Data Science for a Traditional Retail Company

June 10, 2021

ROADMAP

PLAN

Problem Statement
Success Metrics
Data Collection

EXPLORE

Data Cleaning
Data Exploration
Data Visualization
Feature Engineering

MODEL

Clustering
Recommender System
Performance Evaluation

SUMMARY

Conclusion
Recommendation

Learning from experiences of other organizations

Getting it right from the start

Importance of having a clear objective

Be realistic

Don’t take shortcuts

Putting effort into Data Cleaning

Reaping the benefits subsequently

Support at all levels

Getting buy-in across the organization

Self-driven and sustainable (culture)

Setting the context

- Paints the scenario of a traditional retail company
- Transition into a more data-driven organization
- Pilot Data Science team
- Achieve tangible results for management buy-in

Problem Statement with the marketing team:

Despite positive customer reviews gathered by the marketing team, this has not led to an increase in customer growth, it could be attributed towards a shift in data adoption by other retail companies.

- Understanding the customers' profile
- Developing a clustering model using K-Means that allows better targeted marketing campaigns
- Building a recommender system that recommends retail items to upsell, cross-sell, or even lead to product discovery for the customers

Entries in retail dataset

- Invoice: Invoice Number
- StockCode: Product (item) code
- Description: Product (item) name
- Quantity: The quantities of each product (item) per transaction.
- InvoiceDate: Invoice Date and Time when the transaction was generated.
- Price: Unit Price: product price
- Customer ID: Customer number
- Country: Country name, the name of the country where the customer resides.

