STAT 443 Final Project: Predicting Consumer Spending

By Group 5: Sahil Kumar, Jonah Somers

84.51°

1. Introduction

1.1 Project

Data science is an immensely valuable tool for the consumer analytics industry, and it will only continue to rise in importance in the coming decades. Kroger, the United States' largest supermarket chain by revenue, explores insights into consumer behavior with their subsidiary company, 84.51°. Our task was to provide our contact, Matt Zettinger, with a future-based analysis of a large consumer expenditure data set. As our client, 84.51°'s expectations of our project were to understand the consumer landscape from the data set and also to give insights into predicting future consumer behavior. These are broken down by two main questions:

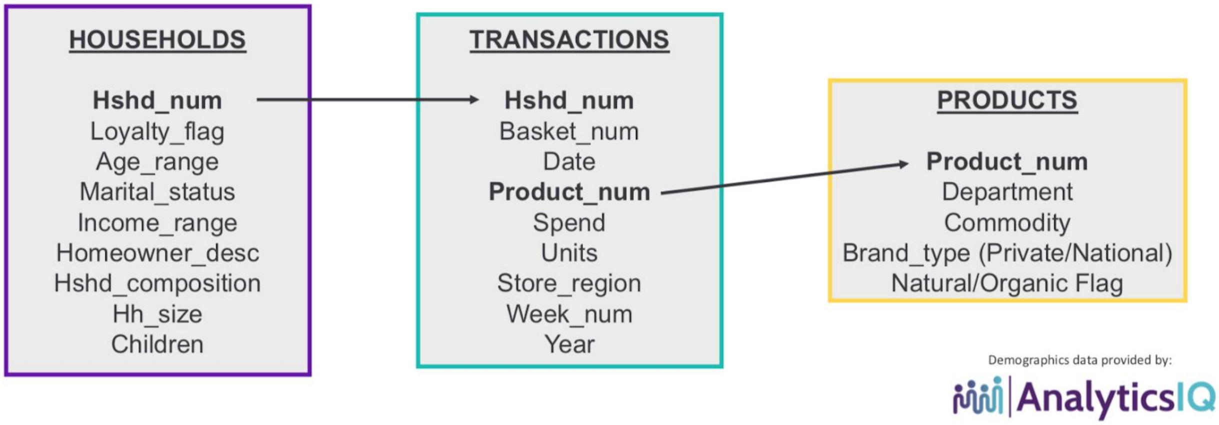
- What are current sales trends?
- How can we increase future sales?

To answer these objectives, our analyses focused on identifying patterns in consumer behavior over time by demographic, geographic, and product features. We also explored predictive analysis. Our goal was to provide 84.51° with the consumer spending insights and recommendations so they are better prepared to adjust Kroger's business plan accordingly.

1.2 Data

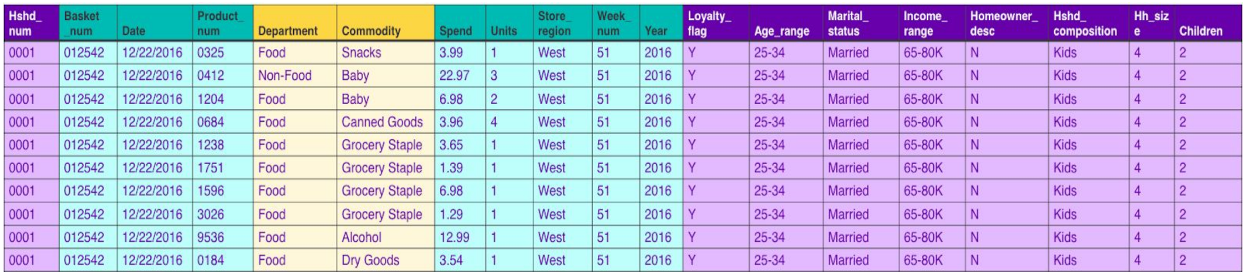
Our data came in the form of 3 distinct files which can be found on 84.51°'s website (See References). Household data in *households.csv* covered 5,000 households (each given a unique household number) and also included other demographic information (*Loyalty_flag, Age_range, Martial_status, Income_range, Homeowner_desc, Hshd_composition, Hh_size, Children*). Product data was contained in *products.csv*, which covered ~150,000 products across 43 categories, and also included other product variables (*Department, Commodity, Brand_type (Private/National), Natural/Organic Flag*). Lastly, the largest data set, *transactions.csv*, contained over 1 million financial transaction records from January 2016 to December 2017. Its variables included both household and product numbers, which are included in the other two data sets, as well as basket numbers, which indicated what products were bought together and in what quantities. All variables were *Basket_num, Date, Product_num, Spend, Units, Store_region, Week_num,* and *Year*.

To merge the data sets, we joined them by the variables *Hshd_num* (contained by *households.csv* and *transactions.csv*) and *Product_num* (contained by *transactions.csv* and *products.csv*), in the manner seen below:



Demographics data provided by: AnalyticsIQ

Each row of the combined data set represents the purchase of an individual item, and multiple rows may make up a basket for a single household. There were ~10 million rows in the raw combined data set, each representing an item purchase. Below are example rows of the merged data set (See Appendix 5.1 for further variable descriptions).



Hshd_ num	Basket _num	Date	Product_ num	Department	Commodity	Spend	Units	Store_ region	Week_ num	Year	Loyalty_ flag	Age_range	Marital_ status	Income_ range	Homeowner_ desc	Hshd_ composition	Hh_siz e	Children
0001	012542	12/22/2016	0325	Food	Snacks	3.99	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	0412	Non-Food	Baby	22.97	3	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	1204	Food	Baby	6.98	2	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	0684	Food	Canned Goods	3.96	4	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	1238	Food	Grocery Staple	3.65	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	1751	Food	Grocery Staple	1.39	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	1596	Food	Grocery Staple	6.98	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	3026	Food	Grocery Staple	1.29	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	9536	Food	Alcohol	12.99	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2
0001	012542	12/22/2016	0184	Food	Dry Goods	3.54	1	West	51	2016	Y	25-34	Married	65-80K	N	Kids	4	2

2. Methods

See Appendix Section 5.8 on reproducibility of methods.

