Project Report

Team Members

- Ciara Malamug: malamug.c@northeastern.edu (Laney's T/F 8am Lecture)
- Jasmine Wong: wong.jasm@northeastern.edu (Laney's T/F 8am Lecture)
- Yu (Diana) Xiao: xiao.yu4@northeastern.edu (Laney's T/F 8am Lecture)
- Yuting (Jane) Zheng: zheng.yuti@northeastern.edu (Laney's T/F 8am Lecture)

Problem Statement and Background

With hundreds of thousands of customer transactions each day, it is immensely difficult to discern which customers a company should prioritize. Creating and managing marketing campaigns requires both time and money, so properly segmenting customer groups is essential to creating better brand experiences for customers and maximizing profits. Effective customer segmentation is extremely valuable to the firm itself in order to streamline costs and increase profits, but it is also valuable for improving customers' experiences. Therefore, in our project, we used analytical and machine learning techniques, in Python, to explore H&M's customer transaction history, and segmented customers into target segments to aid marketing strategy development. However, this brings into question: how should a company segment customers?

We were curious about two segmentation methods—RFM analysis (a pre-established marketing technique) and k-means clustering—and wanted to investigate which algorithm more effectively creates customer segments for H&M. RFM stands for recency, frequency, and monetary value and is a method used to quantify a customer's value based on the transaction history data. It uses the metrics, recency (time since last purchase), frequency (total number of visits), and monetary value (total spending), to compare customers. Because of the recency metric, RFM is best used to drive strategy at a certain point in time. Therefore, our recommendations are based on what the company should do at the “current date” in the data, which is December 31, 2019.

Introduction to your Data

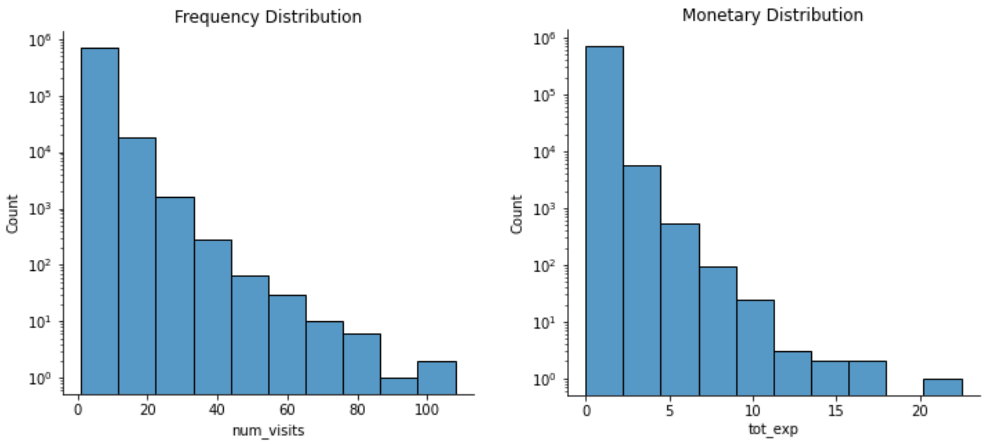
We found our dataset on Kaggle. It is a huge dataset containing sales and customer data within a two-year time span provided by H&M. The dataset contains CSV files (Articles, Customer, Transactions) and product images. Our group decided to focus on the transaction data. The transaction data includes transaction date, customer id, article id, price, and sales channel. Since the original transaction data file is 3.49GB with over 30 million rows, we ran the data through a data cleaning python code to extract a six-month timeframe to a CSV (hm_trans_clean.csv). The six-month period we selected was from July 1, 2019, to December 31, 2019. Each row represented an item purchased by a given customer.

The cleaned data (hm_trans_clean.csv) was 7,751,322 rows by 7 columns. Each row was an individual item purchased. The columns of interest were t_dat, price, and customer id. The column t_dat had the time of each transaction, and we used this to extract recency (current date - last date customer purchased) and frequency (total unique dates a customer purchased). We used the price column to extract the monetary value of each customer. The price column was on a consistent, though very small scale. By comparing the article_id of certain products to the prices on the H&M website, we suspected that the actual monetary value could be calculated by using price * 590. However, since we were unable to verify this, we left price as the original scale in the data.

