Data Science for a Traditional Retail Company

Harnish Shah – Data Scientist
April 2024

01

Introduction & Plan

Setting the Plot

Our company

- A Traditional Retail Company
- Embarking with the new data driven approach in retail world of E-Commerce

Data Involvement

- Transition to more Data – driven E-Commerce Retail
- A Pilot Data Science team to help & achieve evident results to get management buy-in

A Roadmap

01
PLAN
Problem Statement
Success Metrics
Data Collection

02
EXPLORATION
Data Cleaning
EDA
Feature Engineering

03
MODELING
Supervised Models
Performance

04
MODELING
Clusters
Recommender System
Performance Verification

05
SEASONALITY
Time-Series
Performance

06
SUMMARY
Conclusion
Recommendation

A Realization! – from previous projects

Start in Right Way

- Have a Clear Objective
- Take Realistic Approach

Avoid Shortcuts

- Devoting time to data cleaning
- It yields valuable rewards and results!

Support from Stakeholders & Management

- Getting buy-in across the Organization
- Self-driven and sustainable culture

Problem Statment

Despite positive customer reviews, the lack of customer growth may be attributed to changes in data adaptation by other retail companies.

- Understanding the customer behavior
- Developing a model using K-Means to establish a targeted marketing campaign
- Supervised models with countries as the target variable
- Building a recommender system that recommends retail items to upsell, cross-sell or even lead to product discovery for the customers
- Demand Forecasting with Time Series Analysis

The Retail Dataset UC Irvine Repository

1,016,727

Invoices (Rows of data) with **8** Features

> 5,000

Unique Products Sold

43 Countries

Served with > 91% customers in United Kingdom

The Data Dictionary

Fields	Description
Invoice	Invoice Number, 6-digit integral number – Unique transaction ID
Stock Code	Product (item) code
Description	Product (item) name
Quantity	The quantities of each product (item) per transaction was generated.
InvoiceDate	Invoice Date and Time when the transaction was generated. From 01/12/2009 to 09/12/2011
Price	Unit Price, product price per unit
Customer ID	Unique customer number
Country	Country name, the name of the country where the customer resides.

The Data Cleaning

```
========== Item Cleaning ===========
Removed Rows: 6432

===== Negative Values Cleaning =====
Removed Rows: 24062

====== Customer ID Cleaning ========
Removed Rows: 234245

=== Customers' Country Cleaning ====
Removed Rows: 78192

======= Duplicates Cleaning ========
Removed Rows: 24834

========= Overall Cleaning =========
Total Removed Rows: 367765
```

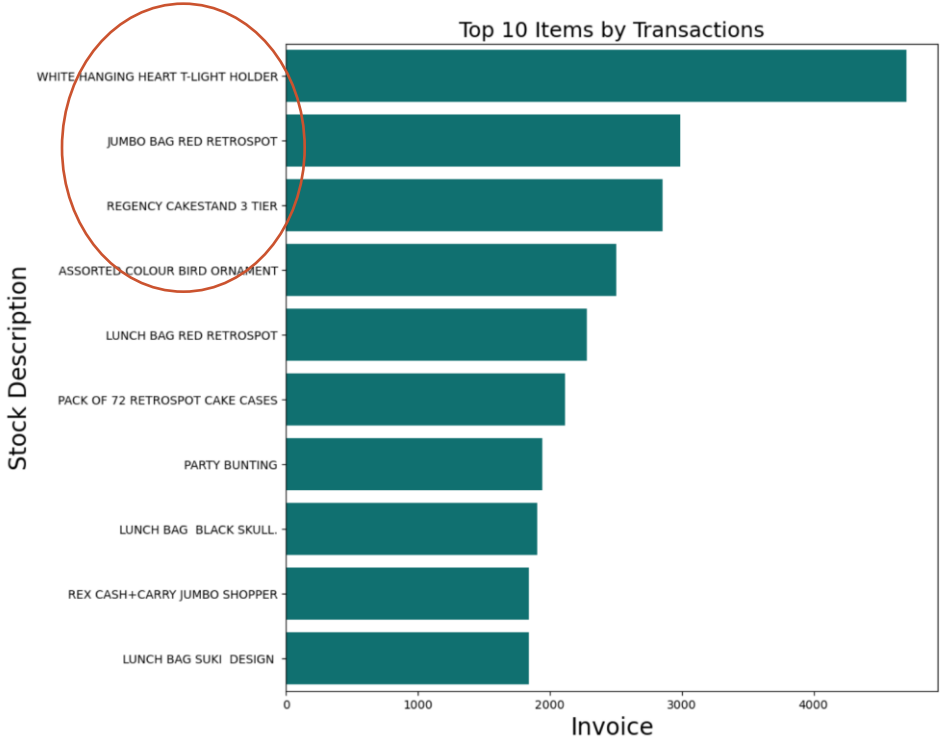
- 'Stock Code' and 'Description' had a mismatch of 5305:5698 unique values
- Addressed the negative entries in numerical columns
- Dropped rows with NaN values
- Focusing on UK only, Dropped rest
- Dropped The Duplicate Rows
- **TOTAL 367765 Rows Dropped**

