Examining Sales Data from an Anonymous Home Goods Company

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

# Abstract

Dianna Sinicrope

Bellevue University

Bellevue, USA dmsinicrope@my365.bellevue.edu

Jontavius Caston

Bellevue University

Bellevue, USA j.caston91@gmail.com

Companies would do well to embrace predictive analytics with their sales revenue data. Embracing predictive analytics by creating reliable forecasting methods allows a company to get a glimpse into the future of the company. With this glimpse, companies can develop the strategies necessary to optimize their sales and marketing efforts with the goal of positively influencing future revenue.

**Introduction**  
For this project, we've explored a dataset that details item sales sold by a UK-based online retailer. The dataset, released by Dr. Daqing Chen with the School of Engineering, London South Bank University, contains all transactions from the retailer between 1/12/2009 and 9/12/2011. The products are specialty gifts, and many customers are wholesalers.

This dataset is highly detailed and will allow us to identify key findings that could help the company optimize sales strategies and increase revenue.

**The Problem**

We would like to determine an optimal sales and marketing strategy for this online retailer. An overall understanding of the data will be particularly insightful, and we expect to be able to answer two different categories of questions in our analysis: data-driven questions and user-driven questions.

Data-driven:

- Which month generates the most sales and the most sales revenue? We expect seasonality to play a part in revenue. Time of day could also be a relevant factor, and we should consider it for online marketing strategies.

- Are there any item-specific sales that we should consider for either our supply chain strategies or pricing strategies?

- Are there any variables that have high correlations? Perhaps quantity sold and price are correlated and can be adjusted depending on supply and demand.

- Are sales increasing or decreasing year over year?

- Based on the data that we have access to, is the business in trouble or thriving?

User-driven:

- Who are our customers, and what kinds of customer segmentation can we perform to determine who they are? Perhaps segmenting customers based on location, items purchased, or money spent could lead us to better customer insights.

- Are there any specific customers or groups of customers that we should focus our sales strategies?

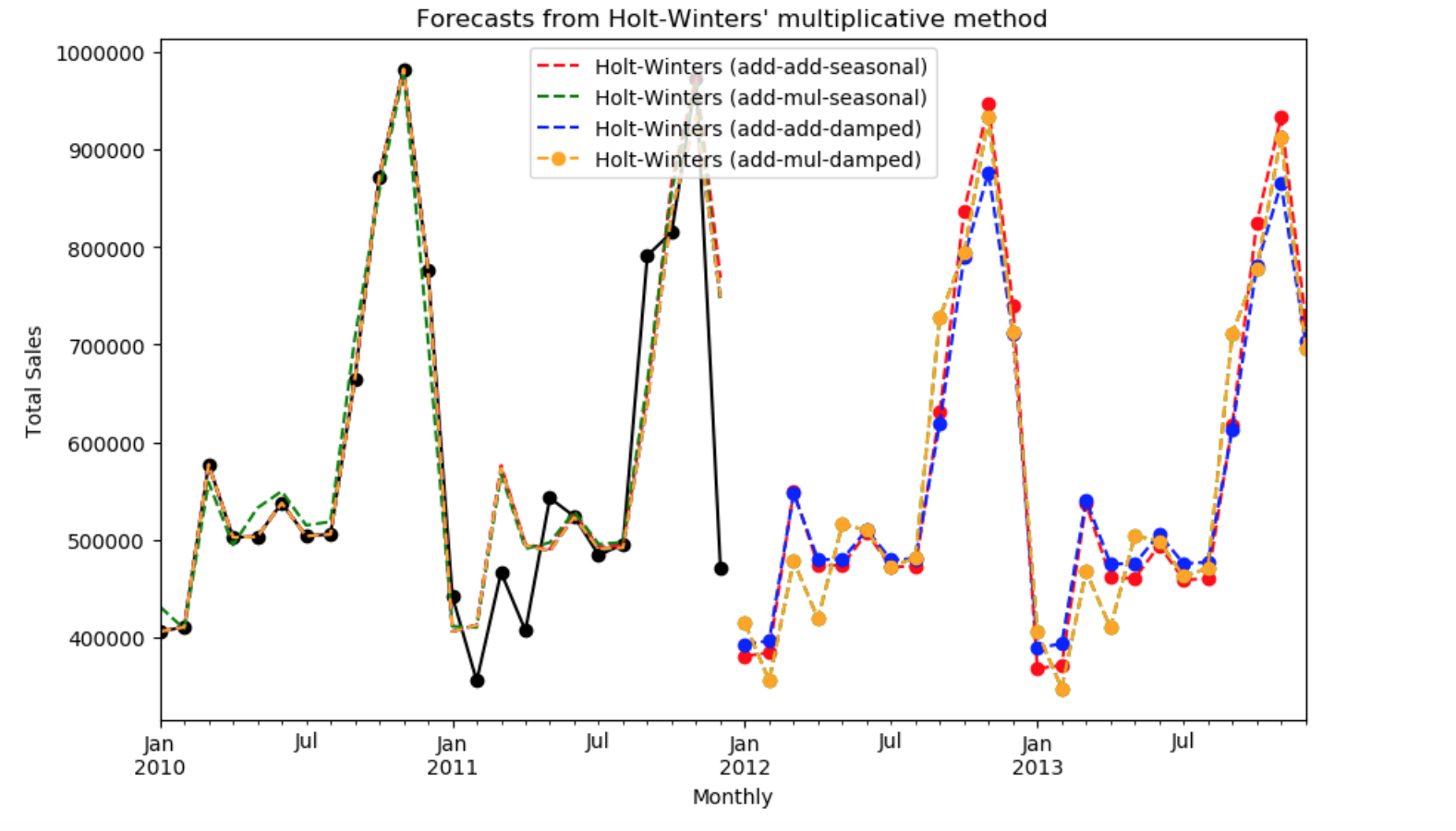
- How can we combine the insights we've learned from our sales-driven analytics with our user-driven analytics to optimize our sales strategy?

