Customer Segmentation

Springboard Capstone Project
Ryan Paik

Problem

A global superstore has transaction orders from 2011 to 2015. The company wants to identify their customer behavior and actions. The focus is to use RFM segmentation and to classify future customers.

Data obtained from DataWorld

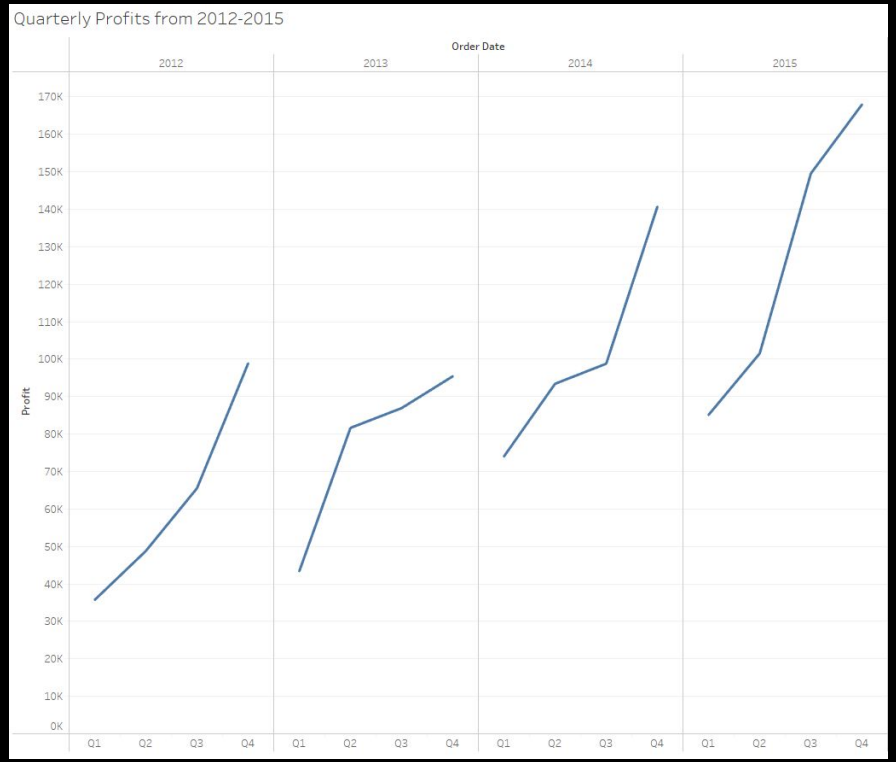
Who does this benefit?

The Global Superstore

- Investors
- Decision Makers
- Marketing Team

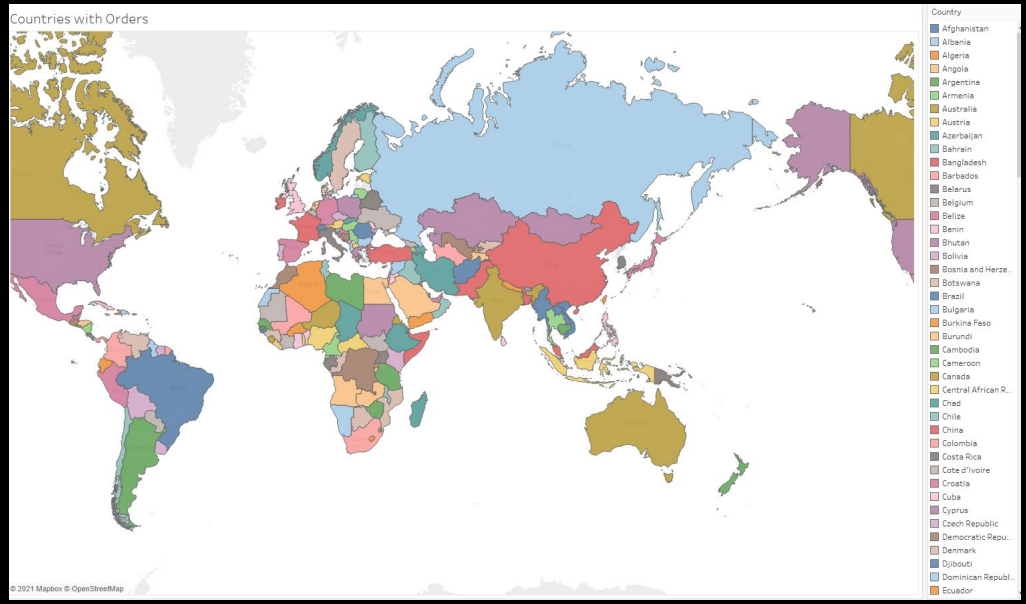
EDA

The Global Superstore is in a positive uptrend with their profits. Q4 has the highest peak each year.