02

Exploratory Data Analysis

Top 10 Items

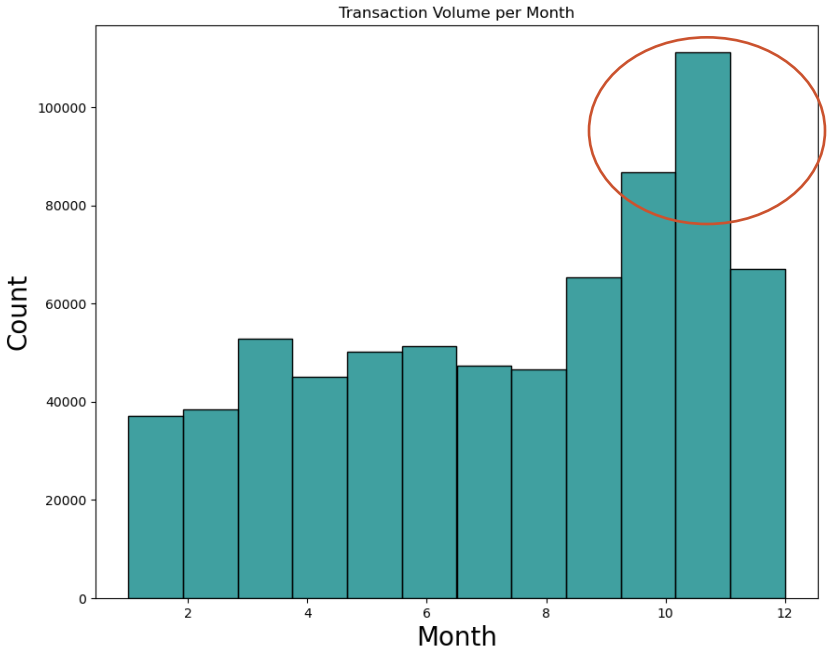
- White Hanging Heart T-Light Holder
- Jumbo Bag Red Retro spot
- Regency Cake stand 3 Tier



Transaction by Volume

- Clear trend of the annual Sales in Retail Industry

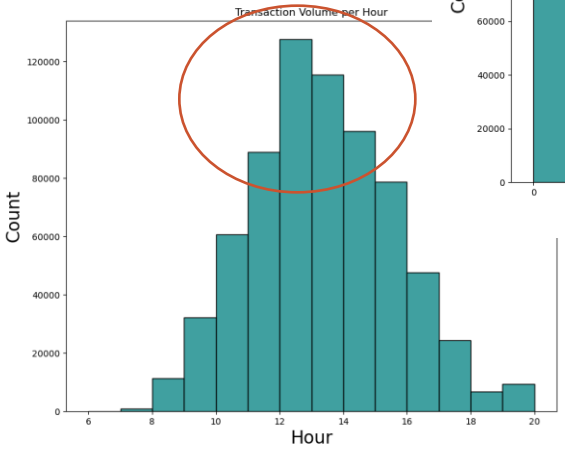
- Q4 Clearly shows substantial volume of sales, indicating purchases for gifting and Christmas festivities

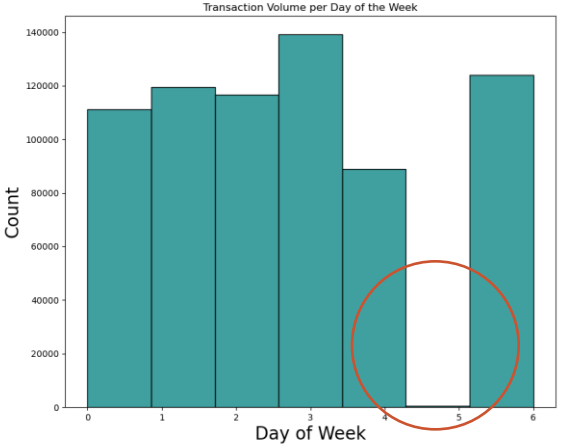


Zoom into Sales Volume

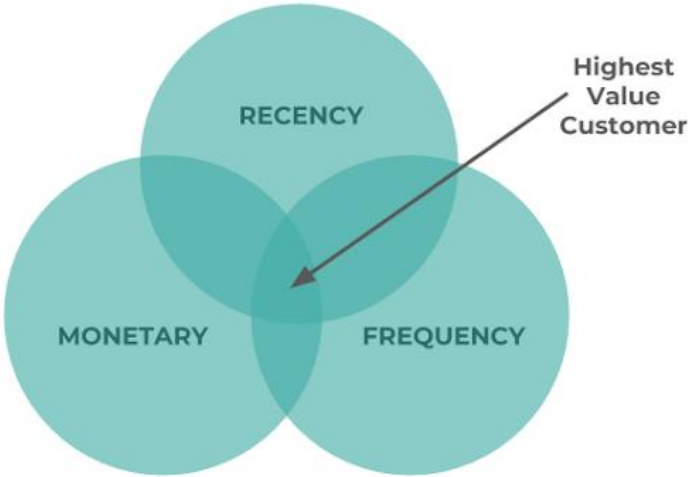
- Clear Indication that business is closed on Saturdays!