**Methods**

Data wrangling became a key factor in our preliminary efforts to understand the data. We were first interested in cleaning and validating the data since many variables have somewhat cryptic datapoints. We removed what we concluded were canceled and returned transactions and settled on a relevant dataset for our first look at the data from a statistical standpoint.

Determining possible correlations between variables in the dataset came next. Our first step was to separate the date column into separate columns. To do this, we used the '.to\_datetime' function to change the column from a string to integers. This allowed us to arrange the data better and conduct feature selection. We then built a correlation matrix using Pearson's correlation.

We also considered performing NLP clustering algorithms on our text variable that described each item sold. Using an LDA model particularly interested us since we've read about how it's a popular algorithm to use in NLP in the data science industry. We encountered some problems within the description variable, though, since the description of each item is rather short and succinct. Thinking that this variable might not have enough information for NLP clustering, we consulted a colleague (an experienced data scientist). Our colleague agreed with our initial thoughts- that this dataset is not a good candidate for clustering algorithms and that it would be a lot of work for little to no results. However, we discovered further through the project that text processing is most likely possible with this kind of seemingly unstructured text data. We had hoped to perform product matching with product databases from Amazon as a way to summarize product categories. However, we did not have time to complete that part of the project for this study.

Figure 1.1: 24 Month Forecasts

Our next step was to use the Holt-Winters method to create a sales forecast for the next 24 months (Figure 1.1). The Holt-Winters method was a logical choice for our dataset because it accounts for seasonality and trends, both of which apply to the data.

Using the statsmodels package, we fit four models using variations of the Holt-Winters method. The first model used an additive approach and the second a multiplicative approach. The third and fourth models also used additive and multiplicative approaches, but also included an adjustment for box cox (an adjustment for non-normal data distributions).

While resulting in slight variations, all the models produced in our analysis showed similar outputs.

**Results**

The most alarming result that we discovered is that each model in the Holt-Winters method showed sales decreases year over year. Based on 2010-2011 data, our predictive analytics show a -4% decrease in U.K.-based sales in 2012 and a -2% year over year decrease in 2013. Without intervention, the company should expect revenue losses rather than revenue gains.

One of the more unsatisfying results that we got was from the Pearson's correlation matrix that calculated the correlation between all of our original variables. The matrix showed only one strong correlation, and the rest were non-significant. Upon further inspection, we realized that the two highly correlated variables- Invoice number and year, only correlated since increasing invoice numbers were created as the year progressed. This isn't a true correlation then since Invoice is essentially an offshoot of the year variable.

It's interesting to note, though, that we did find correlations between extracted features. The next step of our process was to segment customers since we needed a way to organize the types of buyers in our customer base. Before we could begin to segment, though, we had to understand more summary metrics about our customer base. The first question to look into was customer engagement, both in terms of when they most recently visited and how frequently they visited.

For segmenting, we used a k-means algorithm that assigned a value score to each customer based on the customer’s revenue, frequency, and recency. The details of each step of the process are outlined below.

**Recency Clusters**

We created a histogram for recency to understand how recently our customers have visited us:

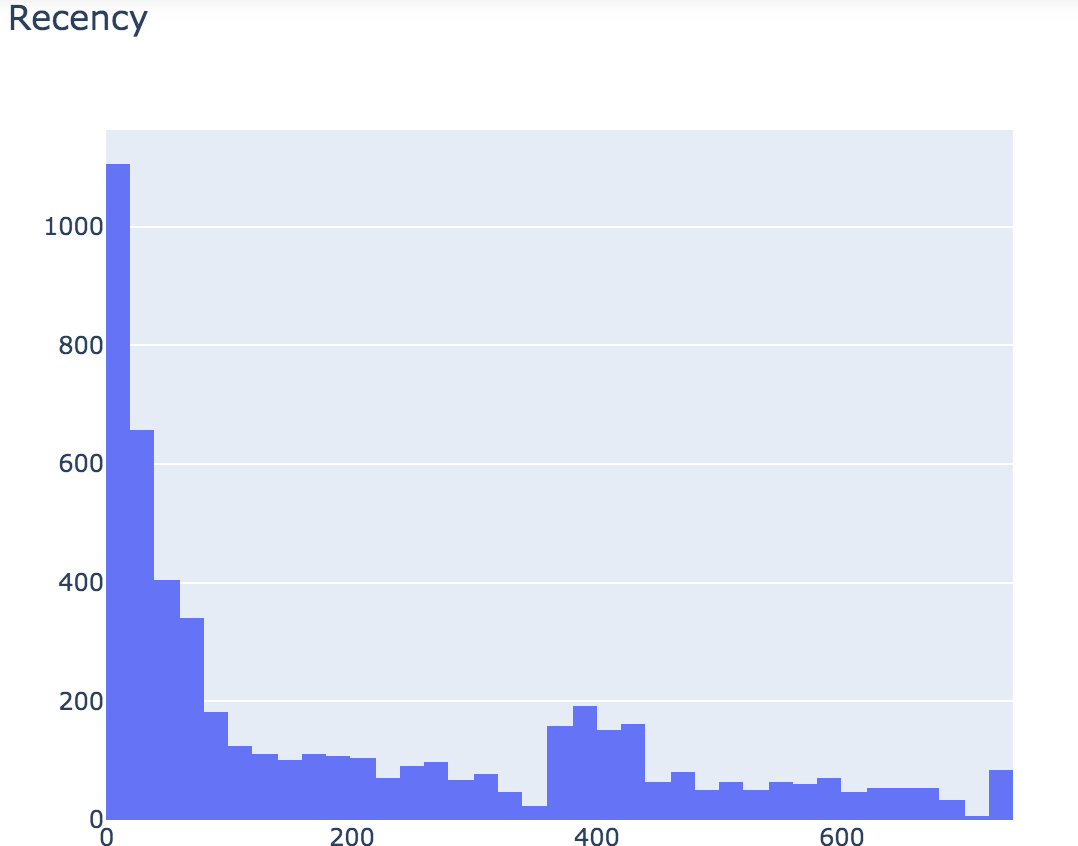


Figure 1.2: Histogram of Customer Recency

Our average customer visited us 201 days ago, but the range was as recently as today (i.e., the most recent day of the dataset) and as early as 738 days ago (which is really the span of our dataset). Median recency is 96 days ago. Here are the summary statistics:

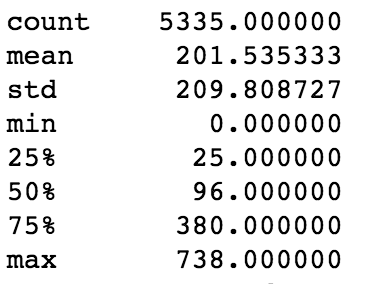
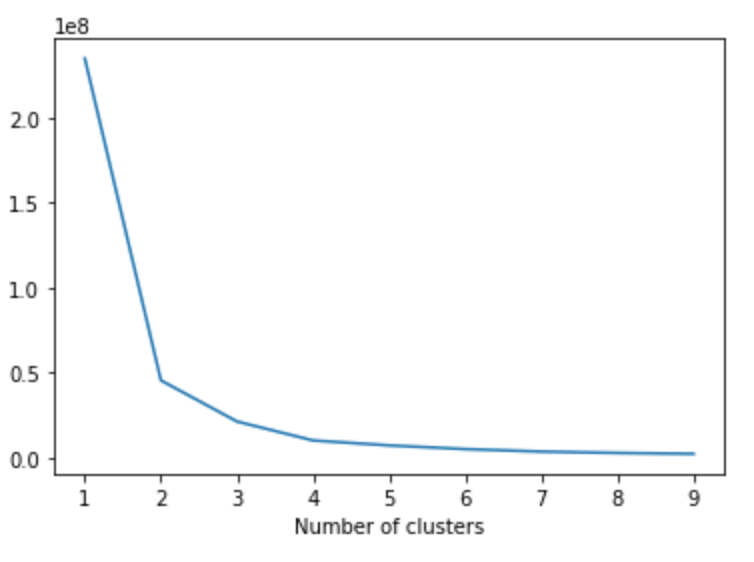


Table1.1: Recency Summary Statistics.

These groupings could be a good starting point for sales tactics based on customer segments, and we’ll address them later in our recommendations section.

We then performed an elbow test to estimate the number of clusters we would need to assign a recency score to each customer when using a k-means clustering algorithm.

Figure 1.3: Recency Elbow Test

We decided on k=4 as a good number of clusters but easily could have split them at k=2 or k=3 as well.

Each customer was then assigned to a recency cluster, based on how recently they had visited the site.

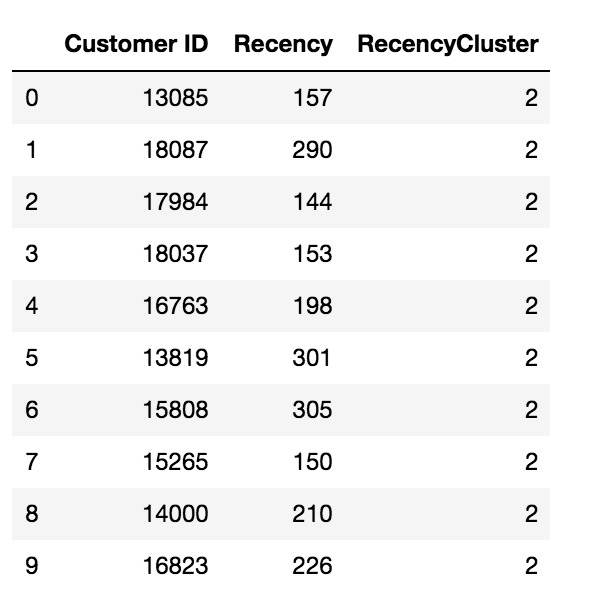


Table 1.2: Customer Assignments to Recency Clusters

We then calculated summary statistics for each recency cluster:

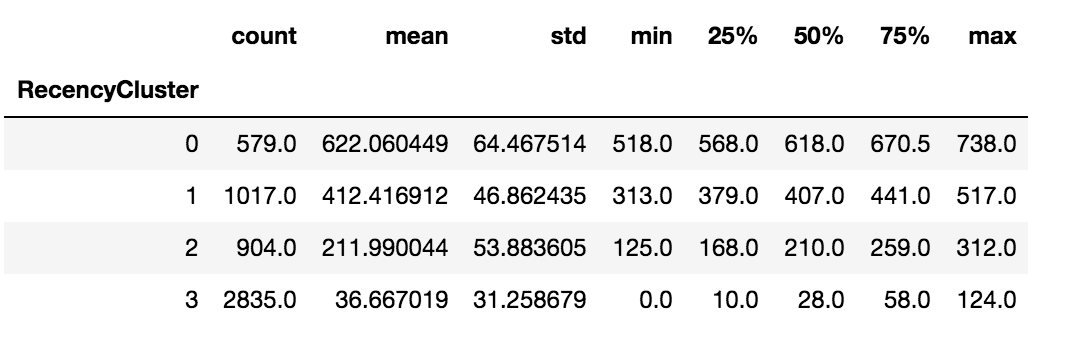


Table 1.3: Summary Statistics for Recency Clusters

**Frequency Clusters**

We created a similar histogram to study frequency. Here is our visualization of frequency:

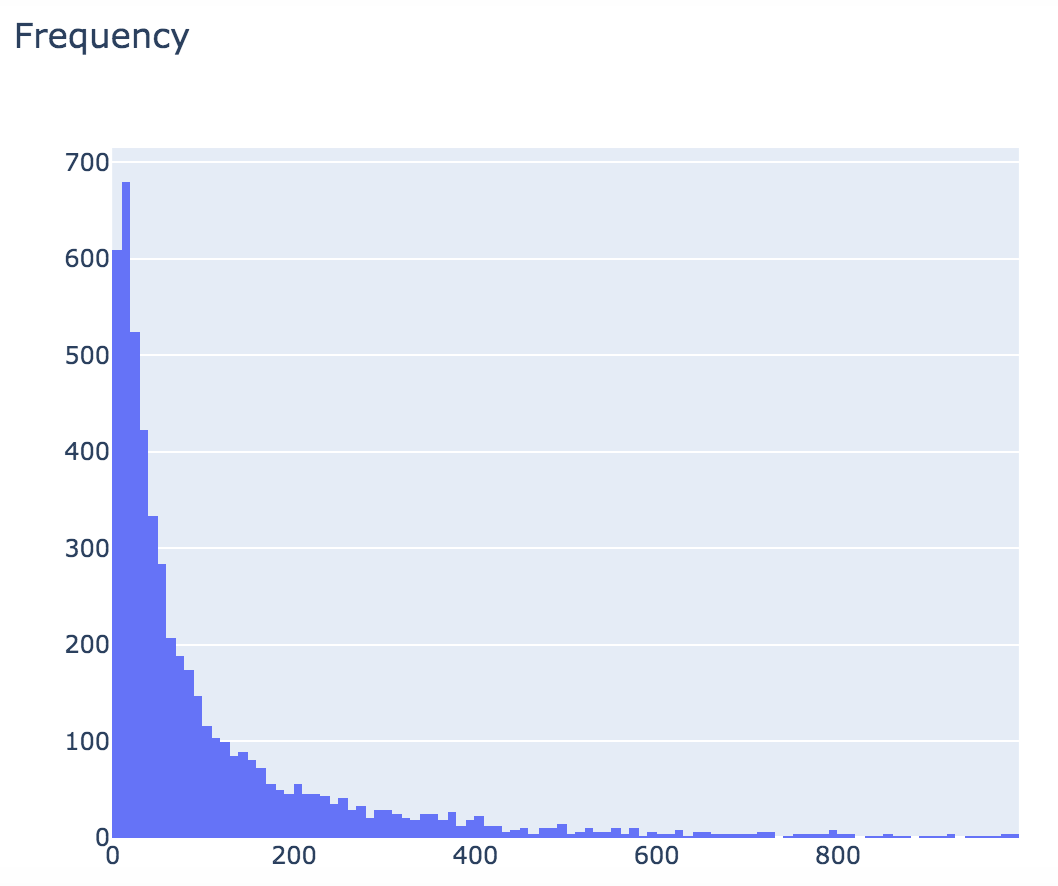


Figure 1.3: Histogram of Customer Frequency

Using the same methodology as we did for assigning a recency cluster to each customer, we also assigned a cluster for each customer based on how frequently they visit our site. Here are the summary statistics for each of those clusters:

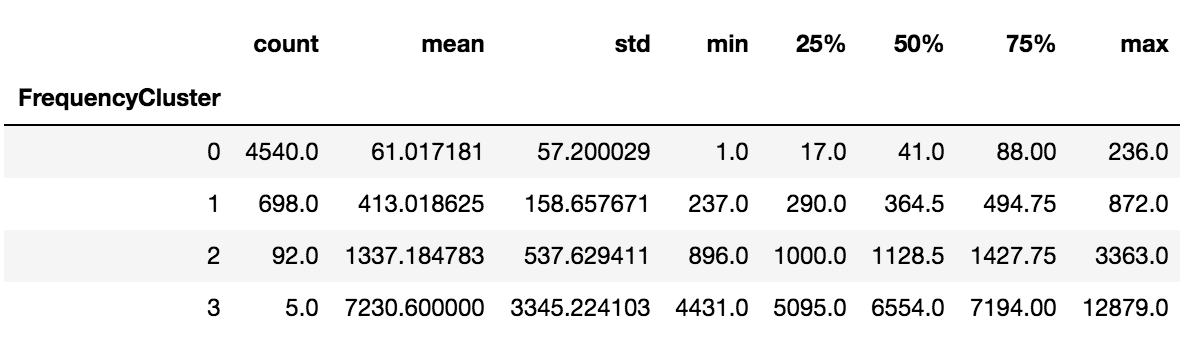


Table 1.3: Summary Statistics for Frequency Clusters

These clusters could also be used as the basis for marketing or engagement tactics within the company.

**Revenue Clusters**

Revenue clusters were one of the aspects we were most curious about since increasing revenue is key to counteracting our projected YOY revenue decreases. We followed the same process of creating a histogram:

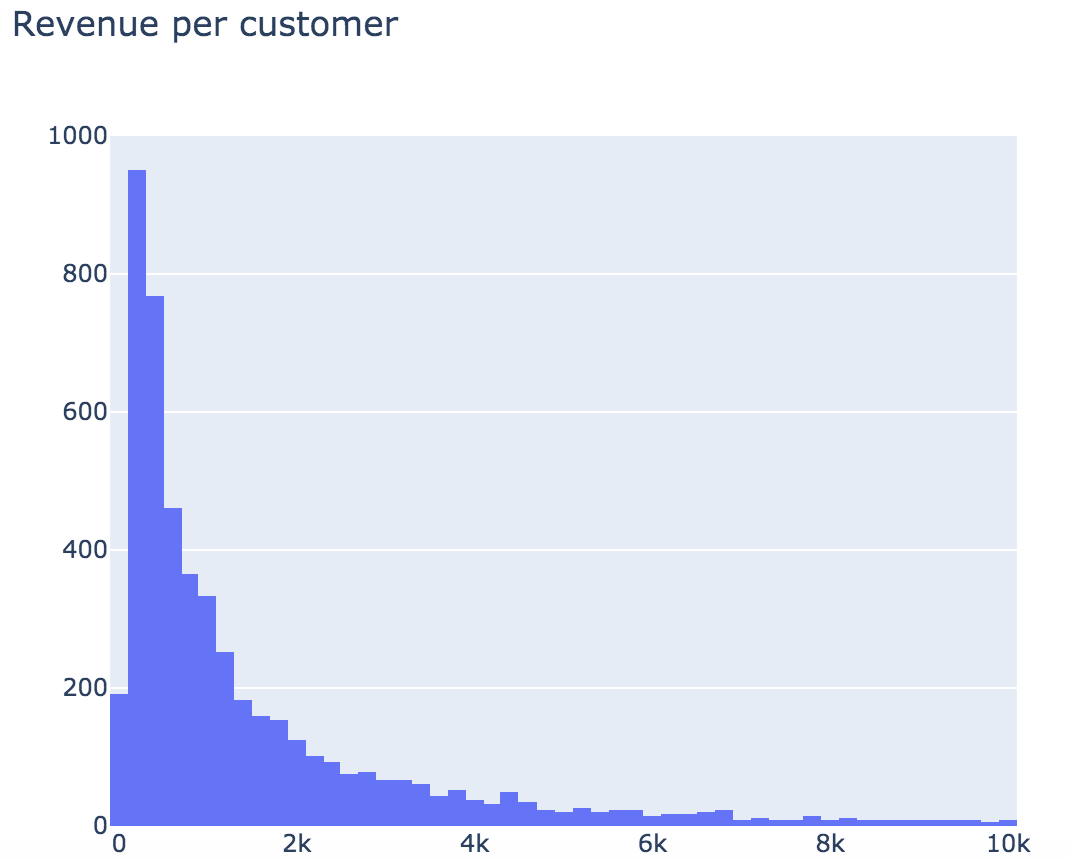


Figure 1.3: Histogram of Customer Revenue

The histogram shows that the largest bucket of customers have spent $100-$300 dollars with the business. We also think it’s important the note the smaller buckets of customers who have spent +$5k with the company. We decided to also cluster customers into four segments based on revenue:

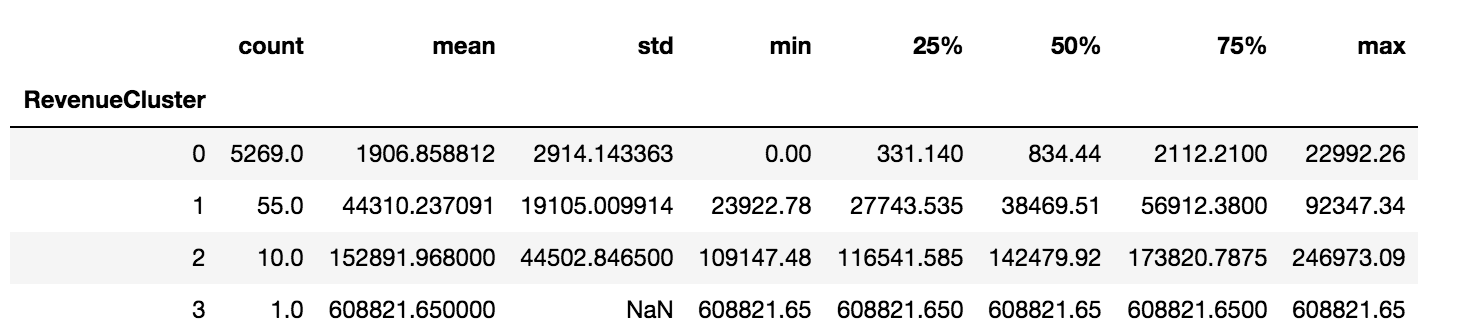


Table 1.4: Summary Statistics for Revenue Clusters

**Combining Recency, Frequency, and Revenue Clusters**

Now that we segmented all of our customers by our three extracted features (recency, frequency, and revenue), we needed to find a way to tie all of our findings together into one score per customer. We calculated a value score for each customer based on the segments of all three extracted features. First, the cluster numbers assigned to each customer were added together to calculate OverallScore. Here’s the list of OverallScores as well as the median for each extracted feature:

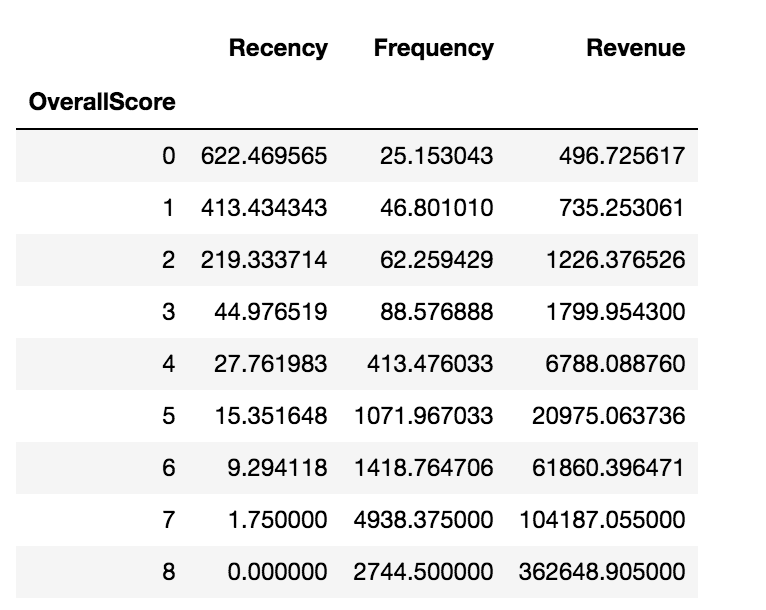


Table 1.5 Overall Scores and Extracted Features Medians

No customer scored above an 8 as an OverallScore.

