Bryan Mahony – Online Retail Customer Segmentation and Sales Analysis

Background

Our client is a UK-based online retailer (which we will refer to as OnlineCo) that sells specialty, seasonal gifts. Given a dataset of 18 months of online product purchases, we aim to deliver the following:

- An analysis and presentation of any underlying patterns in customer behavior
- An understanding of key revenue drivers with respect to product sales and seasonality, and any underlying patterns in product purchases
- Recommendations to the business to maximize revenue based on the dataset and suggestions for further exploration with additional data sources

Executive Summary

After preprocessing and cleaning the dataset, we used an RFM analysis (recency, frequency, monetary value) to identify three unique customer profiles within the timeframe. Based on this analysis, we can make rough predictions of customers' lifetime value, identify predictors of customers with high LTV, and describe purchasing behavior specific to each segment.

We identified 88 products which comprise the top 50 products by revenue for each customer segment, suggesting that we can develop marketing and sales campaigns that appeal to all customer segments and tailor them to segments as appropriate.

Following this analysis, we used Tableau to visualize our data and identified three key trends that have implications for the business:

- Newer customer cohorts are demonstrating lower retention rates than older cohorts. This could indicate a need for more aggressive retention and reacquisition strategies.
- High seasonality in purchase behavior. This could serve to better inform business planning processes related to cash flow and revenue generation.
- Weak December sales across all customer segments. For a highly seasonal business with demonstrated success in capitalizing on other holiday sales windows and a sizeable segment of loyal, frequent customers, December-focused sales efforts could represent a significant business opportunity without the need for additional customer acquisition efforts.

Data

The sales data shared by OnlineCo covers a time period of 18 months between 1/12/20 and 9/12/20. Each of the 525,460 rows in the database represents a product purchased online, grouped by an invoice number and a customer ID (where applicable). Each row contains 9 features describing the product and its sale.

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Item Total
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085	United Kingdom	83.4
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085	United Kingdom	100.8
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085	United Kingdom	30.0

Figure 1: Sample dataset view from OnlineCo

Through our exploratory data analysis, we found the following:

> The dataset covers 28,816 unique transactions, of which 4,592 were returns (we will separate these and treat them differently later in the analysis)
>
> We have 4,383 identified customer IDs, but 20% of the data does not have a customer ID associated with them and **do** have an invoice code. We will group that 20% by invoice code and treat them as unique, guest customers.
>
> Of the 4,383 customers, the top 25 customers performed 2,081 transactions (7% of total volume). Of the total sales volume of $9.46M over the time period, these top customers represent $1.02M – or 10.8% - of the total revenue.

In our data wrangling, we had two main issues to solve for: **guest transactions** and **creating standardized dimensions for the RFM analysis.** We intuited that purchase rows with a transaction id but without a customer id represented guest transactions – e.g. transactions completed without creating an account.

To handle guest transactions, we generated a synthetic customer ID (outside of the range of 12346-18287 of "organic" customer IDs) for each invoice missing a value in the "Customer ID" field. In the RFM analysis, we treated these as customers with a single purchase.

RFM Analysis and K-means clustering

An RFM analysis seeks to assess customers across 3 dimensions: the recency of their most recent transaction, the frequency of their purchases over the specified time period, and the total monetary value (MV) of the purchases. For our purposes, we generated a recency feature indicating the number of days between last purchase and the end of the timeframe; a frequency feature indicating number of purchases during the time period; and a monetary value feature indicating total net revenue generated by each customer (as a reminder, some transactions were returns, which present a negative revenue figure).

We used sklearn's StandardScaler to create a comparable baseline across these dimensions – i.e. scaling each feature to mean zero, with values representing the number of standard deviations away from the mean. This was critical as Kmeans models are sensitive to unit magnitude (i.e., the number of days

between 1/12/09 and 9/12/20 has a different order of magnitude than the total MV of purchases, which would throw off our machine learning model).

After grouping our customers by recency, frequency, and monetary value, we used a cross-validated KMeans clustering approach to identify clustering patterns in customer profiles – i.e. identifying customer segments.

Using yellowbrick's KMeansVisualizer, we can see that a K of 3 produces the optimal distortion score – that is, there are three optimal segments in our customer data.