- More purchases during Lunch hours





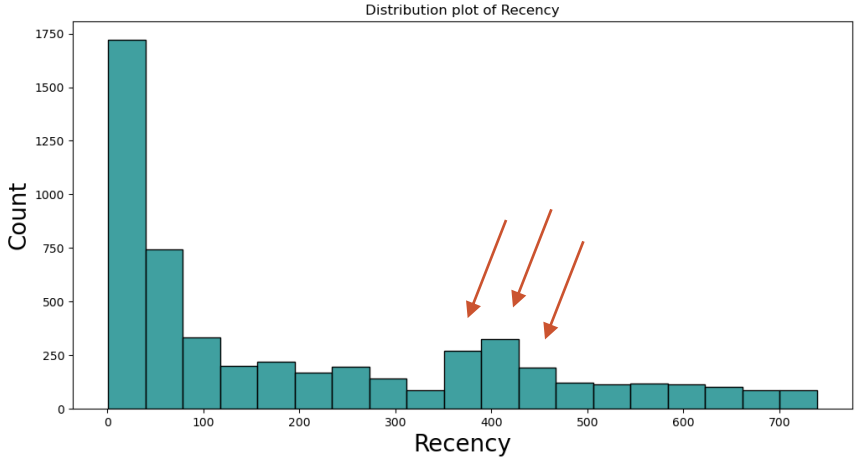
RFM Analysis



Recency, **F**requency, **M**onetary value (RFM) is a marketing analysis tool used to identify a company's best clients based on the nature of their spending habits.

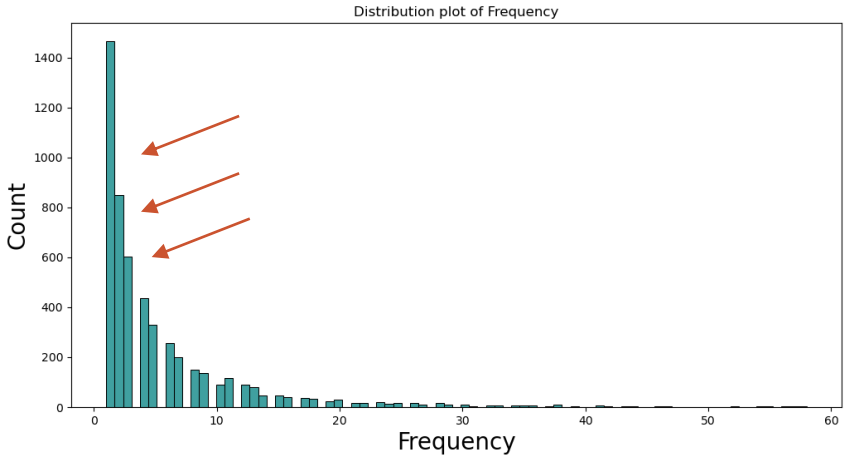
- It helps firms reasonably predict which customers are likely to purchase their products again, how much revenue comes from new (vs. repeat) clients, and how to turn occasional buyers into habitual ones.

RFM – Recency (How Recent?)



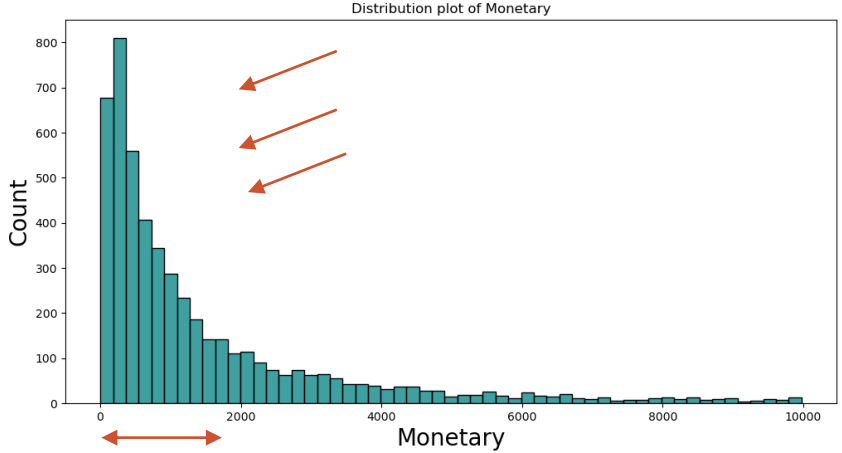
- Most of the customers are active, last purchase within 100 days
- However, there are significant number of customers between recency of 350 – 450 days (~ 500 customers)
- These customers should be targets for making them habitual from occasional

RFM – Frequency (How often?)



