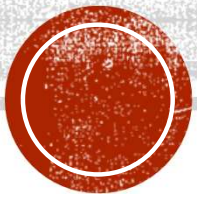
CUSTOMER SEGMENTATION

USING RFM CLUSTERING

Evangelos Tzimopoulos

London, Feb 2020



CONTENTS

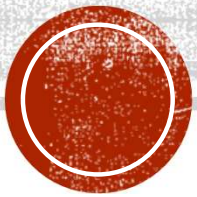
Project Overview

Exploratory Data Analysis (EDA)

Key Performance Indicators (KPIs)

Modelling

Summary



COMING UP NEXT

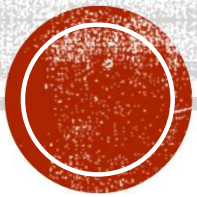
Project Overview

Exploratory Data Analysis (EDA)

Key Performance Indicators (KPIs)

Modelling

Summary



PROJECT OVERVIEW

Objective

Identify customer segments based on the overall buying behaviour of the client

Stakeholders

Product/Services /Ops team

Marketing/Sales team

Dataset

Online E-Commerce Business Data Set in CSV format

http://archive.ics.uci.edu/ml/datasets/online+retail

Outcome

Create an unsupervised model that generates the optimum number of segments for the customer base

Success Criteria

Segments generated can be interpreted and transposed into business actions

COMING UP NEXT

Project Overview

Exploratory Data Analysis (EDA)

Key Performance Indicators (KPIs)

Modelling

Summary

Data Overview

Data Engineering

Data Processing

EXPLORATORY DATA ANALYSIS

```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
InvoiceNo      541909 non-null object
StockCode      541909 non-null object
Description    540455 non-null object
Quantity       541909 non-null int64
InvoiceDate    541909 non-null datetime64[ns]
UnitPrice      541909 non-null float64
CustomerID     406829 non-null float64
Country        541909 non-null object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
dtype: int64
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
299983	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	NaN	United Kingdom
299984	A563187	B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	NaN	United Kingdom

DATA OVERVIEW

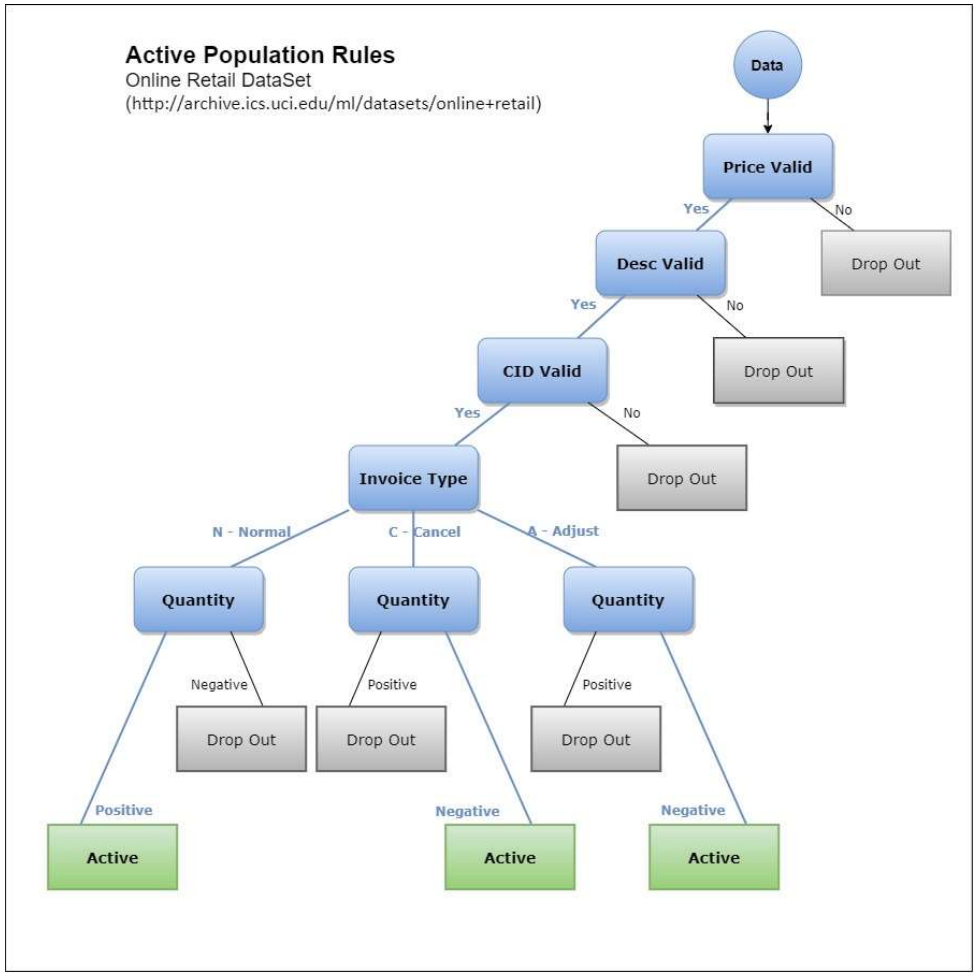
- 8 key features
- Empty rows
  - Description ~1.5k
  - CustomerID ~135k
- Negative values
  - Quantity ~ 10k
  - Unit price ~2
- Invoice
  - C-type – Cancellation
  - A-type – Adjustment
  - Null – Normal
- Country
  - United Kingdom ~ 495k

FEATURE ENGINEERING

- Break down InvoiceNo to
  - InvoiceNumber – New Feature
  - InvoiceCode – New Feature
    - C – Cancellation
    - A – Adjustment
    - Null – Rename to "N" – Normal
- Break down InvoiceDate to
  - InvoiceYearMonth - new feature
  - InvoiceYear - new feature
  - InvoiceMonth - new feature
  - Date – Use as index

Date	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceNumber	InvoiceCode	Year	Month	YearMonth
2011-08-12	A563185	B	Adjust bad debt	1	2011-08-12 14:50:00	11062.06	NaN	United Kingdom	563185	A	2011	8	201108
2011-08-12	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	NaN	United Kingdom	563186	A	2011	8	201108
2010-12-01	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom	536379	C	2010	12	201012
2010-12-01	C536383	35004C	SET OF 3 COLOURED FLYING	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom	536383	C	2010	12	201012

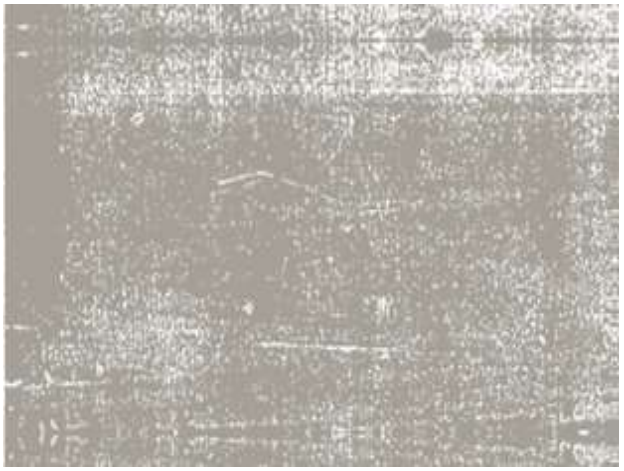
- Date
- Invoice



PRE-PROCESSING & ANALYSIS

- Active population definition based on Business Rules

