



Retail Challenge

Customer Segmentation and Forecasting

Overview

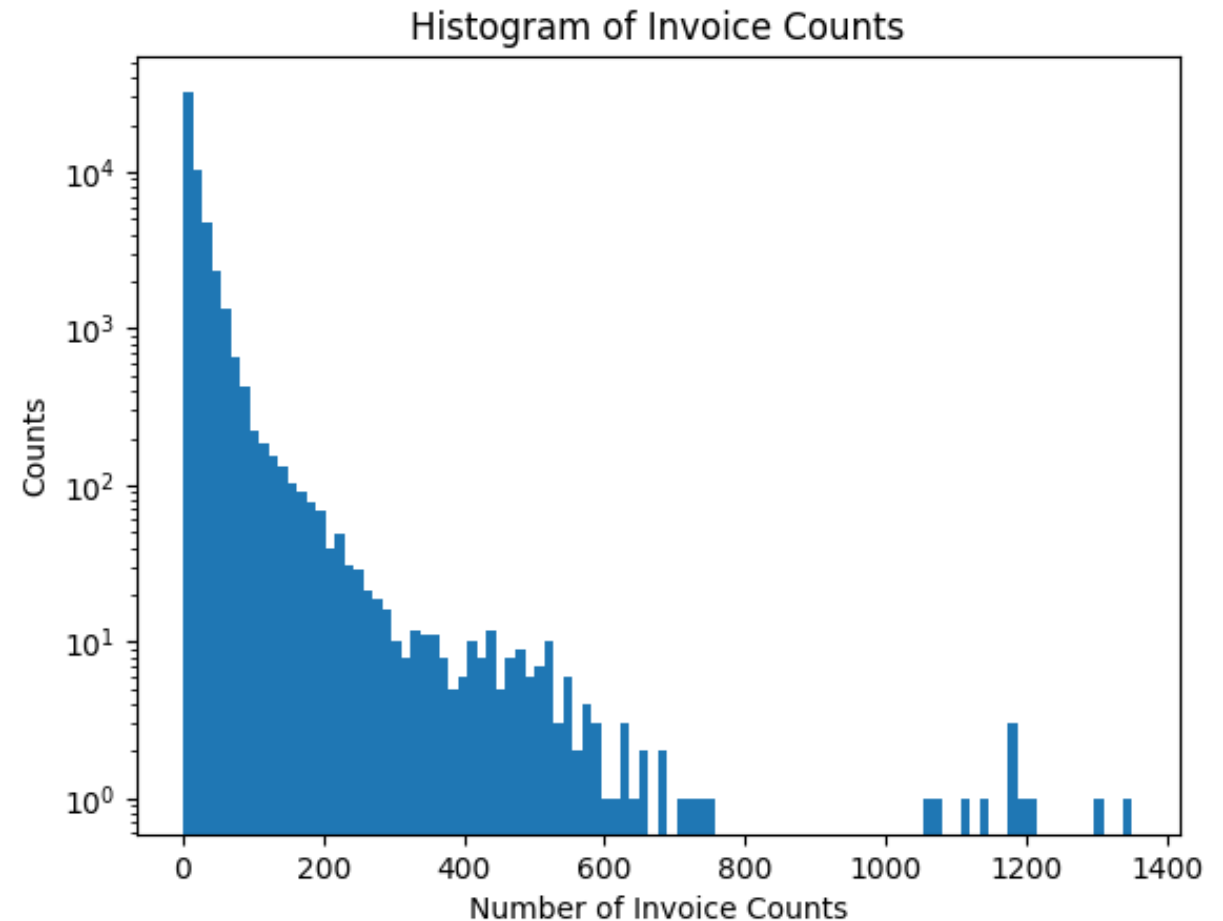
- A retail business specializing in household items.
- Based in the UK, also transacting internationally.
- All transaction data from end of 2009, all of 2010 and 2011 till early December.
- Wish to understand customer behaviour, possible segments, and future forecasts.