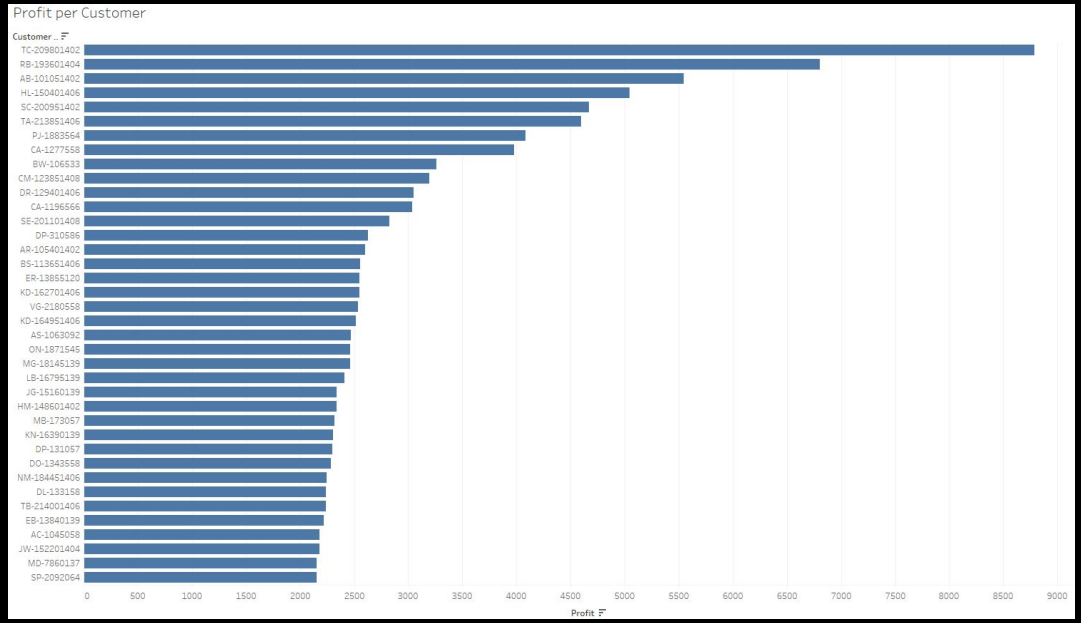
Location

- Well established internationally
- Data contains Region, Country, State, and City



