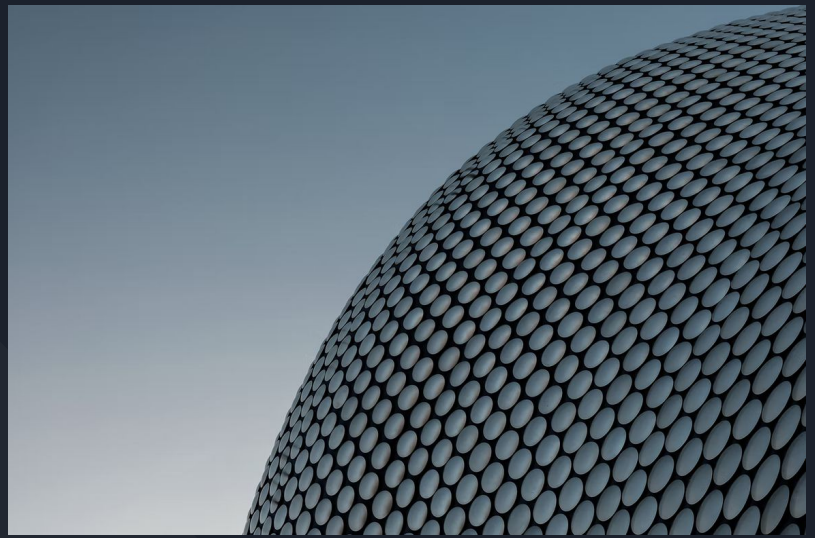
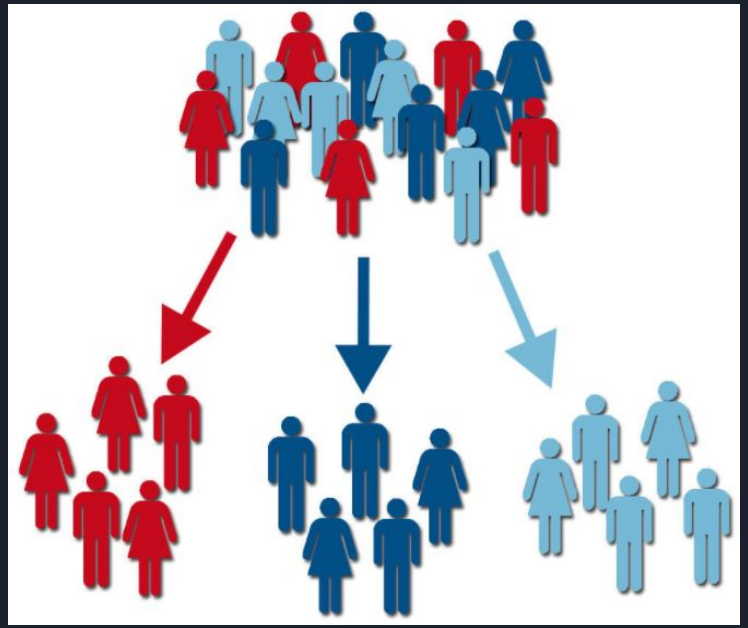
Customer Segmentation



Andrew Wright
July 2022

Introduction



Using past data, analyze how to segment the customers using Cohort Analysis, RFM, and K-Means Clustering

Data Science Approach

1. Examine dataset for missing values, data types, outliers
2. Explore the data
3. Cohort Analysis
4. RFM
5. Prep Data for Clustering
6. Model with k-means

Results Summary

The Data was segmented using:

Cohorts Method: recommended for looking at retention

RFM Method: recommended for using domain knowledge to label the groups

Clustering with K-means: recommended for finding the segments that relate to each other

Data Acquisition

The dataset was acquired from an online source:

https://www.kaggle.com/datasets/jihyeseo/online-retail-data-set-from-uci-ml-repo

Features:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

StockCode: Product (item) code.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invice Date and time.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number.

Country: Country name.

