**On-line Spending Habits: Describing Behavior is Easier Than Predicting It!**

Jolene M. Branch

Data Science Program, Bellevue University

DSC 680: Applied Data Science

Prof. Catie Williams

July 18, 2021

**Abstract**

Before Amazon became a source for nearly all consumer goods (including food), it was a simple site for purchasing books. (Hartmans, 2021). As more categories of items were added and rural families became increasingly dependent upon the site, the number of purchases in some households grew large enough to generate sufficient data for models such as linear regression and possibly even multiple logistic regression. This paper attempts to answer the questions; “Can linear regression be used to predict if I will purchase something based on either one of two derived variables? (Because if I can figure that out, Amazon probably already has too)!”, “Can Amazon predict what categories of product I will pay full price for?”, “Do my husband and I have any duplicate purchases? If we do, who gets the better deal?”, “How much % off does a textbook have to be for me to buy it new?”, “Am I more likely to purchase a book NEW if it is a more recent release?”, “Can Amazon predict how many home deliveries will be required in a month?”, “Hypothesis: I’m willing to pay for shipping if I can get a great deal (such as for a used book)?”, “Are there product category purchases that we 'hide' from each other by using a gift card or reward points?”, “Does Shipment Date correspond with Delivery Date (which isn't a variable)? Because even though my purchases outnumber those of my husband by about 3 to 1, I try and have them all delivered on my Prime day,” and “Who is the ‘better’ customer? (rate of returns) Were returned items not as good of deals? (Can Amazon use that understanding to predict return rates?)?” Data that has been collected by the most reliable source possible, Amazon itself, will be used. This is a sample of convenience, with added benefit of having similar controls between the two subjects. The project required extensive data ‘wrangling’ in Excel and R prior to exploratory data analysis, insight generation, and statistical modeling. The bulk of the coding for this project is done within the data clean-up and visualization parts of the process, using R.

*Keywords:* online shopping, retail, seasonality, Amazon

**On-line Spending Habits: Describing Behavior is Easier Than Predicting It!**

**Purpose**

The purpose of this project started out to retrieve and wrangle the data myself in order to look for predictive models within the purchase history of the family. As the challenges associated with wrangling irregular time series data became known, the nature of the project changed to more of an Explanatory Data Analysis and Storytelling project.

**Data Sources**

Three datasets from amazon.com were used, for a total of six files. The datasets were in the form of .xlsx files retrieved using the ‘Download order reports’ selection under ‘Ordering and shopping preferences’ of both spouse’s accounts. (Vasdev, 2017). One dataset contained items ordered, date, shipping location and recipient, condition, list price, purchase price, payment type. Another contained shipping dates by date ordered and item. A third contained information on returns, of which there were surprisingly few. All were converted to .csv format and imported into R using read.csv. The original file contained 36 columns and 2035 rows.

**Methods**

After importing the datasets into R, they were written back out as .csv files for simpler importation into R. (Either format is incredibly simple to import using Power BI). The files from both spouses were immediately joined, as there was already a category for buyer name. Immediate need for data de-identification became apparent, as the datasets will be accessible to the public via this student’s portfolio. Besides names and email of the two spouses, names and addresses of all recipients needed to be de-identified in such a way that the author would be able to reconstruct all data. Use of relationships to the author and state of residence was sufficient to resolve this challenge. Tracking numbers were removed from the CarrierNameAndTrackingNumber column in case they could be associated with latitude and longitude of recipient addresses. ‘Last 4s’ were removed from the credit card payment type values.

***Variable Selection***

Variables were selected for removal based on singularity of values, blanks, and data of no interest to the author (such as seller’s credentials). Derived variables were created within Excel. SavingsAmount and SavingsPercent were created using the list price and purchase price of items. An additional categorical column was created using bins of percentages off (for possible box and whisker charts). In the event that enough data would be available for predictive analytic modeling, purchases from all time for both spouses was included, going back to 2006, and the variables of year, month, and day of week of purchase were added (for seasonality modeling). Interestingly, the husband, who had about 1/3 of the purchases of the wife, had more returns. The numbers were so few that there was no benefit anticipated from joining that dataset with the other two.

***Data Cleaning and Dataset Trimming***

While creating the ‘ListMinusPurchasePrice’ variable, I noticed many negative numbers in the results. Since Amazon does not show a list price lower than the purchase price on screen while shopping, I went with the assumption that this was an error. Fifty-nine items had purchase price higher than list price. Also, some list prices were entered as $0.00. (425 of them). To overcome this, I created an IF formula so that when list price is $0.00, it automatically entered the purchase price in that cell. =IF(ListPriceCorrected<PurchasePrice, PurchasePrice, ListPriceCorrected). Then I replaced the entire ListPrice column with the corrected list prices.

