**Lakshmiah S Boyini**

**AIT 664-006**

**Professor Hashmi**

**13 December 2022**

**Final Project: Is Product Price Determined by Stock?**

**Abstract:**

This experiment examines if a product’s price is a factor for stock volatility and identifies product categories that are subject to the stock market fluctuations. Our hypothesis assumes cost increases/decreases to a selection of products reflect how a company's stock price is assessed.

**Introduction:**

Three research questions were utilized to confirm or disprove our hypothesis. Our first question was centered on data consistency, and if we could visualize the amount of randomness in our datasets. Large eCommerce and Stock Market datasets were required for our experiment. If our Stock Market dataset produced largely Brownian patterns, certain model features would have to be cut. One feature was chosen to provide better clustering: Dividends. Dividends are a good measurement because they highlight a company’s health. They require that investors are paid regularly for their investment and are more likely to be paid out by large, stable companies. Using K-means clustering, strong groups could be identified using this feature. This poses our second question.

If any groups were identified and a stock market change occurs is there a correlation between stock prices and item prices? Our abstract touches on this, but this second question requires consistent positive, negative or neutral trends to be identified between stock price differences and item prices. Stock differences are calculated here because for ROI and stock investments, a return of 10% is seen as significant (Speights, 2022). This volatility can be positive or negative depending on the timeframe, but if certain item prices are affected by a stock volatility it could yield informative data. The item price operates as our independent variable for univariate analysis and stock price differences operate as our dependent variable. The inverse of this relationship was also tested, with the caveat that there were filters on item price.

Our models for this second question utilize Linear Regression to determine correlation. MSE and R2 evaluations will be used to determine how "off" our metrics were. MSE determines the total error associated with the model by comparing our actual stock price differences to our predicted stock price differences. R2 is used to ensure that the dependent variable's variance (i.e. the Stock Difference) is *[sic]* "predictable from the independent variable(s)." (Rowe, 2018). If R2 is 1, we can generally predict the variance associated with both variables. Pearson correlation interprets the correlating features of our stock difference data and item prices. It also provides p-values for determining how significant these results are. Results can be seen in the Implementation section of this paper. This question's conclusion led to our third question: What are some identifiable and comparative characteristics about the listed products, their consumers and their branding?

This question came as a result of two sub questions: Can you identify different types of customers based on buying patterns, and can certain brands be associated with a given product? The first subquestion is relevant because product price is determined by what a customer is willing to pay (Stobierski 2020). Certain customers may correspond to certain groups, so identifying and providing data is informative to the end goal of determining what makes an item's price relevant. The second subquestion identifies if brands have an enduring effect on identification. If certain brands (e.g. Apple) are associated with one or more electronics (e.g. smart watches, smart phones, etc.), it could affect how products are bought.

**Dataset, Preprocessing, and Analysis:**

The Ecommerce Purchase History dataset contains information about products that have been purchased online. It includes several features for each product: its category, brand, price, and event time. Product IDs, Order IDs, and User IDs were included as well, but were unneeded and subsequently cleaned. To join this data set with stock information, the item's brand had to be extracted. The brand acts as a key to a given ticker symbol pulled from the *finance.yahoo.com* site. To avoid ingesting static html information, ticker information had to be parsed through regex and searched by a product's brand association. This ticker information would be written to a file called *TickerInformation.json*, which would make loading the data easier for later re-runs.

The next step utilized python's yfinance library to obtain information about the stock prices from these companies at the time of the purchase. All yahoo finance data was scraped in between the years of 2010-01-01 and 2020-12-31. This served two purposes: Dividends tended to be sparse, and more information could lead to better results when analyzing. The dividend data and stock prices were then scaled. KMeans Clustering was performed to determine if there were any similar groupings that could be identified. These results can be found under Figure 1, to which a large yellow cluster of data can be identified among the other outliers.