```
# Valid Price
valid_price = retail.UnitPrice >= 0

# Valid Description
valid_desc = retail.Description.notnull()

# Valid CID
valid_CID = retail.CustomerID.notnull()

# Invoice type-N (Normal)
inv_N = retail.InvoiceCode == "N"

# Invoice type-C (Cancellation)
inv_C = retail.InvoiceCode == "C"

# Invoice type-N (Amendment)
inv_A = retail.InvoiceCode == "A"

# Quantity Negative
q_neg = retail.Quantity < 0

# Quantity Positive
q_pos = retail.Quantity >= 0
```



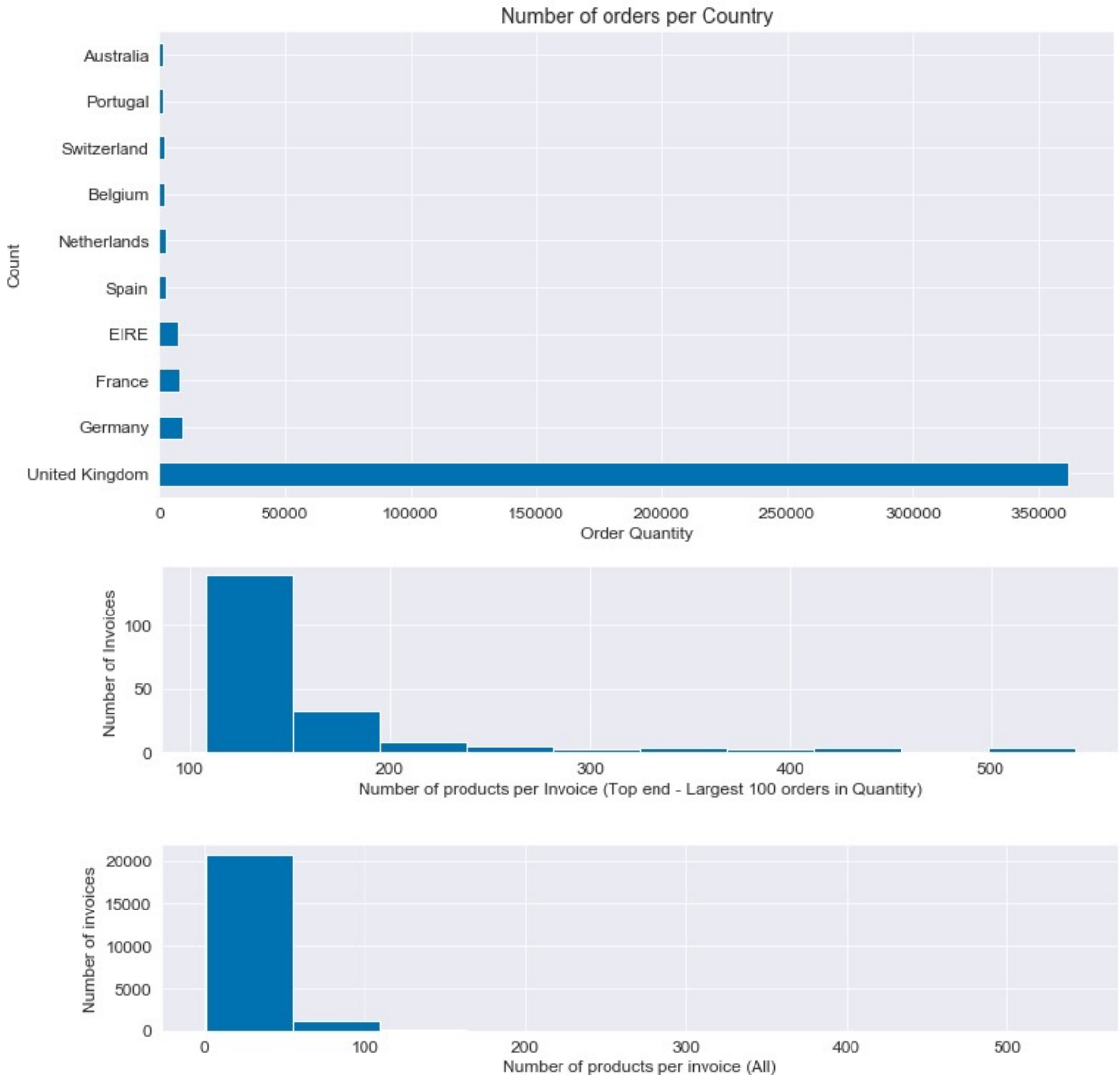
Active Population - Paths to Nodes

```
# Path1 - Filter population down to include all
# valid Customer IDs with Valid Price and Description
p1 = valid_price & valid_desc & valid_CID
```

```
# Path2 - Filter population down to include all
# Normal (type-N) transactions with Positive Quantities
p2 = inv_N & q_pos

