**Springboard Capstone 3 Final Report:**

**RFM+ Customer Segmentation**

Overview

In this project, I compared several different unsupervised clustering models to find the best customer segmentation for an online retailer based on customer transaction history. The three best models all used K-Means clustering: a 4-cluster model using only RFM features for each customer, a 3-cluster model using RFM plus 4 other calculated features (described below), and a 6-cluster model that used RFM plus one additional feature describing the average time between orders for each customer.

Each of these models offers a reasonable segmentation of customers based on their purchasing behavior, though the results could lead to different strategies in marketing for certain customers. All three models will be presented, highlighting some similarities and differences between the resulting segments, so that company leadership can make an informed decision on the best strategy to pursue.

Background

As a business sets a goal to expand its customer base, there are two clear and complementary questions to be considered: how do we attract new customers, and how do we keep existing customers coming back? There might be many different answers to these questions, each with its own pros and cons to be debated. Some might rely on conventional assumptions about the business, the industry in which it operates, and traditional ideas about how customers are expected to behave.

For example, a retail business may decide to hold a sale at a certain time of year, or take out an advertisement in the same place they and/or their competitors have always advertised. This strategy can definitely lead to success, yet the world—and consumer habits along with it—is constantly evolving. The best strategy is one that uses what we know about our customers—not just what we think we know, but what we can actually observe about them.

By analyzing data related to customer identity and behavior, a company can make more informed decisions about what the customer wants, and how the company can best provide it in order to increase revenue. But, even in specialized industries, not all customers behave equally. This is where an effective system of customer segmentation can help. By identifying different patterns present in a customer base, a company can begin to understand what types of customers they have and react to each segment’s needs. In other words, customer segmentation is a way to organize your customer base, identifying the most valuable areas for investment, and markets ripe for growth.

There are many different methods of customer segmentation, using one or more different types of characteristics. This could include demographic information (such as age and gender), behavioral information (such as products purchased and frequency of purchase), psychographic information (related to customers’ beliefs and motivations), and other identifying characteristics. For an online retailer, it can sometimes be difficult to learn much about your customers. Asking them to self-report identifying features may not always be effective, and other methods to learn about them could raise ethical questions around privacy. One type of information that can easily be examined, however, is a customer’s transaction history.

A reliable method of analyzing customer transactions, which has been in use for decades, is RFM analysis, which summarizes a customer’s behavior over a period of time. RFM is:

1. **Recency-** the length of time since the last purchase (where low recency indicates more recent purchases, associated with currently active customers),
2. **Frequency-** the total number of purchases over a time period (where high frequency is associated with loyal repeat customers), and
3. **Monetary Value-** the total amount of revenue generated by that customer (where high value is associated with customers who spend the most).

In a simple analysis, we can split all customers into subgroups based on the quartile they belong to for each metric. However, we can also go a step further by using machine learning clustering models to find appropriate boundaries for each segment based on the data.

Furthermore, this project attempts to improve on the standard RFM analysis by using additional features calculated from each customer’s transaction history. These features include:

1. **Average Order Time-** the average number of days between sequential orders from a single customer
2. **Maximum Price-** the highest individual unit price for any product purchased by the customer
3. **Average Number of Items-** the average number of different types of items purchased in one order
4. **Country-** the geographic location of the customer