Our goal was to make this data science process easily digestible for anyone in the company. So while 8 segments of customers could be beneficial if we wanted to dig deep into customer segmentation, it might be confusing for some employees. Instead, we summarized our findings into three groups: high-value, medium-value, and low-value. Customers with an OverallScore of 0-2 were categorized as low-value, 3-5 as medium-value, ad 5+ as high-value. Here’s an example:

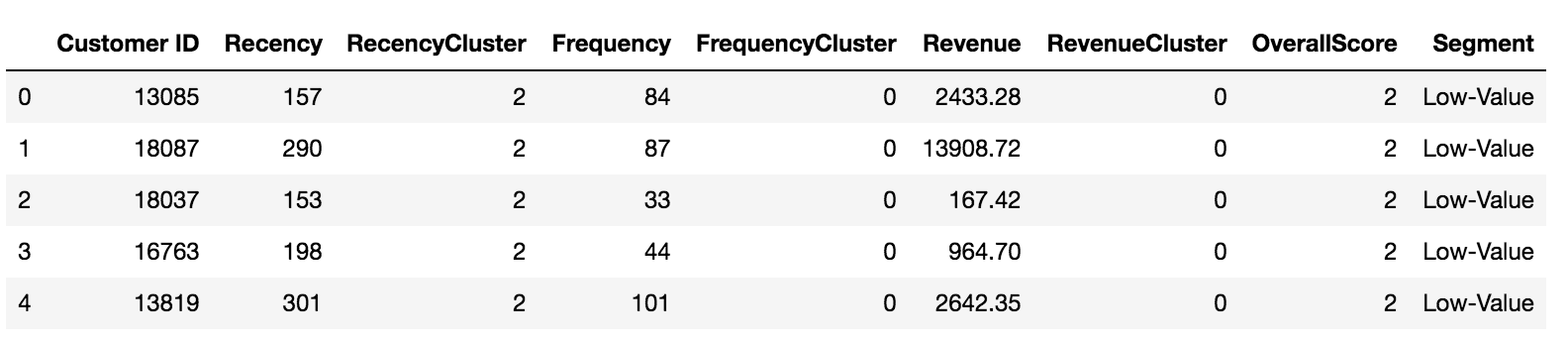


Table 1.6: Examples of Calculated OverallScore and Segment

We think visualizing these three easily digestible groups leads to a better business understanding of who the company’s customers are in terms of revenue and brand engagement. They also allow the business to easily identify similar customers based on purchasing habits. These clusters can then be used to form sales and marketing initiatives based on similar customers within a cluster.

We also used scatterplots to study the relationships between all three extracted features. Each datapoint is one customer, and each customer is color-coded by customer segment.

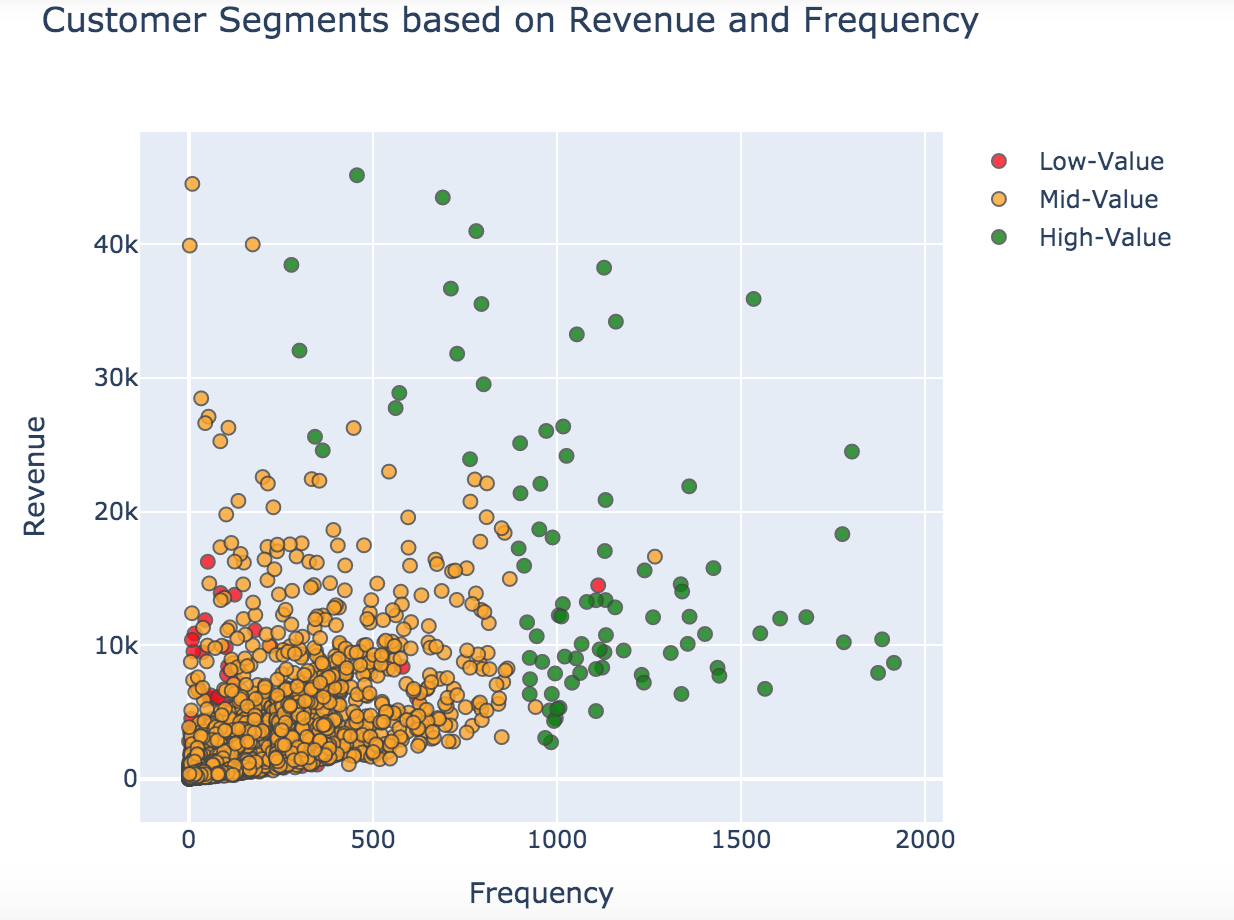


Figure 1.4 Revenue and Frequency Clusters

Revenue and frequency, as extracted features, showed the strongest correlation in our dataset. In general, the more frequently a customer returned, the more likely that customer was to spend more money. This is a key point, which, while it was probably an assumption within the company before this study, is now confirmed.