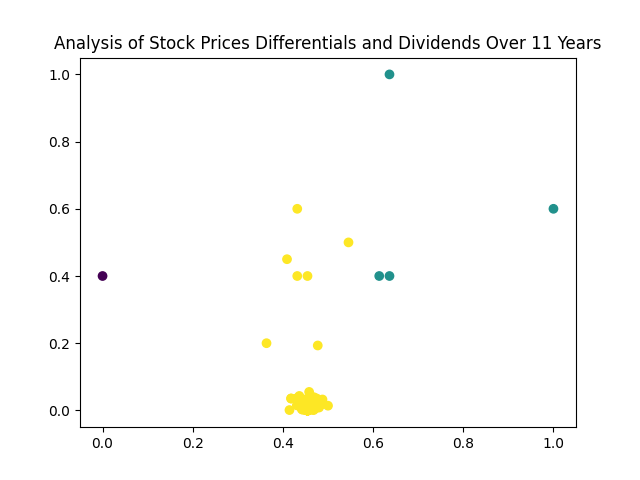


Figure 1

This provides an answer to our first analytical question because we can visibly see how different each company is with the clustered data provided. KMeans clustering (and clustering in general) are great tools for identifying which groups are where based on their centroids. Figure 1 highlights this concept, because you can visually identify a strong grouping near the bottom of the group. There is also an acknowledgement of outliers based on the centroids seen.

This stock data (with and without dividends) was then joined to our eCommerce data, and run through a Linear Regression algorithm. The Linear Regression resulted in the following statistics. When running, correlation was positive at around 0.037 with a p-value of 0.016. A total MSE was listed at around 2.71e-05 and R2 at 0.016. The graph can be viewed under Figure 2:

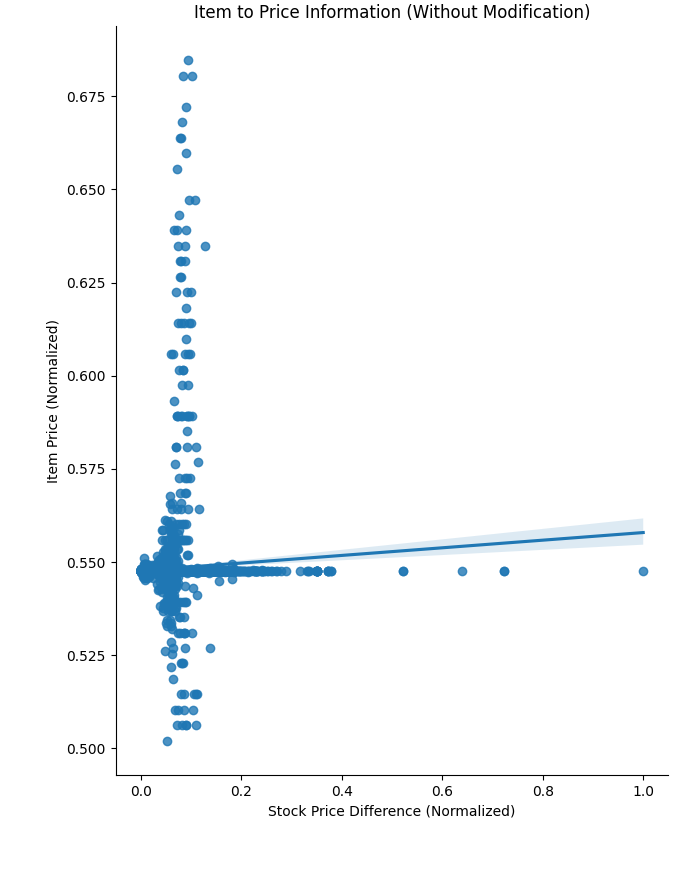


Figure 2

The Item Price is labeled on the y-axis, and the Stock Price Differential is listed on the x-axis. You may notice that the item price is limited to the ranges of 0.5 and 0.7. This is because the majority of prices remain below $1000 from the raw data we used for discovery. This is an intentional design decision for the purposes of reducing skew and properly analyzing the regression taking place. It should be noted this is the rawest form of the data provided. With an incredibly low p-value and R2 score, it's likely this model wouldn't fly. It's too random for our tastes, so it's necessary to modify it for our purposes. The following includes three charts that have been filtered by their Stock Price change rate to highlight different values within each, starting with Figure 3.