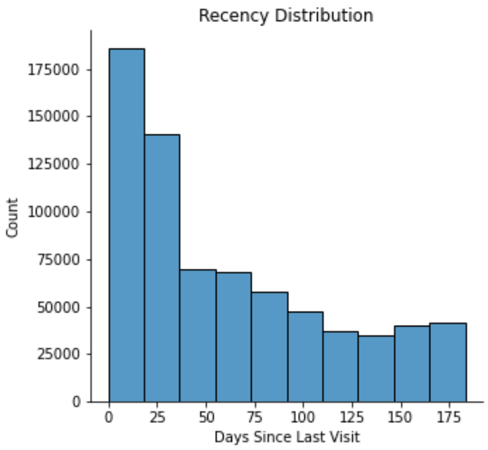
Data Science Approaches

Although RFM is a pre-established customer segmentation method, we wanted to see if using unsupervised learning through k-means clustering would be more efficient in segmentation. We decided to implement and visualize RFM and clustering in order to compare the two models.

In order to conduct an RFM analysis, we needed to extract some data from what was given. As mentioned in the data section, each row was an item purchased by a customer. However, we wanted each row to be a customer. Using pandas, we consolidated the data for each unique customer and stored RFM raw data for each. For recency, we took the difference between the last transaction date and the ending date of the data (December 31, 2019), denoted as the last_trans_td column. For frequency, we counted the number of unique days each customer bought at least one item, denoted as the num_visits column. For monetary, we took the sum of the price of all transactions for a given customer, denoted as the tot_exp column. This resulted in 723,630 unique customers.



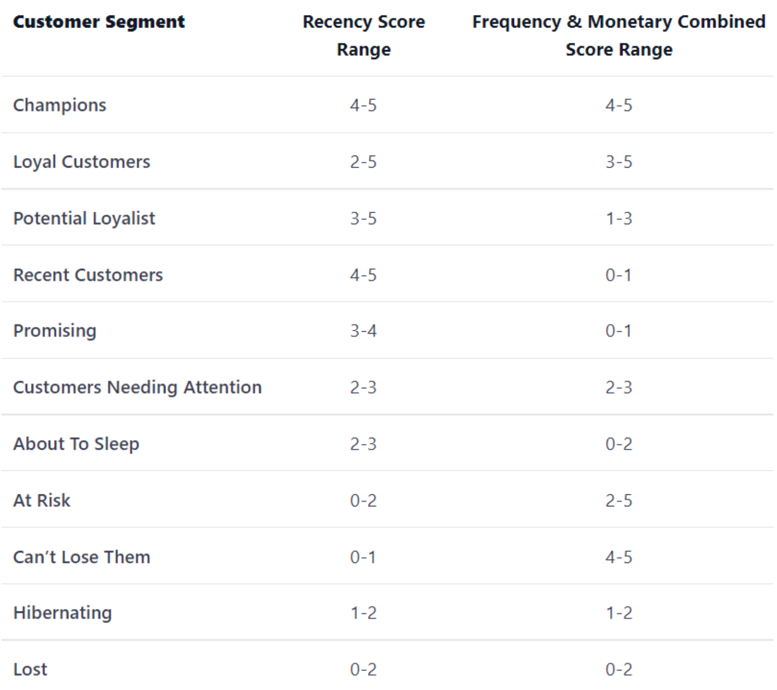
Frequency (left) and monetary distributions of all customers. Count is logarithmically scaled.



Recency distribution of all customers. Scale is normal.

Due to the large amount of data and the fact that there is considerable skew in our data, we believed that a completely random sampling of customers would not be representative of our data. Therefore, we decided to take a stratified random sample from all unique customers. This was implemented by dividing each RFM column (last_trans_td, num_visits, and tot_exp, respectively) into 10 percentiles and taking a random sample of 20 from each percentile. This resulted in a total of 600 customers in our stratified random sample. For the rest of the analysis, we based it on this random sample.

In order to achieve customer segmentation, we sorted each of the RFM raw data columns into ranks from 1-5, 1 being the worst and 5 being the best. From here we compiled the three numbers into one RFM score (e.g., 125 = worst tier recency, second-lowest tier frequency, and best tier monetary). We then divided the customers into 11 segments, based on marketing theory.