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom

Exploratory Data Analysis

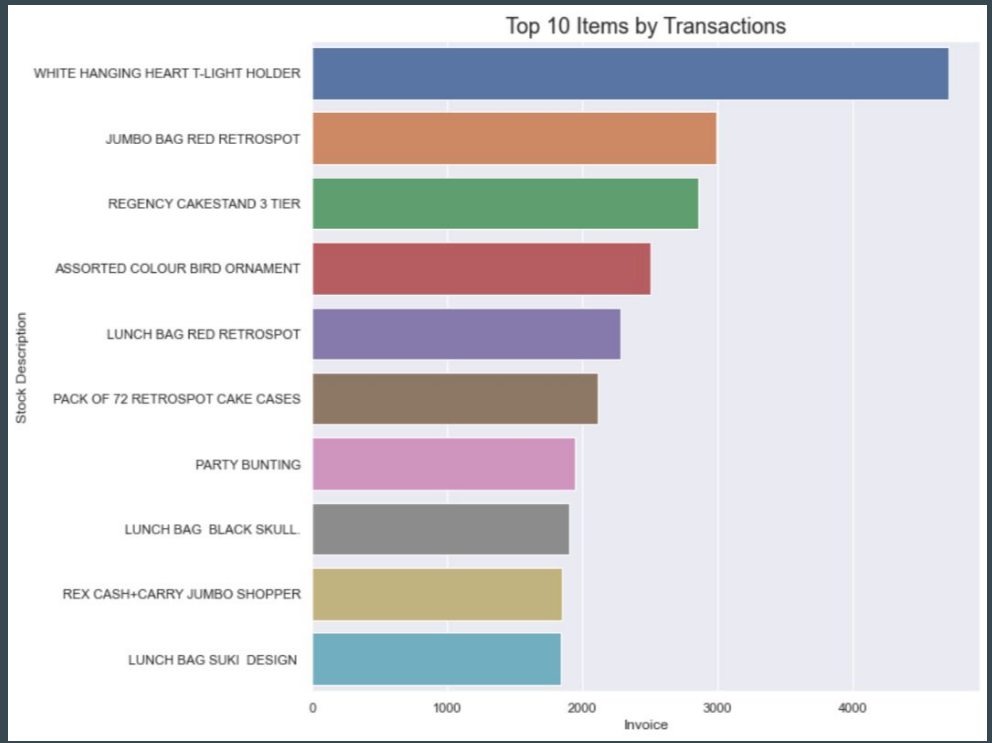
Top Items

Findings

White Hanging Heart T-Light Holder

Jumbo Bag Red Retrospot

Regency Cakestand 3 Tier

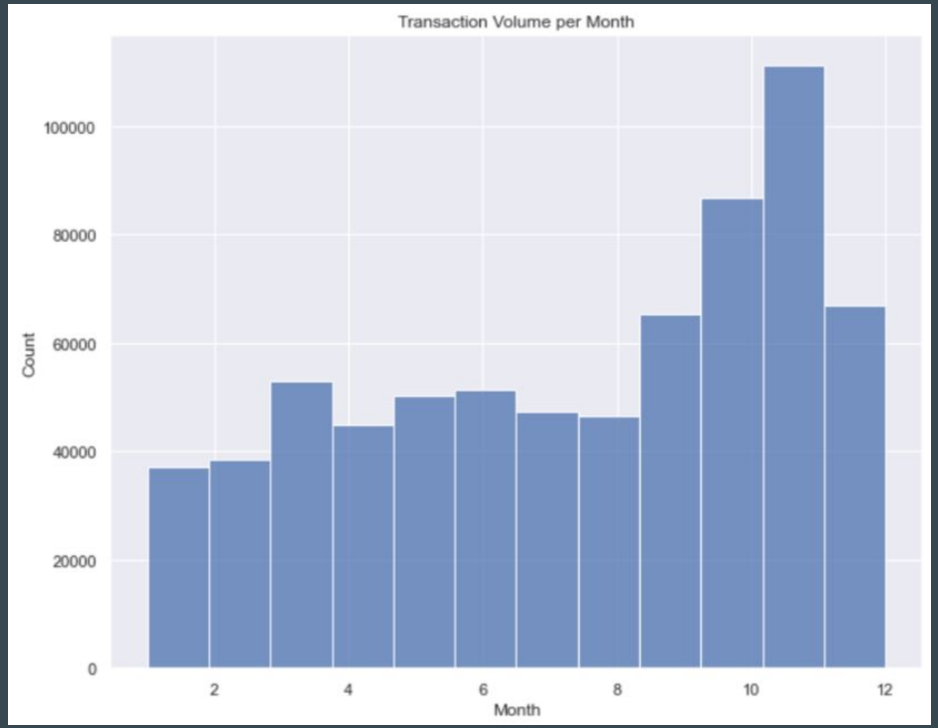


Transaction Volume

Findings

Clearly this is a retail dataset.