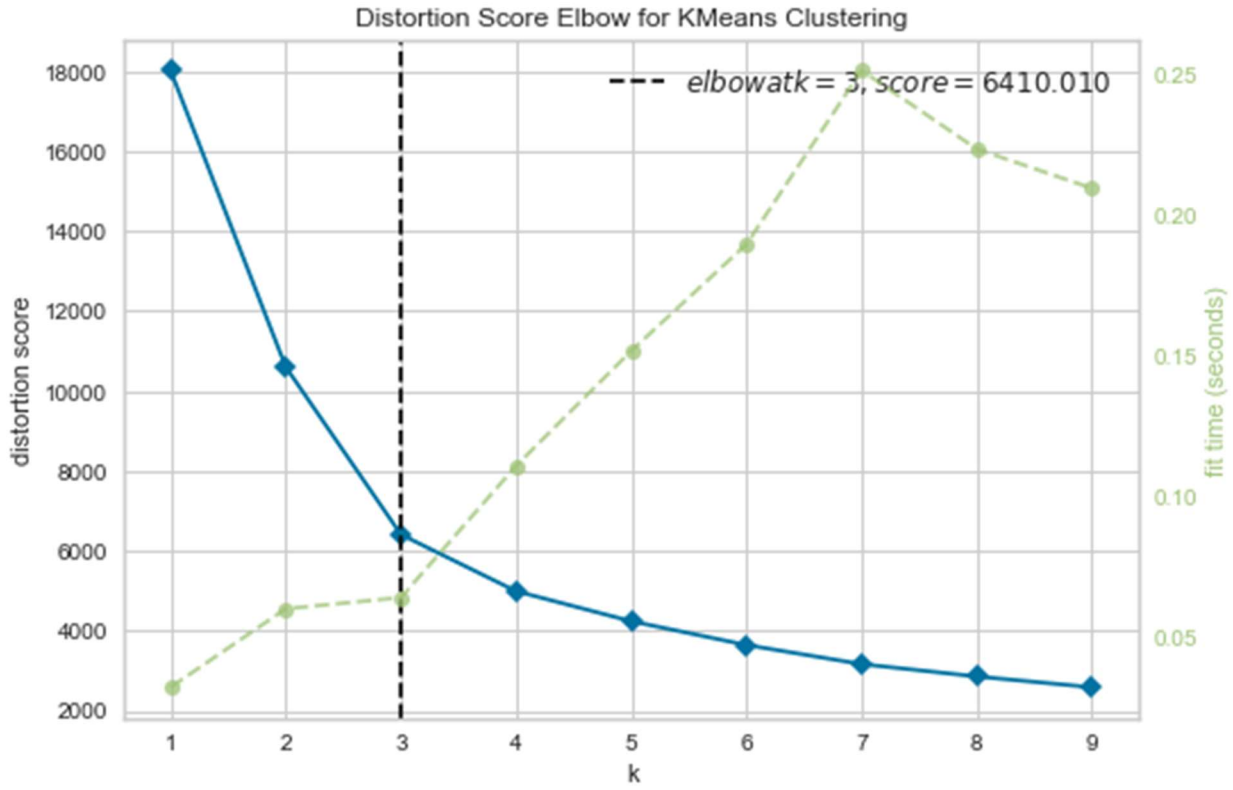


Figure 2: Distortion score comparison for different k-values

Using a k of 3, our KMeans model can assign each of our customers to one of the 3 segments.

In the visual below, you see the segments represented by **color,** with (scaled) frequency and recency on the axes and size of the datapoint indicating monetary value.