- Most of the customers have bought less than 3-5 times

RFM – Monetary (how much expenditure?)



- Most of the customers have spent less than $1,500 to $2,000

Supervised Models

- Models are built with countries as the target variable

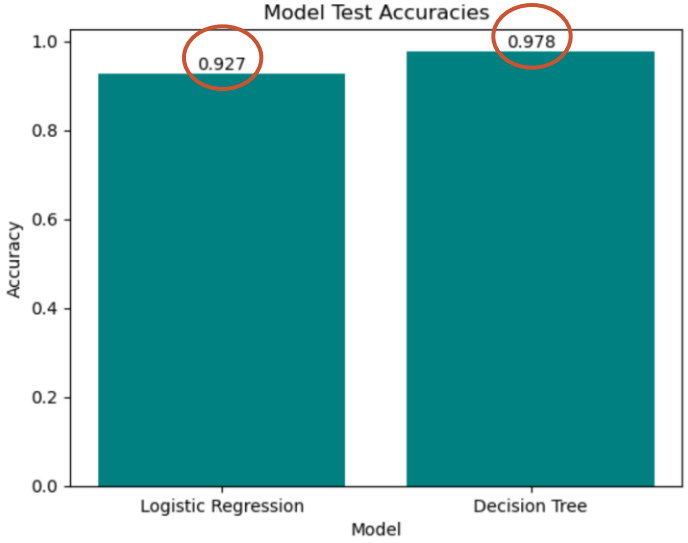
Logistic Regression: 92.74%

- **'C = 1' best params in Logistic Regression,**

Decision Tree Classifier: 97.8%

- **The best max depth found is 8**

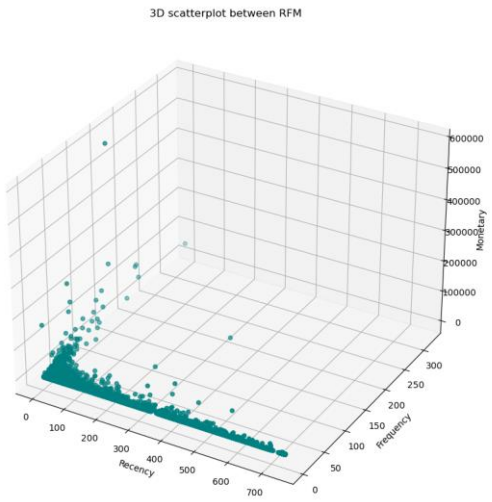
- **The best split found is 6**



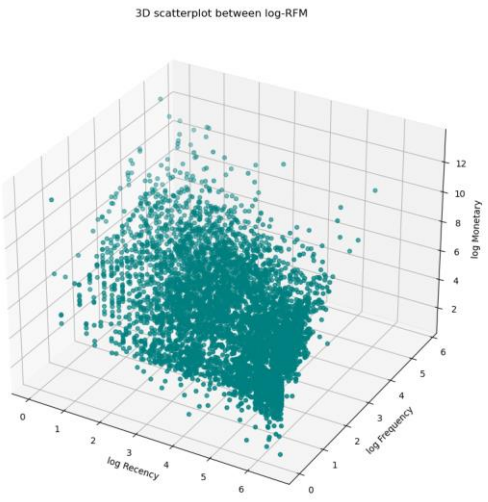
04

RFM Analysis with K means Clustering