Customer Segment	Recency Score Range	Frequency & Monetary Combined Score Range
Champions	4-5	4-5
Loyal Customers	2-5	3-5
Potential Loyalist	3-5	1-3
Recent Customers	4-5	0-1
Promising	3-4	0-1
Customers Needing Attention	2-3	2-3
About To Sleep	2-3	0-2
At Risk	0-2	2-5
Can't Lose Them	0-1	4-5
Hibernating	1-2	1-2
Lost	0-2	0-2

Marketing framework for customer segments created using RFM analysis.

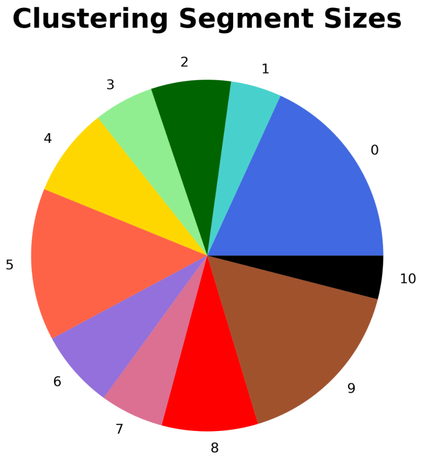
We then segmented the customers into the same number (11) of groups using sklearn's k-means clustering.

Results and Conclusions

We ended up with two models to segment the data: RFM segmentation and clustering (k-means) based on these RFM values. We learned that the pre-established theory (RFM) was much better at segmenting the customers than clustering.

It was really interesting how clustering seemed to depend on the recency factor (last_trans_td) the most when segmenting (resembling a layer cake), as displayed in figure 3. Different colors represent different clusters as explained in the pie charts (figure 1 corresponds to figure 3 and figure 2 corresponds to figure 4). The inertia of the clustering came out to be about

15272.10, signifying a high variance in cluster groups, which makes sense as the groups were widespread across the frequency (num_visits) and monetary value (tot_exp) features.



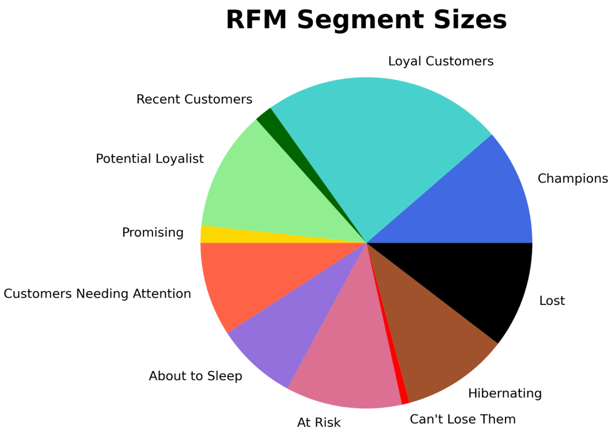
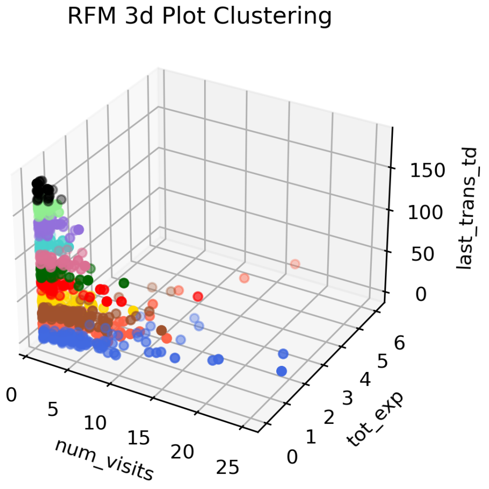