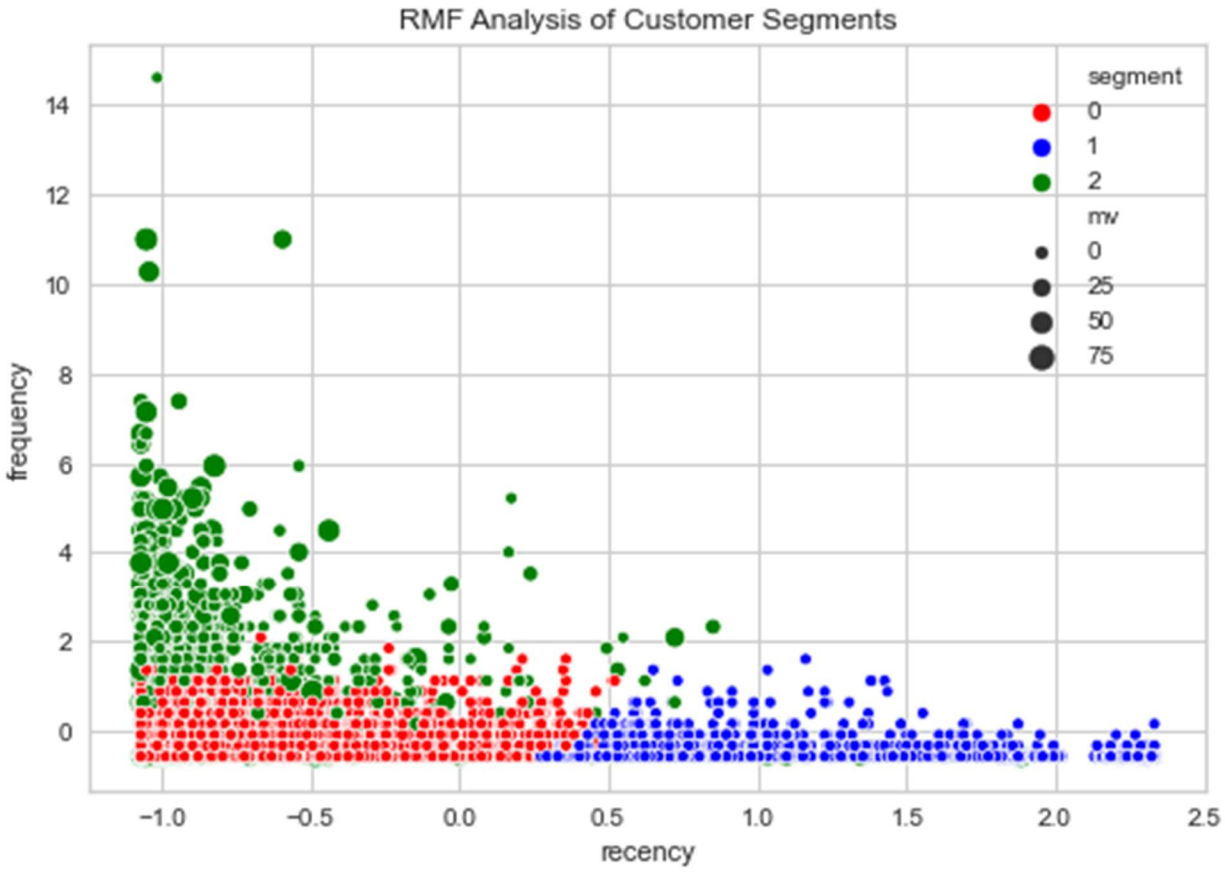


Figure 3: K-means analysis identifies 3 unique customer segments

We can intuit our three customer profiles as follows based on the graph and the cluster centers provided by our KMeans model:

- **Low frequency, low recency, low MV customers** - customers who have purchased recently, have a small number of overall purchases, and make relatively lower total purchases (indicated in **red**)
- **Low frequency, high recency, low MV customers** - customers who have not purchased recently with a small number of overall purchases, and a low total purchase volume (indicated in **blue**)
- **High frequency, high recency, high MV customers** - customers who purchase relatively often and recently, and have a high purchase volume (indicated in **green**)

Our strategic recommendations will consider these profiles in light of their overall value and their purchasing habits.

Below you will see a Jupyter output of two product sets: the products that generated the max gross revenue, and the products most often purchased with them.

Description	Quantity	Item Total
REGENCY CAKESTAND 3 TIER	13092	163051.46
WHITE HANGING HEART T-LIGHT HOLDER	57427	155825.52
ASSORTED COLOUR BIRD ORNAMENT	44917	72454.12
PAPER CHAIN KIT 50'S CHRISTMAS	17083	57870.20
JUMBO BAG RED RETROSPOT	30306	54332.97
PARTY BUNTING	10075	49645.52
ROTATING SILVER ANGELS T-LIGHT HLDR	13675	47672.49
POSTAGE	2154	46092.36
JUMBO BAG STRAWBERRY	19818	35854.59
VINTAGE UNION JACK BUNTING	4120	35819.71
EDWARDIAN PARASOL NATURAL	7318	35748.00
STRAWBERRY CERAMIC TRINKET BOX	26562	33834.70