Festive season (Q4) has higher volumes

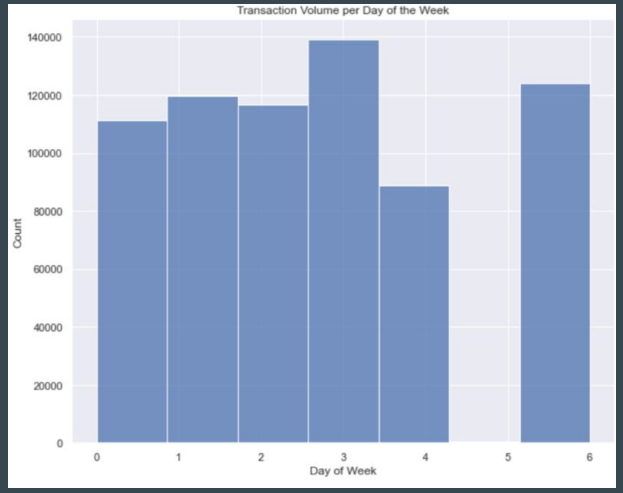


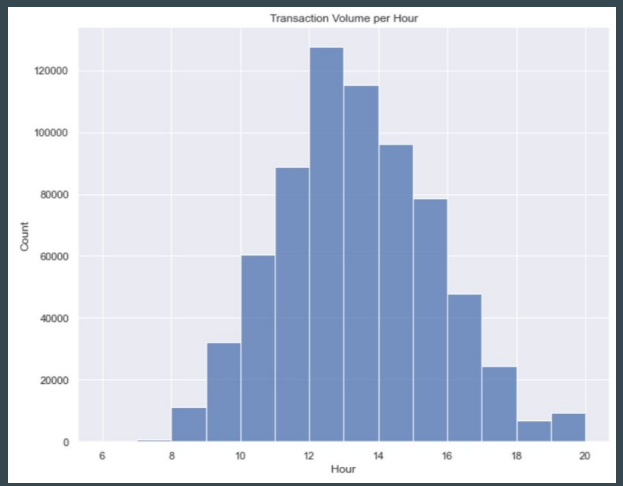
Transaction Volume

Findings

Closed on Saturday

More crowd during lunch hour





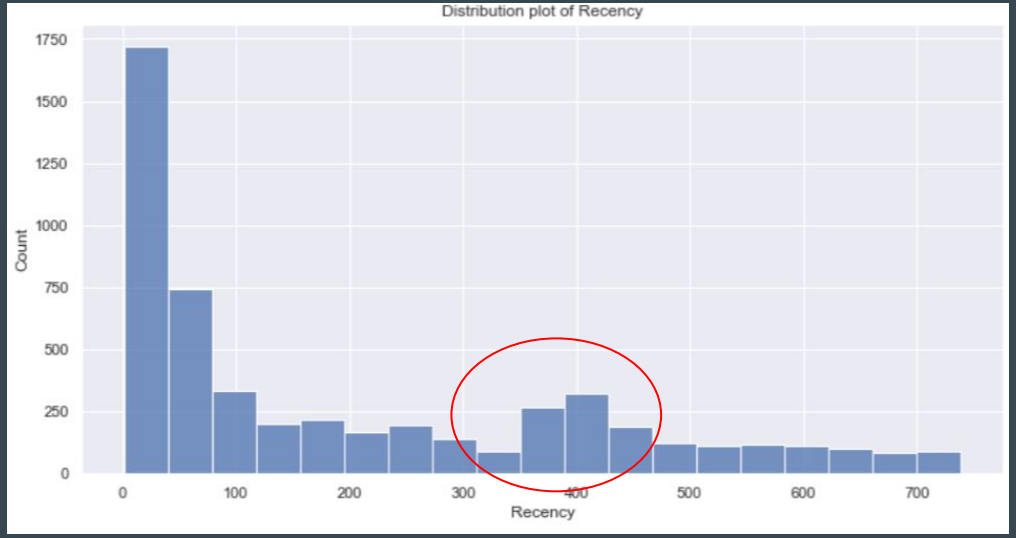
RFM Analysis

RFM - Recency - How Recent?

Findings

Most of the customers are still rather active (<100 days)

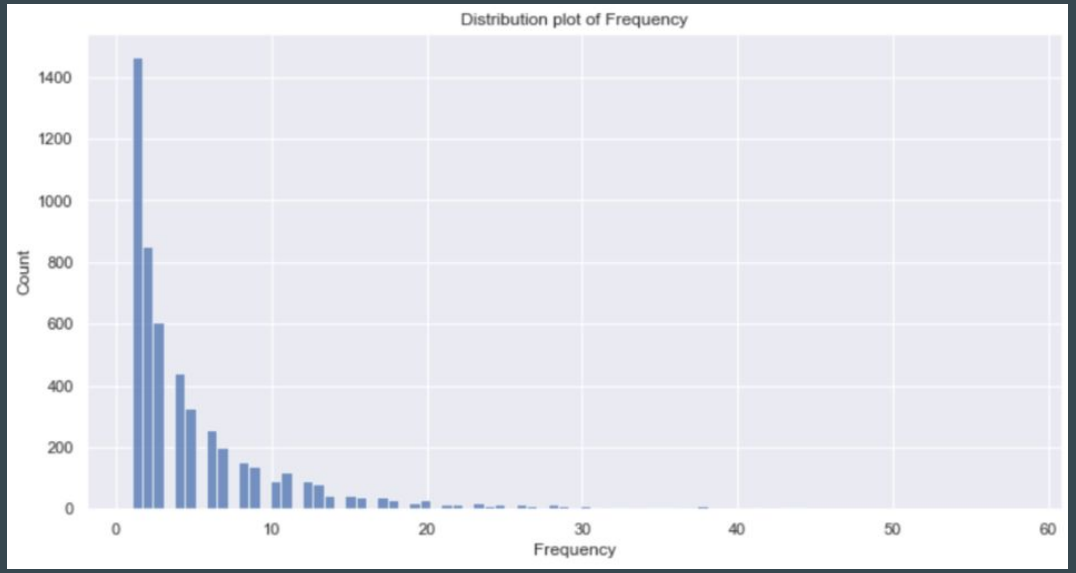
At recency between 350 to 450 days, there is a rather significant number of customers.



RFM - Frequency - How often?