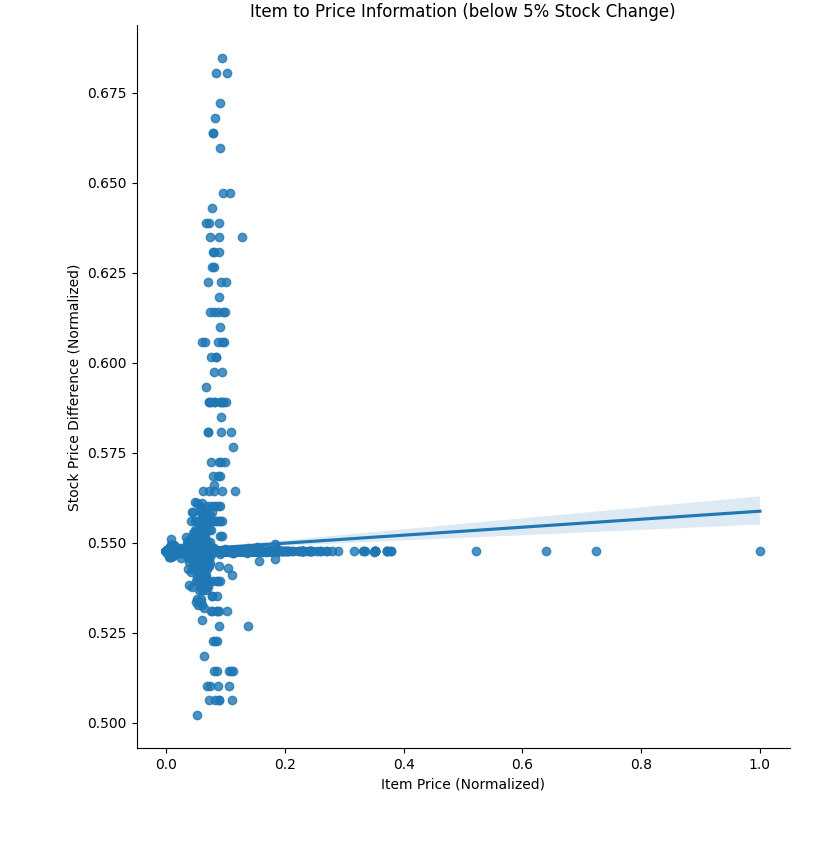


Figure 3

Figure 3 doesn't have too many differences from Figure 1. The correlation is 0.05, p-value is 0.0008, R2 is 0.001, and MSE is 3.28e-05. This information is being heavily skewed by the data provided.