Data Wrangling: Summary

	Quantity	UnitPrice
count	401604.000000	401604.000000
mean	12.183273	3.474064
std	250.283037	69.764035
min	-80995.000000	0.000000
25%	2.000000	1.250000

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

```
df=df[(df['Quantity']>0) & (df['UnitPrice']>0)]
df.describe()
```

	Quantity	UnitPrice	CustomerID
count	392692.000000	392692.000000	392692.000000
mean	13.119702	3.125914	15287.843865
std	180.492832	22.241836	1713.539549
min	1.000000	0.001000	12346.000000

Data Wrangling: Summary

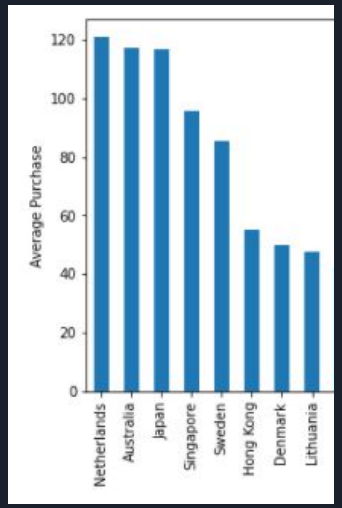
	Quantity	UnitPrice
count	401604.000000	401604.000000
mean	12.183273	3.474064
std	250.283037	69.764035
min	-80995.000000	0.000000
25%	2.000000	1.250000

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

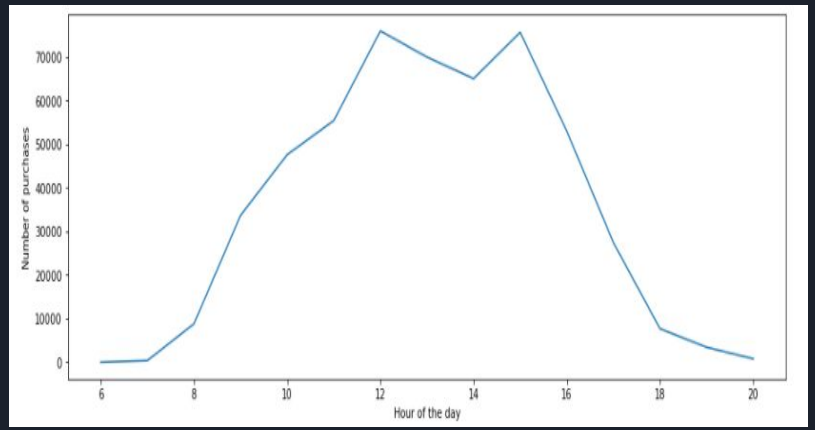
```
df=df[(df['Quantity']>0) & (df['UnitPrice']>0)]
df.describe()
```

	Quantity	UnitPrice	CustomerID
count	392692.000000	392692.000000	392692.000000
mean	13.119702	3.125914	15287.843865
std	180.492832	22.241836	1713.539549
min	1.000000	0.001000	12346.000000

Exploratory Analysis



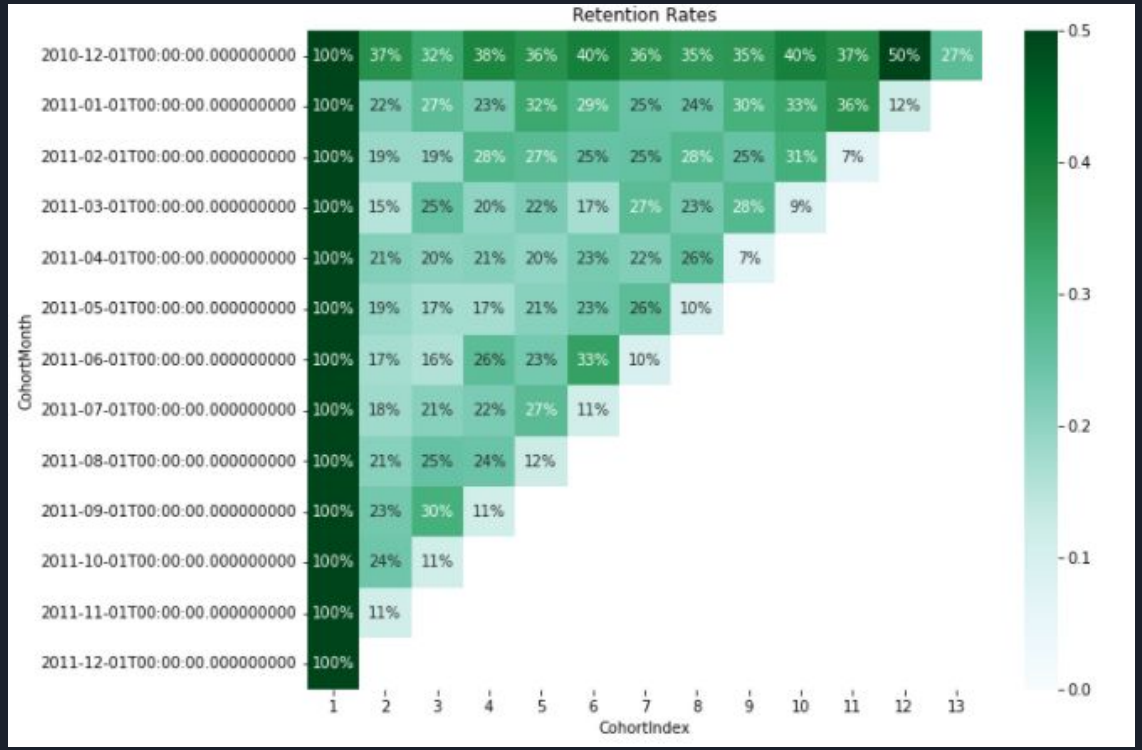
Netherlands is highest average purchased