Findings

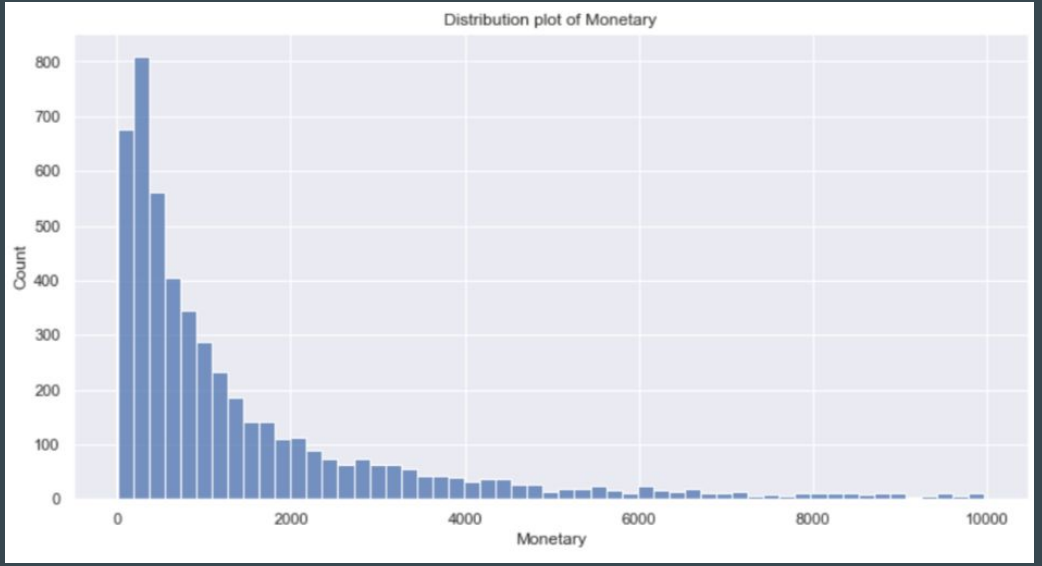
Most of the customers bought less than 3 times.



RFM - Monetary - How much?

Findings

Most of the customers spent less than $2,000

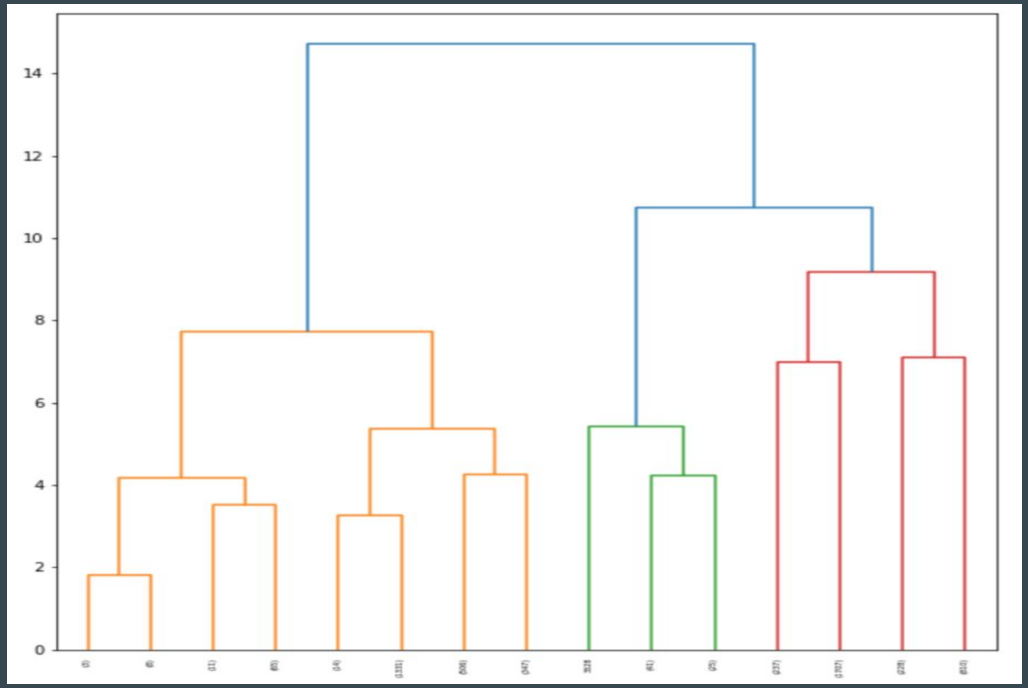


RFM Analysis with K-Means Clustering

K-Means Clustering - Dendrogram

Findings

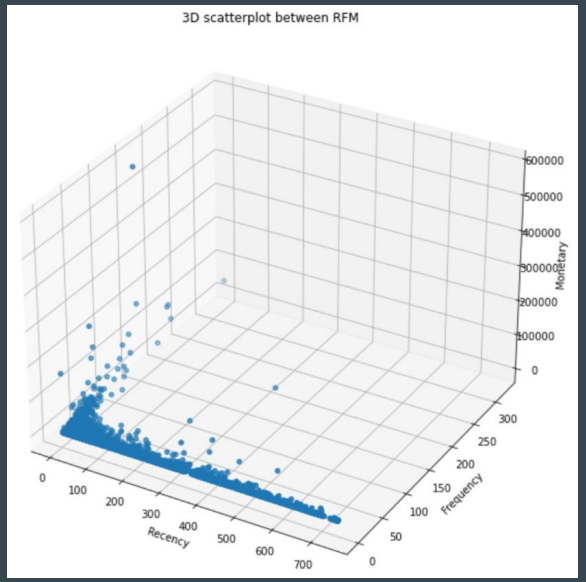
Mainly split the data into 3 (or 4)