Image src: <https://clv.h-cdn.co/assets/17/20/1280x640>

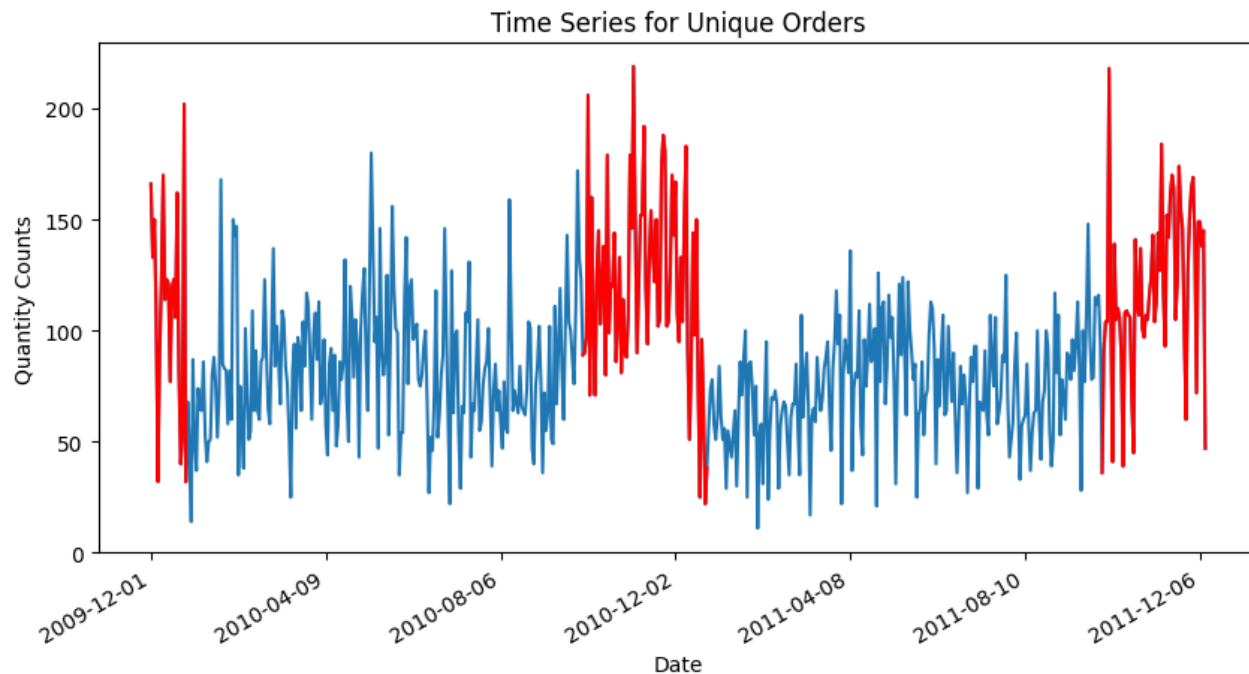
Data Section - EDA

- Most invoices have one transaction, some have over a 1000.
- ~2% of all invoices were cancellations
- Cancellations are also entered as a row with negative quantity for items and price.
- ~0.5% of all invoices were returns.




Data Section - EDA Continued


- Holiday months for Oct, Nov, and Dec (shown in red) do show higher number of invoices.
- ~90% of all invoices were for transactions within the UK.
- The business isn't open all days, so some days have zero transactions.






Data Section – Data Quality

- Data contains mixed transaction types; this can be problematic.
Ex: StockCode: TEST001, AmazonBill payments added as a transaction row,
 - Some transactions are missing Customer ID values, they are stored as 'nan'.
 - Not all StockCode values are 5-digit integers.
 - Significant orders and cancellations made within minutes.
- 