# Path3 - Filter population down to include all
# Cancel (type-C) or Adjust (type-A) transactions
# with Negative Quanitities
p3 = (inv_A | inv_C) & q_neg

# Path to Leafs: Combine Paths 1, 2 and 3:
# ************** CREATE A COPY ************
retail_pp = retail.loc[p1 & (p2 | p3)].copy()

# Inspect new pre-processed and clean dataset
retail_pp.head()
```

PRE-PROCESSING & ANALYSIS

- Masks implementing Business Rules
- Active population derivation



PRE-PROCESSING & ANALYSIS

Visualising features:

- Countries
- Number of Products per invoice

COMING UP NEXT

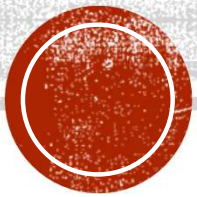
Project Overview

Exploratory Data Analysis (EDA)

Key Performance Indicators (KPIs)

Modelling

Summary



Transactional KPIs

Customer KPIs

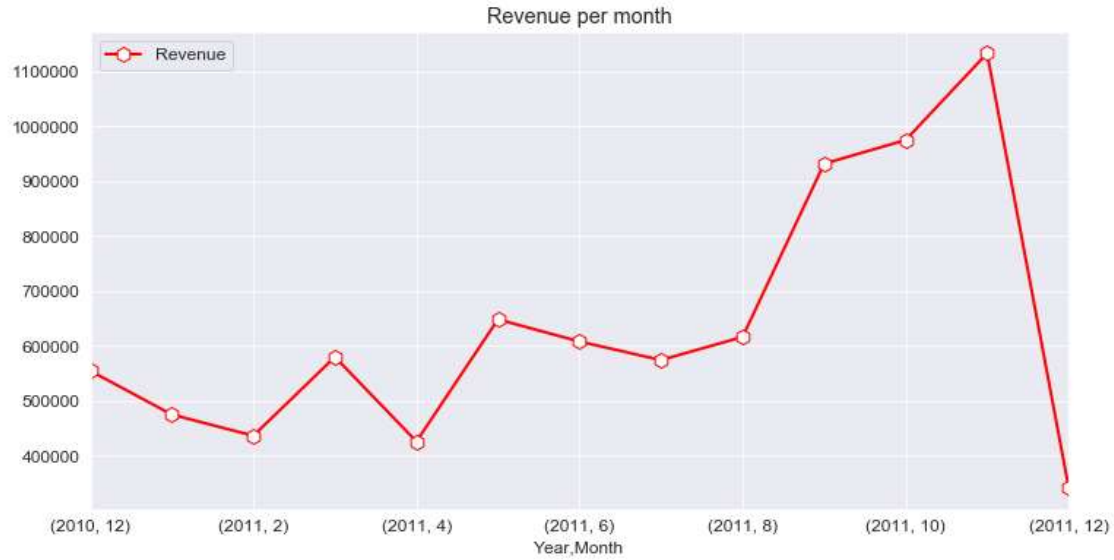
Product KPIs

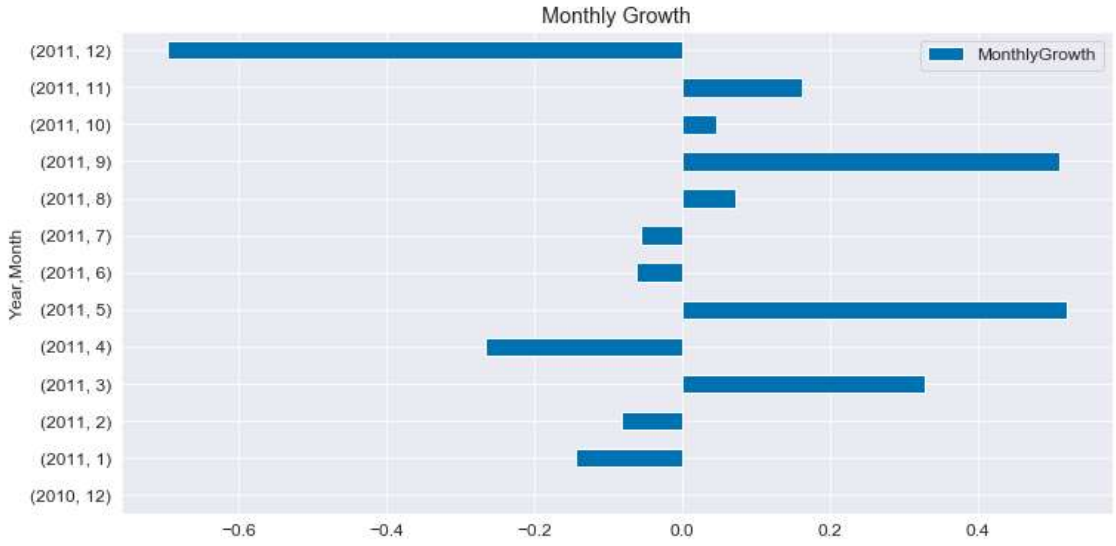