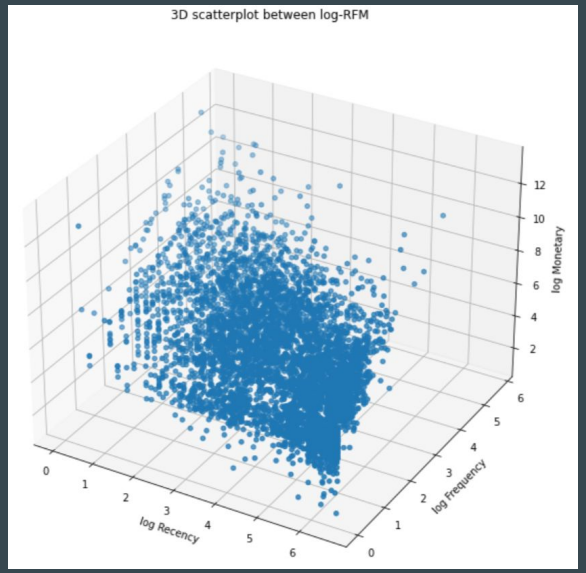


K-Means Clustering - Before vs After transformation

Findings

Data transformation to help us visualize the data better

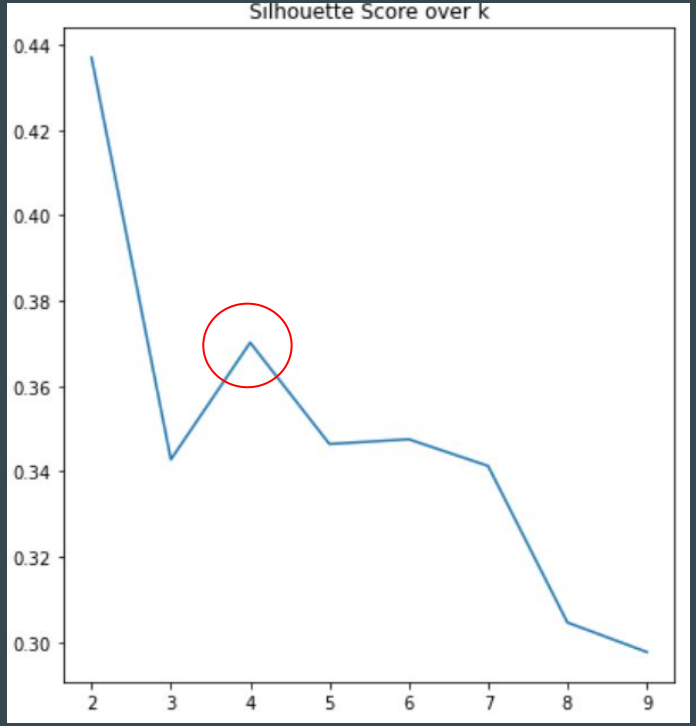




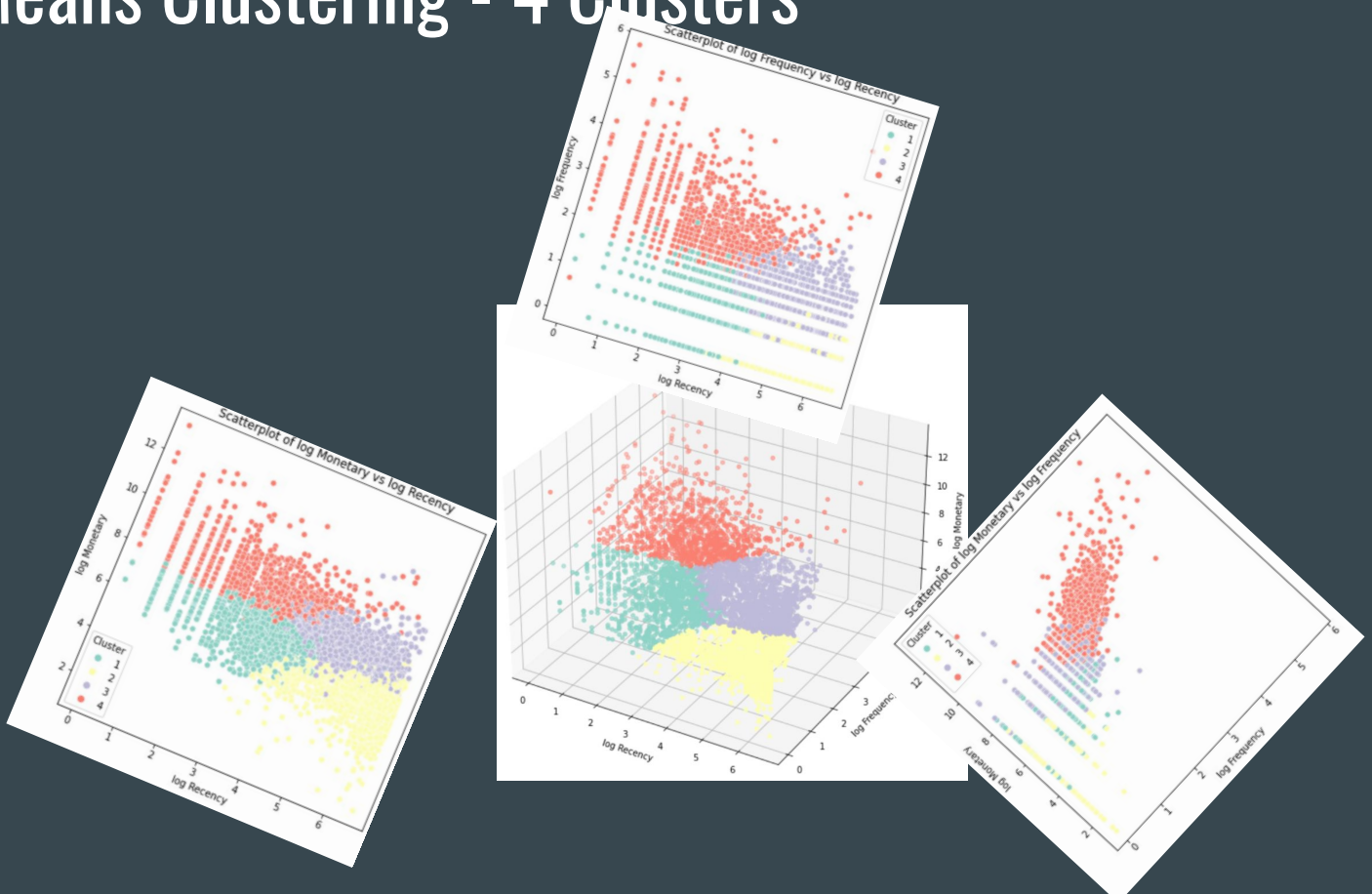
K-Means Clustering - Silhouette Score over Cluster

Findings

Cluster = 4 performs better than Cluster = 3.



K-Means Clustering - 4 Clusters



K-Means Clustering - 4 Clusters

Findings

Cluster 1 (Green): New Customers?

- Low R, Low F and Low M

Cluster 2 (Yellow): Lost Customers?