2.1 Data Cleaning

Our first task was to inspect the combined data set after matching the household information to the transactions for each individual product. When we initially combined these data sets, we found that our systems could not handle running predictions or graphing on the full data set due to memory limitations. We discussed our options with the client who informed us that while the missing data could provide valuable insight, we could focus our results on data that was non-missing. By looking through the data manually we saw that most of the missing information came from the *households.csv* data file, where ~3,300 of the total 5,000 households had some form of missing data in at least one of the demographic variables. While our data set shrinking from ~10M to ~3M rows seemed preferable, we tried to keep as much data as possible by checking what missing data we could impute while keeping our assumptions to a minimum.

We started our data cleaning by standardizing the names of the entries in all 3 data sets, where some entries had extraneous spacing and others required changing all “null”, “NOT AVAILABLE”, and “Unknown” entries to be NA. We then tried to identify variables we could impute with minimal assumptions. For example: we found that *Children* was a commonly missing variable, while the households had still reported their *Household Composition*. Based on *Household Composition*, if a household reported either “1 adult”, “2 adults”, “Single Male”, or “Single Female”, we imputed the number of children as 0. A variable that we also found to have many missing values across households was *Marital Status*. By looking at the *Household Composition* again, if a household reported themselves as “Single Male” or “Single Female”, we imputed their *Marital Status* with “Single”.

While we were now able to retain ~3,100 households with this imputation, we decided to take this a step further with a few assumptions to retain another ~300 households. For households where *Marital Status* was missing and *Household Composition* was listed as “1 Adult” or “1 Adult and Kids”, we imputed their *Marital Status* with “Single”. For households where *Marital Status* was missing and *Household Composition* was listed as “2 Adults” and their *Age Range* was listed as “75+” or “2 Adult and Kids”, we imputed their *Marital Status* with “Married”.

Along with this, there were a few households we spot removed due to issues in modeling or trend analysis, such as one household being the only household to report *Household Composition* as “3” but *Children* listed as “0”. Another house we removed after sorting through trend analysis was a household with a reported “Single Male” but average monthly spending of $1,500 on alcohol, as this greatly deviated from the average monthly spending for most households and could influence our results to report a focus on increasing alcohol sales due to a single household. Finally, another household that had an average monthly spending of $2,500 on groceries seemed like an outlier for monthly spending, so we removed this to try and keep our analysis unbiased by strong outliers. Our final household size was 3,474 households that we could analyze.

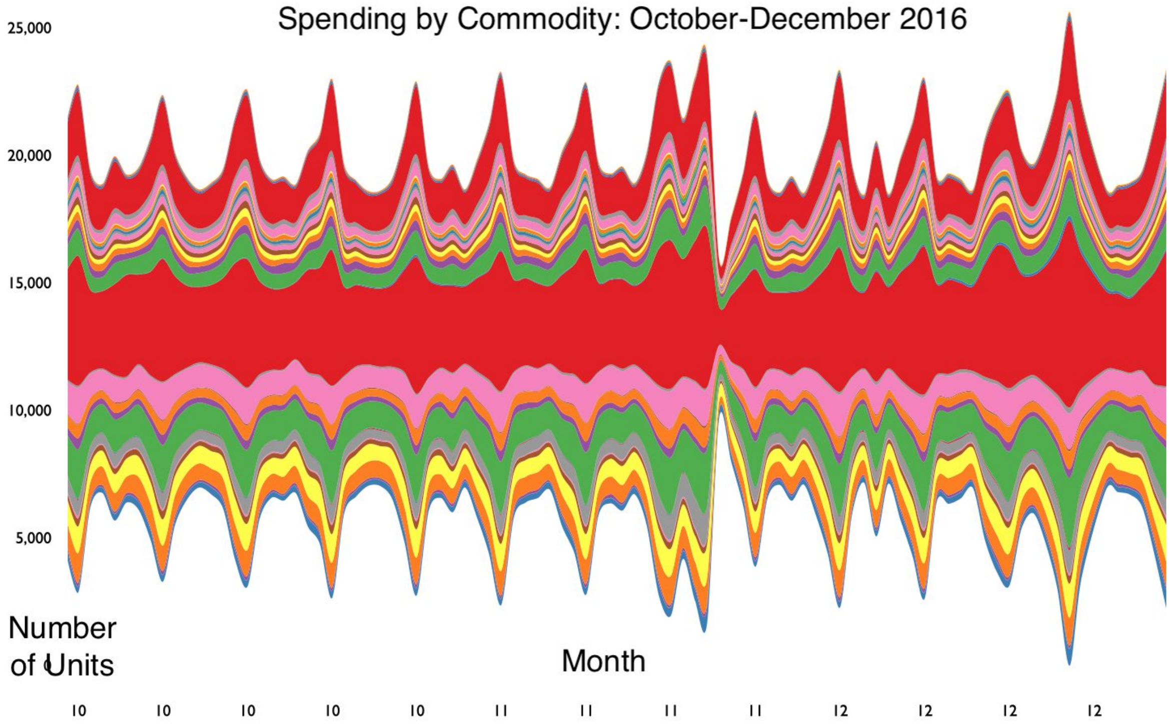
We moved our data cleaning to look at the transactions alone, and identified that some transactions reported 0 or negative amounts spent or units purchased. After again conferring with the client, we were informed that these data points could be transaction errors or reports of merchandise being returned, and we could eliminate these transactions to further clean our analysis. This reduced the number of transactions by 100,000, which could be seen as very significant but we decided to focus on what we could guarantee was a purchase made by a customer. We also extracted date information that we could use in our analysis, such as the specific transaction month or day.

Finally, we looked at the product data for cleaning. From our initial trend analyses, we saw that average monthly expenditure on “Pharma” products across all households was extremely low (See Appendix 5.2). Our client noted that this was something they were familiar with, but told us that we could remove these entries because this was not an area that Kroger wanted to focus on. We agreed that analysis on “Pharma” would probably not be useful for general shopping trends, and while it may be an influential factor on why individuals go to Kroger, it would be better to remove it from further analyses.

By the end of our data cleaning, we had a total of 7,265,112 rows of combined data across 3,468 households. From cleaning on transaction data and product data, we actually saw the number of households decrease from 3,474 to 3,468. While further analysis on those households would have provided some useful insights, we moved forward on our remaining households for general shopping analysis.