KEY PERFORMANCE INDICATORS

TRANSACTIONAL KPIS

Monthly Revenue | **Monthly Revenue Growth** | **Active Customers** | **Monthly Order Count** | **Average revenue per order**

Year	Month	Revenue	MonthlyGrowth	ActiveCustomers	MonthlyOrderCount	MonthlyOrderAverage
2010	12	554604.020	NaN	948	296362	20.655643
2011	1	475074.380	-0.143399	783	269379	21.681014
	2	436546.150	-0.081099	798	262833	21.438204
	3	579964.610	0.328530	1020	344012	20.845540
	4	426047.851	-0.265390	899	278585	18.365715
	5	648251.080	0.521545	1079	367852	22.424626
	6	608013.160	-0.062072	1051	356922	21.842691
	7	574238.481	-0.055549	993	363418	20.879881
	8	616368.000	0.073366	980	386612	22.282120
	9	931440.372	0.511176	1302	537496	22.817118
	10	974603.590	0.046340	1425	569666	19.224846
	11	1132407.740	0.161916	1711	669915	17.262839
	12	342506.380	-0.697541	686	203836	19.393374





TRANSACTIONAL KPIS

- Revenue per month
- Revenue growth

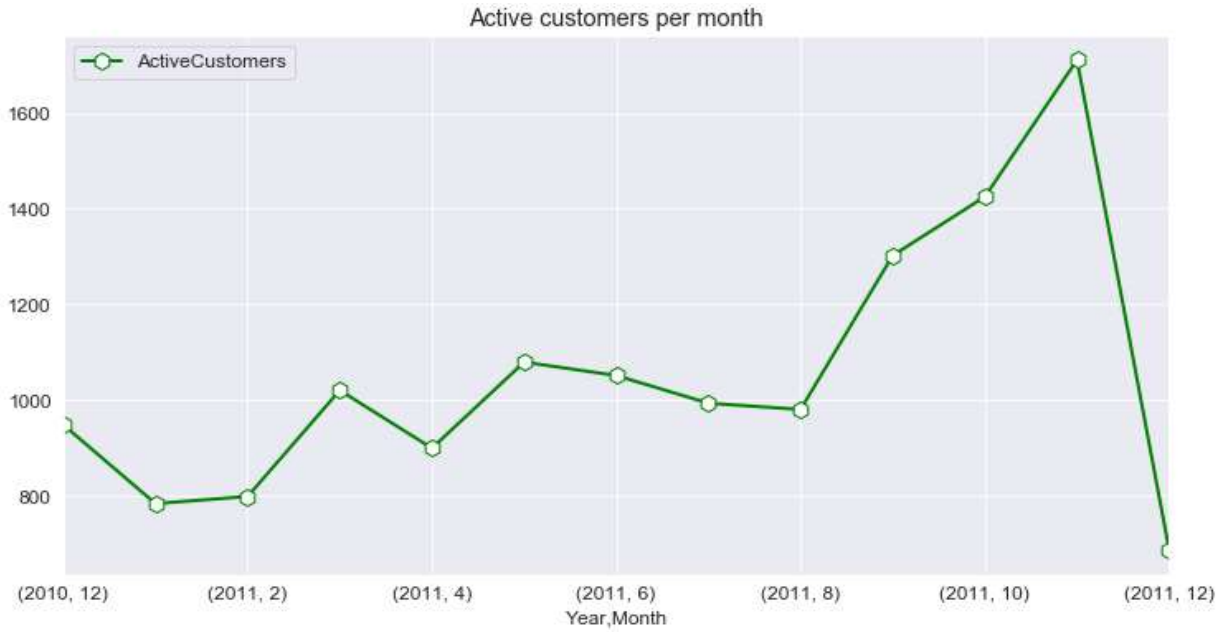
TRANSACTIONAL KPIS

- Number of orders per month

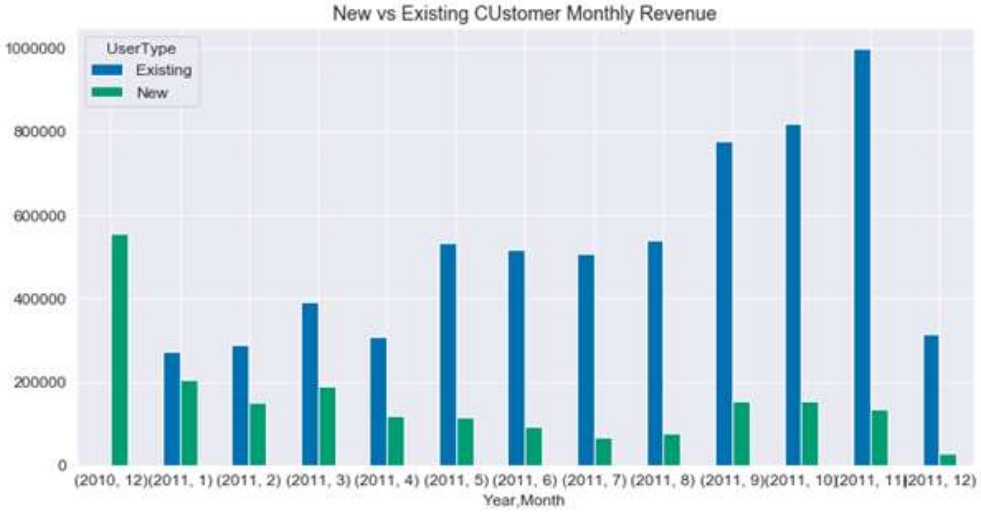
- Monthly order value

CUSTOMER KPIS

- Active Customers

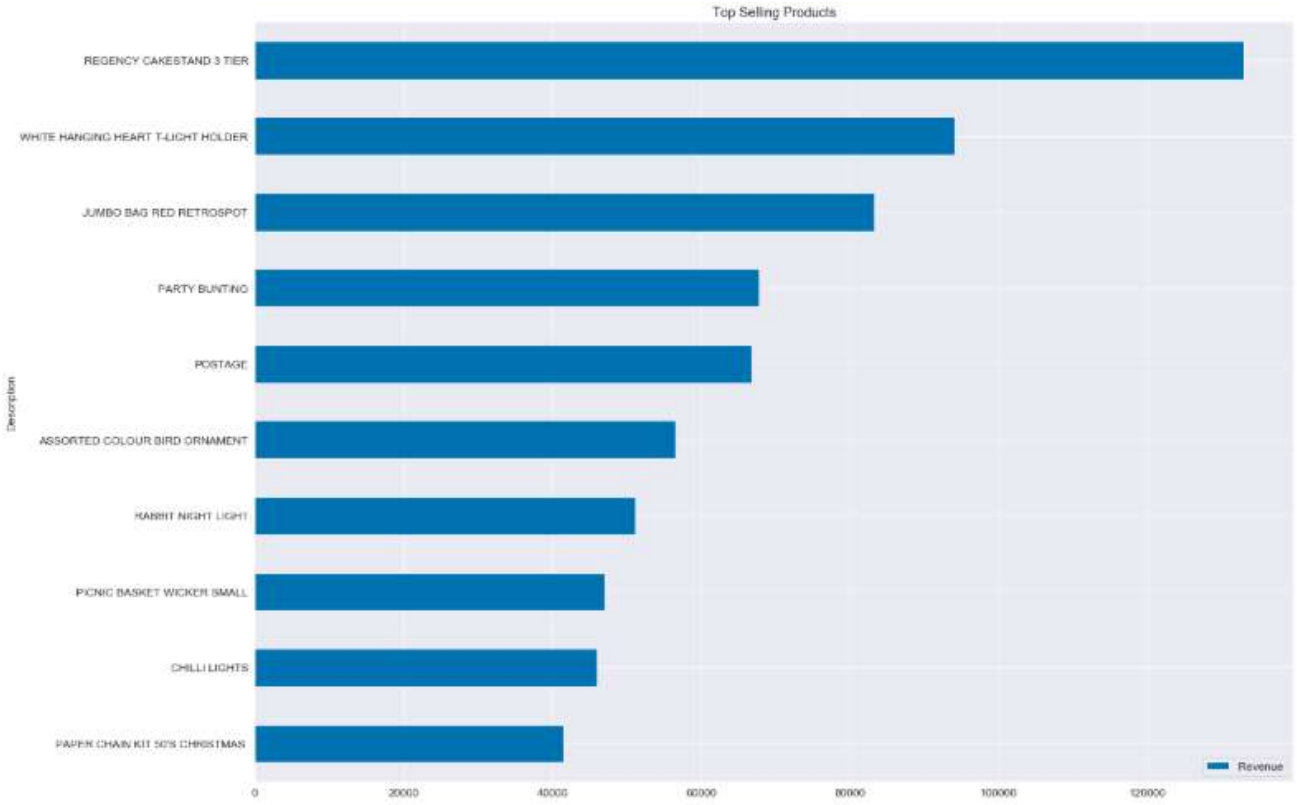


Year	Month	UserType	Revenue
2010	12	New	554604.020
2011	1	Existing	271616.520
		New	203457.860
	2	Existing	287024.770
		New	149521.380
	3	Existing	390034.530
		New	189930.080
	4	Existing	306283.600
		New	119764.251
	5	Existing	532392.340
		New	115858.740
	6	Existing	515486.650
		New	92526.510
	7	Existing	508355.610
		New	65882.871
	8	Existing	538709.770
		New	77658.230
	9	Existing	778161.781
		New	153278.591
	10	Existing	819672.900
		New	154930.690
	11	Existing	998176.360
		New	134231.380
	12	Existing	315729.630
		New	26776.750



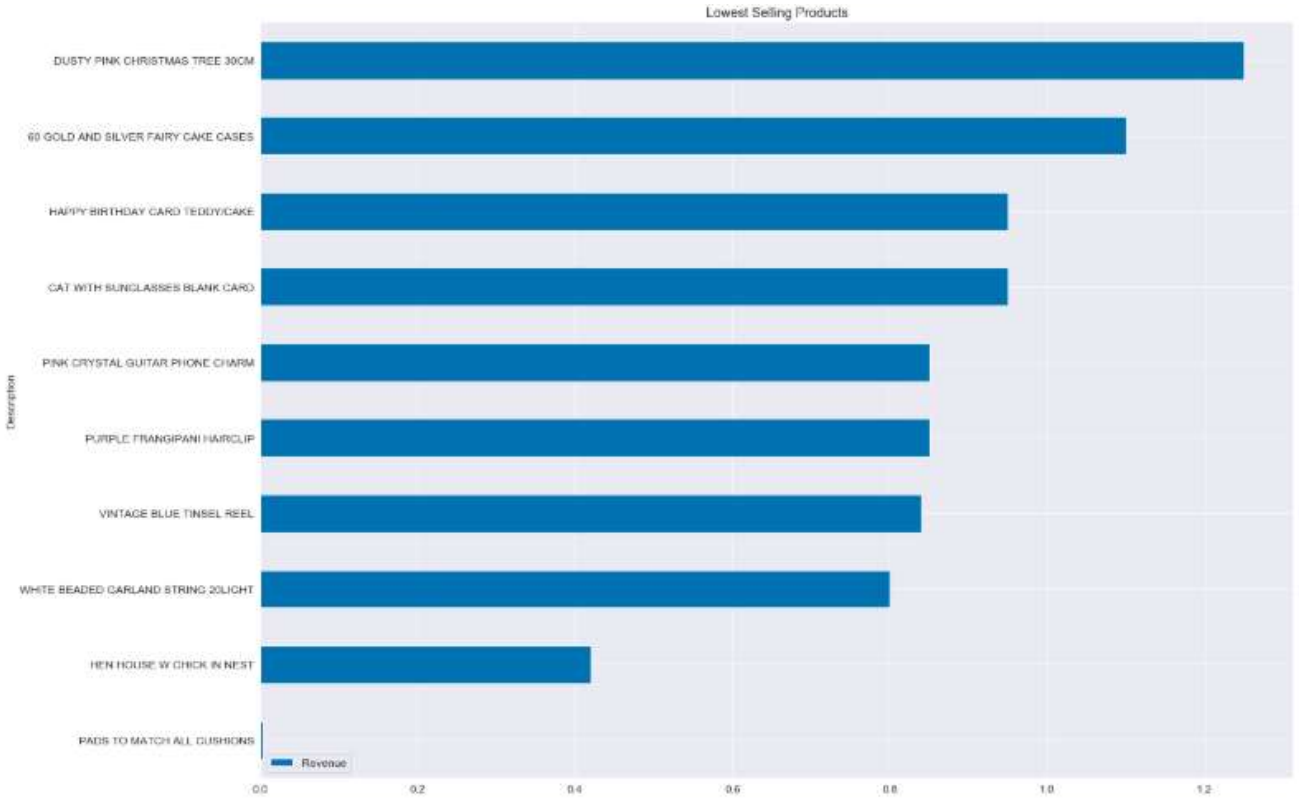