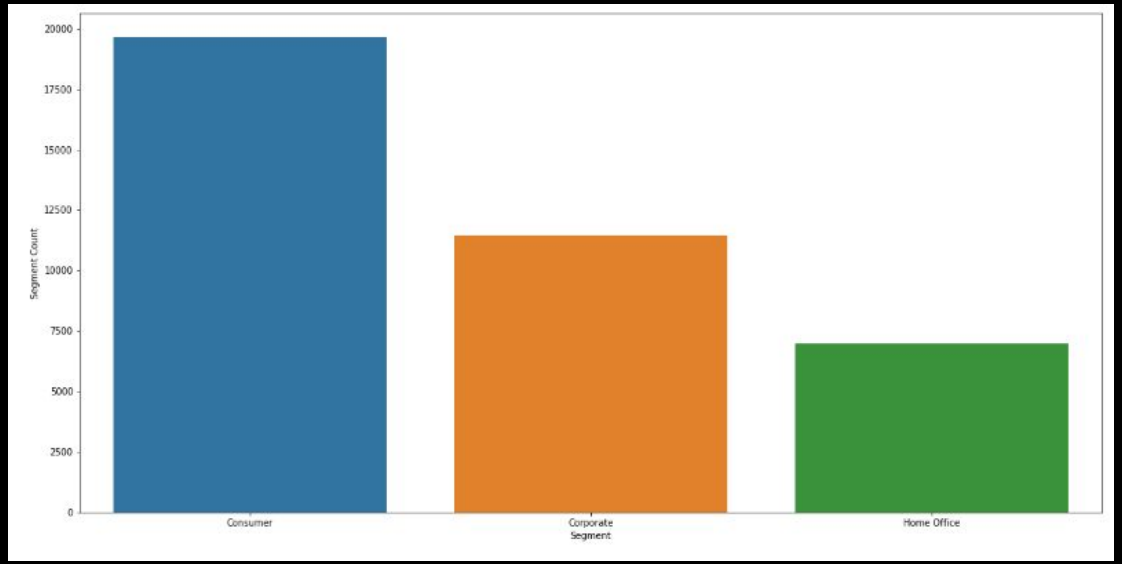
Customers

The highest profiting customer is almost $9,000.



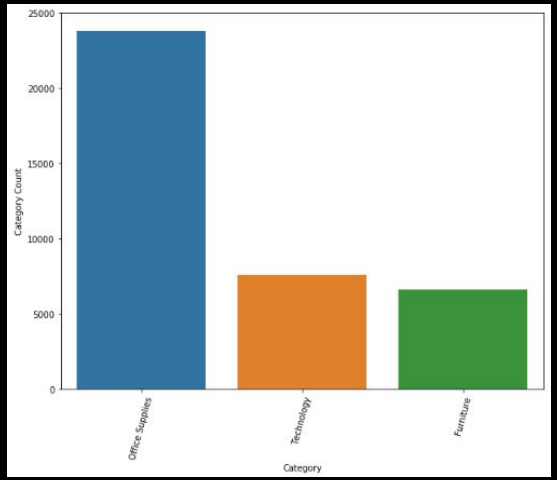
Customers

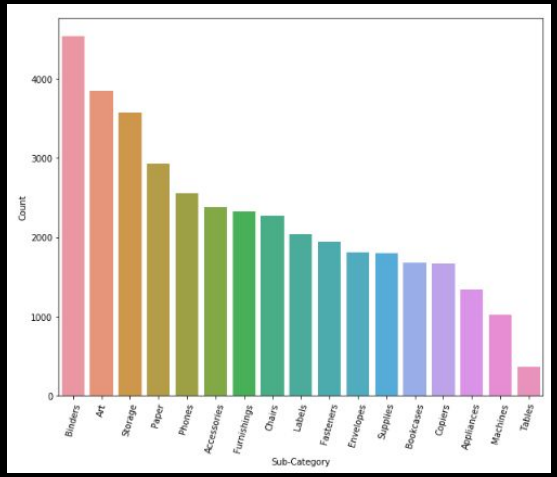
Consumers are the largest type of customers.



Products

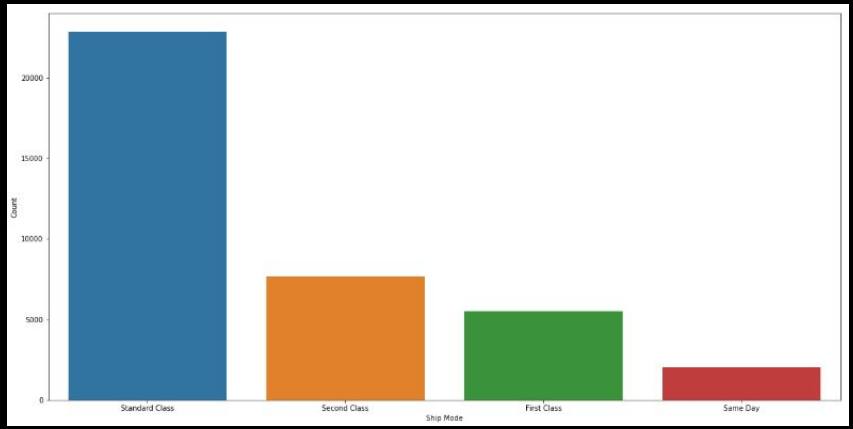
- Office Supplies is largest category based on orders
- Blinders and Art sub-categories have the largest number of orders