2.2 Trend Analysis

Once the data set was cleaned, our initial exploratory analysis through visualizations like line and stream graphs gave us a better idea of the data we were working with. A example stream graph shown below gave us a taste of consumer spending analysis to come.



This nicely illustrates the cyclical nature of consumer spending at Kroger, showing expenditures rise on weekends no matter the commodity category of the product. This also led us to the discovery that Kroger is not very busy on Thanksgiving and closed on Christmas, resulting in bottlenecks near those dates. The timeline of this stream graph is limited to show more closely the behavior of overall spending around the end of the year holidays.

Next, we began to investigate general consumer spending trends in terms of many different categorical variables. We investigated consumer spending aggregated by *Store_region* by converting aggregated data tables into time series objects so they could then be easily plotted. Next, we looked at spending by *Commodity* using the same method. In fact, that method of visualization was utilized many times in this trend analysis section. From the resulting plots, we also decided to pursue exploration of notable niche trends within *Commodity* categories. These included spending by region on grocery staple, floral, outdoor, and international food. This was followed by an in-depth look at the alcohol *Commodity* category which explored trends in alcohol sales over time by *Store_region* and demographic factors. We paid specific attention to those groups in each aggregation that made up a large part of the rows in the data set (grocery staple for *Commodity* and East for *Store_region*)

We then turned our focus to analysis of individual consumer baskets. To get an idea of the amount of products each customer was purchasing during every trip to the store, we aggregated the data to find the average number of items per basket for each week in the data set. We also focused on what items were commonly found together across visits to the store, and aggregated the data to count the frequency of specific pairings.

2.3 Modeling

We performed a number of traditional modeling techniques, including linear regression, LASSO (least absolute shrinkage and selection operator) / Ridge / Elastic Net regression, Random Forest, Gradient Boosting Machine, and Support Vector Machine in order to obtain a predictive accuracy for the monthly expenditure per household. We also attempted Time Series modeling like Seasonal Naïve, Time Series Linear Models, and ARIMA (autoregressive integrated moving average), but we found that this was not very accurate nor appropriate given limitations in the data (further explained in our Results section).

As a preliminary preprocessing step for our regression models, we aggregated the total amount spent per month for each household by the region they shopped. Since we wanted to avoid time series within the standard regression models, we did not categorize the months by the specific years but we left it as a single factor of the month name. For our time series data, we did account for the difference between the months for each year and used it to split our data into training and testing sets. We then matched up the remaining demographic information per each aggregate expenditure for both model types.

2.3.1 Regression

For regression, we decided to stick with Elastic Net Regression and Random Forest Regression for their ease of usage and variable importance interpretability. These models are traditionally more inclusive of the results due to random effects, which is very apparent in the analysis of the trends.

For the Elastic Net model, all of the demographic factors, store regions, and months were converted to factor variables and split into dummy variables to create one hot encoded “continuous” variables. For any factor with only 2 unique responses (such as *Loyalty, Marital Status, and Home Ownership)* we removed the extra variable made due to the encoding method as the information from a single variable would incorporate both responses. We also converted the numbers in the individual months to 3 letter names for those months to make the results easier to read. The assigned household numbers were dropped for both modelling algorithms, but we attempted to incorporate these once for the Elastic Net model, which is shown below in the results.

For the Random Forest model, we only needed to convert all the same variables we used in the Elastic Net model to factor variables since R’s “randomForest” package would properly one-hot-encode the variables to the algorithms specifications.

After preprocessing for both methods, we split the datasets into training data and testing data sets, with the training data using 3,000 households worth of transactions to build the model and the testing sets having the remaining households. This split was chosen over the more common 80-20 split because we wanted to make sure there were as many households as possible in the training set to draw further analyses.

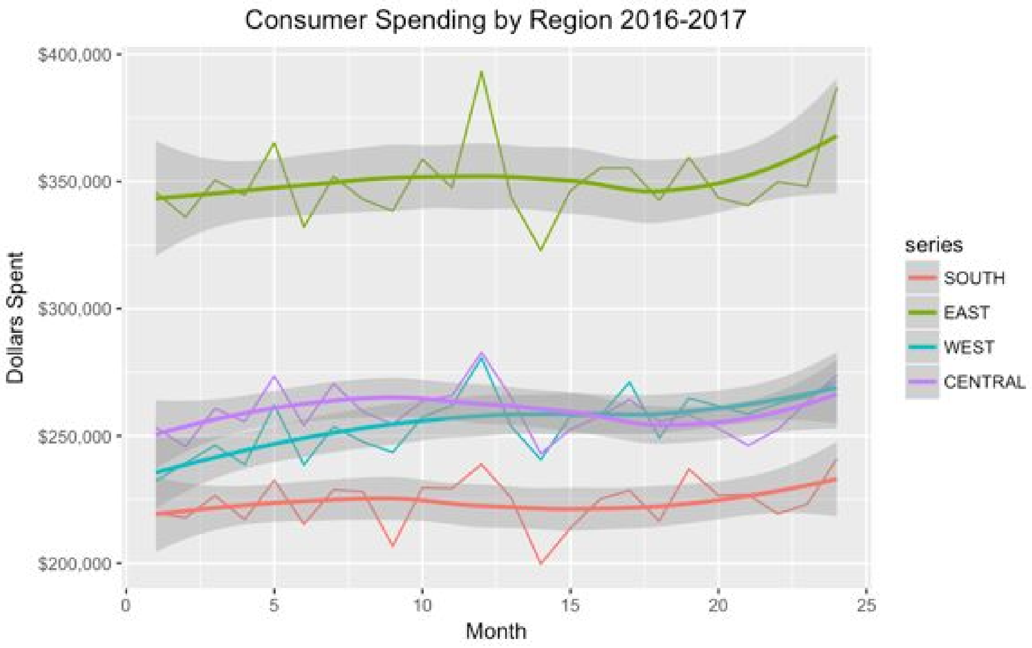
2.3.2 Time Series

For the time series modeling, we took the aggregated data set and split them into training and test sets, where the train set had purchases in 2016 and the test set contained purchases in 2017. We also performed additional training with the time series using the entire data set for training and plotting a predicted trend for 2018. We stuck to Time Series Linear Models for forecasting as they provided the only real insight due to the short timespan available for modeling, and the difficulties of using these techniques on non-stationary time series. The Time Series Linear Models were made with and without seasonality, based on how many years we had available for training.