Figure 1. Cluster Size

Figure 2. RFM Segment Size



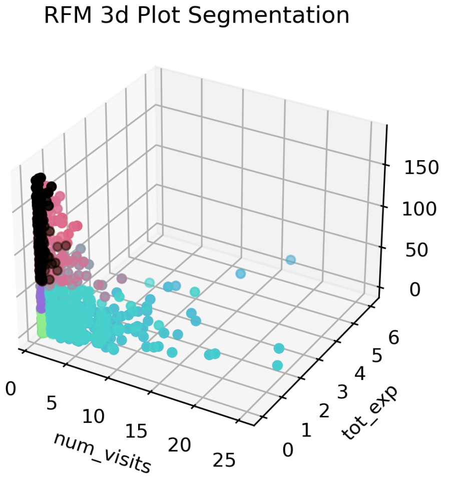
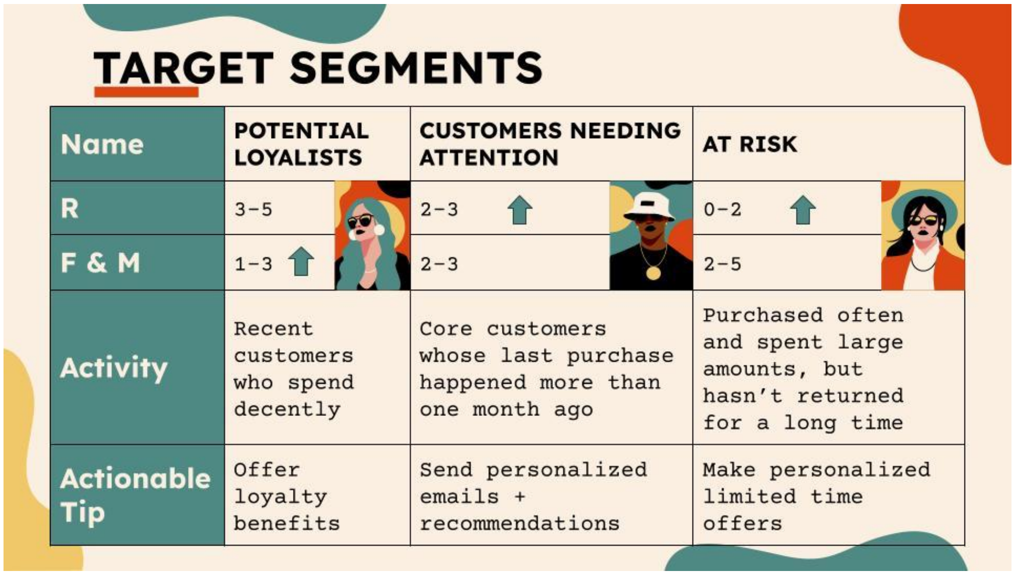


Figure 3. RFM Clustering Results

Figure 4. RFM Segmentation Results

Based on the better segmentation created by RFM, our group recommends that H&M focus on targeting three segments of substantial sizes: “potential loyalists,” “customers needing attention,” and “at-risk” customers. In figure 4, the pink group represents the “at-risk” customers. They are customers who used to frequent the store and purchased decently but have not returned in a substantial amount of time. Therefore, if they keep moving up, they will become "lost customers," a group represented by the black dots. We recommend sending limited-time coupons to these customers to encourage them to come back into the store. On the graph, the goal is for the pink dots to move down and towards the right to become potential loyalists and loyalists (represented by blue dots). Our recommendations for the other two target segments, "potential loyalists" and "customers needing attention," can be found in the table below:



TARGET SEGMENTS

Name	POTENTIAL LOYALISTS	CUSTOMERS NEEDING ATTENTION	AT RISK
R	3-5	2-3 ⬆	0-2 ⬆
F & M	1-3 ⬆	2-3	2-5
Activity	Recent customers who spend decently	Core customers whose last purchase happened more than one month ago	Purchased often and spent large amounts, but hasn't returned for a long time
Actionable Tip	Offer loyalty benefits	Send personalized emails + recommendations	Make personalized limited time offers

Over the course of this project, our group learned that RFM segmentation created better customer segments than k-means clustering. Furthermore, we reached a deeper understanding of H&M's customer base during this time. Our results pertaining to how we segmented customers and which groups should be targeted are valuable for marketing managers to formulate their marketing strategies.

Future Work

Some future work we can do to further expand our project is to make our machine learning method supervised. We would be able to use precision, accuracy, and recall scores to verify if the model that we constructed is a valid system that can accurately predict customer segmentation based on demographic characteristics and spending trends. By using KNN Classification, the model can automatically learn and then classify each customer based on the characteristics they learned from the already segmented customer groups. This way, marketers don’t need to go through the whole process of manual work and self-segmenting thousands of customers into different segments to identify their target groups. This system can save time, and marketers can work more on how to effectively communicate with each customer segmentation and build better customer relationships instead of trying to find customers that are most profitable to the business. Additionally, incorporating other CSVs from the Kaggle data would be helpful. For instance, using the article IDs could allow us to build personalized recommendations. We could also identify further segment details based on demographic markers to develop a more detailed segmentation and to learn more about each segment in terms of what they usually buy and what their age ranges would be, which can lead to their lifestyle preferences and unique customer personas.