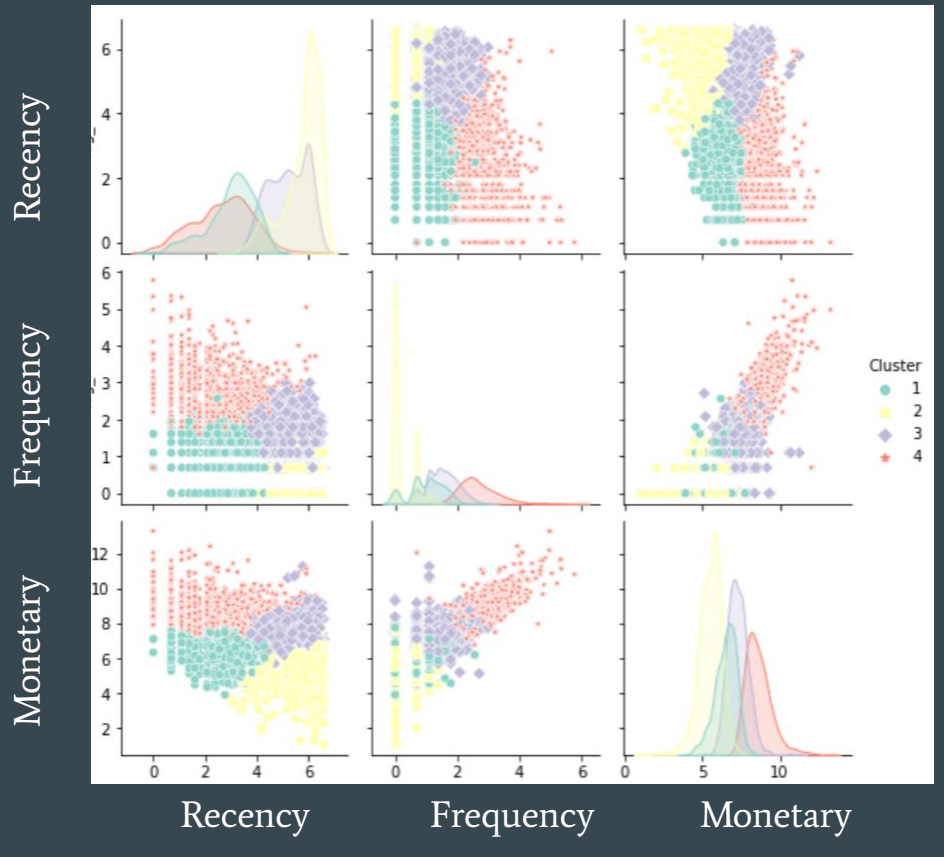
- High R, Low F, Low M

Cluster 3 (Purple): Lost Customers?