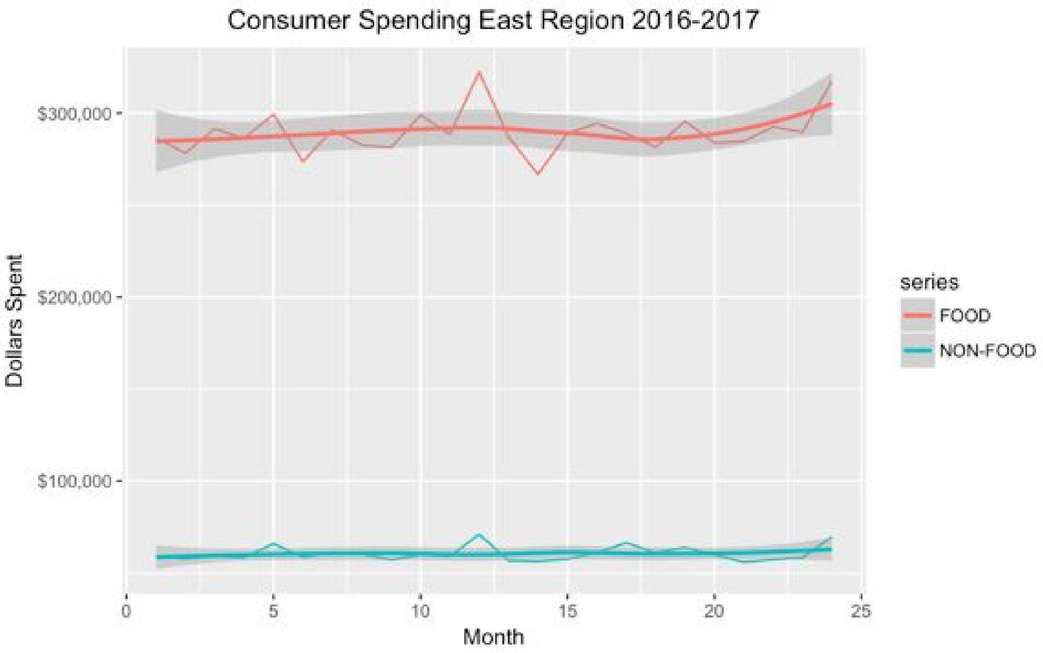
3. Results

3.1 Trend Analysis

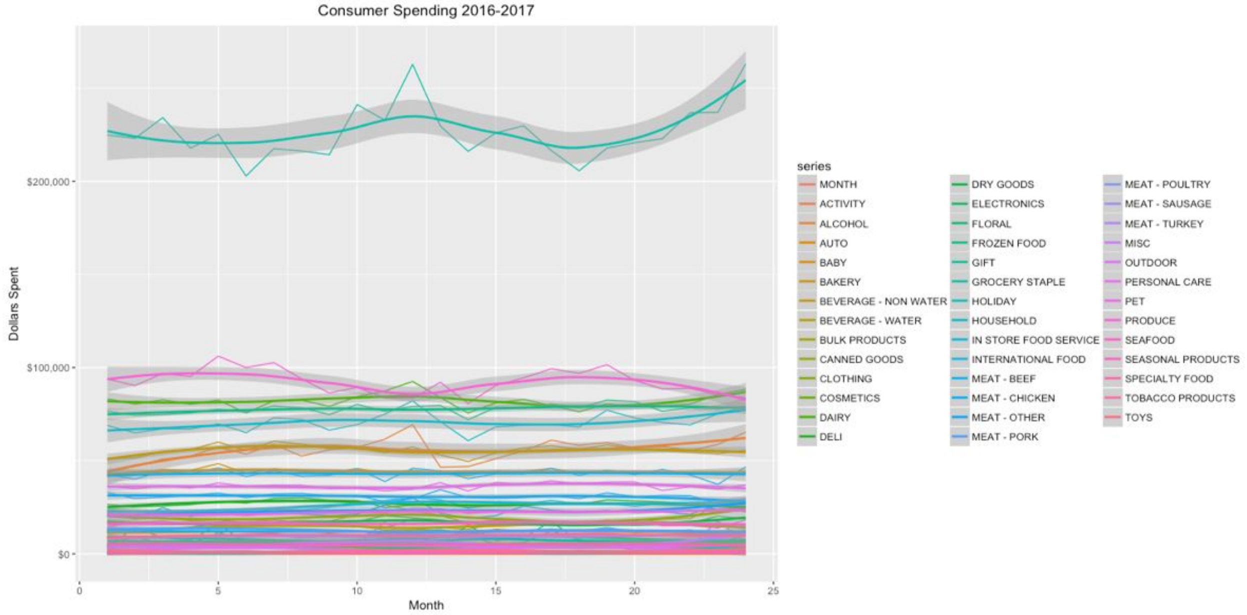
Following our trend analysis, many important discoveries were evident. The consumer behavior not easily seen in the initial stream graph shown above was revealed in more detailed plotting techniques. The following plot which breaks down spending by region gave us new geographic consumer information to dissect.



The East region has the most consumer spending by month of the four regions, and by a large margin of about $100,000 over the next highest, the Central region. And when we broke down that spending in the East by *Department* we found that, as is the case in every region, the Food department outsells Non-food by a large amount in dollars spent per week, as seen below (Pharma was even lower and was dropped from the analysis - See Appendix 5.2).