12p to 3p are the highest purchasing hours

The last quarter of the year is the highest for monthly sales

Cohort Analysis



Monthly Cohorts

Conclusion:

recommended for looking at retention and

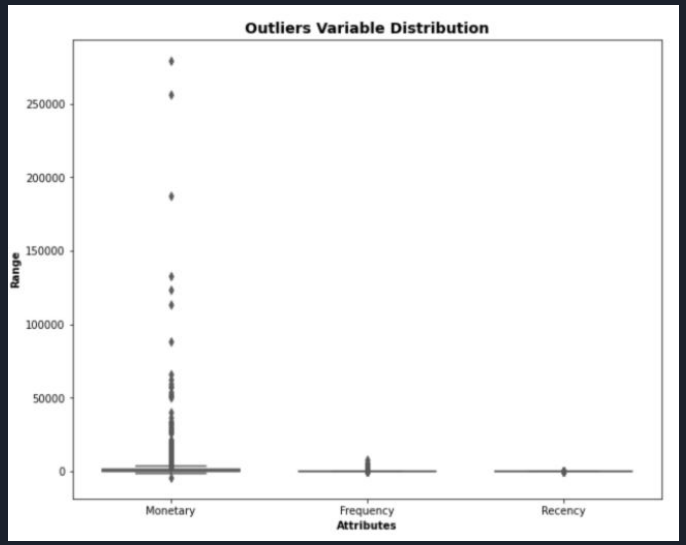
RFM

CustomerID	Recency	Frequency	MonetaryValue	R	F	M	RFM_Score	RFM_Segment	RFM_Level
0	1	132186	1754901.91	3	3	3	9	3.03.03.0	Top
12346	326	1	77183.60	1	1	1	3	1.01.01.0	Low
12347	2	182	4310.00	3	3	3	9	3.03.03.0	Top
12348	75	31	1797.24	2	2	2	6	2.02.02.0	Middle
12349	19	73	1757.55	3	3	3	9	3.03.03.0	Top
...	...	...	...	...	...	...	...	...	...
18280	278	10	180.60	1	1	1	3	1.01.01.0	Low
18281	181	7	80.82	1	1	1	3	1.01.01.0	Low
18282	8	12	178.05	3	1	1	5	3.01.01.0	Low
18283	4	721	2045.53	3	3	3	9	3.03.03.0	Top
18287	43	70	1837.28	2	2	2	6	2.02.02.0	Middle

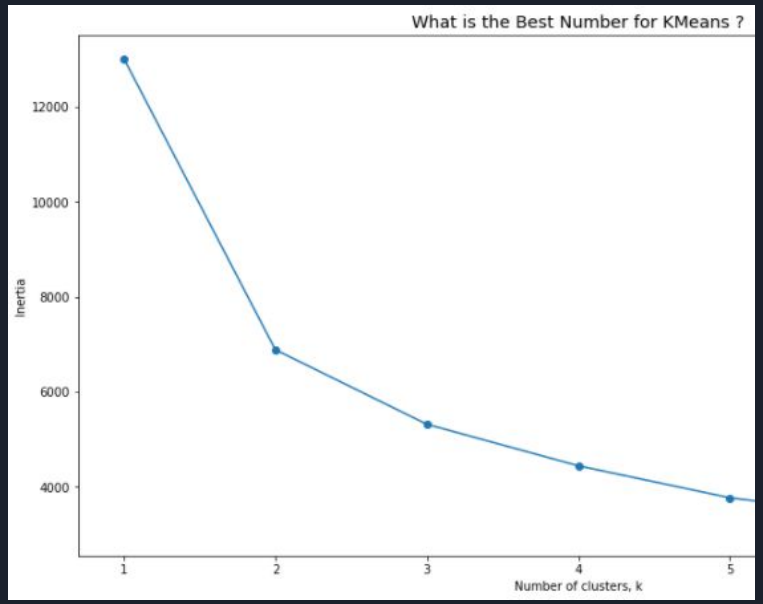
Conclusion: recommended for using domain knowledge to label the groups