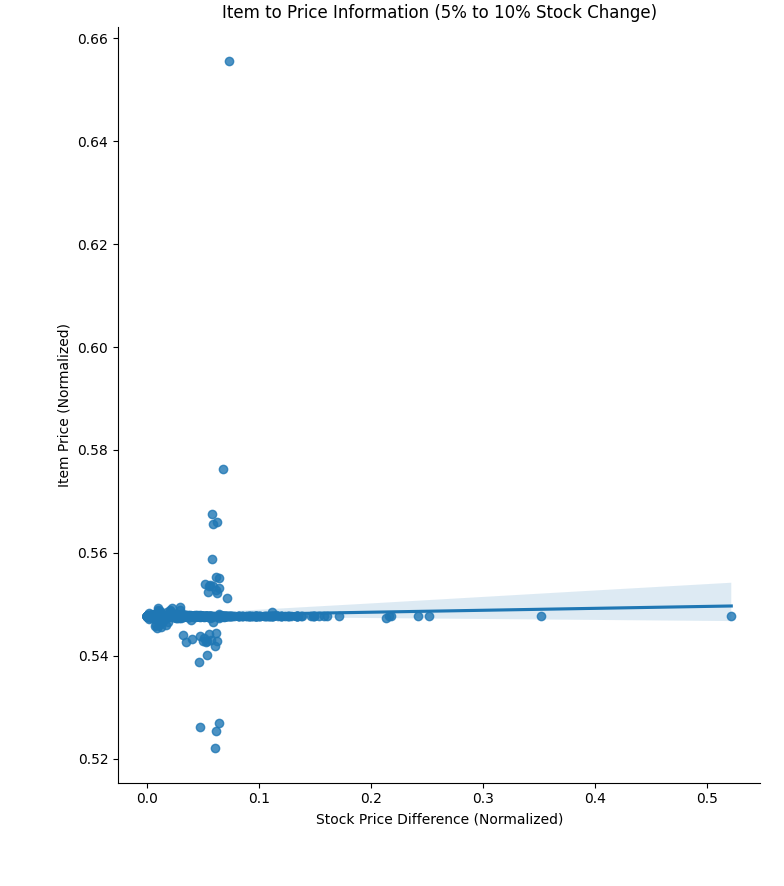


Figure 4

Figure 4's results are considerably better. They represent a 5% to 10% stock change, which is correlative to the prior findings in what we can expect from stock changes. The correlation is slightly negative at -0.009, with a p-value of 0.829, and an MSE of 3.266e-06. However, the R2 does float into the negative territory at -0.02, which harms the credibility of the graph.

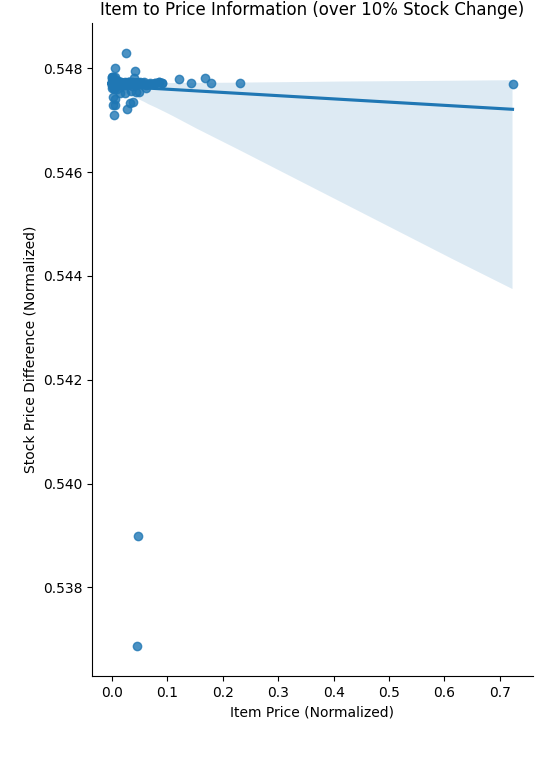


Figure 5

Figure 5 has much more handicapped values than Figure 4, and the results appear to be worse than expected. The R2 value of this graph is at -1, indicating bias which does throw the graph off. The MSE and P-Value are still noticeably better than Figures 2 and 3 at 1.77e-08 and 0.49 respectively, but there is not a lot of confidence here. The correlation is listed as -0.069, making it negative. When swapping the dependent and independent variables, the issue makes a little bit more sense.