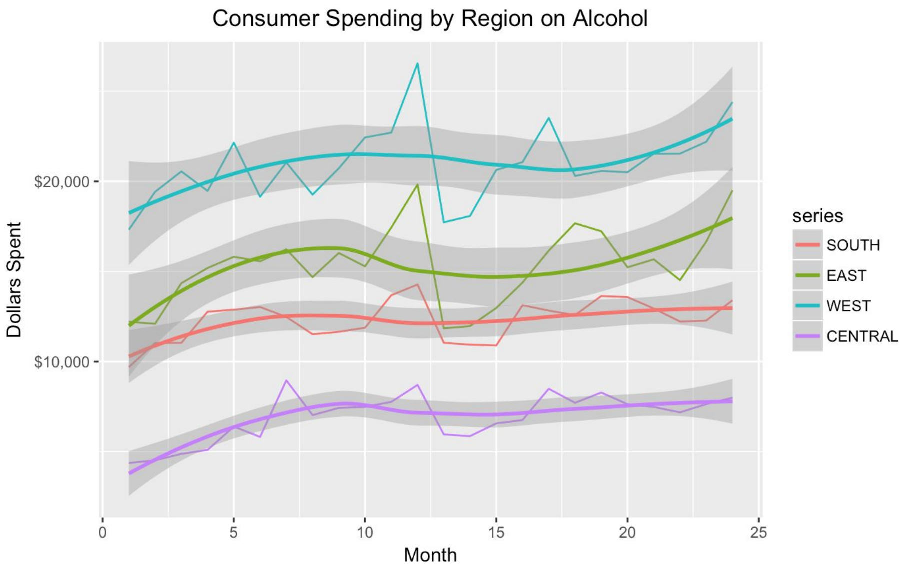
Splitting spending by Commodity next, we found several points of interest in this plot:



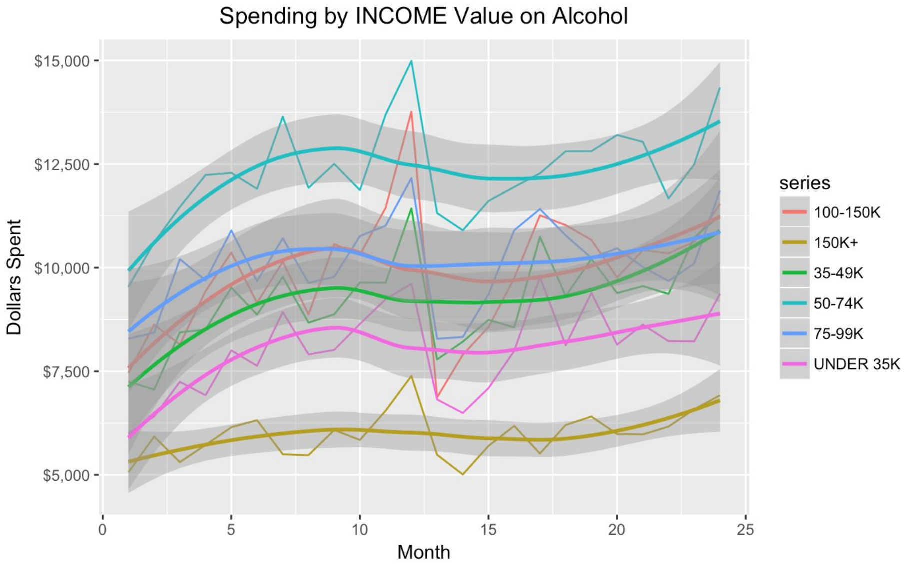
We see that grocery staple is a notable leader in sales among all other commodities. Produce comes in a distant second, and it is interesting that produce follows a cycle of increased sales during the warmer period of the year across the country, showing a notable dip from about October to March. This is due to the seasonality of produce's availability and price fluctuations. Grocery staple actually follows the opposite trend than that of produce, seeing increases in the span from October to March. The remaining *Commodity* sections appear to have few noticeable patterns in this plot and remain somewhat constant in sales throughout the year. However, there may be a general spike at the end of the year evident in other commodities.

Grocery staple, since it represented such a huge part of total sales, was also explored individually (See Appendix 5.3). Other niche trends of interest were also explored like floral, outdoor, and international food (See Appendix 5.4). These are all smaller spending categories with monthly sales by region never exceeding $5,000.

One niche sales market that we took particular interest in was for alcohol. It represented a larger amount of total sales than the three other niche markets detailed in the appendices. The West region had the highest amount of alcohol sales, followed by the East. Notable spikes included the holiday seasons in both years and also early summer (May and June) for the West and East. This is seen in the plot below.



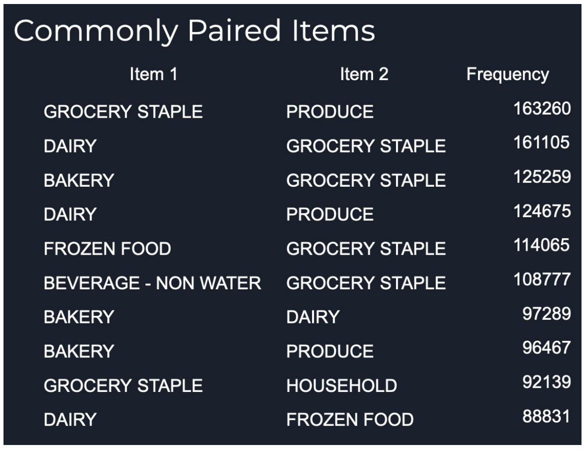
The largest subset of alcohol sales were to those earning a medium-level income at $50-74,000. The highest salaried customers did not constitute the most alcohol sales, but actually the least. Even the lowest earners (<$35,000) always spent more on alcohol. We believe this may be because the highest-earning customers may buy their alcohol at specialty stores, or for other reasons just don't purchase as much. The second highest subgroup ($100-150,000) actuallys buys a very similar dollar amount of alcohol as do those in the third highest subgroup.



All of these revelations from the plot hold business value, so we continued to explore alcohol sales in terms of different consumer variables (*HSHD_Composition, Age*) (See Appendix 5.5).

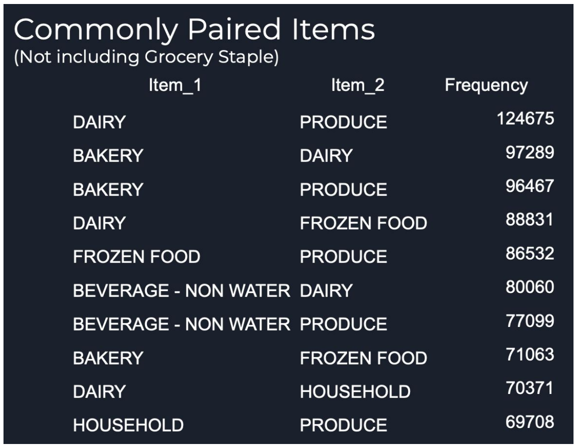
The basket analysis showed that the average number of items in a customer's basket stayed quite constant throughout the year, with the exception of the end of the second year in the data set. This is most definitely the result of a computational error, as it is not plausible that the jump in average number of items per basket would occur in the second and not the first year. For reference, the plot is included in Appendix 5.7.