Data Wrangling

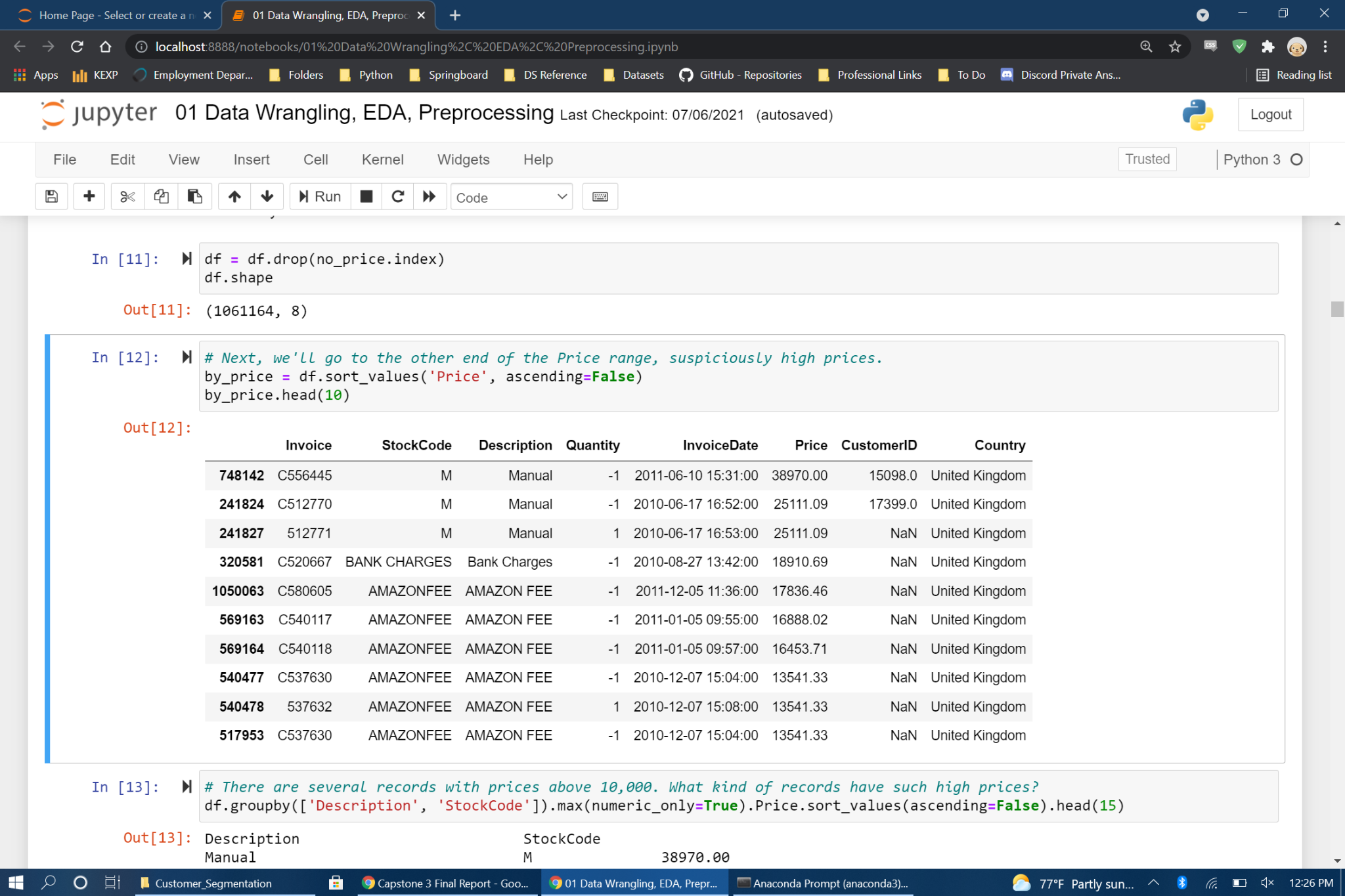
The Online Retail II Data Set from UCI’s Machine Learning Repository consists of data on real online transactions made for a UK retailer over a period of two years. However, the dataset contains many rows of irrelevant, non-customer related transactions, as well as apparently erroneous entries and transactions missing Customer IDs.

In order to ensure the best quality data for analysis, I decided it was necessary to thoroughly examine which rows were useful, and which were not. First, I deleted rows with a price ≤ 0, which appeared to be adjustments made by the company and therefore not actual customer purchases.

Other rows were removed based on the Stock Code column. These included bank charges, Amazon fees, postage, commissions, free samples, and debt adjustments. Even though some of these rows did have a corresponding Customer ID, I determined that they were not relevant to the desired segmentation. Labeled discounts, however, were saved with the data as relevant to a customer’s overall monetary value.

Several other rows were marked as ‘Manual’ transactions in the description. After careful consideration, I decided that the manual rows were too unreliable to keep in the dataset. For instance, look at the rows below, which represent the three highest prices in the entire dataset.

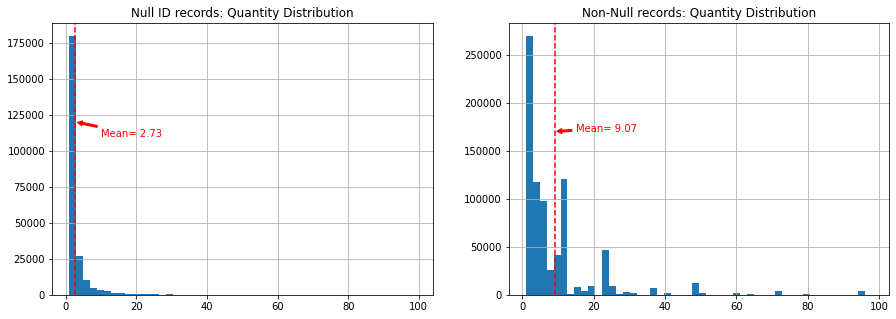
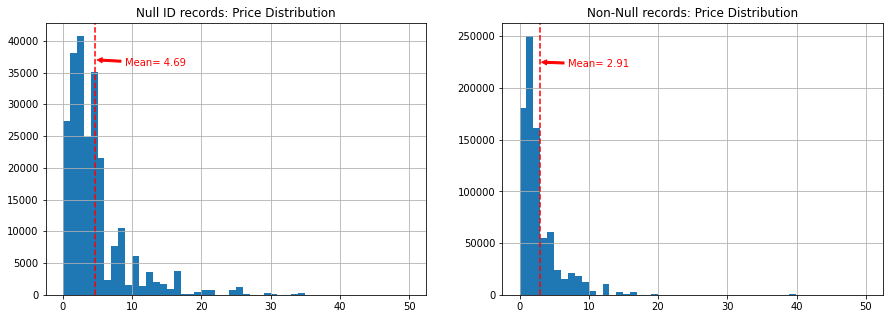
The second row, with invoice number C512770, is a cancellation in the amount of -£25,111.09 (calculated as quantity \* price), attributed to customer 17399. One minute later, invoice 512771 was entered manually as a purchase of the same price, yet this purchase does not have an associated Customer ID. It is difficult to believe that these two transactions are not for the same customer, and inspection of other rows containing the description ‘Manual’ revealed similar ID-mismatched purchase/cancellation pairs. Consequently, if I were to keep manual records in the data, the total value for some customers would be significantly inaccurate.