Figure 5: Max gross revenue generating products

Description	Quantity	Item Total
WOOD S/3 CABINET ANT WHITE FINISH	2713	20638.08
ROUND SNACK BOXES SET OF4 WOODLAND	7047	20363.21
NATURAL SLATE HEART CHALKBOARD	6539	20350.89
RECYCLING BAG RETROSPOT	9162	19765.67
VICTORIAN GLASS HANGING T-LIGHT	15884	19515.48
WOODEN PICTURE FRAME WHITE FINISH	7965	18946.72
JUMBO BAG OWLS	9746	18341.58
3 HEARTS HANGING DECORATION RUSTIC	5906	18335.00
60 TEATIME FAIRY CAKE CASES	36326	18128.25
GIN + TONIC DIET METAL SIGN	8675	17307.11
ZINC METAL HEART DECORATION	14240	17295.10
COOK WITH WINE METAL SIGN	9292	16794.44

Figure 4: Most frequently co-purchased products

Strategic Recommendations

	Segment 0	Segment 1	Segment 2
Description	**Low frequency, low recency, low MV**	**Low frequency, high recency, low MV customers**	**High frequency, high recency, high MV customers**
Total sales over time period	**$5.87M**	**$2.72M**	**$.86M**
Total customer #	**3,615**	**755**	**1957**
Avg transactions/customer	**3.95**	**11.38**	**1.34**
Median transaction value	**$229.57**	**$247.60**	**$139.34**
Strategy type	**Upselling**	**Reacquisition**	**Retention**

The 88 top-grossing products across customer segments represent $2.46M of the $9.46M total sales. Targeted campaigns promoting sales of these 88 products (representing the common products across

the top 50 purchased products by segment) could boost customer retention across all segments and generate significant additional revenue through increased purchasing.

These 88 products also represent items usually copurchased - that is, things like cupcake wrappers and decoration that are usually bought among other items. Through our analysis we have also offered an itemwise list of highly copurchased items – the below representing the top 5 products that are not themselves top sellers but are frequently copurchased with top sellers.

	Desc	Count
42	KENSINGTON COFFEE SET	26
63	ENGLISH ROSE DESIGN QUILTED THROW	14
32	PAPER CHAIN KIT 50'S CHRISTMAS	13
68	POLKADOT CUTLERY 24 PCS IN TRAY	5
34	MIRROR, ARCHED GEORGIAN	4

Figure 6: Frequently co-purchased products

Switching to Tableau, we can begin to do some additional visualization and exploration in the data. You may view the following Tableau views in story form here.

First, to perform a basic business analysis, we can view customer churn by time cohort of first purchase in the dataset. We can see, using a standard goal of 20% customer retention, that newer cohorts are underperforming in terms of retention.

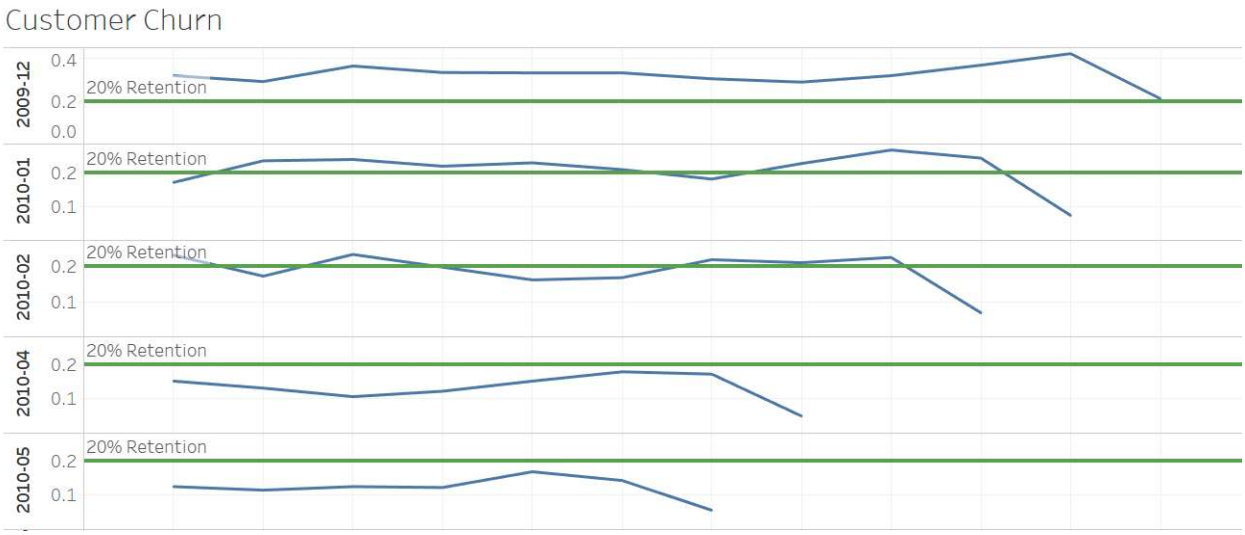


Figure 7: Customer retention by month by time cohort

The absolute number of customers has increased over the past ~4 months at a more rapid pace than we've seen before. Does increased customer base = lower touch/engagement = lower retention?