K-Means Clustering
Before vs After Log Transformation

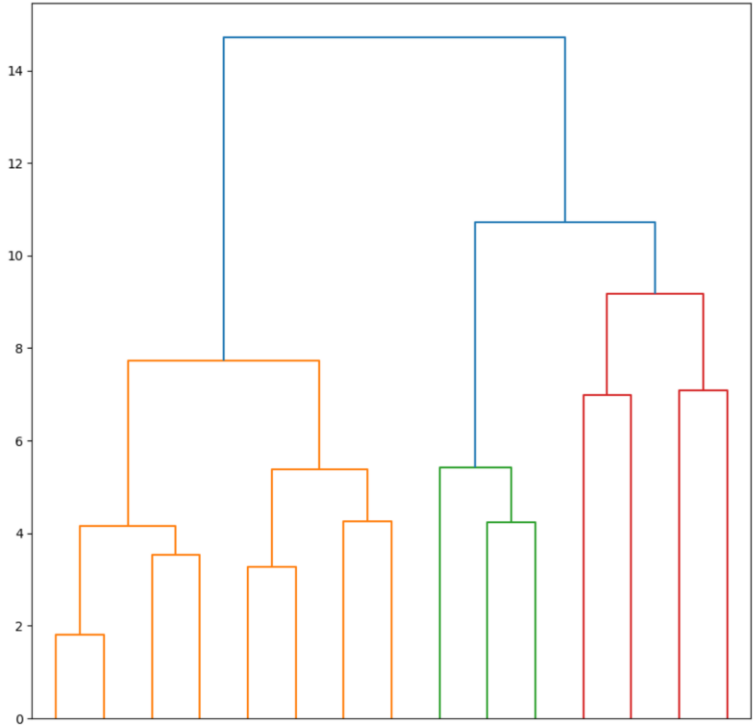


Log Transformation



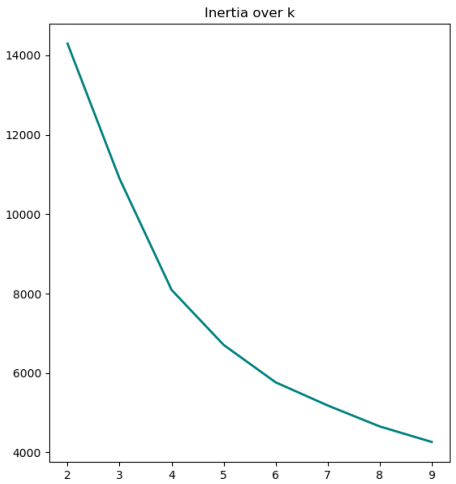
K-Means Clustering - Dendrogram

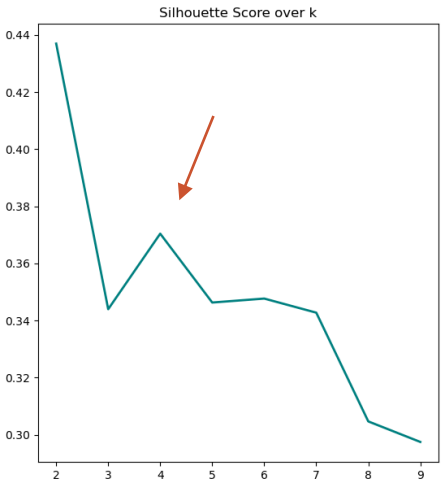
- This Dendrogram clearly suggests to split the data into 3 - 4 Clusters



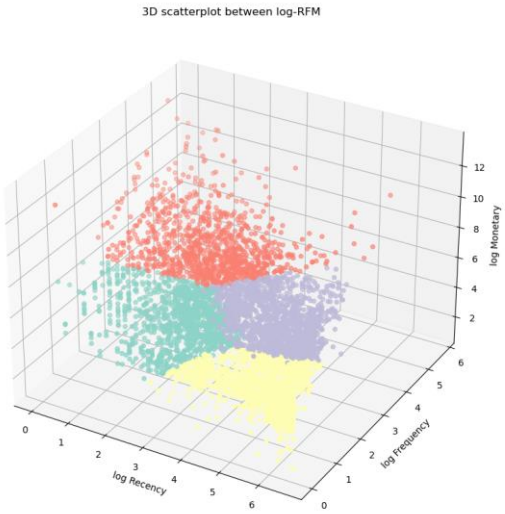
K-Means Clustering Silhouette Score over Cluster

- Here In the left graph, lower inertia indicates better clustering, line shows the inertia decreasing as k increases.
- Also, Cluster 4 is clearly performing better than Cluster 3