- High R, Decent F, Decent M

Cluster 4 (Red): Loyal Customers

- Low R, High F, High M

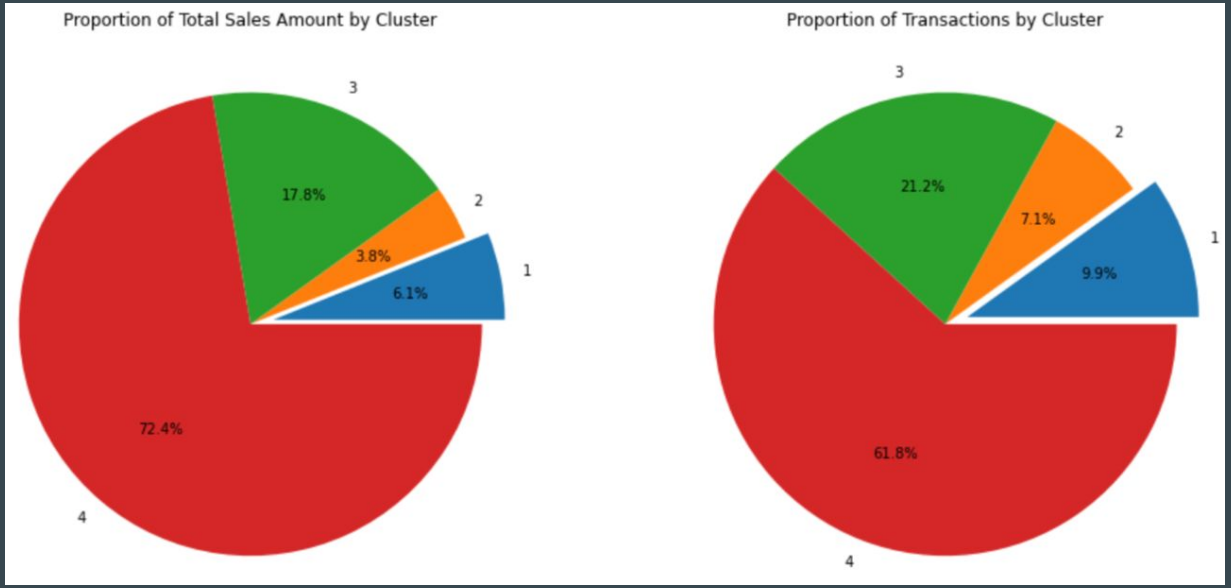


K-Means Clustering - Proportion

Findings

Cluster 4 -> ~70% of Sales Amount, and ~60% of transactions

Cluster 4 > 1 > 3 >2



Recommender System

Recommender System - User-Based Collab

Steps/Findings

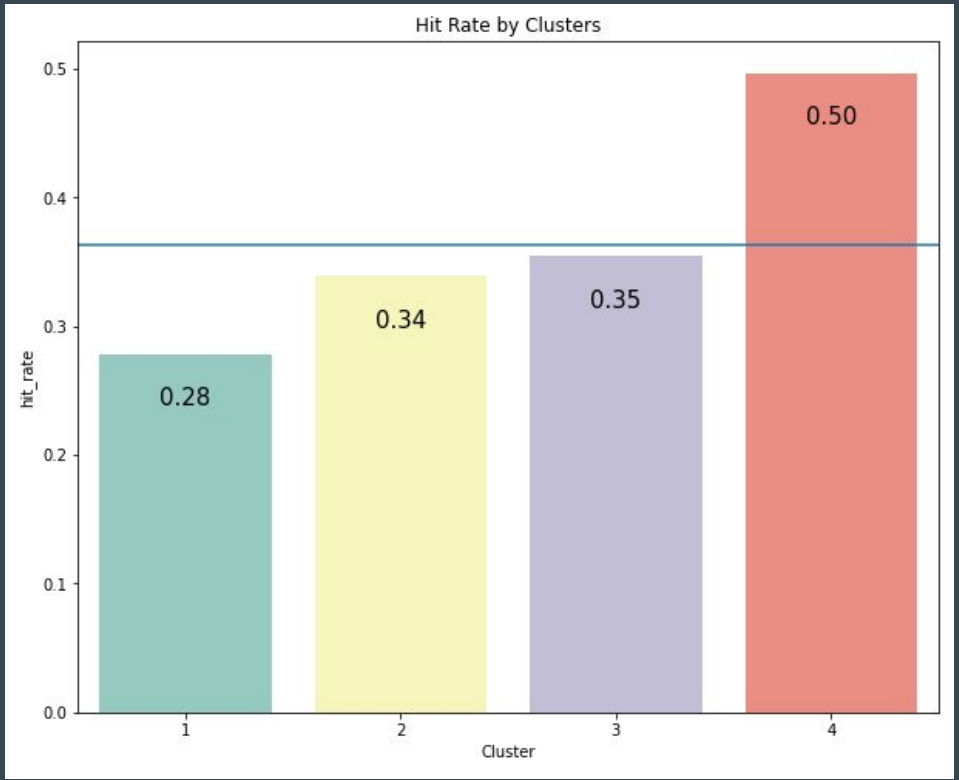
Split the customers into 4 clusters.

Performed collaborative filtering

Recommended top-10 items for each user based on their clusters.

Evaluated using Hit Rate (in-house)

If the user purchased 1 of the top 10, then it is considered as a “hit”



Recommender System - Example

StockCode	Stock Description	Hit
85123A	WHITE HANGING HEART T-LIGHT HOLDER	1
82494L	WOODEN FRAME ANTIQUE WHITE	1
82482	WOODEN PICTURE FRAME WHITE FINISH	1
21754	HOME BUILDING BLOCK WORD	1
21755	LOVE BUILDING BLOCK WORD	0
82486	WOOD S/3 CABINET ANT WHITE FINISH	1
72741	GRAND CHOCOLATECANDLE	0
22457	NATURAL SLATE HEART CHALKBOARD	0
82483	WOOD 2 DRAWER CABINET WHITE FINISH	0
21135	VICTORIAN METAL POSTCARD SPRING	0

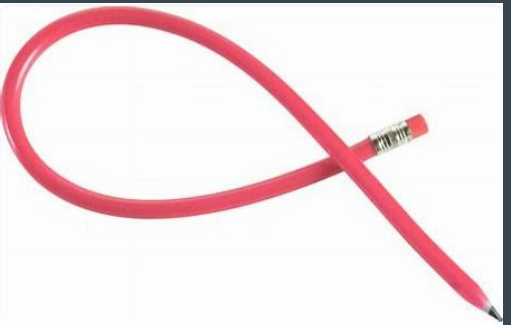
Recommender System - Item-Based Collab

Steps/Findings

Recommended top-10 items for each item