CUSTOMER KPIS

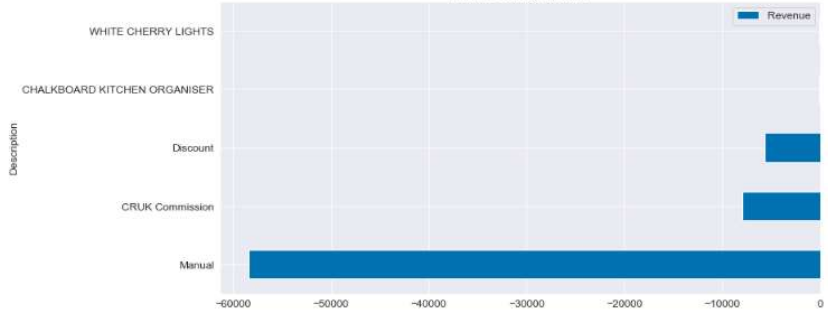
- New vs Existing Customers



PRODUCT KPIS

- Top selling product





PRODUCT KPIS

- Lowest selling product

COMING UP NEXT

Project Overview

Exploratory Data Analysis (EDA)

Key Performance Indicators (KPIs)

Modelling

Summary

MODELLING & EVALUATION

RFM Analysis

K-Means Clustering

Optimization & Metrics

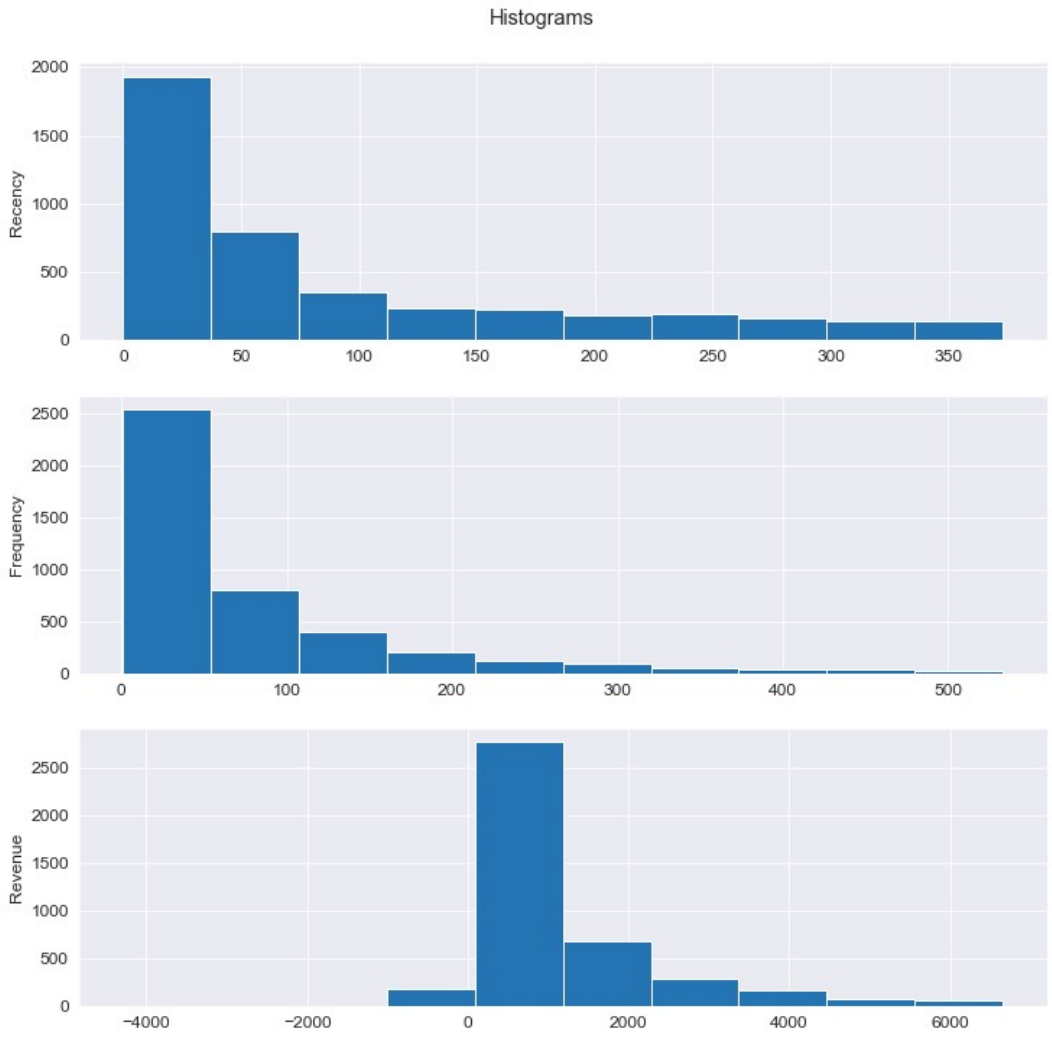
Segments Mapping & Business Actions

Recency - Given a current or specific date in the past, when was the last time that the customer made a transaction

Frequency - Given a specific time window, how many transactions did the customer do during that window

Monetary Value or Revenue - Given a specific window, how much did the customer spend

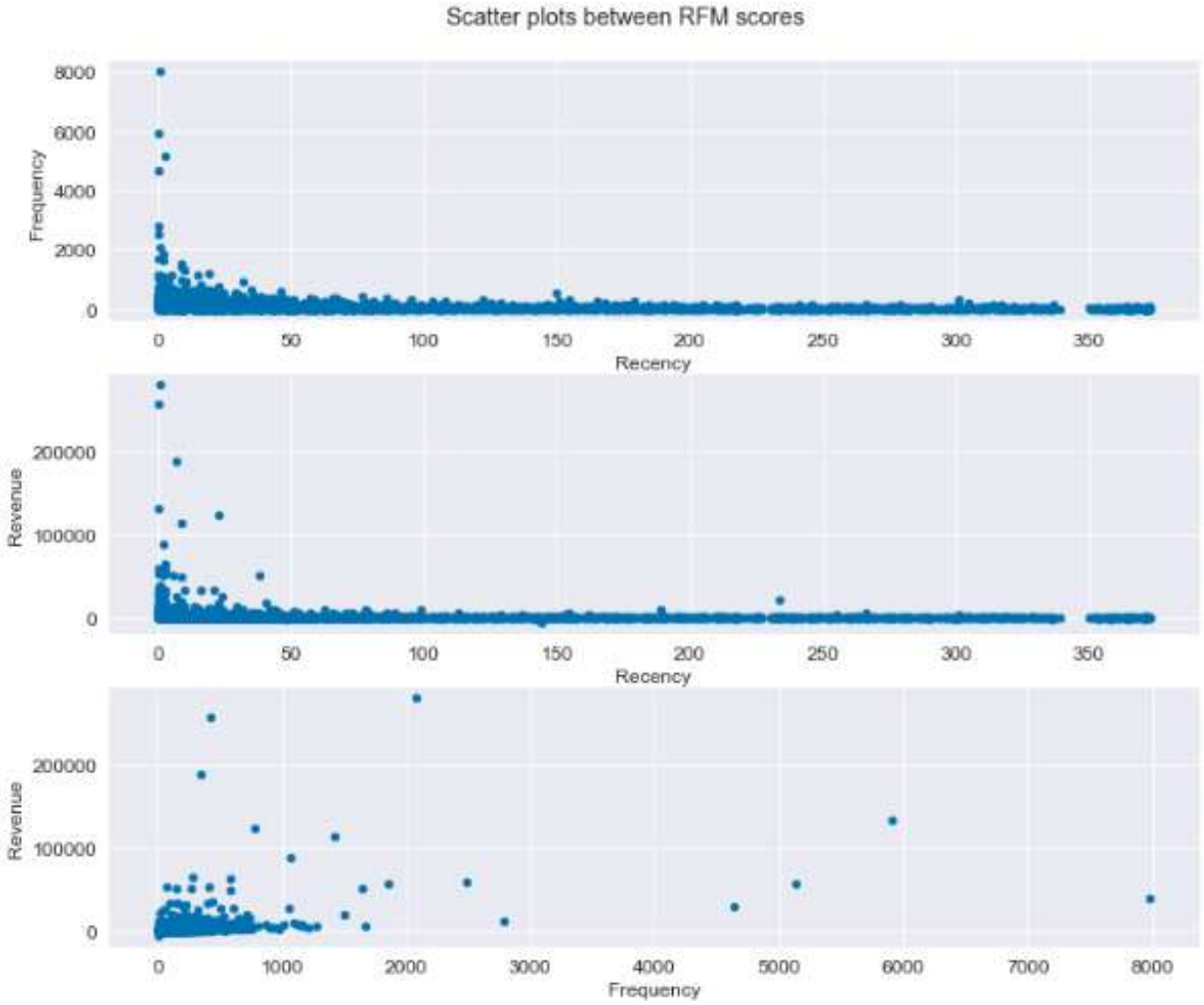
RFM ANALYSIS



RFM ANALYSIS

Histograms

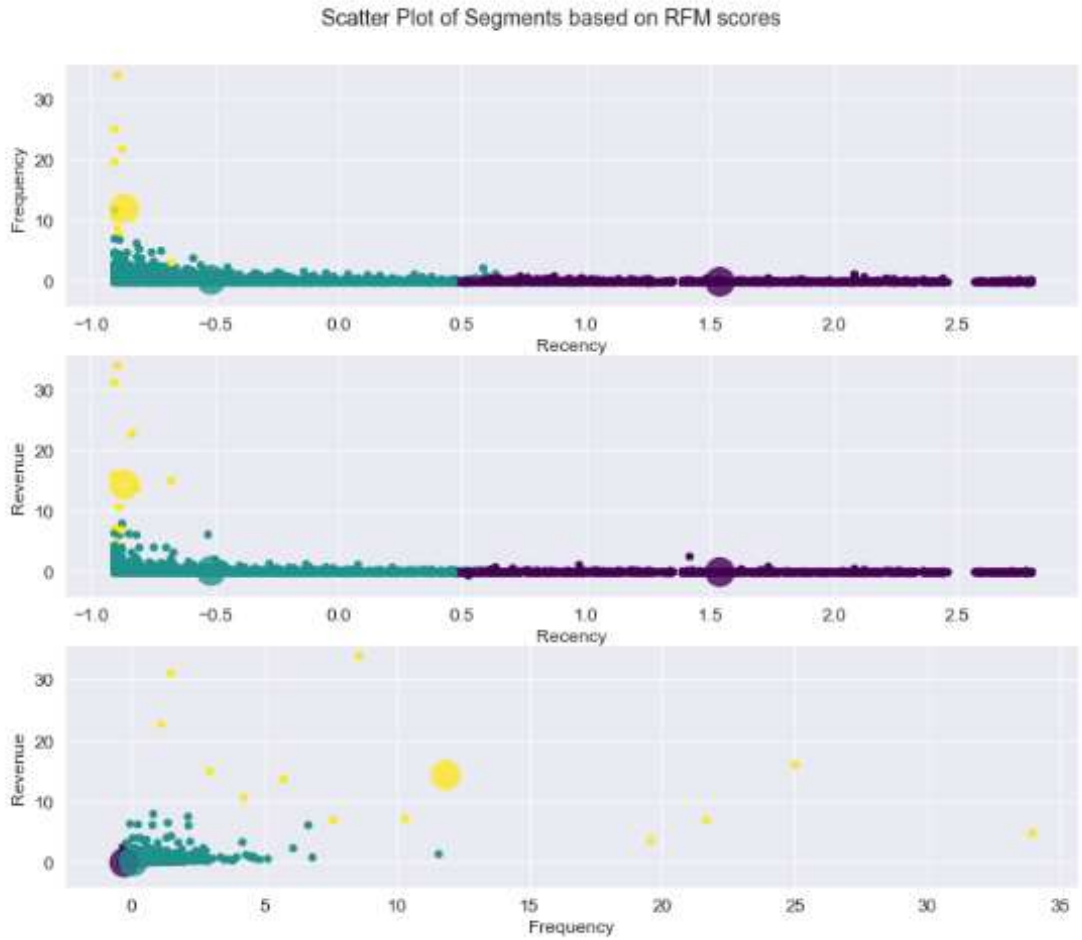