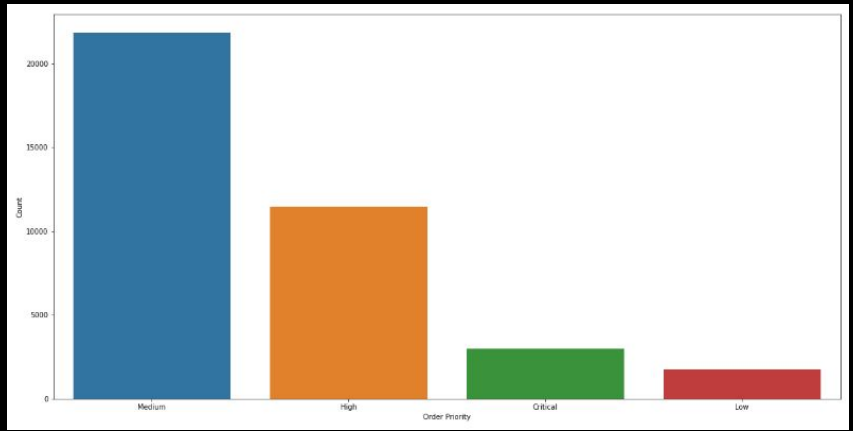




Products

- Majority of customers use standard class for shipping method
- Medium is the highest used order priortiy





RFM Segmentation

- Recency: most recent activity from customer
- Frequency: how often the customer makes a transaction
- Monetary: the amount of spent by customer

Segmentation

Each customer score was created based on RFM. The highest score possible is 9 and lowest is 3. Based on their score, the customer is divided into Elite, High, Medium, and Low.

	Customer ID	Recency	Frequency	Spending	R	F	M	RFM_Score	Customer_Score
0	AA-10315102	358	1	348.51	3	1	1	5	Medium
1	AA-10315120	959	1	18993.87	2	1	1	4	Medium
2	AA-10315139	149	12	9707.91	3	2	1	6	High
3	AA-103151402	184	5	2702.18	3	1	1	5	Medium
4	AA-103151404	818	3	1507.02	2	1	1	4	Medium
...	...	...	...	...	...	...	...	...	...
14307	ZD-2192548	750	3	478.32	2	1	1	4	Medium
14308	ZD-2192564	1409	1	490.86	1	1	1	3	Low
14309	ZD-219257	1198	1	239.76	1	1	1	3	Low
14310	ZD-2192582	196	2	1946.32	3	1	1	5	Medium
14311	ZD-2192596	749	2	1476.33	2	1	1	4	Medium

Pre-Processing

Test data was made and added more categorical data to have a higher specificity in classification.

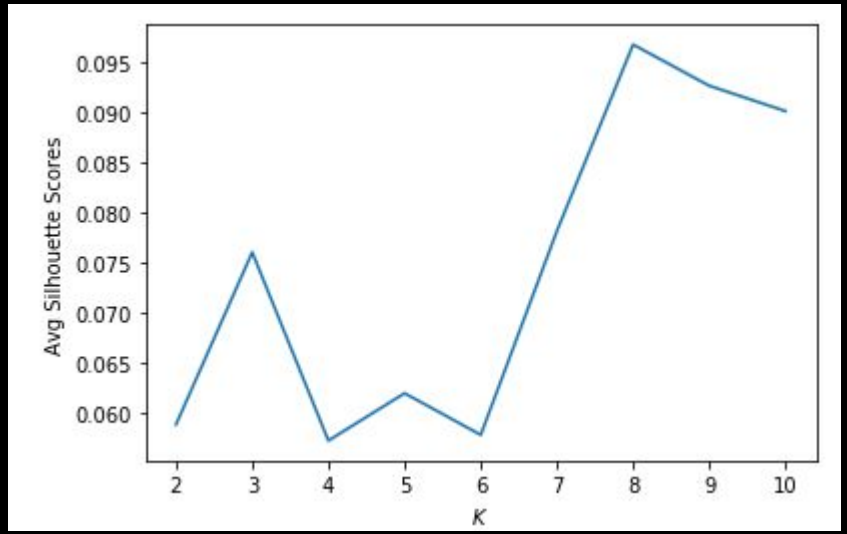
- Log was used to unskew the data
- Get_dummies was performed on categorical data
- StandardScaler was used to scale the data