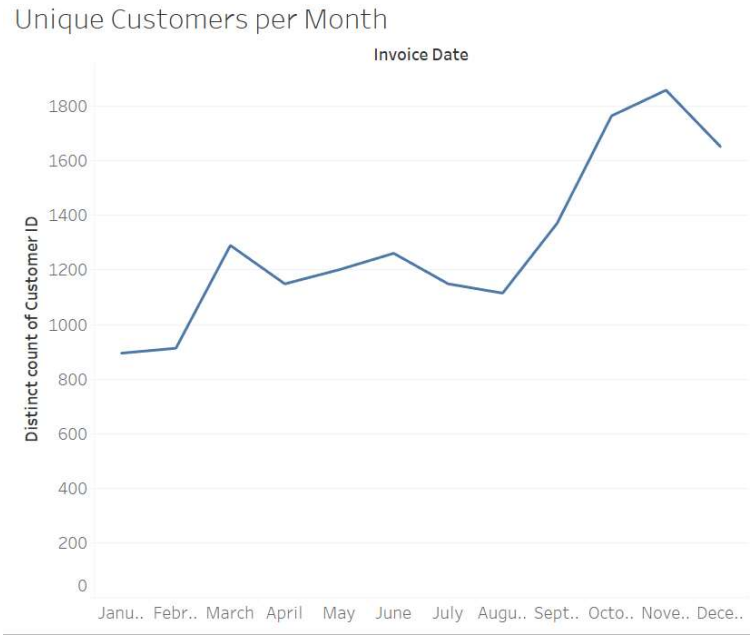


Figure 8: Unique customers purchasing by month

We can see that there is a strong seasonality in our sales - to be expected as the client is a seasonal gifts webstore. Importantly, we can see that revenue per customer goes up dramatically as well, indicating that our customer acquisition strategies are working effectively.