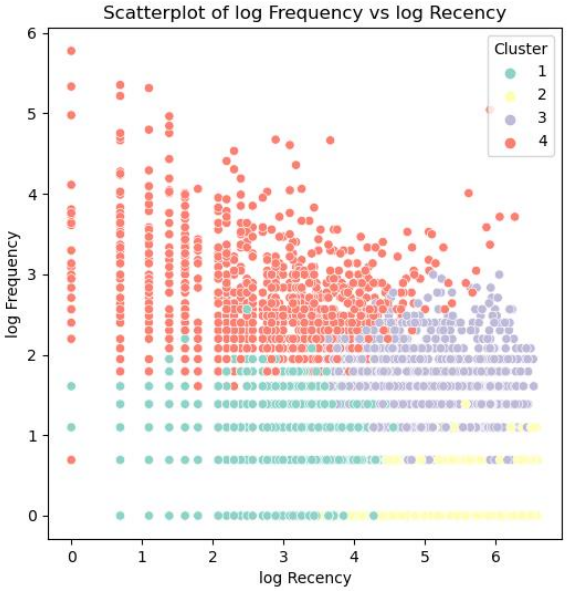




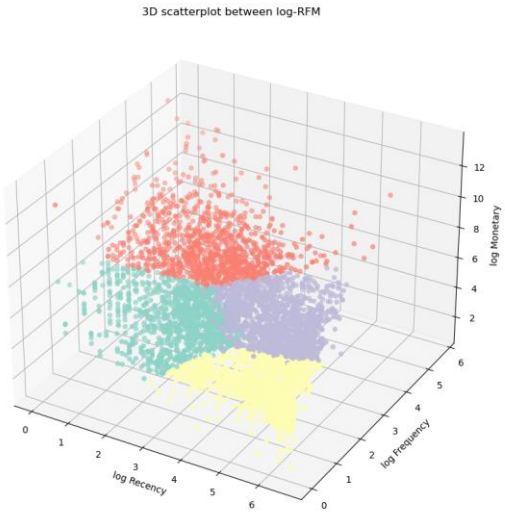
K-Means Clustering – 4 Clusters



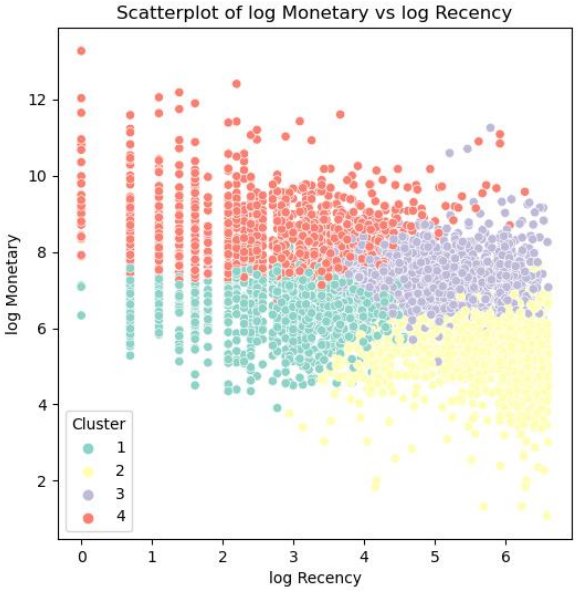
Top-Down View



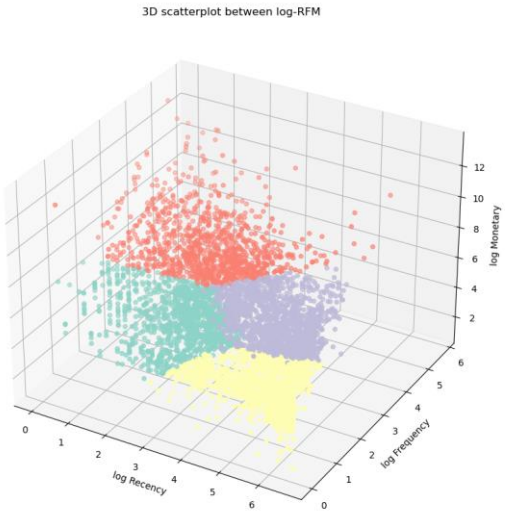
K-Means Clustering – 4 Clusters



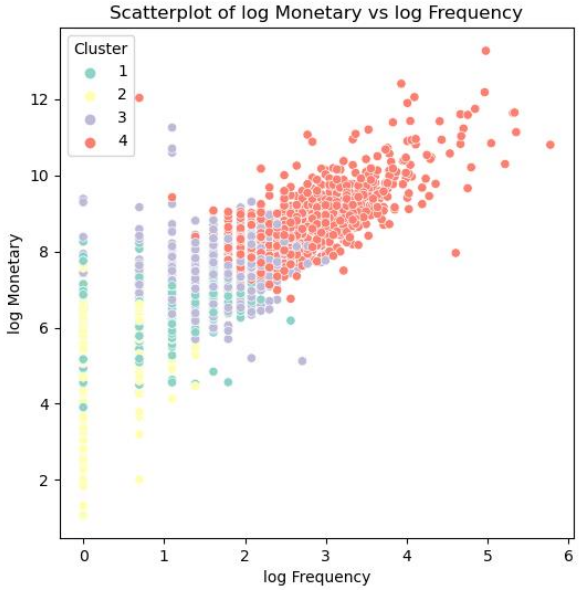
View from Left



K-Means Clustering – 4 Clusters



View from Right



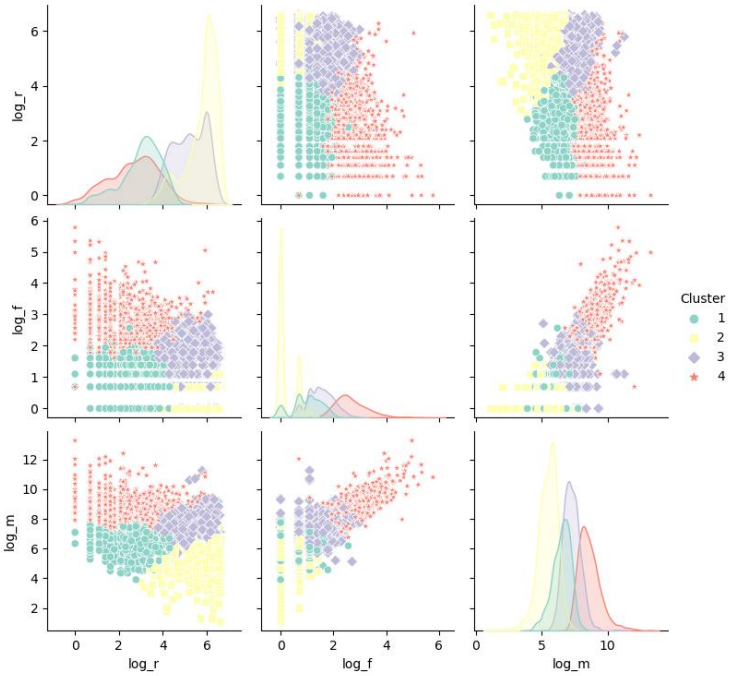
K-Means Clustering – 4 Clusters

Cluster 1 (Green): New Customers
- low R, low F & low M

Cluster 2 (Yellow): Lost Customers
- high R, low F & low M

Cluster 3 (Purple): Lost Customers with moderate Spending?
- high R, average F & average M

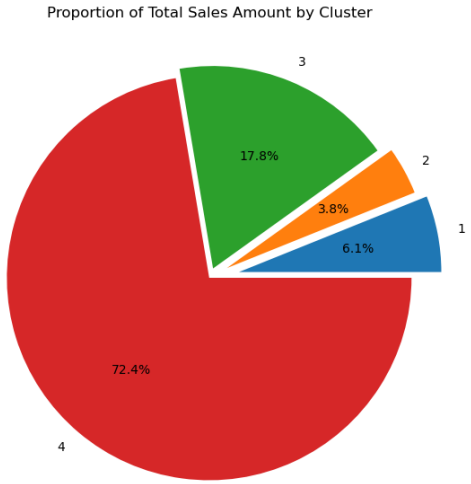
Cluster 4 (Red): Loyal Customers
- low R, High F & High M

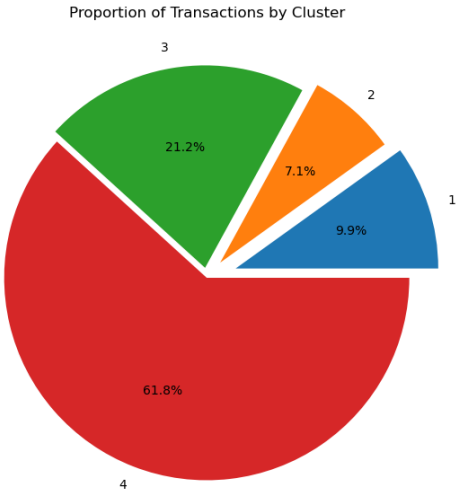