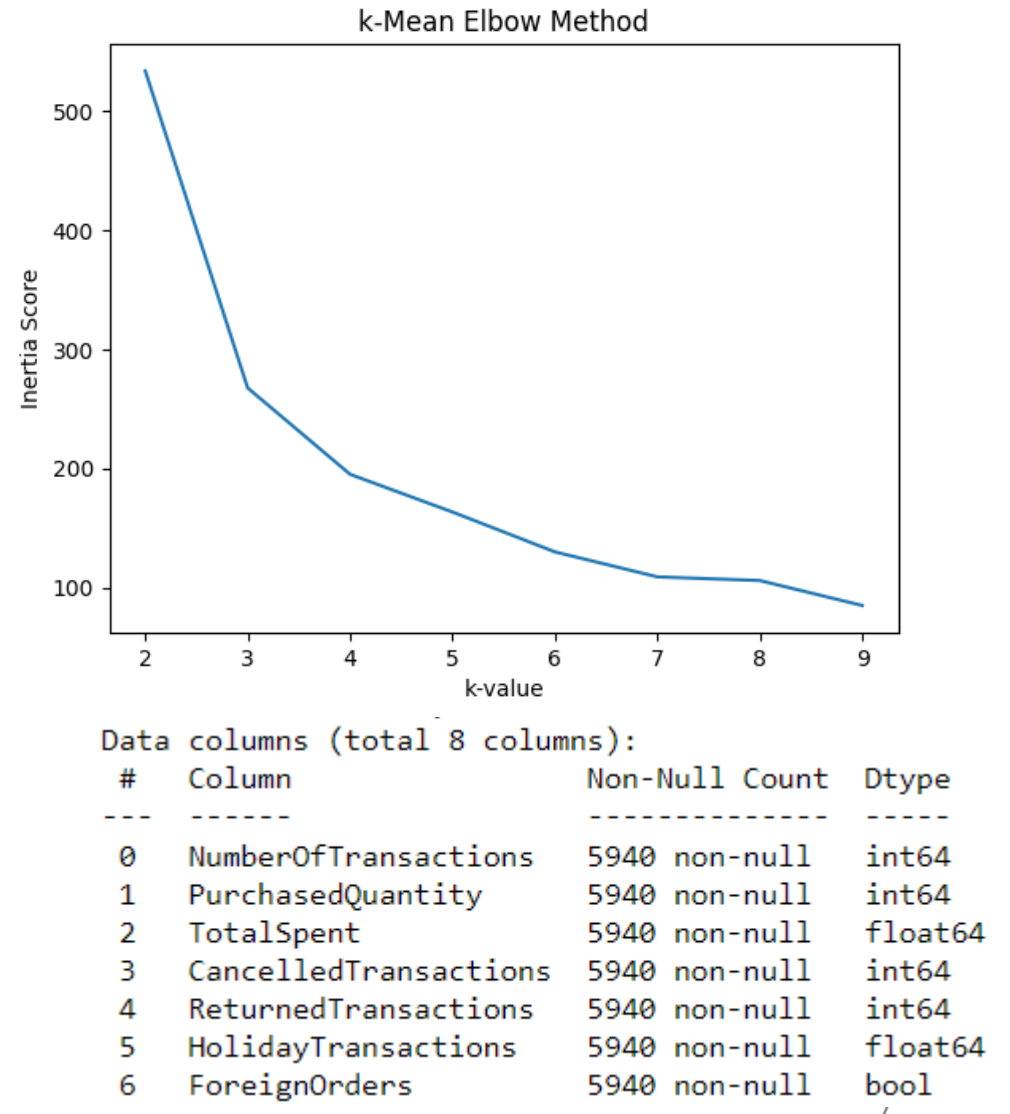


Data Section - Preprocessing

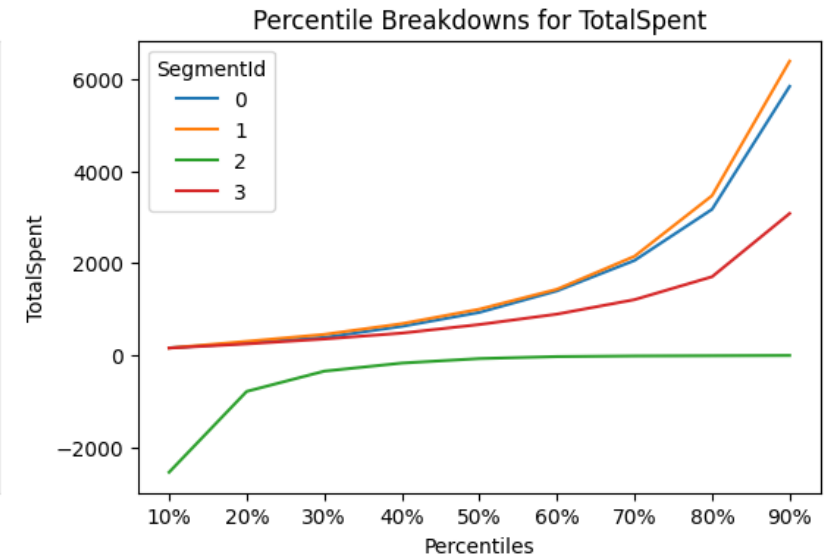
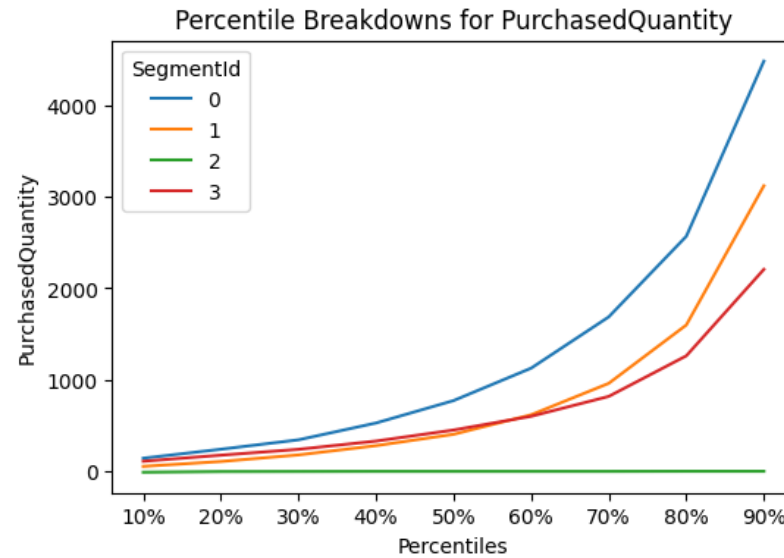
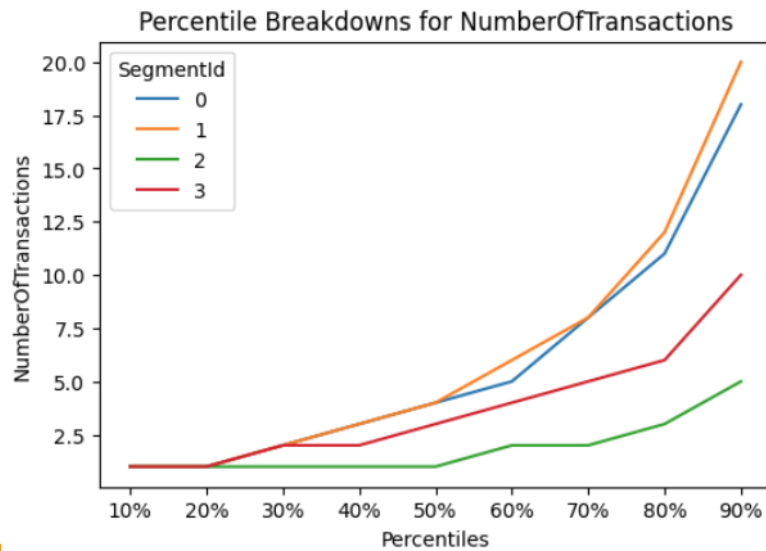
- Clean up the data based on data quality findings.
 - Group by customers and generate additional features from the data:
 - Number of cancellations or returns for a customer.
 - If the customer transacted during the holidays.
 - If the customer is an international customer, has had international orders.
 - Total transaction amount based on customer profile.
- 

Model Section - Clustering

- Used aggregated data for customers, along with generated features.
- Preprocessing for normalization, passed through k-values between 2-10.
- Went with a k-value of 4 to compute the clusters.