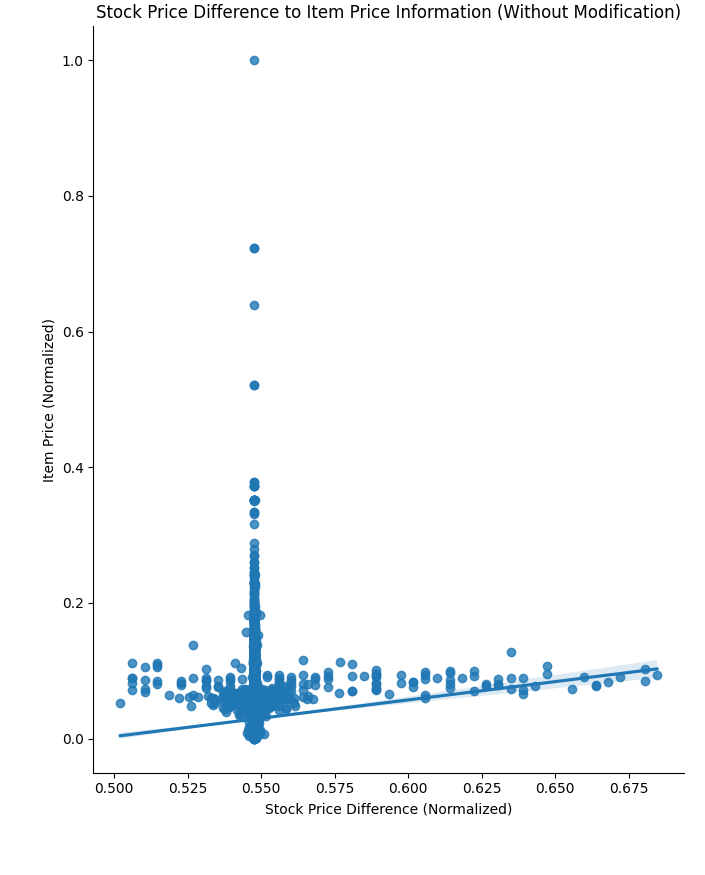


Figure 6

There is a noticeable linear correlation that can be identified, but it is incredibly distracted by a skew based on item price. The correlation sits at 0.03, it's p-value 0.016, R2 at -0.003, and MSE at 2.71e-05. When this data is limited by it's Item Price, the following occurs:

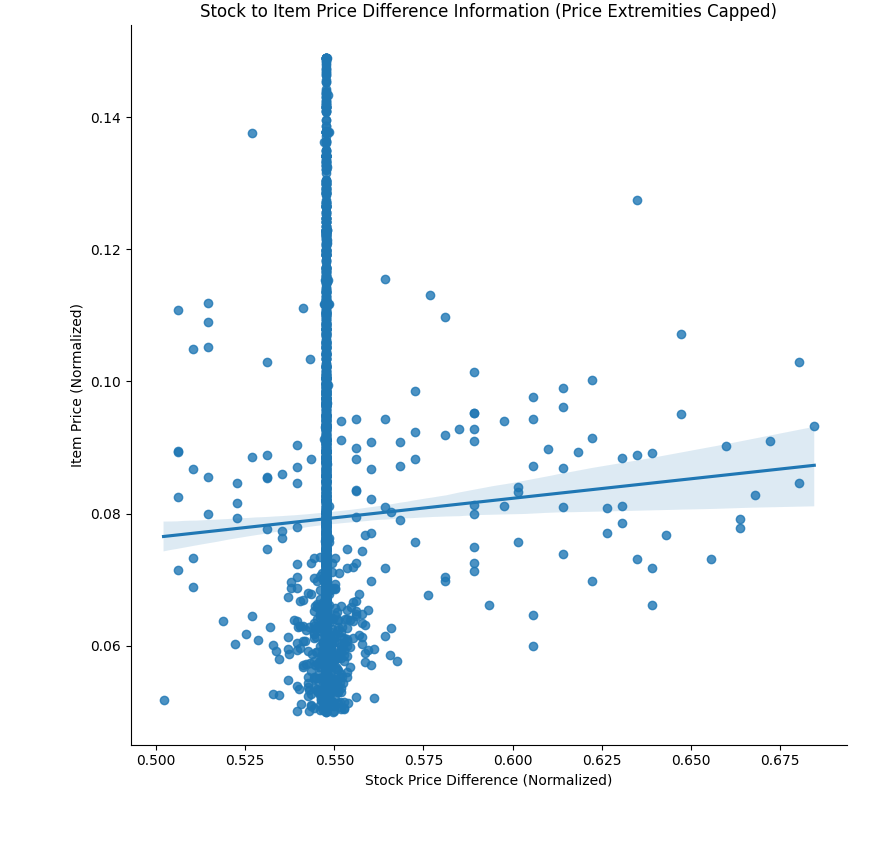


Figure 7

This is the moment when we realized that there was no discernable pattern that could be derived from utilizing Item Price and Stock Price Differentials. The values associated with these price extremities included correlation being at 0.039, p-value at 0.27, R2 at -0.006, and MSE increasing by 0.0005. The correlation and R2 identify that there is no real positive or negative pull when considering all stock pricing. The p-value's assessment also tips off that this is not significant, meaning that it can't be applied to all circumstances. There is a high likelihood that it might be applied to some under Figure 4, but it's not consistent enough to happen all the time. These issues could arise due to a number of factors. Temporal bias is a concern in this situation, and much of the eCommerce pricing data was averaged by its event time and brand.

Our exploratory data when assessing customers, branding, and item prices yielded interesting results. Customers could be grouped into 6 different categories based on our dataset. A frequency chart of these different classifications has been produced in Figure 8. This data could be grouped into 8 different categories, but we have chosen the most frequent. All results were based on code that we developed for this program.

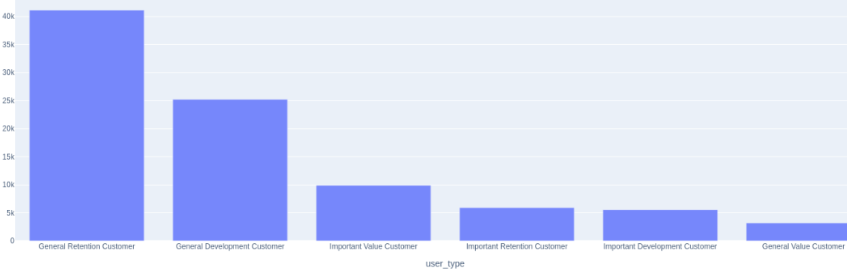


Figure 8

RFM (Recency, Frequency, Monetary) is a measurement used to identify how long ago a purchase was made, the frequency of a purchase, and how much revenue is generated from the purchase. It was assumed that these customer groups listed above would generate at least some good insights to what made a product valuable. KMeans clustering was provided to explain the relationships between customers and their buying habits. Figure 9 displays an elbow graph used to mark the best number of k-iterations possible to reduce the amount of error found.

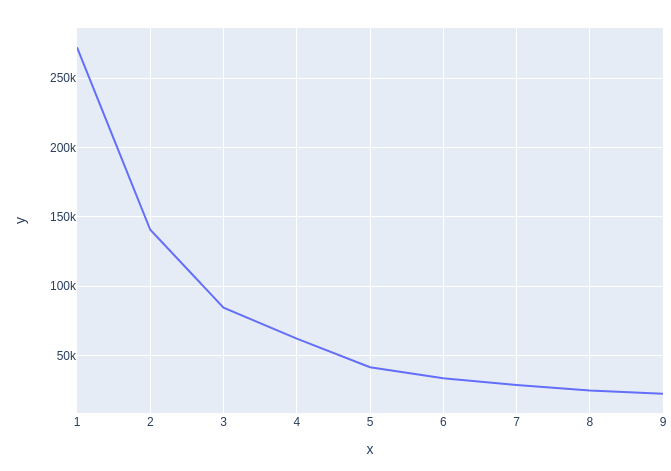


Figure 9

Figure 10 provides the resultant unsupervised values of our KMeans Clustering. It highlights the Frequency, Recency, and Monetary results from worst (1) to best (5). The data indicates that the best values found include Frequency/Monetary values that have a rating of 0 to 5, and negative Recency values which must be around 0 to -2.