- Recency
- Frequency
- Revenue



RFM ANALYSIS

Scatter Plots

- Recency vs Frequency
- Recency vs Revenue
- Frequency vs Revenue



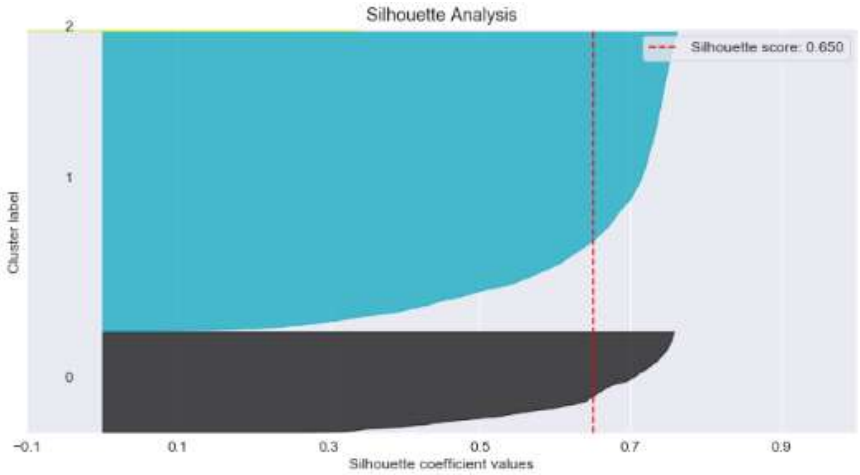
K-MEANS CLUSTERING

Input X

- Recency
- Frequency
- Revenue

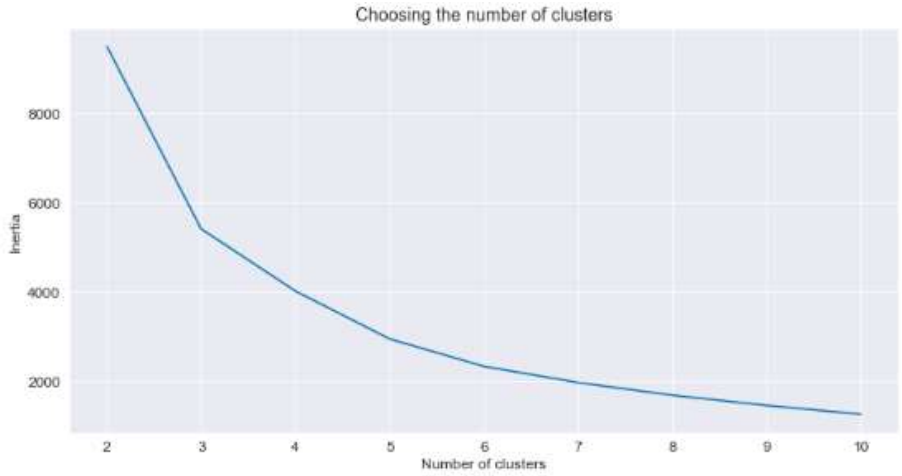
Initial configuration

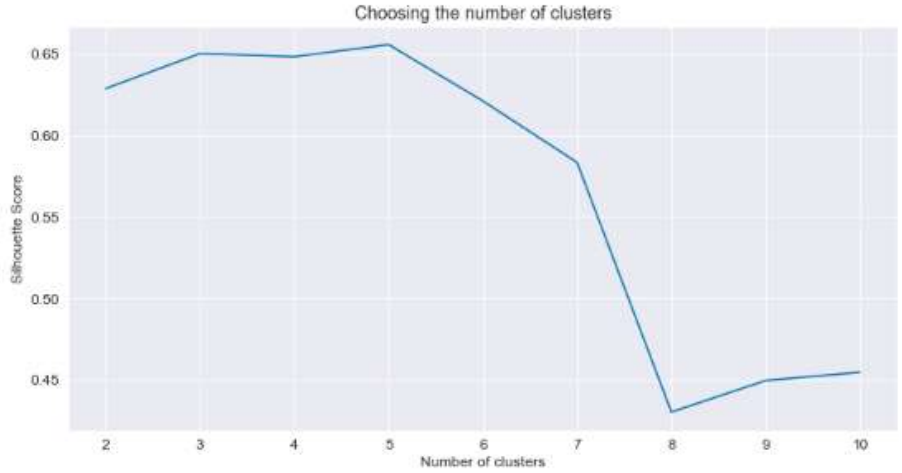
- K=3



K-MEANS CLUSTERING

- Inertia Score - 5408.4046
- Silhouette Score - 0.650





OPTIMIZATION & METRICS

Minimal Scores

- Inertia – 1255.5644
- Silhouette – 0.454
- K = 8

Optimal K

- K = 3-5

SEGMENTS MAPPING

Initial 3 Segment approach

Low value
Mid value
High value

Optimal 8 Segments Approach

Low Segment: 0-2
Mid Segment: 3-5
High Segment: 6-8

BUSINESS ACTIONS

High Value:

Improve Retention of these customers as they are the most valuable asset

Mid Value:

Increase Retention and Frequency and bring them closer to the brand and the product so eventually they become High Value

Low Value:

Increase Frequency and understand if there are any potential issues around the product or service

COMING UP NEXT

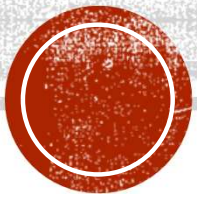
Project Overview

Exploratory Data Analysis (EDA)

Key Performance Indicators (KPIs)

Modelling

Summary



SUMMARY

Data Review & Understanding

Feature Engineering & Processing

KPIs & Visualizations

RFM analysis & new features

Unsupervised learning K-Means clustering

Optimization & Evaluation

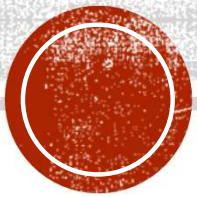
Business Actions & Segments Mapping

Next Steps

NEXT STEPS

Customer Lifetime Value (CLTV) Prediction

- Supervised learning
- Classification based on RFM analysis (input features) and Revenue (output label) on specified windows




QUESTIONS?