We also visualized revenue and recency:

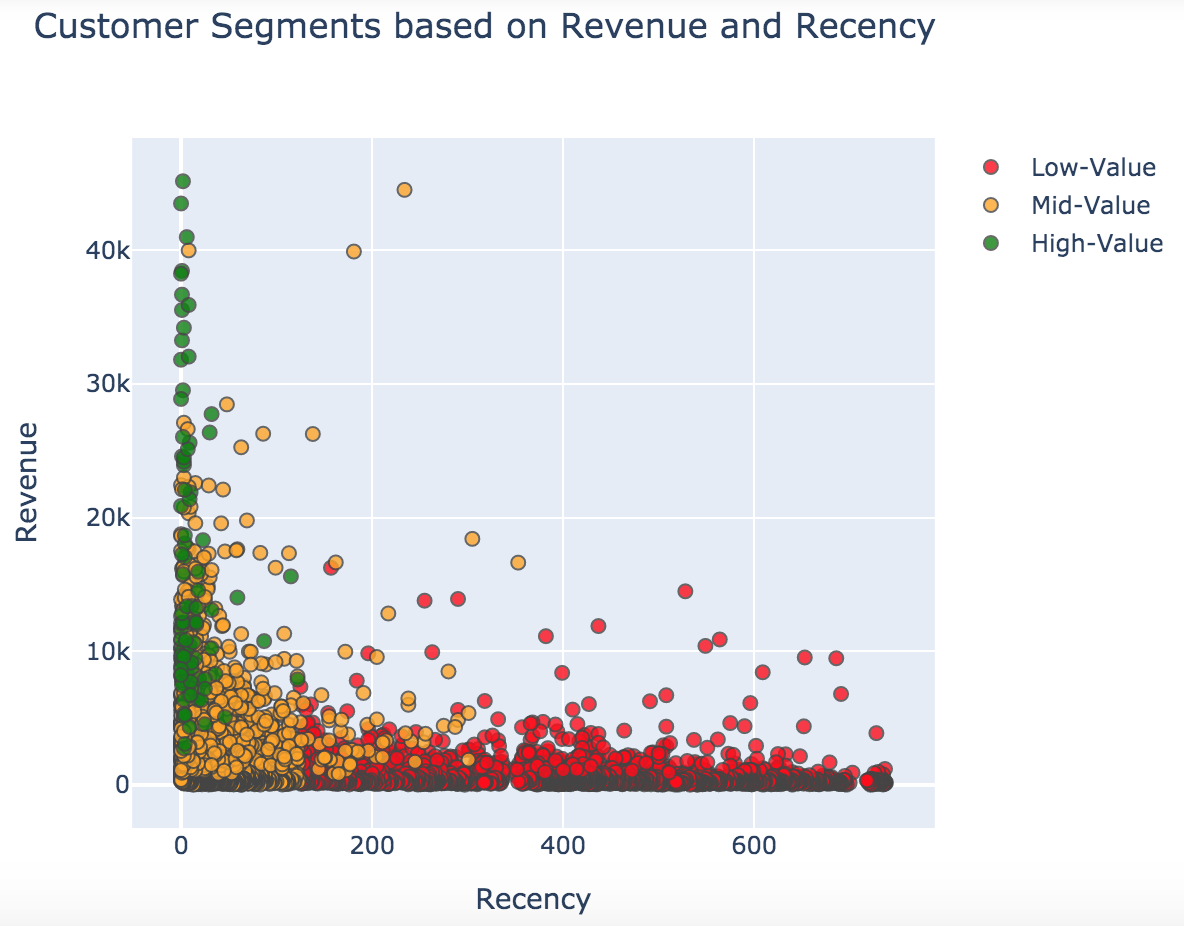


Figure 1.5 Revenue and Recency Clusters

And we visualized frequency and recency:

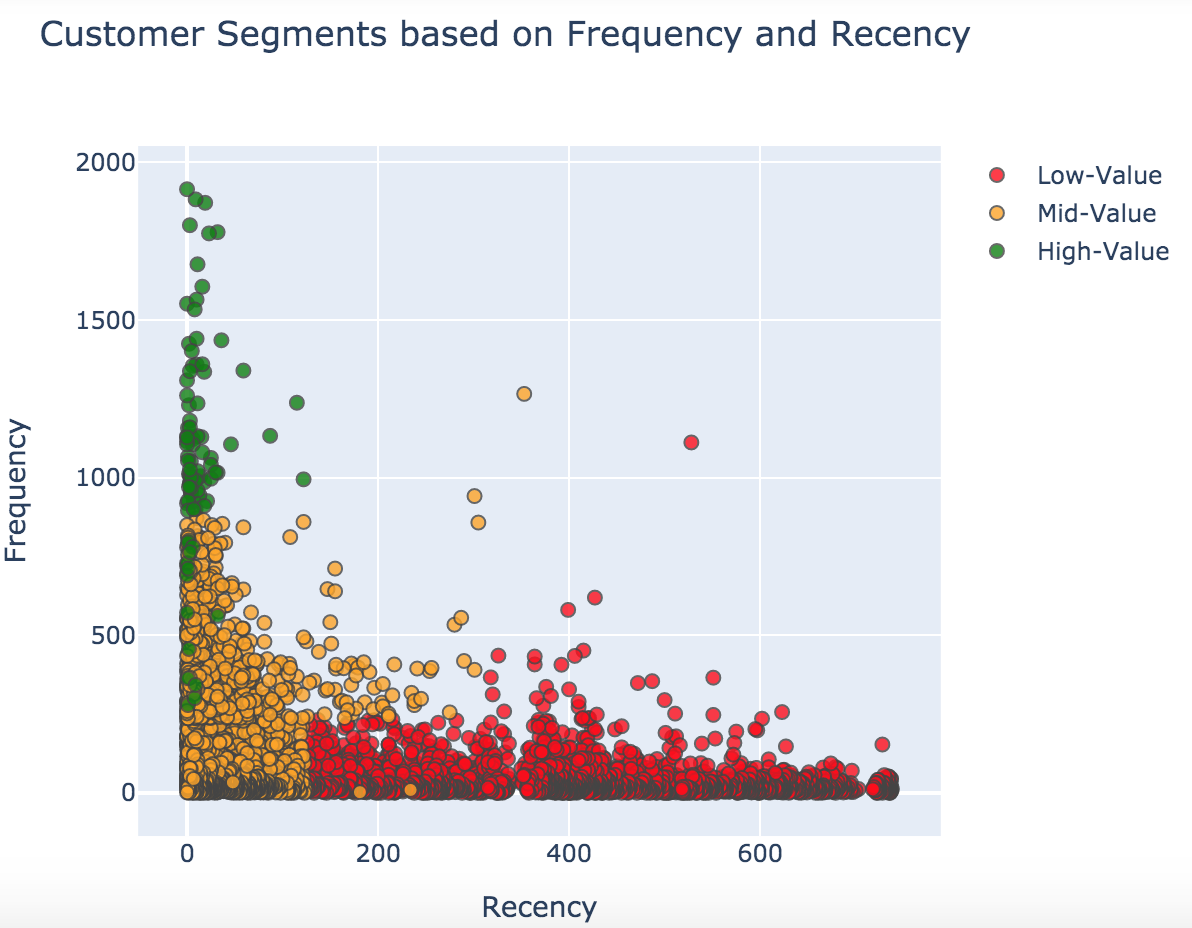


Figure 1.6 Frequency and Recency Clusters

For this project's purpose to increase sales revenues, the relationship of revenue based on frequency and recency is most important to us. However, we can see the company studying the relationship between recency and frequency for other projects such as customer engagement insights.

**Strategy Recommendations**

Our top finding in this study is that customer segmentation is key to reversing the current revenue decline that this company is seeing. Based on customer segmentation, we created five strategy recommendations for the company:

**Incentive Programs**

We suggest that the company start programs to incentivize customers to purchase more. These could include:  
- BOGO programs  
- Customer loyalty or points systems  
- Discounts over $x spent  
- Free shipping on orders over $x

These programs shouldn’t be a blanketed approach, though, and should be targeted to specific customer segments depending on initial segment testing results.

**Email Marketing**

Email marketing is our second strategy recommendation for this company. Since this is an online retailer, the company already has a database of all customer emails, and email marketing could serve as another customer touchpoint and customer engagement strategy to increase frequency and revenue. Email marketing also pairs well with the aforementioned incentive program initiatives.

**Social Media Branding**

Social media branding would also serve as another customer touchpoint, as well as open a two-way line of communication with customers. Branding in this way would not only remind people that the company exists, but it would also be likely to attract new online customers. So while engaging existing customers is important, the company can’t ignore the fact that they need to bring in new business whenever possible.

**Monitor Customer Segment Movements**

It’s important for the company to monitor when individual or even groups of customers shift from one value segment to another. For example, customers who show signs of shifting from one value segment to a higher one can be offered incentives to purchase more. For customers who downshift to a lower value segment, the company needs to play defense in order to retain the customer’s business. This is a great opportunity for the company to use predictive analytics in order to increase sales revenue.

**Increase SEO**

The fifth and final strategy recommendation that we offer is for the company to increase their SEO strategies as a competitive necessity. There’s huge competition in the online sales world (think Amazon and Walmart, for example), and the company needs to have an aggressive strategy to compete for customers in the world of Google. Increasing SEO best practices so that the company remains on top of Google Search indexes, and so that their product remain in Google Shopping searches, will be key in competing in such a fierce online marketplace.

**Conclusion**

Having identified a likely decrease in year over year sales led us to a profitable path in this data science project. We found that providing custom marketing strategies through customer segmentation is key to this company turning around their decreasing revenue trendline. We hope that the company has a newfound perspective on how to interpret and use their data to create advantageous sales and marketing strategies.

**Acknowledgments**

We would like to thank our professor, Catie Williams, for teaching this predictive analytics course. We’d also like to thank the authors of the materials below who have helped us learn more about machine learning and improve upon our project results.

**References**

1. Box Cox Transformation. (2018, May 20). Retrieved from https://www.statisticshowto.datasciencecentral.com/box-cox-transformation/.
2. Dr. Daqing Chen, Course Director: MSc Data Science. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK. <http://archive.ics.uci.edu/ml/datasets/Online+Retail+II>
3. Navalani, A. (2018, September 26). Introduction to Customer Segmentation in Python. Retrieved from https://www.datacamp.com/community/tutorials/introduction-customer-segmentation-python
4. Statsmodels Tutorial. (n.d.). Retrieved from <https://www.statsmodels.org/dev/examples/notebooks/generated/exponential_smoothing.html>.
5. Tomar, A. (2017, August 1). Customer Segmentation Using RFM Analysis in Python. Retrieved from https://medium.com/@tomar.ankur287/customer-segmentation-using-rfm-analysis-in-python-bb6f1bd2fce5