The format of the money variables was changed to ‘currency’ in Excel This allowed R to recognize the values as numeric upon import. Since there was no unique identifier (because multiple items purchased at the same time shared an OrderID), and ID column was added as an index in R. Then it was moved from last to first in the dataframe.

The irregularity of the data cause significant obstacle in attempting to re-create others’ time series project ideas. I was able to create a time series of all the order dates, but as soon as I attempted to work with them, such as putting them into bins, they changed to numeric date-times, like 14400.

***Outlier Detection and Missing Value Decisions***

The normal process of outlier detection was viewed differently within this project than within most data science projects. Traditionally, each variable in a dataset is examined, often via visual representations such as histograms, cumulative distributive function (CDF), or probability mass frequency (PMF) charts. However, due to the extremely personal nature of this dataset, a type of ‘own recognizance’ process was utilized in the validation of individual and aggregate values within the datasets. For example, while most other outlier detection methods or duplicate data triggers would have highlighted the three-time purchase of the Calvin and Hobbles book ‘Attack of the Deranged Mutant Killer Monster Snow Goons,’ all shipped to the home address, as a data accuracy error, the mother of two untidy teenage sons knows that this is simply reflective of the poor item storage habits within the household. Of note; the other pleasure-reading book with duplicate purchases is titled ‘What Do You Call a Sociopath in a Cubicle? Answer: A Coworker.’ It seems that individual data at this level of granularity within a family unit could be easily misconstrued! “*Who are these people*?!”

In anticipation of looking for outliers in savings percentages using box plots, I wrote a nested IF formula in Excel and created another derived column, PercentOffCategory;

=IF(R2>0.9,"91% or more",IF(R2>0.8,"81-90%",IF(R2>0.7,"71-80%",IF(R2>0.6,"61-70%",IF(R2>0.5,"51-60%",IF(R2>0.4,"41-50%",IF(R2>0.3,"31-40%",IF(R2>0.2,"21-30%",IF(R2>0.1,"11-20%",IF(R2>0.01,"Up to 10%","0"))))))))))

Items missing the item name but containing all other data elements were retained. Nine rows with ItemTotal of $0.00 were deleted. All belonged to the husband.

A large gap in days without purchases was seen in early 2021 by both spouses.

A picture containing timeline

Description automatically generated

Nothing like this was seen in the years since ‘order everything from Amazon’ became the family shopping strategy.

**Analysis and Findings**

Challenges associated with creating an appropriate time series for this irregular data were not overcome. The only time series chart able to be constructed is a simple ‘Item Cost in USD’ based on ‘date’ chart.

Chart, bar chart, histogram

Description automatically generated

This chart alone is not able to detect and predict seasonal variation, but it is a clue that this project is attempting to head in the right direction.

The chunk of time in early 2021 with no items purchased by either spouse is likely attributed to COVID-19 movement restrictions. The family simply was not ‘doing anything’ that required purchasing new items.

***Question 1: “Can linear regression be used to predict if I will purchase something based on either one of two derived variables? (Because if I can figure that out, Amazon probably already has too)!”***

Chart, scatter chart

Description automatically generated

It appears that there is a positive correlation between the Cost of Item and Dollars Saved. However, the strength of relationship weakens as the Cost of Item increases due to small number of instances at the higher price levels. This might also be skewing an Item Total and Percent Savings chart, subsetted by spouse:

Chart, scatter chart

Description automatically generated

***Question 2: “How much % off does a textbook have to be for me to buy it new?” AND  
Question 3: “Am I more likely to purchase a book NEW if it is a more recent release?”***

Having purchased about half of the textbooks for a graduate program used but seeing only one used book purchase listed in recent months leads to the unfortunate conclusion that Amazon no longer tracks the condition of the books that it sells through outside vendors or itself. This results in an inability to predict purchase of a used book based on date of publication or purchase price.

***Question 4: “Can Amazon predict how many home deliveries will be required in a month?”***

This is likely the case, however, the ShipmentDate is only a proxy variable for the ReceiptDate.

***Question 5***: ***“Hypothesis: I’m willing to pay for shipping if I can get a great deal (such as for a used book)?”***

Having purchased about half of the textbooks for a graduate program used but seeing only one used book purchase listed in recent months leads to the unfortunate conclusion that Amazon no longer tracks the condition of the books that it sells through outside vendors or itself. This results in an inability to predict purchase of a used book based on date of publication or purchase price.

***Question 6: “Are there product category purchases that we 'hide' from each other by using a gift card or reward points?”***

No such pattern was found. However, an interesting finding of very few items costing more than $100 being ordered by the wife indicates that at least one of the married couple remembered their pre-marital agreement to always check with the other prior to any item purchase of over $100.