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
RFM_Level				
Low	160.3	18.8	642.0	1926
Middle	52.9	84.5	1576.6	1591
Top	10.4	430.8	8390.9	822

K-Means Clustering

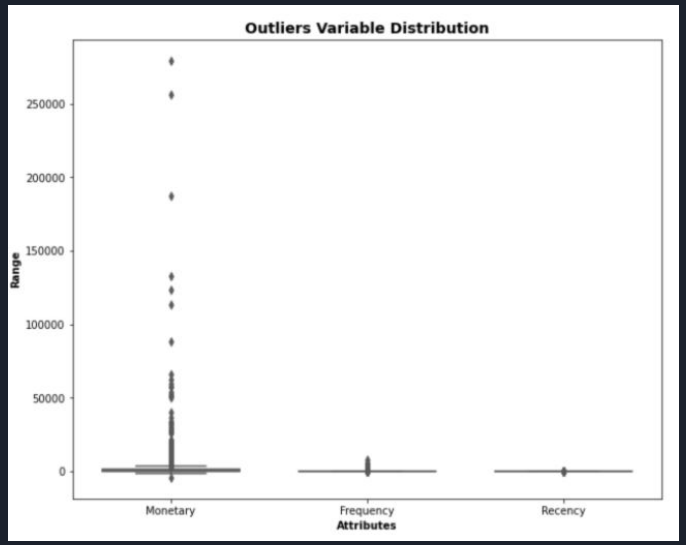


Normalize Data using StandardScaler

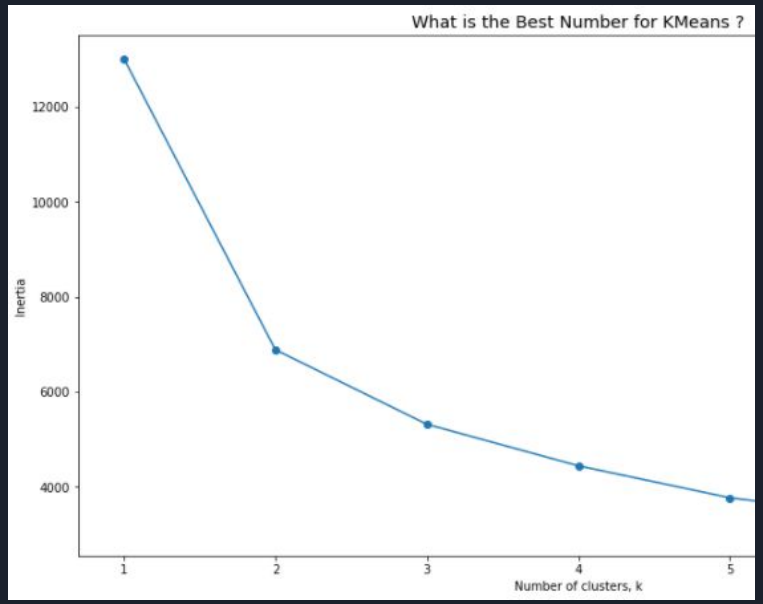


Optimal clusters is 3

K-Means Clustering

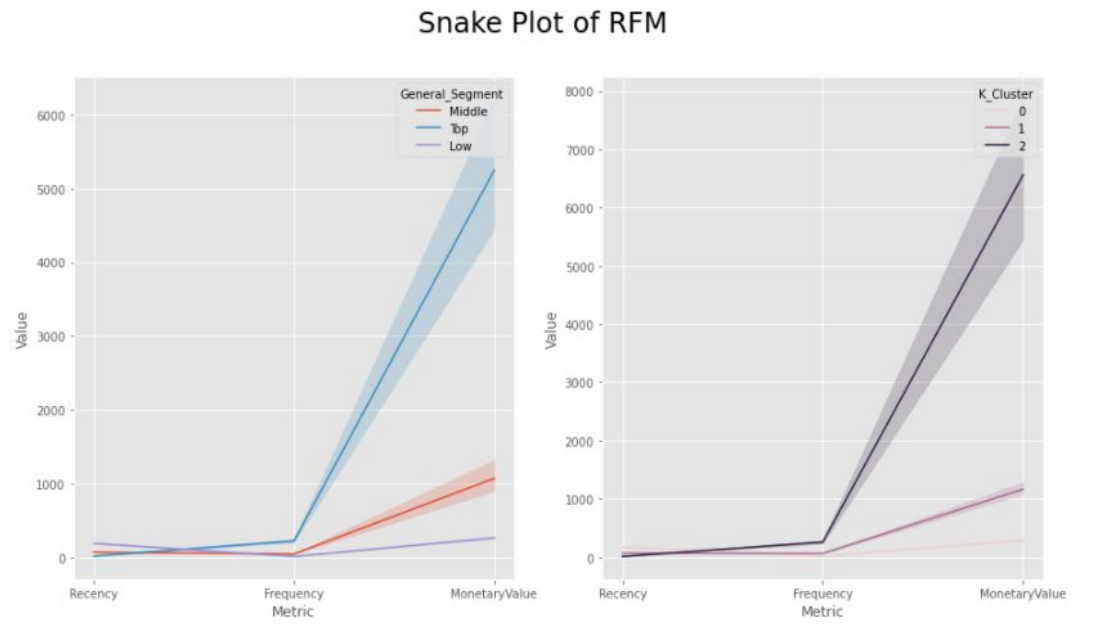


Normalize Data using StandardScaler



Optimal clusters is 3

Comparison

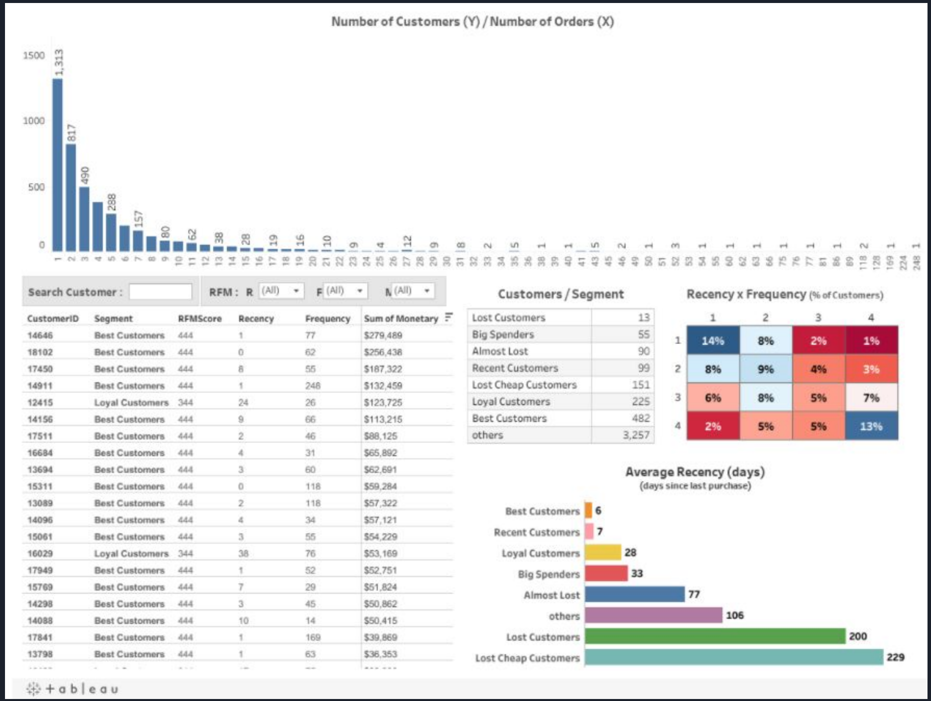


RFM_Level	Recency mean	Frequency mean	MonetaryValue mean	MonetaryValue count
Low	160.3	18.8	642.0	1926
Middle	52.9	84.5	1576.6	1591
Top	10.4	430.8	8390.9	822

K_Cluster	Recency mean	Frequency mean	MonetaryValue mean	MonetaryValue count
0	171.0	15.0	293.0	1523
1	69.0	65.0	1167.0	1859
2	13.0	260.0	6559.0	956

Integration:

Tableau Dashboard:



Recommendations

Cohorts Method: recommended for looking at retention

RFM Method: recommended for using domain knowledge to label the groups

Clustering: recommended for finding the segments that relate to each other

Future Work:

Attempt other types of clustering:

Centroid-based
Density-based
Distribution-based
Hierarchical Clustering

Questions?