Interestingly, the few rows marked as manual adjustments by a specific person (for example, “Adjustment by john on 26/01/2010”) didn’t seem to have the same problem of mismatched cancellations, although a couple rows were indeed missing Customer IDs. As a result, I decided to keep these rows in the data.

Having gone through all different transaction types and decided which to keep and which to discard, it was time to take a look at the remaining records and deal with null values. Customer ID was the only column that contained any null values, with about 22% of the full dataset having no ID. Unfortunately, there is no way to fill in these missing values, which are essential to find patterns in the behavior of individual customers. For the purposes of this project, these records had to be discarded.

However, it is worth considering why these records are missing information, and whether they have any special characteristics. If this field is null due to some error (say, accidental deletion of the data or software malfunction), then perhaps we can safely ignore them. But if these purchases don't have IDs because they were made as a 'Guest' at checkout or some other customer action, the information here may tell us something about a distinct segment of customers (i.e., those who do not want to create an account).

Below are histograms of both price and quantity columns for null and non-null records. Extreme outliers were omitted to better compare the distributions. There does seem to be a noticeable difference in the two groups, with transactions missing a Customer ID having a higher mean price and lower mean quantity.



Further analysis should be considered on this group of null IDs. The differences between this subset and the rest of the data indicates that there may be a systematic reason why some or most of these records are null, which could tell us something about a different segment of customers. Imagine, for example, that these are customers who checkout as 'Guest'. Since this group tends to purchase fewer items at a higher price, you could perhaps offer a discount as an incentive to have the user create an account.

Feature Engineering & EDA

Before exploring the data for trends, I had to create the desired features from the cleaned transaction data. The first step was to aggregate the data by Invoice number, in order to get a summary of each individual order. Then, the orders were aggregated again into summary statistics for each unique Customer ID. Features were calculated as follows:

**Recency** is defined as the number of days since the most recent purchase. As a reference date, I used the day after the latest date present in the dataset, which was 2011-Dec-10. As a result, this column ranged from 1 to 739, the total number of days included in the data. It is important to remember that for this feature, a low value is a more desirable trait than a high value, as a low value represents a customer that has ordered more recently (fewer days ago), and thus is still actively engaged with the company.

Another important note is that cancellations were **not** included in calculating recency, to make sure it is a measure of the most recent purchase, not merely the most recent invoice. For example, if a customer placed an order 50 days ago, then returned some items 10 days later, their recency would still be 50 days.

**Frequency** was calculated as the sum of distinct orders for each Customer ID. Again, cancellations were not included to calculate this feature. Frequency ranged from 1 to 376, with a median of 3 orders.

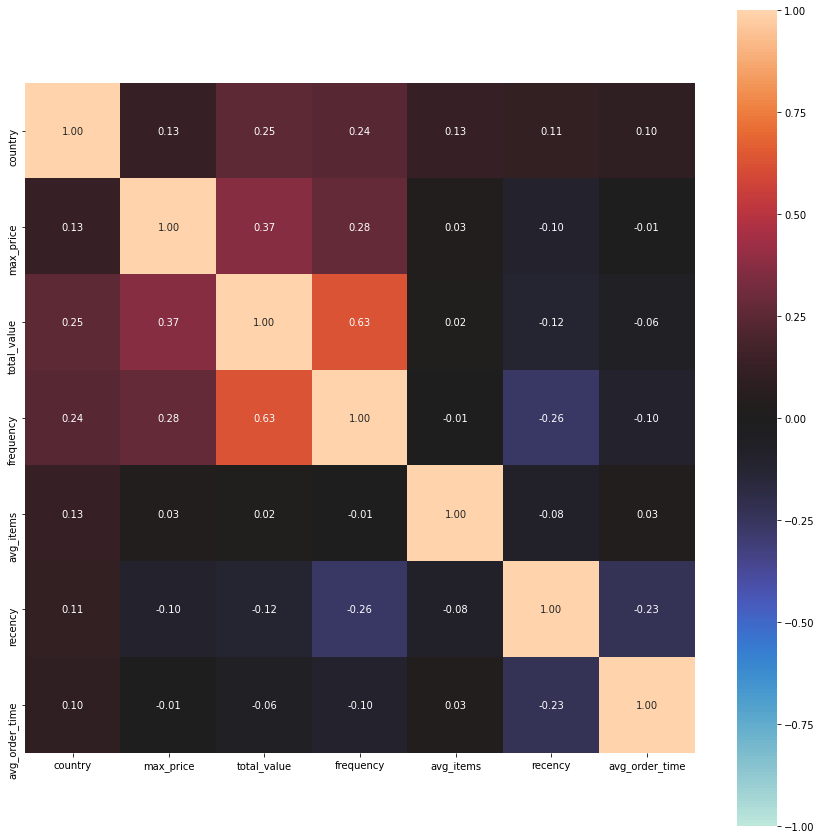
**Monetary Value**, or **Total Value**, was found first by calculating the value of each line item in the original dataset, as quantity \* price. Since all cancellations are recorded with a negative quantity, this resulted in negative values for those records, which is desired. The total value does include cancellations, and is the sum total of all orders for each customer. The median total value in our dataset was £865.17.

**Average Order Time** (AOT) was calculated using the aggregation by order (using invoice #). For each separate customer ID, I recorded the date of their first order as 0. If the customer had any subsequent orders, the order time was calculated as the number of days since the previous order. These order times were then averaged together for each customer.