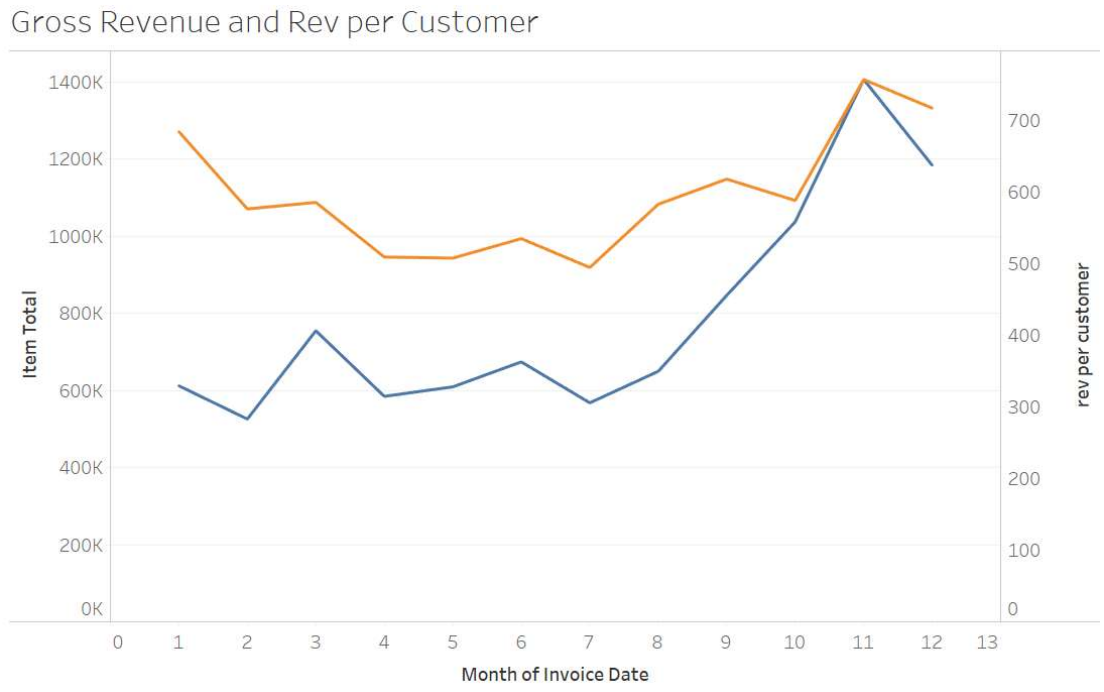
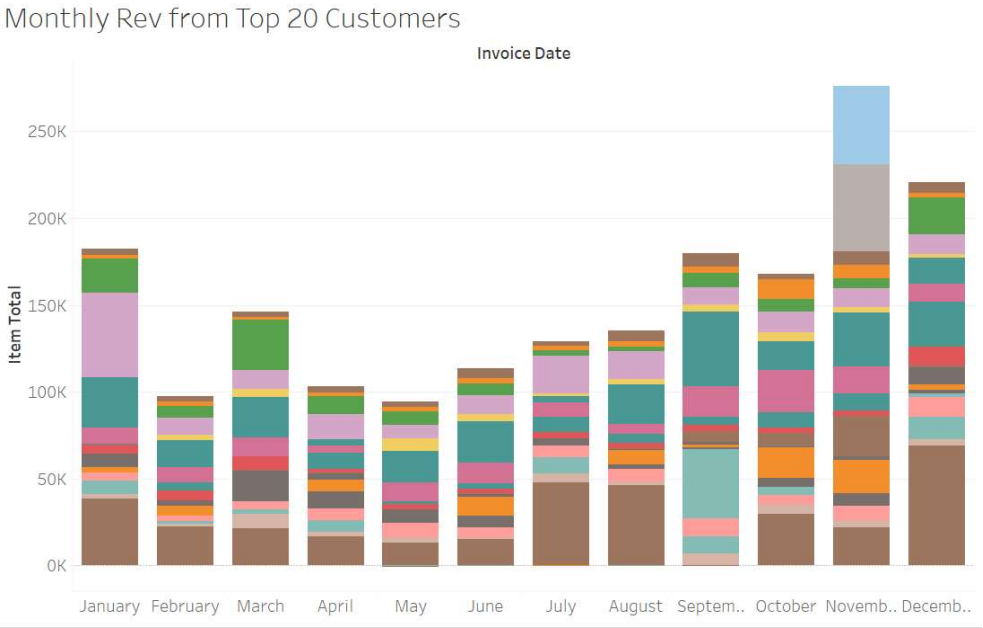
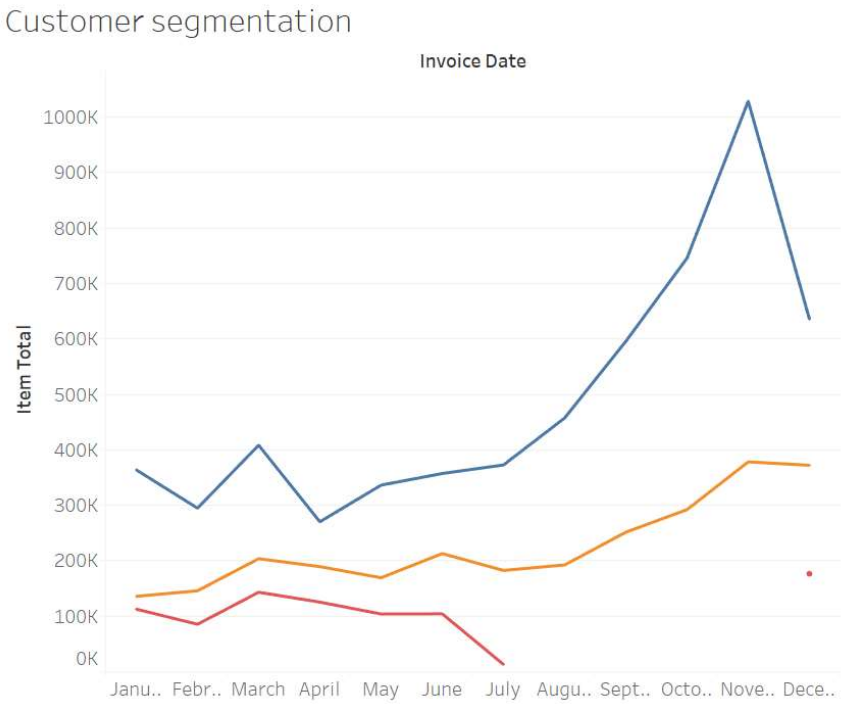