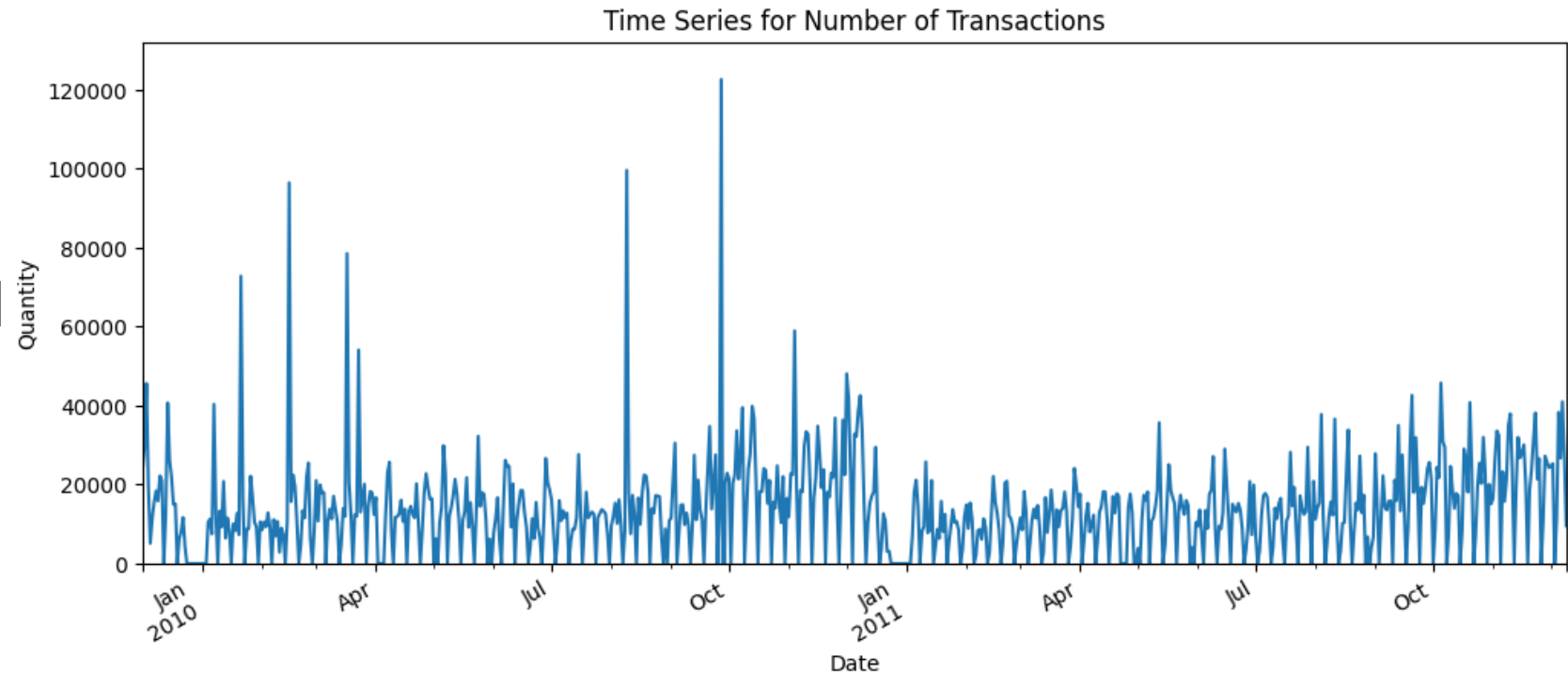


Model Section – Cluster Descriptions



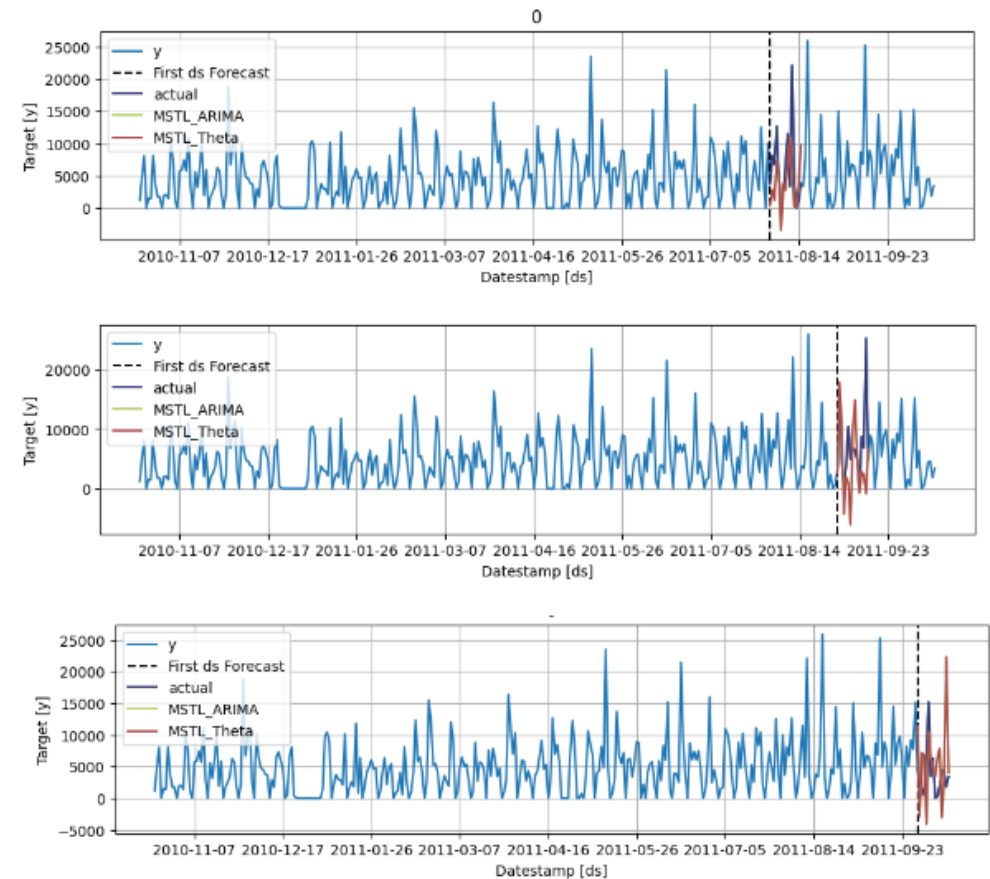
Model Section - Forecasting

- Reindexed the dates all existing values.
- Clear pattern of missing transactions during weekend and Christmas.