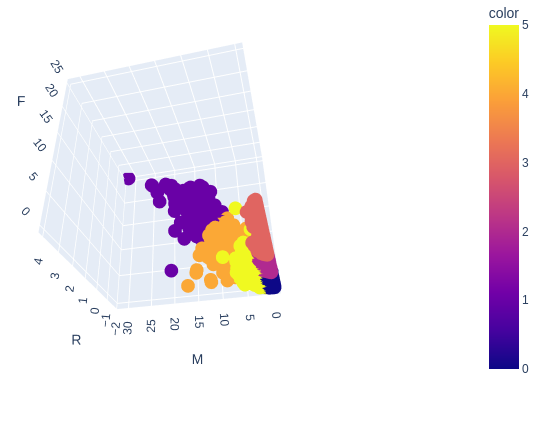


Figure 10

To find if brands contain select items, apriori and association rules were utilized. Figures 11, 12, and 13 highlights that it is possible to retrieve the support. This is helpful because it describes how items are used by each brand.



Figure 11

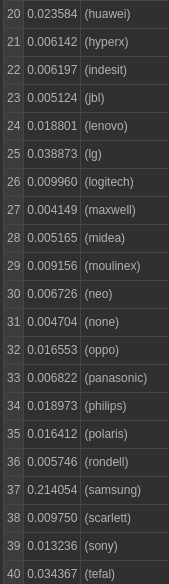


Figure 12

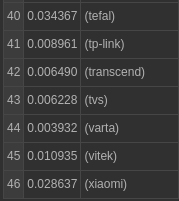


Figure 13

Apple and Samsung clearly have the highest levels of support, but when a min confidence of 0.007 is provided, there is no connection that can be seen (Figure 14).

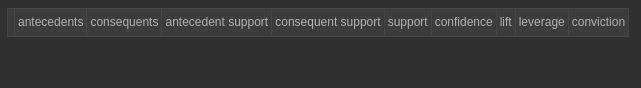


Figure 14

This indicates that no one brand has an overwhelming influence on a particular device, and this can be seen even with the brands listed.

**Conclusion:**

Many of the questions asked within the introduction were able to be answered, but some yielded unsatisfactory conclusions. We were able to visualize the randomness of our datasets utilizing dividends as a way to show company stability. The R2 scores from all product pricing and stock price differences yielded inconsequential or wrong data. The highest predictability of our Linear Regression data suggests that this is not a consistent occurrence, but that it does occur with some frequency (approximately 82% at best). Brands are not likely to be connected to any given items based on the confidence levels provided, but the RFM results for the customers listed do highlight a potential field of research. All RFM values have neatly clustered data, and present possibilities to better tune this information based on the customer groupings listed in Figure 8. Future research into customer RFM would be beneficial to those who are interested in studying financial data. Next steps would be to identify which products customers feel most aligned towards, but overall the data provided should be better filtered to one specific domain.

**References:**

O'Shea, A., Lam-Balfour, T., & Durana, A. (2022, November 21). *What is a dividend and how do they work?* NerdWallet. Retrieved December 13, 2022, from https://www.nerdwallet.com/article/investing/what-are-dividends

Murphy, C. (2022, November 19). *Recency, frequency, monetary value (RFM) definition*. Investopedia. Retrieved December 13, 2022, from https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp

Speights, K. (2022, November 8). *What is a good return on investment?* The Motley Fool. Retrieved December 13, 2022, from https://www.fool.com/investing/how-to-invest/stocks/good-return-on-investment/

Rowe, W. (2018, July 5). *Mean square error & R2 score clearly explained*. BMC Blogs. Retrieved December 13, 2022, from https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

Stobierski, T. (2020, October 20). *Willingness to pay: What it is & how to calculate*. Business Insights Blog. Retrieved December 14, 2022, from https://online.hbs.edu/blog/post/willingness-to-pay