Due to the way it is calculated, AOT has a unique interpretation. Any customer with only one order (or multiple orders on the same day, with no subsequent orders) has an AOT of 0, which includes 28.6% of all customers. An ideal customer would have an AOT greater than 0, which indicates they have, at some point, returned to the site for another order. However, a very high AOT is not ideal. For example, the max value for AOT was 714 days, which suggests a customer made two orders, nearly two years apart. This is clearly not a very engaged customer. So, it seems that there is probably some ‘sweet spot’ of AOT that describes the best customers (perhaps near the median value of 47 days).

**Maximum Price** is the largest product price (unit price, not value of purchase based on quantity) among all items purchased by each customer. This feature was recorded in the hope that it might tell us something about the customer that can’t be reflected in monetary value alone. In other words, a customer that purchases many low-cost items might be quite different from a customer that buys a small number of high-cost items, even if their total value is the same. Maximum price ranged from £0.17 to £1867.86, indicating a large difference in the prices some customers are willing to spend.

**Average Items** is a feature indicating the typical size of an order for each customer. However, it is based on the count of distinct products in each order, **not** the quantity of each product purchased. This may separate customers who only buy a few different types of products from customers who are interested in a wider range of offered products. In our dataset, this feature ranged from 1 to 300, with a median of 17.5 products per order.

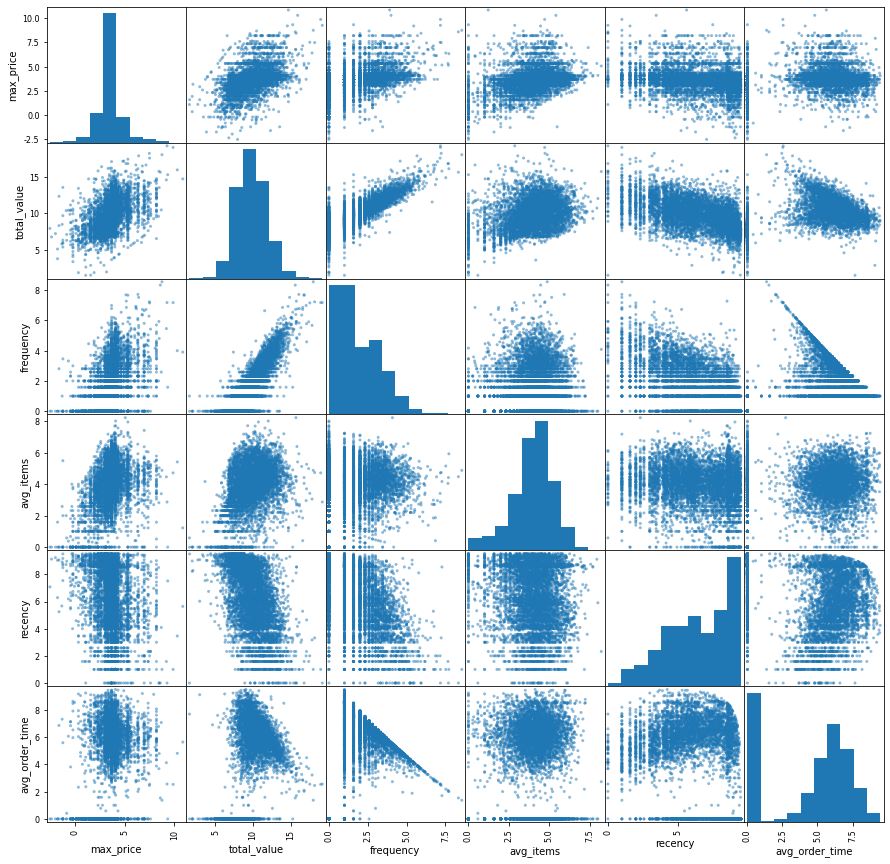
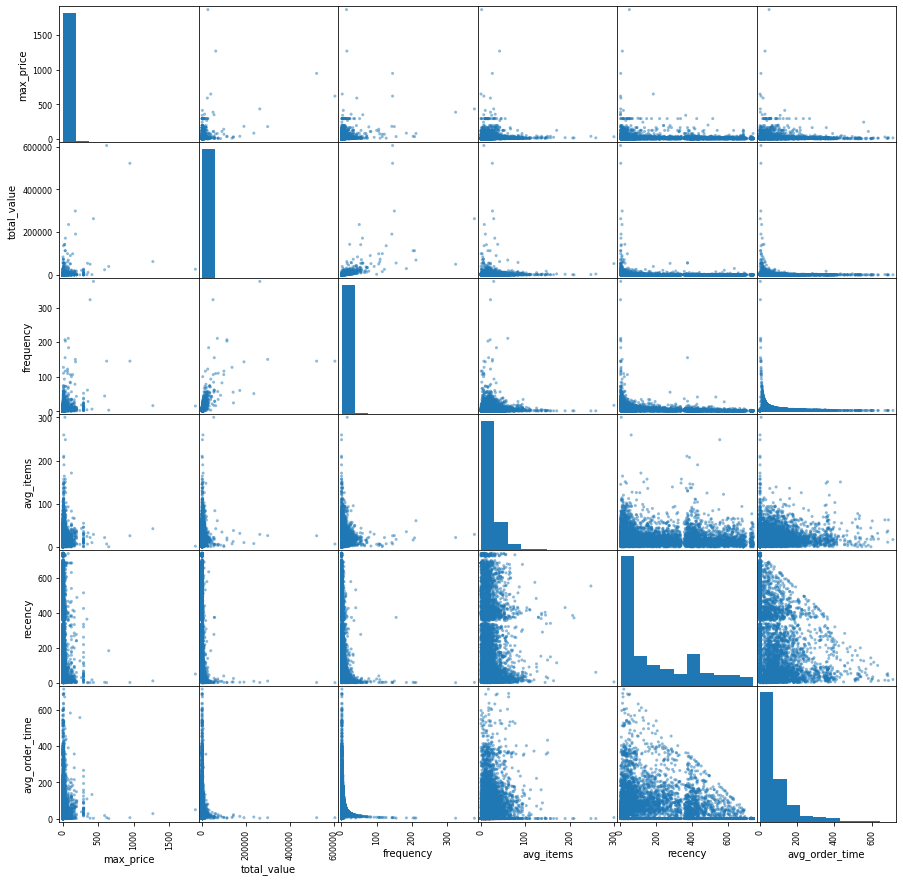
Finally, the **Country** column from the original dataset was saved as our only customer demographic information. Although the vast majority (91%) of customers in the data are in the UK, I hoped this information might be a useful segment divider. Unfortunately, due to many of the 41 countries in the dataset having a very small number of customers, I had to encode the country column into 3 ‘region’ categories: UK, Europe (8%), and Other (1%), then remove the UK column as a redundant category. Ultimately, the region columns did not seem to be helpful to the model.