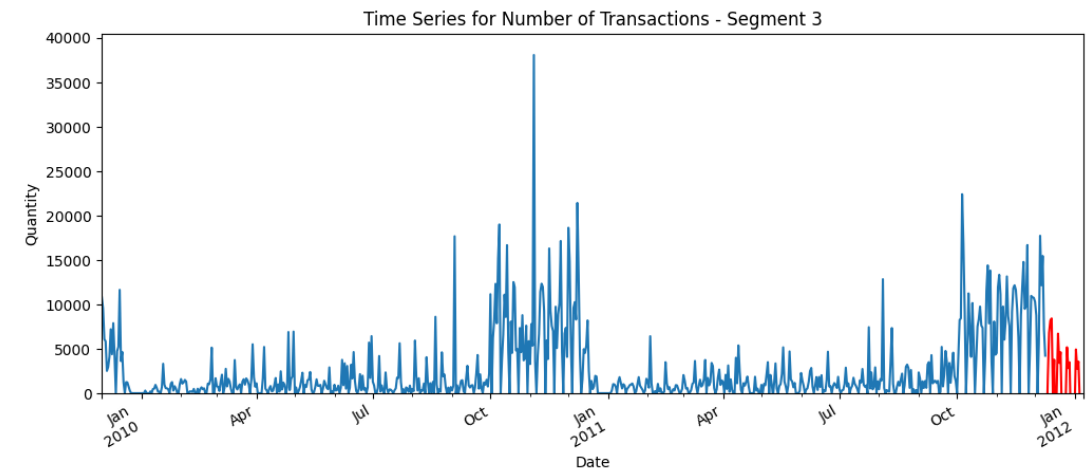
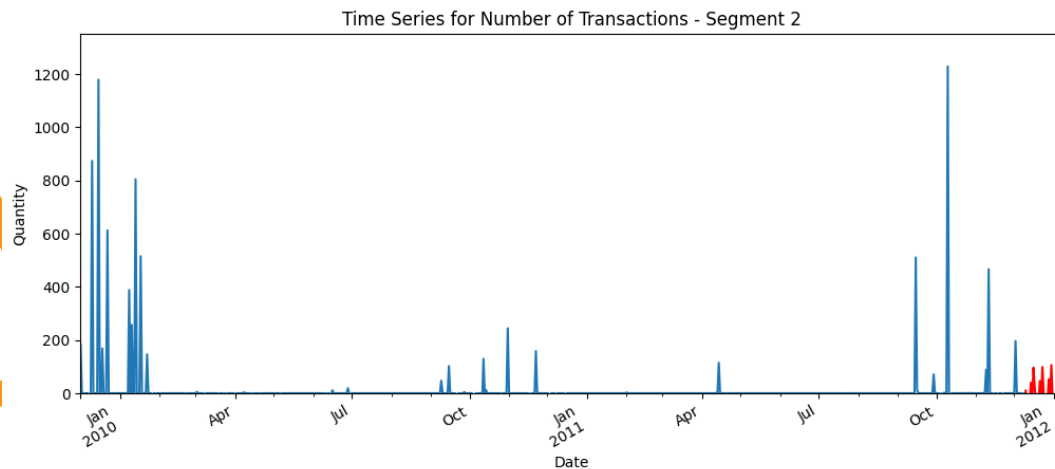