The second basket analysis on item pairings is shown below:



Commonly Paired Items

Item 1	Item 2	Frequency
GROCERY STAPLE	PRODUCE	163260
DAIRY	GROCERY STAPLE	161105
BAKERY	GROCERY STAPLE	125259
DAIRY	PRODUCE	124675
FROZEN FOOD	GROCERY STAPLE	114065
BEVERAGE - NON WATER	GROCERY STAPLE	108777
BAKERY	DAIRY	97289
BAKERY	PRODUCE	96467
GROCERY STAPLE	HOUSEHOLD	92139
DAIRY	FROZEN FOOD	88831



Commonly Paired Items
(Not including Grocery Staple)

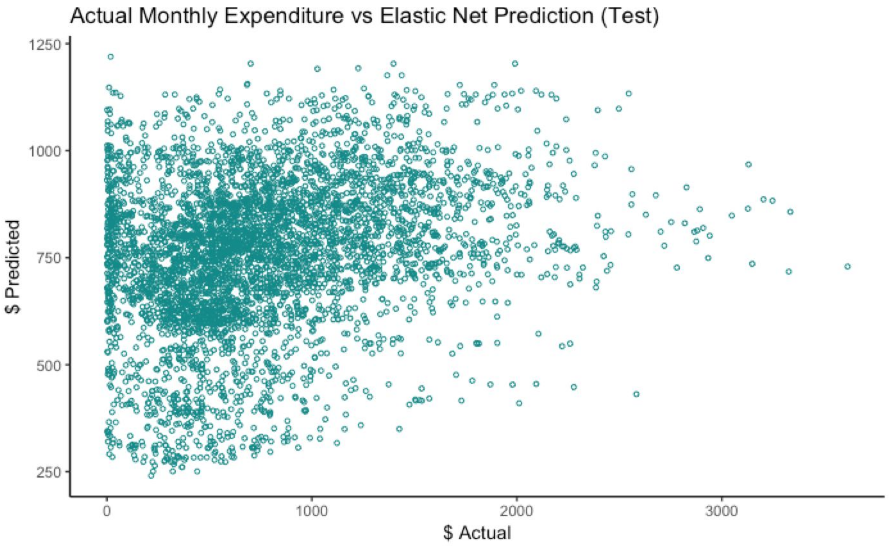
Item_1	Item_2	Frequency
DAIRY	PRODUCE	124675
BAKERY	DAIRY	97289
BAKERY	PRODUCE	96467
DAIRY	FROZEN FOOD	88831
FROZEN FOOD	PRODUCE	86532
BEVERAGE - NON WATER	DAIRY	80060
BEVERAGE - NON WATER	PRODUCE	77099
BAKERY	FROZEN FOOD	71063
DAIRY	HOUSEHOLD	70371
HOUSEHOLD	PRODUCE	69708

From these pairings, we see that "Grocery Staple" came in 5 of the top 6 baskets when paired with other items. Given that it is difficult to determine what "Grocery Staple" entails from the basket analysis alone, we also paired the items after removing entries with "Grocery Staple". Here, we see "Produce" and "Dairy" are the most commonly paired items, and down the list we see that common pairs usually contain one of these items. "Bakery" and "Frozen Food" also appear often on the list, followed by "Beverage - Non Water" and "Household".

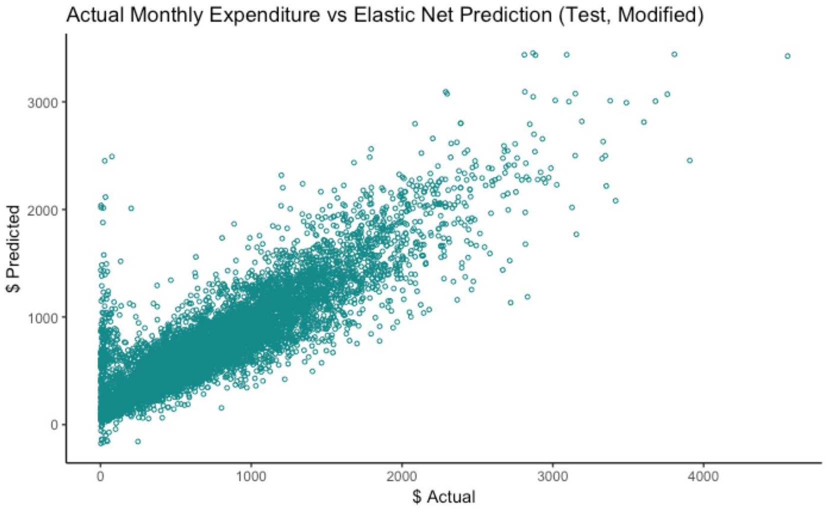
3.2 Modeling

Our models generally performed very weakly due to a number of limitations to the dataset and our increased focus on trend analysis once we determined that predictive modeling may not be the best solution at this time. While we were able to reduce prediction errors by utilizing more data cleaning and could reduce prediction errors by marginal amounts by splitting the data in different ways, such as focusing solely on certain groups of individuals, a complete model that accounts for all variables was infeasible at this time.