Once all eight desired features had been calculated, I looked for correlations between them (there are 7 here because country had not yet been encoded while calculating correlations). There did not appear to be any very strong correlations present that would indicate collinearity. The strongest correlation found was between frequency and monetary value (r=0.63), an understandable positive relationship between the number of orders and the total value of all orders.

There was a smaller correlation between total value and max price, and negative correlations between recency and both frequency and AOT. None of these seemed to be enough justification to drop any features, however.

Next, I plotted scatterplots of the relationship between all pairs of variables, along with histograms of each variable’s distribution. The results showed that due to the few customers with extremely high values, all of the 8 variables were skewed to the right, and relationships were difficult to determine. By converting each of the features to a logarithmic scale, I was able to normalize the data to an acceptable degree, as you can see in the plots below.

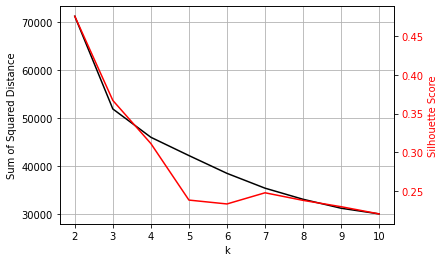
Unscaled Data Distributions Log Transformed Data



Note that in order to convert AOT to a log value, I had to add 1 day to all rows, thereby avoiding log(0). Another observation to note is the relationship between avg\_order\_time and frequency (which you can see in the third plot on the bottom row of the above scatter plots). Thinking about these features, it's clear that if log\_frequency is 0—meaning actual frequency 1—avg\_order\_time must also be 0. For frequency > 1, the avg\_order\_time has an upper bound: the total number of days of the dataset (739) divided by one less than the frequency. It's impossible, for example, to have a frequency of 10 with average order time greater than 74.

Modeling

By using unsupervised machine learning techniques, I tried to find natural patterns in the data to determine the boundaries of each customer segment. Without the benefit of labeled data (or even a specific number of segments we wish to find), I cannot say if the results of one particular model is more or less “correct” than another. For that reason, I tried and compared several different models to find one that seemed to best capture the differences between customers.



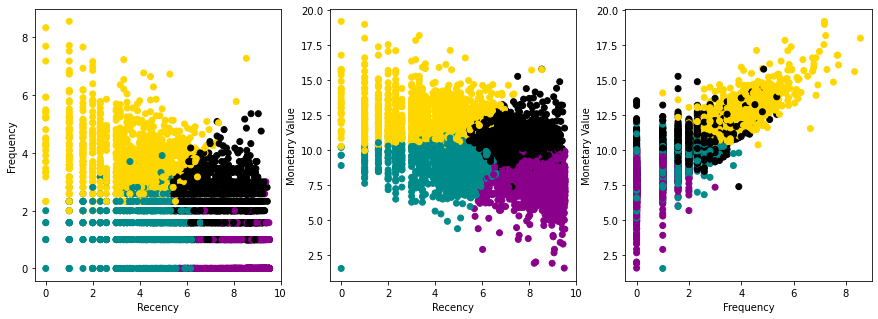
Starting with K-Means clustering, I calculated the sum of squared distance and silhouette score for models over a range of k values. I looked for an ‘elbow’ in the decline of SSD, which would indicate a diminishing reward for increasing the number of clusters, as well as a silhouette score of at least 0.30, showing that the clusters are reasonably well-defined enough to be reflective of an actual underlying pattern in the data. Using both of these techniques, the best value for k was 3, with SSD = 51900 and silhouette score = 0.367.

In order to better visualize the clusters found by the model, I used a PCA transformation on the data to reduce it from 8 dimensions to 2, then produced a scatterplot of the first vs. second principal components, with points colored by cluster label.