Cluster Summary

Cluster	1	2	3	4
Count Invoice	3310.00	2372.00	7078.00	20601.00
Count Product	3792.00	3801.00	4295.00	4506.00
Count Customer	1061.00	1802.00	1380.00	**1091.00**
Count Days	533.00	548.00	577.00	604.00
Total Amount	869,837	544,034	2,536,642	**10,337,692**
Mean Recency	28.11	389.30	233.56	27.27
Mean Frequency	3.12	1.32	5.13	18.88
Mean Monetary	819.83	301.91	1838.15	9475.43
Avg Sales Amt per Trans	262.79	229.36	358.38	501.81
Avg Sales Qty per Customer	518.11	181.11	1158.02	5543.76

K - Means – Distribution between Clusters

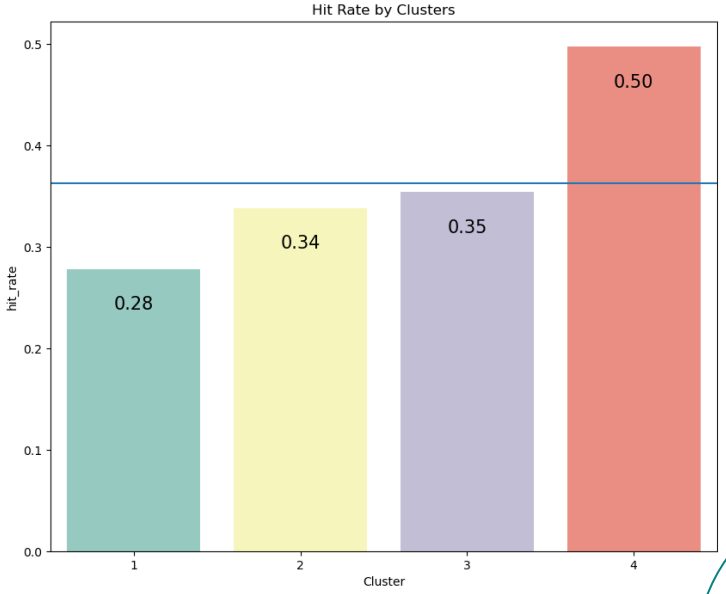




- **Cluster 4** takes **~70%** of the sales amount
- **Cluster 4** also takes **~60%** of overall transactions
- It can be evidently recommended to focus on basis of Importance as below:

4 > 1 > 3 > 2

Recommender System – User based Clusters filtering



- Recommended top 10 items for each user based on their clusters
- Evaluated using Hit-Rate Model, built in-house

- If the user purchases 1 of the top 10 products, then it is considered and labeled as **Hit!**

Recommender System
Example: User-based Cluster filtering

Recommended based on Customer 14440

	StockCode	Stock Description	Hit
0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	1
1	82494L	WOODEN FRAME ANTIQUE WHITE	1
2	82482	WOODEN PICTURE FRAME WHITE FINISH	1
3	21754	HOME BUILDING BLOCK WORD	1
4	21755	LOVE BUILDING BLOCK WORD	0
5	82486	WOOD S/3 CABINET ANT WHITE FINISH	1
6	72741	GRAND CHOCOLATECANDLE	0
7	22457	NATURAL SLATE HEART CHALKBOARD	0
8	82483	WOOD 2 DRAWER CABINET WHITE FINISH	0
9	21135	VICTORIAN METAL POSTCARD SPRING	0

- If a customer buy the product in Subject here, the customer is more likely to buy the listed top 10 products

Recommender System
Example: Item-based Cluster filtering

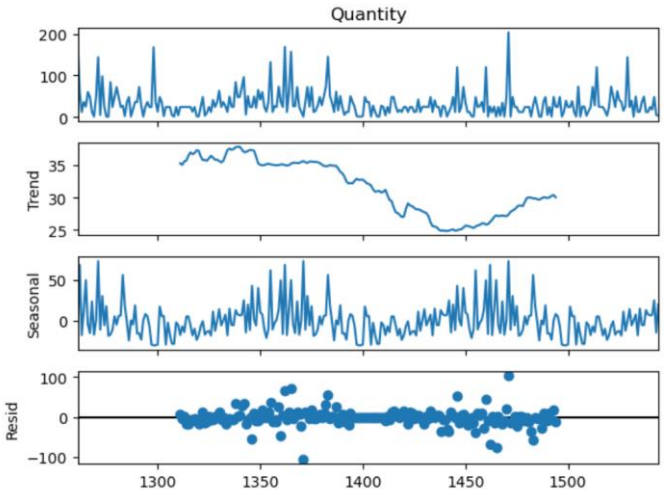
- Recommends top 10 Items for each item
- Evaluated using Hit Rate model, built in-house

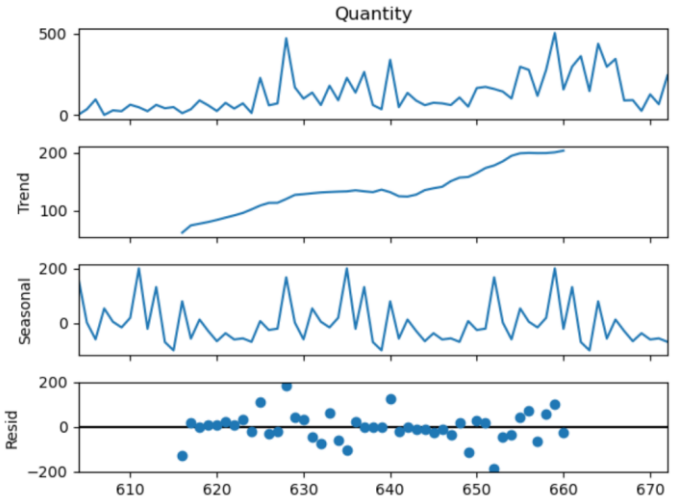
- Select an item, If the customer in a order has both items (selected *'Bendy Color Pencils'* & 1 from the recommended top 10 items for the selected item, then it is considered as a **Hit**!

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

Seasonality Analysis





Daily Analysis

- Grouped Description, daily, and dropped all rows below 200 sale days in attempt to see the trend

Monthly Analysis

- Grouped 'description' monthly, and dropped all rows below 200 sale days to see the trend

Conclusion & Recommendation

Clear business objective has to be set right from the start
- Models can be improved & tuned along the way

Stake holders supper and commitment to implement change is important
- Make realistic data strategy/roadmap

Diverse Expertise
- To gain domain knowledge / useful insight so that the modeling process can be catered accordingly, especially for the unsupervised machine learning (Like RFM Analysis)

A/B testing could be performed further to evaluate the recommender system

And, certainly with more data, the clusters could be more discrete and distinguished
The Time series(Seasonality) testing would have been more descriptive.

Thanks!

Do you have any questions?

harnishshah25@gmail.com
+1 732 351 3241