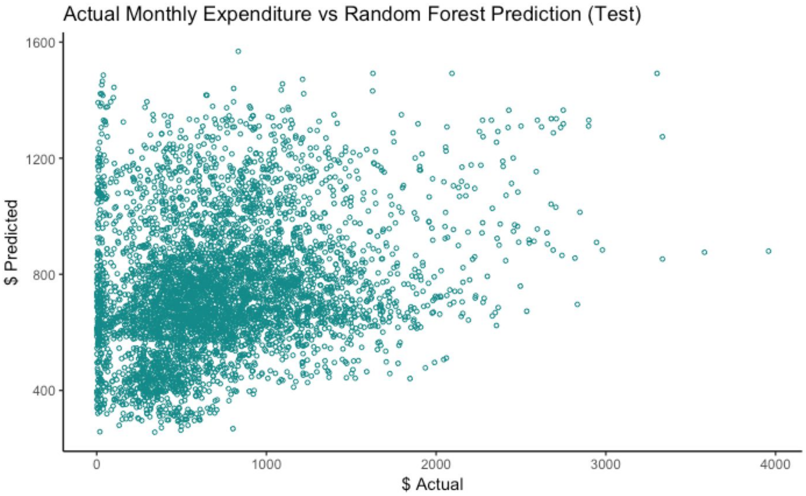
3.2.1 Regression



The results of running the Elastic Net algorithm show that the model was able to make predictions on household monthly expenditure within ~$468 of the actual monthly expenditure on average for the test set of households. While this number is very large (given that a large number of households spent between $500-$1500 in any given month from looking at the graph above) the model was still able to help us discern some useful features that contributed to the variation in the data. The full list of these variables is provided in the Appendix 5.6. We also attempted to model the data using each household as a specific predictor, and obtained the following graph of the actual monthly expenditure vs. the prediction:



The results of running the Elastic Net algorithms with each household being included as a specific predictor show that the model was able to make close predictions. These predictions on household monthly expenditure were within ~$260 of the actual monthly expenditure on average. Rather than making predictions on a test set of households, these predictions were on a sample of random months across different households. These results helped us confirm more of the meaningful variables that contributed to the variation in predictions on the data, which are also provided in full in the appendices.

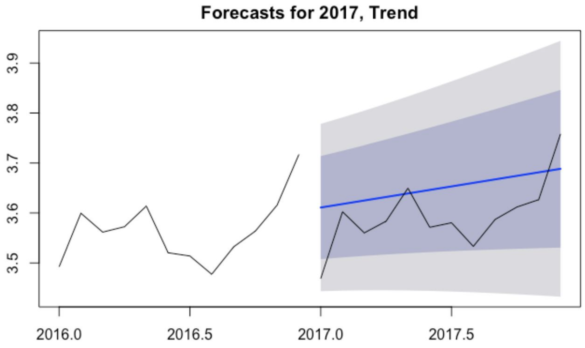


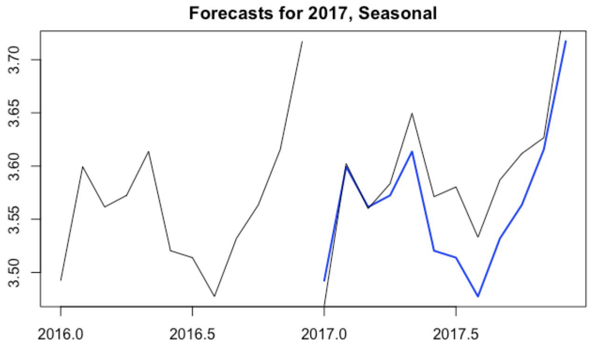
The results of running the Random Forest algorithm show that the model was able to make predictions on household monthly expenditure within ~$495 of the actual monthly expenditure on average for the test set of households. Again, this number is still very large compared to the median range of monthly spending per household, but this model was also used to determine which features played the greatest importance in making the model. The full list of these variables is provided in the appendices.

Looking at the trends of the first Elastic Net graph and the Random Forest graph, we see that predictions tend to stay within a certain range, regardless of actual amounts spent. The range of predictions also shifted based on the modeling method, where Random Forest made higher predictions than Elastic Net. In all 3 graphs, it is noticeable that when a household spent close to $0 on any given month, predictions would wildly vary between the ranges of predicted monthly expenditure. Using each household for fitting the data in the Elastic Net modified model showed a much stronger correlation between the model predictions and actual values, while the other models showed much weaker correlations between the model predictions and actual monthly expenditure.

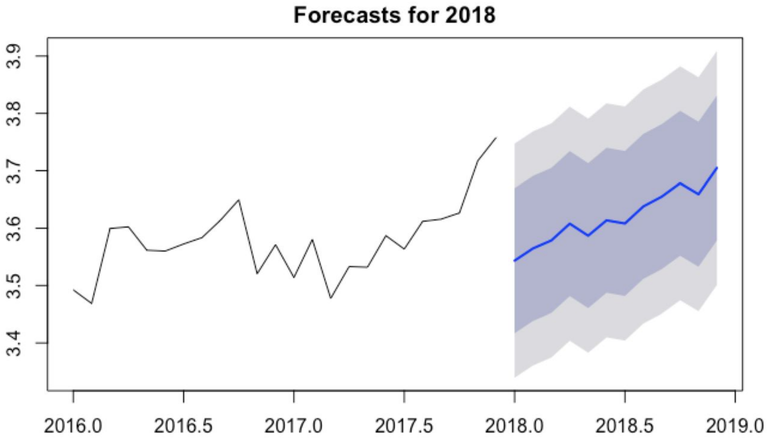
The variable importance methods for Elastic Net tended to list “Month” as being either the most significant variable or close to the most significant, which matched our expectations on the trends being time series dependent, but our Random Forest model designated “Month” as a poor predictor to base the splitting algorithm. “Region” was listed as a very significant factor to prediction for all models, which supports our expectations from some of the trend analysis where we saw different commodity and spending habits between the different regions. “Loyalty” proved to be an extremely significant factor for Random Forest and also appeared high on the list for the Elastic Net models, which suggests that some sort of focus on Loyalty membership may help improve sales (For another Loyalty membership plot, see Appendix 5.9). “Age Range” and “Income Range” were highly relevant for splitting in the Random Forest algorithm, but they appeared near the middle of the list for Elastic Net. It was surprising to find was that demographic factors related to the number of individuals in a given household (such as “Marital”, “Household Composition”, “Children”, and “Household Size”) did not seem to be very useful in making predictions, as they trended near the bottom of all feature importance lists. Our initial hypothesis was based on the thought that more individuals would mean more money spent per household, but this did not seem to match our expectations from the data. While we do think that looking into trends related to families for commodities and shopping might reveal some connection between certain products being more popular (such as baby-related items) it seems that focusing on items for for this type of demographic split may not yield much fruit.