I will refer to this model as the RFM+ Model, since it uses all 8 features discussed above rather than just the conventional three RFM features. The most noticeable characteristic of this plot is the long, thin cluster to the right side of the x axis, which the K-Means model clustered together as the black segment. There does indeed seem to be an underlying pattern in the data that our model has captured. Remember, however, that the 2D PCA projection is just a means of easily visualizing the data, and the PC1 and PC2 values were not used to train the model. Only the 8 features above were used.

Aside from K-Means, I also trained an Agglomerative Clustering model to see if the performance improved. The resulting clusters were similar to the K-Means model, but the silhouette scores for the Agglomerative models were lower for each value of k when compared to K-Means. Spectral Clustering and DBSCAN models did not produce meaningful clusters (see associated jupyter notebooks for additional model results and visualizations).

The next step was to try a different model using only the three conventional RFM features. I wanted to make sure that the results of the RFM+ model actually outperformed the model with fewer features. Again, the K-Means algorithm produced better results compared with an Agglomerative Clustering approach. Using the same methods as with the RFM+ model, I found the best value of k of 4 for the RFM only model, achieving a similar silhouette score of 0.364, and graphed the results. (Here, it is not necessary to use a PCA for visualization since there are only 3 dimensions. We can see the clusters in all dimensions using 3 scatterplots.)



In these graphs, there is no obvious (to the eye) clustering of the data, though there are some distinct lines of points in the lower ends of the axes that are an artifact of the logarithmic scales. Still, this model achieved a good silhouette score.

With two viable segmentation models, achieving similar scores, I had to decide how to evaluate and compare them. The RFM model finds 4 segments instead of 3, which could help the company create more specific targeted marketing strategies for their customers. However, I also wondered if there was still some benefit to the other features available. Remember the long, thin cluster in the PCA of the full dataset - could that segment add anything to our understanding of our customers? To better compare the two models, we can see whether the RFM only model captured the thin cluster as well, by using the RFM cluster labels on the same PCA of the full dataset.

We see in this graph that the RFM model does not seem to make a distinction between the thin cluster to the right and the rest of the data. In other words, there was some pattern present in the other features that we lost when we used RFM columns only.

On inspection, I determined that the division of data we can see in the PCA is mainly due to Average Order Time. The thin cluster we see is almost exclusively made up of customers with an AOT of 0. Remember that these are customers who have made only one order, or perhaps multiple orders on the same day with no subsequent purchases.

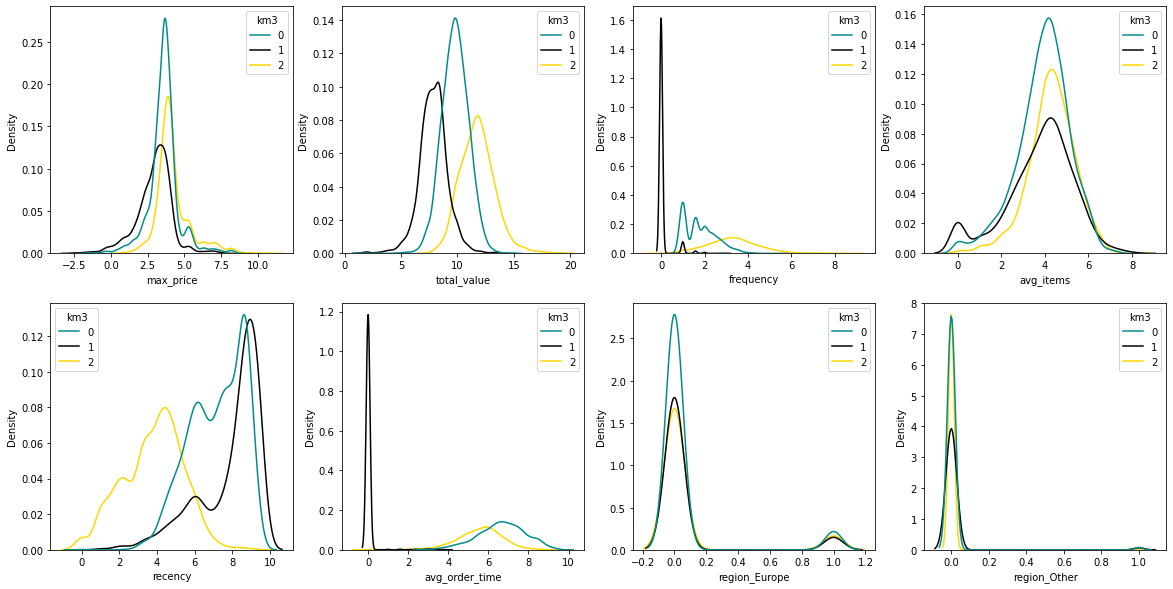
It seemed that the best model would be one that can combine the specificity of the RFM model having at least 4 well-defined clusters, while also capturing the patterns from the AOT feature. The result is a K-Means clustering of just 4 features: Recency, Frequency, Monetary Value, and Average Order Time, which I will refer to as the RFMO model. This model achieved higher silhouette scores over a range of k values, which made it possible to choose a model with 6 clusters that scored close to the other models, at 0.356. 

Here, we see the resulting labels on a new PCA transformation of just the four RFMO features. This model distinguishes between the thin cluster and the rest of the data, but also identifies different segments within each side of the division.

With these three models—the 3-dimensional RFM model, 4-dimensional RFMO model, and 8-dimensional RFM+ model—we can begin to interpret what they tell us about the different segments of customers in order to ultimately choose an appropriate marketing strategy for each group.