Chart, scatter chart

Description automatically generated

***Question 7: “Does Shipment Date correspond with Delivery Date (which isn't a variable)? Because even though my purchases outnumber those of my husband by about 3 to 1, I try and have them all delivered on my Prime day,”***

Yes, these two variables correspond, but have random variation even within themselves.

***Question 8: “Who is the ‘better’ customer? (by rate of returns) Were returned items not as good of deals? (Can Amazon use that understanding to predict return rates?)?”***

Neither spouse had enough items returned to attempt to use this variable for prediction modeling.

***Question 9: “Are item prices normally distributed?”***

Histogram

Description automatically generated with medium confidence

No. Item prices are not normally distributed. Even without outliers greater than $100 removed from the chart, a right skew is still seen among the data points.

***Question 10: “Who orders the more expensive items?”***

Chart, scatter chart

Description automatically generated

While the wife ordered about three times as many items, the number of items that cost more than $100 was at least two times more apparent within the husband’s data.

**Conclusions**

Initial hypotheses (in the form of questions) proved to be as likely to be correct as incorrect. This is due to inability to validate hypotheses based on irregular time series data. The number of items that the husband purchased that cost more than $100 comes as a complete surprise, especially given how few the wife purchased.

That shoes and books are the two named categories is not a surprise, as that is how the company got its start, and it has had many years to perfect the business model.

Based on the very close linear regression line of percent off purchases between the husband and the wife, it is unlikely that a prediction model would find much predictive power in the difference between the spouses, but with multiple regression based on category and cost or purchase, it is possible that this could be done in the general sense for the whole family unit.

Due to the unsuccessful attempts to wrangle this irregular time series data within R in manner to be useful for modeling, the descriptive nature of this data is all that can be used to make inferences of future behavior.

**References**

Custer, C. (2021, June 9). *How Much Have You Spent on Amazon? Analyzing Amazon Data*. Dataquest. <https://www.dataquest.io/blog/how-much-spent-amazon-data-analysis/>

Editors, T. (2021, April 26). *The 100 Best Shopping Sites to Browse Through*. Marie Claire. <https://www.marieclaire.com/fashion/g1898/best-online-shopping/>

Federal Trade Commission. (2021, May 1). *Online Shopping*. Consumer Information. <https://www.consumer.ftc.gov/articles/online-shopping>

Ganley, M. (2021, April 2). *How to check the total amount you've ever spent on Amazon*. KSAT. <https://www.ksat.com/features/2021/04/02/how-to-check-the-total-amount-youve-ever-spent-on-amazon/>

Hartmans, A. (2021, July 2). *Jeff Bezos originally wanted to name Amazon 'Cadabra,' and 14 other little-known facts about the early days of the e-commerce giant*. Business Insider. <https://www.businessinsider.com/jeff-bezos-amazon-history-facts-2017-4>

Jusoh, Z. M., & Ling, G. H. (2012). Factors Influencing Consumers' Attitude Towards E-Commerce Purchases Through Online Shopping. *International Journal of Humanities and Social Science*, *2*(4), 223–230.

Muenchen, B. (2017, March 23). *The impact of R's dplyr package and its tidyverse companion packages on R's learning difficulty*. r4stats.com | Data Science Services. <http://r4stats.com/2017/03/23/the-tidyverse-curse/>

Safdari, N. (2018, December 2). *If You Can't Measure It, You Can't Improve It !!!* Medium. <https://towardsdatascience.com/if-you-cant-measure-it-you-can-t-improve-it-5c059014faad>

Sapegina, S. (2018, March 29). STDS Vignette : Regression Modelling with Time Series Data in R. <http://rstudio-pubs-static.s3.amazonaws.com/374830_ab4ac8951c94411aa9abfe5f561cef91.html>

Shork, J. (n.d.). *R ggplot2 Error: Discrete Value Supplied to Continuous Scale (2 Examples)*. Statistics Globe. <https://statisticsglobe.com/error-discrete-value-supplied-to-continuous-scale-in-r>

*Spend Analysis 101: Comprehensive Guide for Beginners*. Sievo. (n.d.). <https://sievo.com/resources/spend-analysis-101>

Team, T. I. (2021, June 23). *The Pros and Cons of Buying on Amazon*. Investopedia. <https://www.investopedia.com/articles/personal-finance/070215/how-buying-amazoncom-works.asp>

Vasdev, S. (2017, May 15). *How to find and visualize your Amazon order history data (Tutorial)*. TechChange. <https://www.techchange.org/2017/05/15/how-to-find-and-visualize-your-amazon-order-history-data-tutorial/>

Wang, J. (2020, April 8). *How To Model Time Series Data With Linear Regression*. Medium. <https://towardsdatascience.com/how-to-model-time-series-data-with-linear-regression-cd94d1d901c0>

**Appendix**

Chart, treemap chart

Description automatically generated  
Source: Power BI, Amazon Dollars Spent, by Category