Figure 9: Gross revenue against revenue per customer

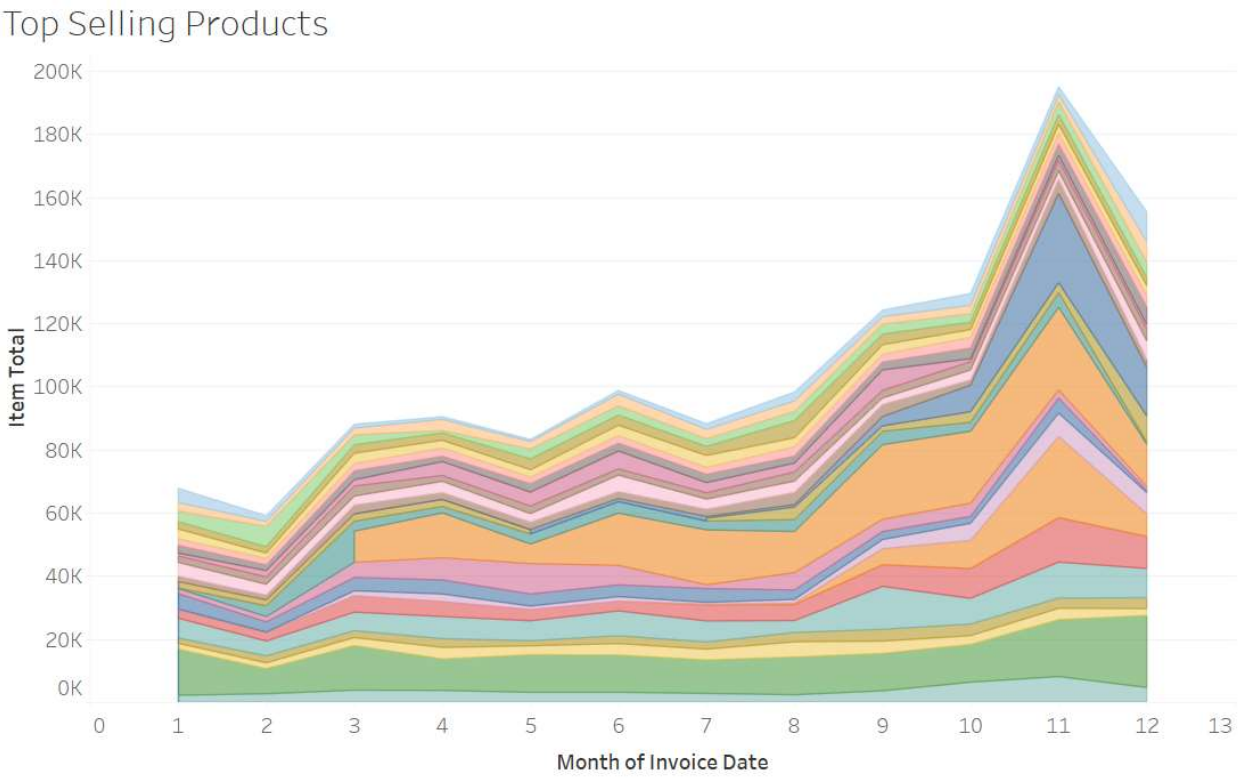
We see a similar trend in our top 20 customers - a strong seasonality but also a dip in this group in December. We know we increased both the number of customers and the revenue per customer during the last month of the year, so we may want to focus our engagement/upselling tactics with this top segment.



From our previous segmentation exercise, we know we have three distinct segments - and can see below their various purchasing patterns. The dip in December is still noticeable for our top revenue generating group.



Our 25 top selling products also see this dip. And we can see that, while sales of some increase over time, our largest sellers tend to vary revenue over time. Developing a more refined understanding of what purchases may be correlated and the underlying factors playing into their seasonality could help us generate more revenue in the future.



Summary Recommendations

- OnlineCo has the opportunity to better target customer acquisition and retention strategies based on newly-identified customer profiles.
- OnlineCo has significant opportunities to improve sales and customer retention in light of seasonality identified in purchasing behavior
- OnlineCo now has a framework to approach revenue generation by cross-promoting frequently copurchased products, focusing campaigns during peak potential purchase times, and driving customer retention of larger numbers of newly-acquired customers to the same levels as previously seen