Interpretation

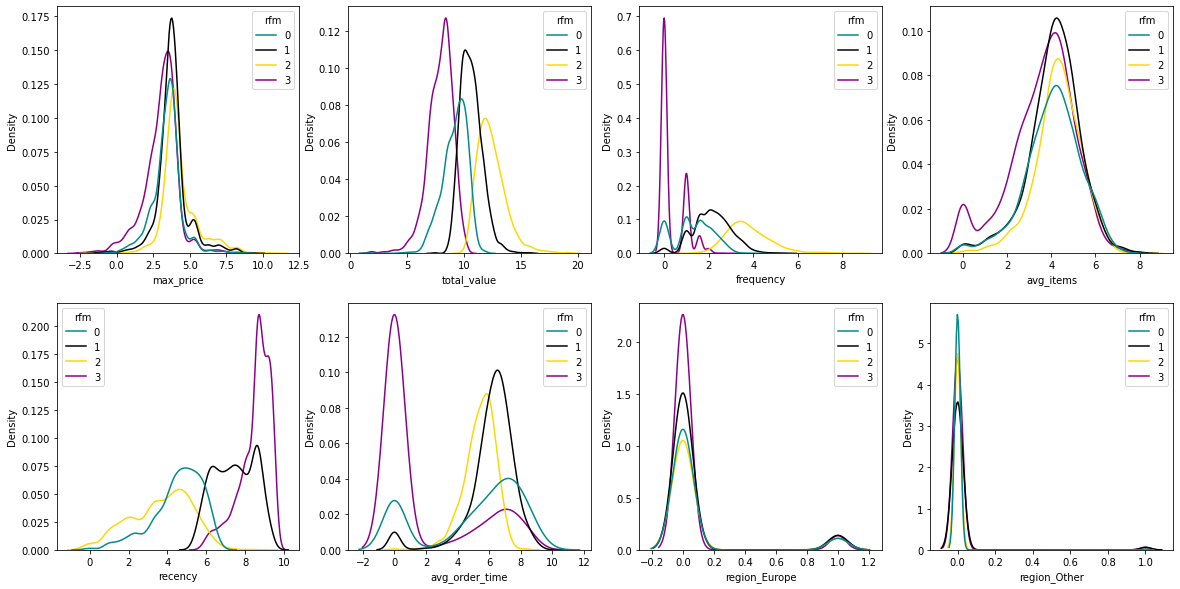
To find contrasting characteristics between segments, I plotted KDE distributions of each of the eight features, separated by segment labels for each model. First, we can look at the results of the RFM+ model, which had three different segments.



The big takeaway here is the plot of AOT (row 2 column 2), where we see the black segment primarily consists of customers with AOT of 0, as previously mentioned. These customers have high recency (meaning they are older orders), and low monetary value. I have labeled this group the “One & Done” segment, since they mostly stopped after one order.

The teal group, in contrast, have a higher monetary value and a bit higher frequency, but a high (meaning slow) AOT and not much better recency. These could be customers that used to order regularly but have since lapsed, so I have labeled them “Old Friends”. The yellow group are our most valuable segment, and could be considered “Loyal Customers”, having high frequency, low recency, and high value. The last takeaway from this segmentation is that even though all of these features were used to determine the segments, there does not appear to be a meaningful difference in max price, average items, or region.

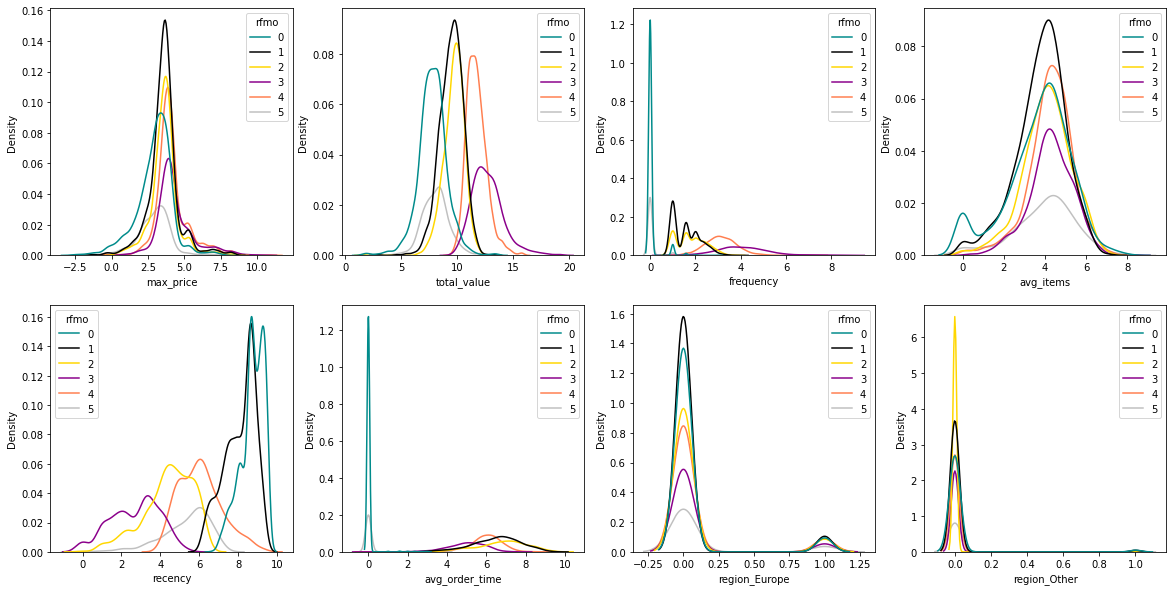
Next, we will look at the results of the RFM only model.