Evaluated using Hit Rate (in-house)

If the the same receipt has both the item and 1 of the recommended top 10 items, then it is considered as a “hit”



Recommender for Item 10109 - BENDY COLOUR PENCILS

	StockCode	Stock Description
0	16215	FUNKY GIRLZ MAGNETIC TO DO LIST
1	16245A	PINK MINI STATIONERY SET W CASE
2	81953B	ROUND BLUE CLOCK WITH SUCKER
3	81953P	ROUND ARTICULATED PINK CLOCK W/SUCK
4	23185	FRENCH STYLE STORAGE JAR JAM
5	84455	SET OF 3 RABBIT CARROTS EASTER
6	47552A	DOTS IRONING BOARD COVER
7	84925C	FAIRY CAKES WALL THERMOMETER
8	84340	LARGE FIBRE OPTIC CHRISTMAS TREE
9	20673	STRAWBERRIES PRINT BOWL

Conclusion and Recommendation

- Clear business objective has to be set right from the start.
  - Models can be enhanced along the way.
- Stakeholder support and commitment to implement change is important.
  - Make realistic data strategy/roadmap
- Diverse expertise
  - To gain more domain knowledge/useful insight so that the modeling process can be catered accordingly, especially so for unsupervised machine learning.
- A/B testing could be performed to further evaluate the recommender system.
- More data could be gathered to use for Clustering