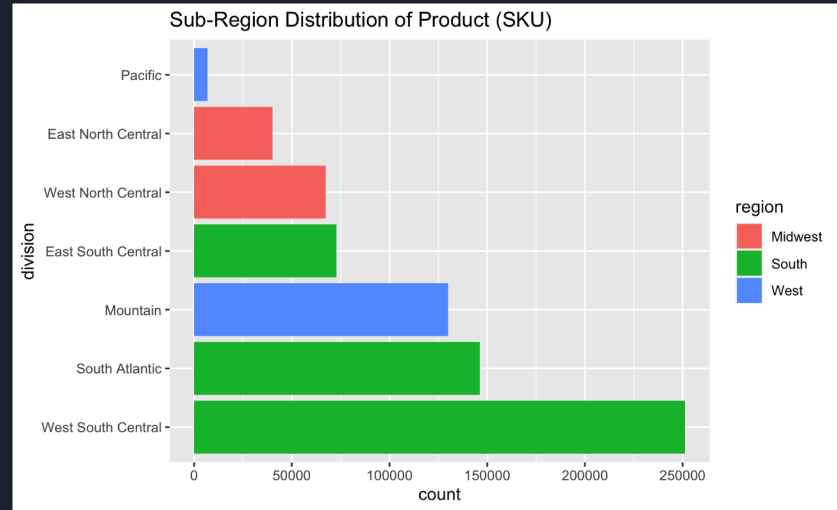
- Color:
  - Over 200 used colors in SKU, after one hot encoding too computationally intensive
  - Reduced to 17 most-common color groups
  - Each sku assigned to a color



# Features of Choice

- Location:

- People in different regions might have entirely different preferences
- For each product, find the state that the item was sold the most
- Used a mapping table from US census to get the "Region"





# Cluster Technique - 1

## Principal Component Analysis



- ❑ Breaks the features down into principal components such that each component is linearly independent of the other component
- ❑ Can explore it if more computational resources were available

## K-Means Clustering



- ❑ 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 - 2

## Gaussian Mixture Modeling



- ❑ Assumes each data point from a gaussian distribution. There is a latent variable  $\gamma$  for each data point that determines which type of gaussian distribution was used.
- ❑ Similar to k-means, we assumed 7 clusters.

## Hierarchical Clustering



- ❑ Computationally expensive. Cannot handle 680,000 rows. We ran it on 10000 rows and it gave 2 clusters
- ❑ We stop when the inertia by clusters elbows. We used 7 clusters in our analysis

# Results - Illustrated

- 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 Department Stores

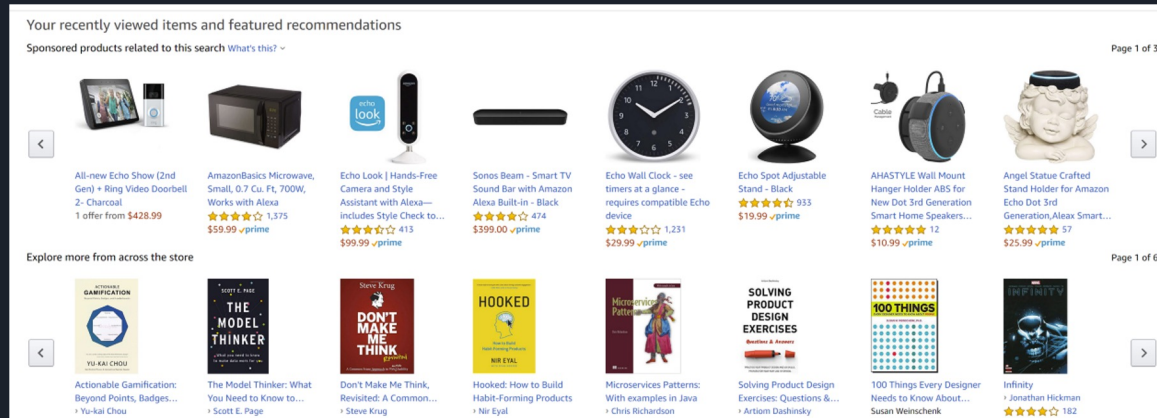
- Sales reps scan QR code to get sku when customer takes an item to fitting room or checkout line
- System automatically recommends items from corresponding cluster
- Sales reps select items from clusters to recommend based on knowledge
- Record customer's decision for further data collection



# Use Case 2

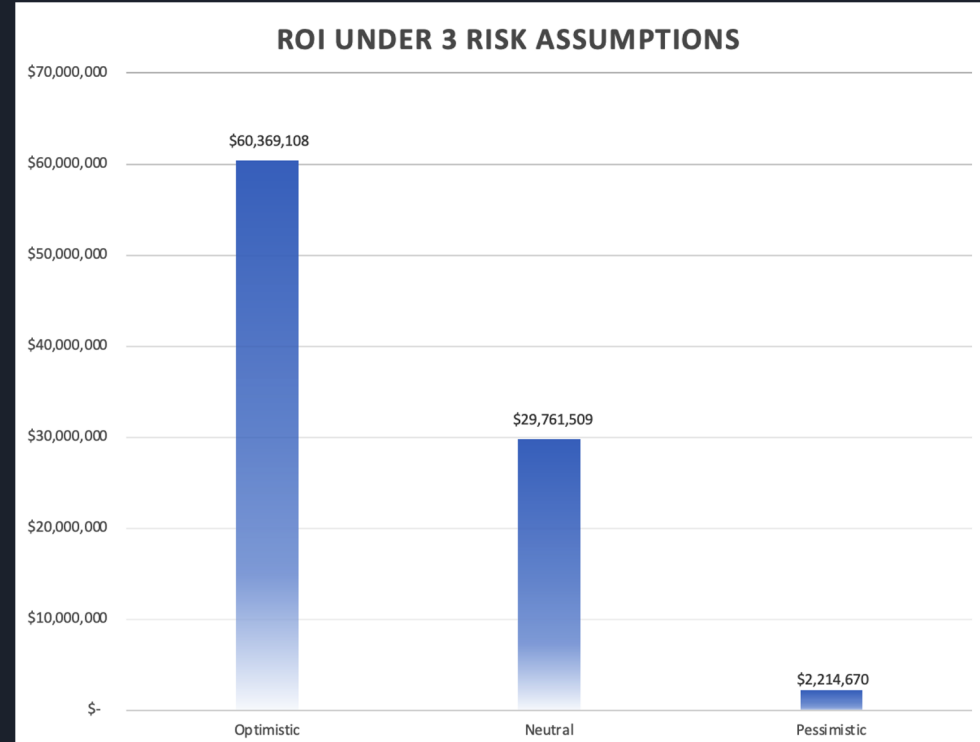
## Dillards.com Online Store

- Replace current recommendation system 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 filter by the same brand to achieve at least similar success 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