The purple and teal segments both have low frequency, but are divided by recency, with the teal group being more recently active than the purple group. We can call these groups the “Moved On” segment and the “New/At Risk” segment. Neither is very valuable, because the Moved Ons are no longer interested in ordering while the New/At Risk segment are too new to determine if they will continue to order or not.

Again, the yellow segment is comprised of the most valuable and frequent customers, earning the label “Loyal Customers”, while the black segment are our “Old Friends” who seem to be previously loyal customers that have lapsed.

Finally, let’s examine the six segments identified in the RFMO model.



With the benefit of the Average Order Time feature, we once again see a large distinction between customers with AOT of 0. The teal and gray groups represent these customers, but compared with the One & Done group from the RFM+ model, there is a distinction made here between the older customers and the more recent ones. Perhaps there is still more of a chance to recapture the gray group and convert them to regular customers, instead of letting them join the teal group. So, I’ve labeled the teal group “One & Done”, while the gray group is “One & At Risk”.

As the most recent, frequent, and highest value customers, the purple group earns the label “Loyal Customers”. The black group have high recency and low frequency, so they are the “Moved On” segment. The pink group has high value, high frequency, but high recency and thus are the “Old Friends”, while the yellow group is still new and could develop into either loyal customers or moved-ons, which means they are “At Risk”.

Business Insights & Next Steps

Armed with the segments found by these models, company leadership can work to develop targeted strategies focused on increasing revenue from each segment. The model used can depend on what level of detail the company decides to act upon. The three segments found by the RFM+ model provide a simpler basis for strategizing, while the RFMO model is a more complex division of customer types.

While the company’s leaders, including marketing experts and customer service managers, should spend time brainstorming and developing specific strategies, some ideas could include the following suggestions based on the six RFMO segments:

* Incentives to encourage customers to return after a lapse in orders, such as discounts/gifts for new orders or highlighting popular new products that customers may not have seen before. These strategies would be most appropriate for segments with high recency, including the One & Dones, Moved Ons, and Old Friends.
* Survey customers to see whether they plan to order again, and why they may have stopped placing orders. This can help to identify customers who lapsed due to some problem leading to dissatisfaction. Perhaps customers who lapse will come back if some problem is fixed. It could also help with newer customers, to see if their first purchase worked out well and determine if they can also be converted into loyal customers. This strategy would be especially helpful for the One & At Risk and At Risk segments.
* The loyal customers tend to have slightly lower Average Order Time compared with the others (excluding the 0 AOT groups), while the At Risk group had the highest AOT. A resulting strategy might be to target the At Risk group more frequently, hopefully converting them to more regular buyers.
* Membership benefits for the high-frequency segments could help our Loyal Customers feel appreciated, and hopefully entice the Old Friends back. An example could be exclusive deals or gifts for frequent buyers.
* For our most Loyal Customers, offering referral bonuses could help us find new customers on the word-of-mouth recommendations of high-value customers.
* Comparing popular products across groups can help identify which items might excite lapsed or low-frequency customers to return.

Once a collection of strategies has been developed and agreed upon, the next step is to decide how to implement them. One way to observe whether the new strategies are effective is to do A/B testing within each segment. For example, with the One & Done segment, we could choose half of the customers at random to receive an email featuring discounts on new products. If a significant number of the test group do go on to make another purchase, we would consider the test to be successful and follow up with emails to the rest of the customers in that segment.

Meanwhile, the company can ensure that they are catering to the needs of the most valuable customers, which are the Loyal Customer segment. We do not want to lose these customers, and should therefore make sure they are getting what they need. That might include making sure to have stock of the most popular products in that segment, and that any issues that arise with orders from these customers are carefully handled.

Hopefully, these actions will help the company to grow revenue from all segments, retain more customers, and reach new customers as well. Whether or not the new strategies are successful, however, it will be necessary to review the segment model after some time. For various reasons, including implementation of new strategies as well as overall trends in market behavior, the characteristics of our customer segments will likely change over time. In the future, we can also use our new segment labels to train a supervised machine learning model to predict the behavior of new customers.

Other engineered features might be explored to see whether they can improve the definition of customer segments. Based on just the data present in this project, some possible features not explored include: the number of unique products purchased, the product each customer purchased most, whether customers purchased large or small quantities of specific products, or the time of day when orders were placed.

Additional data collection can also help to learn more about our customers and the best ways to invest in marketing strategies seeking new customers. In this dataset, most customers are businesses that buy stock for their own retail stores. If we can find out more about our customers, such as the size and location of the business, how many customers they have on average, who is in charge of ordering product, etc., we can develop even more in-depth targeted strategies to find new customers that match the profiles of our Loyal Customer segment.

The customer segments identified in this project are just the beginning stages of understanding our customer base and reacting to their needs and behaviors. By harnessing the descriptive power of the data, company leadership is more informed to make the best decisions to enhance business performance.

Referenced Works

Chen, Daqing, et al. Online Retail II Data Set, available from UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml/datasets/Online+Retail+II>

Reddy, Rahul. “Who’s who: Understanding your business with customer segmentation.” <https://www.intercom.com/blog/customer-segmentation/>