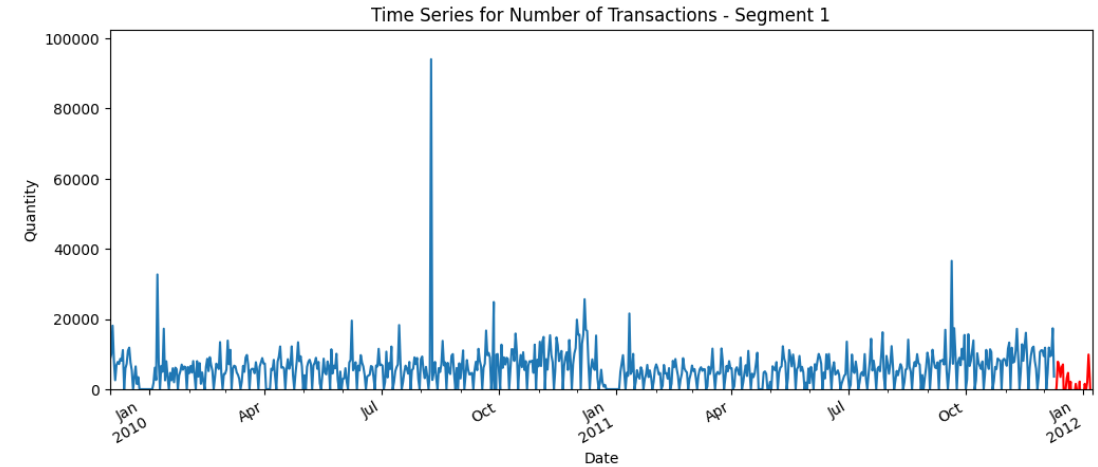
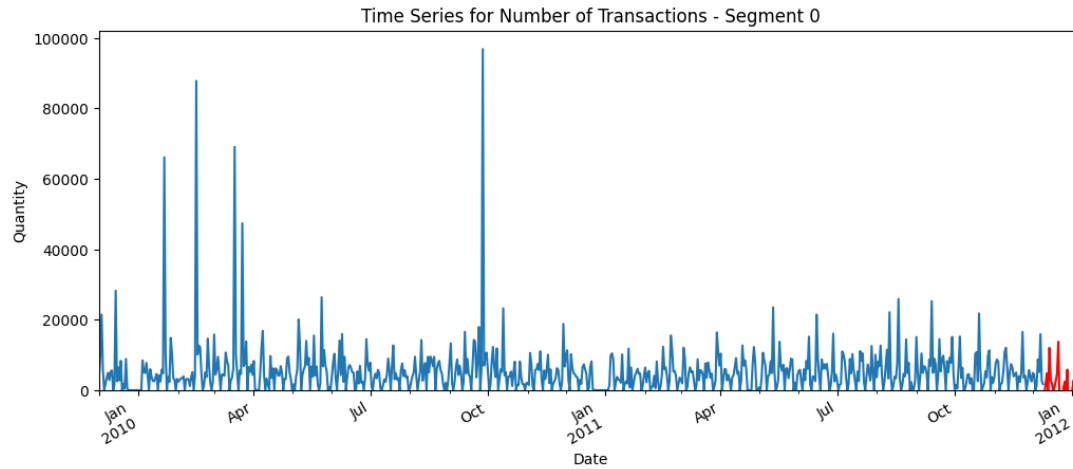