	Recency	Frequency	Spending	R	F	M	RFM_Score	Ship Mode	Segment	Region	Market	Category	Sub-Category	Order Priority
0	358	1	348.51	3	1	1	5	Standard Class	Consumer	Southeastern Asia	Asia Pacific	Technology	Machines	Medium
1	959	1	18993.87	2	1	1	4	Standard Class	Consumer	Southern Europe	Europe	Furniture	Bookcases	Medium
2	149	12	9707.91	3	2	1	6	Second Class	Consumer	Northern Europe	Europe	Technology	Phones	High
3	149	12	9707.91	3	2	1	6	Second Class	Consumer	Northern Europe	Europe	Technology	Phones	High
4	149	12	9707.91	3	2	1	6	Second Class	Consumer	Northern Europe	Europe	Furniture	Bookcases	High
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38073	1198	1	239.76	1	1	1	3	Second Class	Consumer	Oceania	Asia Pacific	Office Supplies	Fasteners	Medium
38074	196	2	1946.32	3	1	1	5	Second Class	Consumer	Central America	LATAM	Technology	Copiers	Critical
38075	196	2	1946.32	3	1	1	5	Standard Class	Consumer	Central America	LATAM	Furniture	Furnishings	Low
38076	749	2	1476.33	2	1	1	4	Standard Class	Consumer	Northern Europe	Europe	Technology	Accessories	Medium
38077	749	2	1476.33	2	1	1	4	Standard Class	Consumer	Northern Europe	Europe	Office Supplies	Paper	Medium

Silhouette Method

Silhouette method was performed to ensure the correct optimization of clusters.

- 8 is the optimal number of clusters



Machine Learning Models

Type: Supervised Learning - Classification

Train-Test: Train 80% and Test 20%

Models: Decision Tree, Random Forest, Naive Bayes, Gradient Boosting

Hyperparameter tuning: GridSearchCV

Decision Tree

Train Score: 0.83

Test Score: 0.83

Avg Validation Score: 0.85

Accuracy Score: 83.2%

Run Time: 0.2s

	precision	recall	f1-score	support
0	0.91	0.95	0.93	1529
1	0.97	0.59	0.74	1854
2	0.97	0.73	0.84	256
3	0.63	0.36	0.45	337
4	0.75	1.00	0.85	1884
5	0.85	0.83	0.84	550
6	0.61	0.86	0.71	518
7	1.00	1.00	1.00	688
accuracy			0.83	7616
macro avg	0.83	0.79	0.79	7616
weighted avg	0.86	0.83	0.82	7616

Random Forest

Train Score: 1.0

Test Score: 1.0

Avg Validation Score: 0.0

Accuracy Score: 100%

Run Time: 5.14s

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1529
1	1.00	1.00	1.00	1854
2	1.00	1.00	1.00	256
3	1.00	1.00	1.00	337
4	1.00	1.00	1.00	1884
5	1.00	1.00	1.00	550
6	1.00	1.00	1.00	518
7	1.00	1.00	1.00	688
accuracy			1.00	7616
macro avg	1.00	1.00	1.00	7616
weighted avg	1.00	1.00	1.00	7616

Naive Bayes

Train Score: 0.99

Test Score: 0.99

Avg Validation Score: 0.85

Accuracy Score: 100%

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1529
1	1.00	1.00	1.00	1854
2	1.00	1.00	1.00	256
3	1.00	1.00	1.00	337
4	1.00	1.00	1.00	1884
5	1.00	1.00	1.00	550
6	1.00	1.00	1.00	518
7	1.00	1.00	1.00	688
accuracy			1.00	7616
macro avg	1.00	1.00	1.00	7616
weighted avg	1.00	1.00	1.00	7616

Gradient Boosting

Train Score: 1.0

Test Score: 1.0

Avg Validation Score: 1.0

Accuracy Score: 100%

Run Time: 76.17s

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1529
1	1.00	1.00	1.00	1854
2	1.00	1.00	1.00	256
3	1.00	1.00	1.00	337
4	1.00	1.00	1.00	1884
5	1.00	1.00	1.00	550
6	1.00	1.00	1.00	518
7	1.00	1.00	1.00	688
accuracy			1.00	7616
macro avg	1.00	1.00	1.00	7616
weighted avg	1.00	1.00	1.00	7616

Conclusions

Model: Random Forest

Future Improvements:

- Increase more features - unable to currently due to lack of computational power
- Target Marketing for superstore to focus on. For example, create a reward system based on RFM segmentation