Model Section - Forecasting

- Transaction counts for all segments were passed through a multiple seasonality model for a weekly and yearly seasonality trend.
- 3-fold cross-validation for 15 day periods.
- Best performing model still had an MAE of over 40% after removing negative predictions.

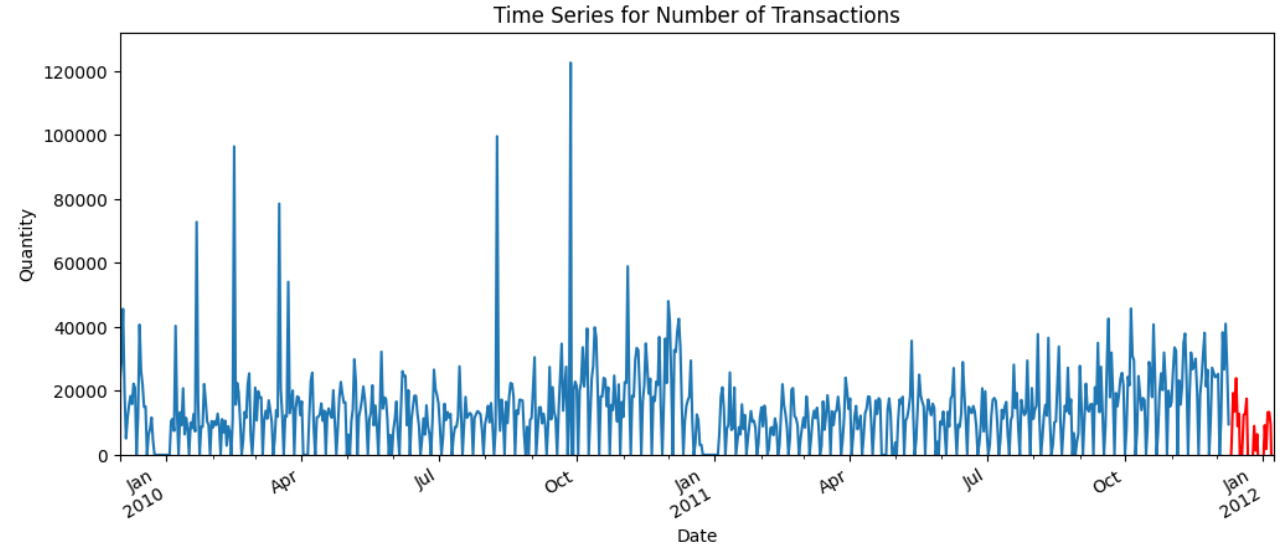


Results – Forecasts for Clusters



Conclusion

- Room for improvement with additional tuning and better segmented data.
- Segments usable for marketing initiatives.





Thank you