3.2.2 Time Series





The results of the time series forecasting on the 2017 monthly sales data using 2016 monthly sales data is shown in the two graphs above, one using trend to forecast for 2017 and the other using seasonal relationship to forecast for 2017. The model predictions are in blue for both graphs. The trend graph shows a general increase over the course of 2017, which does seem to be somewhat true given the comparison between the actual 2017 data and predicted 2017 data from the seasonal relationship based model. However, the limitation here is that with a single year of data the prediction from a time series linear model will just be the change in mean over the previous year, and the season relationship based model will simple be a copy of the previous years time series. Neither of these results really capture the trend that is occurring year after year since it is biased to the single year data alone.



The results of the full data being used for training the linear model showed a significant drop in sales at the beginning of 2018 that would gradually increase over the year. This result may seem like a general trend of the two years used for training, but this is the result of time series modeling done on a non-stationary time series. For a time series to be used for forecasting, a few constraints limit what kind of data can be used for accurate predictions, one of them being the need for a stationary time series. A stationary time series is a time series of data that has approximately constant mean and variance over the time in the data. From our training data, we can see that the mean and variance are shifting quite a bit between the successive years, and the trends seem very dissimilar. Because of this, an adjusted time series needed to be made, which can be found in Appendix 5.8.

4. Conclusion and Discussion

4.1 Conclusions

In general, our project objectives were best explored through trend analysis. Modeling conclusions were found, but may not be the most appropriate for this data set.

4.1.1 Trend Conclusions

From our trend analysis, we have several recommendations based on geographic findings. The East region by far brings in the most sales overall, so 84.51° should know that any experimental business decisions made in that region will have greater returns or losses. That said, it may be the best target for specific campaigns and product promotions to have the largest impact on sales. Similarly, the Food department (vs. Non-food and Pharma) dominates sales in the same manner. Conversely, sales or marketing campaigns designed to bolster overall sales may be appropriate for the South region, the lowest spending region.

Grocery staple dominates all other commodities, so once again, business practices targeted toward this commodity will see the largest losses or returns. The seasonality of this commodity presents opportunity to bolster sales during its “on” season, and focus on them less during its “off” season. Every commodity, in general, shows increased sales during the end of year holiday season. 84.51° undoubtedly is aware of the importance of this season to overall spending, and we would like to reinforce this fact.

Floral departments should focus on marketing to customers during the periods leading up to Valentine’s and Mother’s Day, paying particular attention to the latter. In a similar fashion, outdoor items should be marketed accordingly and placed appropriately in the store layout during it’s warm season selling window.

Due to international food's unproportional popularity in the West region, Kroger needs to market promotional materials for those items in that area. This is presumably related to the large and diverse customer bases along the west coast of the country. As mentioned, due to the smaller market share of these items, Kroger should expect slimmer returns for varying business practices for these families of items.

Our alcohol analysis led us to several notable recommendations. The obvious spikes in general alcohol sales around Christmas and New Year's eve should continually be focuses of marketing campaigns. However, during other parts of the year, patterns aren't as discernible. There were alternating spike during summer 2017 across regions, which could have numerous explanations, such as weather patterns or sales in different regions.

An interesting trend was also found between private brand labels and national brand labels. Within the products dataset, there was 6 times as many national labels as there were private labels, but national label sales only accounted for a little more than 2 times as many sales. While the manufacturing costs should be assessed before making any transitions, we suggest that Kroger look further into expanding its private label offerings to increase sales. A graph of the supporting information can be found in Appendix 5.10.

As noted before, since the West region beats out the East in alcohol sales, despite the East being the overall sales leader, alcohol marketing and product placement strategies in western stores should be adjusted to exploit this fact. In terms of demographics, those customers with mid-range salaries spend the most on alcohol, while those with the smallest salaries seem to buy a fair amount of alcohol in proportion to their income level. Marketing toward these groups is recommended. While we did not show graphics related to homeownership, we concluded that those who rent buy about 10 times less alcohol than homeowners. Finally, it is notable that the largest buyers of alcohol include those households of two adults with at least 2 children. They beat out the next highest group (2 adults) by a considerable amount. Also, middle-aged customers (45-64) spend the most on alcohol when compared to older and younger customers. This is an interesting finding that we found in contrast to our own intuition. Placing parenting products closer to alcohol sections of stores may allow for this demographic group to buy more of both item types.

We conclude that targeted marketing, product placement, in-store product arrangement, sales, and other business techniques carefully targeted with the above recommendations in mind will prove to be effective in impacting sales at Kroger stores.

4.1.2 Modeling Conclusions

From our regression modeling, the key takeaways include a focus on the different regions for different types of sales, as heavily compounded by the results from the trend analysis. While region has already been a focus for Kroger to market specific items, a change in store layout that may promote certain areas to shop that provide more revenue may help with increasing sales. Another takeaway is to focus on loyalty membership and how sales from loyalty members differ from non-loyalty members, and what membership looks like for individuals who are not already loyalty card owners. A focus on targeting customers within older age ranges may also prove useful as spending trends could be associated greatly with them. Finally, a focus on lower income and higher income type foods may also be beneficial based on the variable importance of income ranges, as the demographics split on income showed very different levels of spending.

For the time series modeling, even by adjusting the data for what looked like a better fitting, time series analysis is not recommended for this dataset. Due to the time series of the given data having very erratic trends, this makes prediction very generalized and indiscriminate of the different factors that might have affected the large fluctuations. Time series is usually recommended to have at least 10 years worth of data for training, and a smaller set of years to compare short term predictions. With non-stationary time series, even more is recommended to be used for training to capture a large portion of the variance and covariance over time. When very little data is used for modeling, the predictions are biased towards averages or trends of very little data, which most likely would not reflect the general trends that occur over greater periods of time.

4.2 Discussion on Further Analysis

We found that when looking forward to additional analyses, several alterations should be considered, especially in regards to predictive analysis.

- The category of grocery staple is large, and an important set of products to Kroger due to the high volume of sales. It would be beneficial to decompose grocery staple into several subcategories to further explore what sections of it behave differently in terms of sales.

- A continuous variable for salary categories would be useful in analysis, but it may not be possible to get this specific information on consumers at this time.

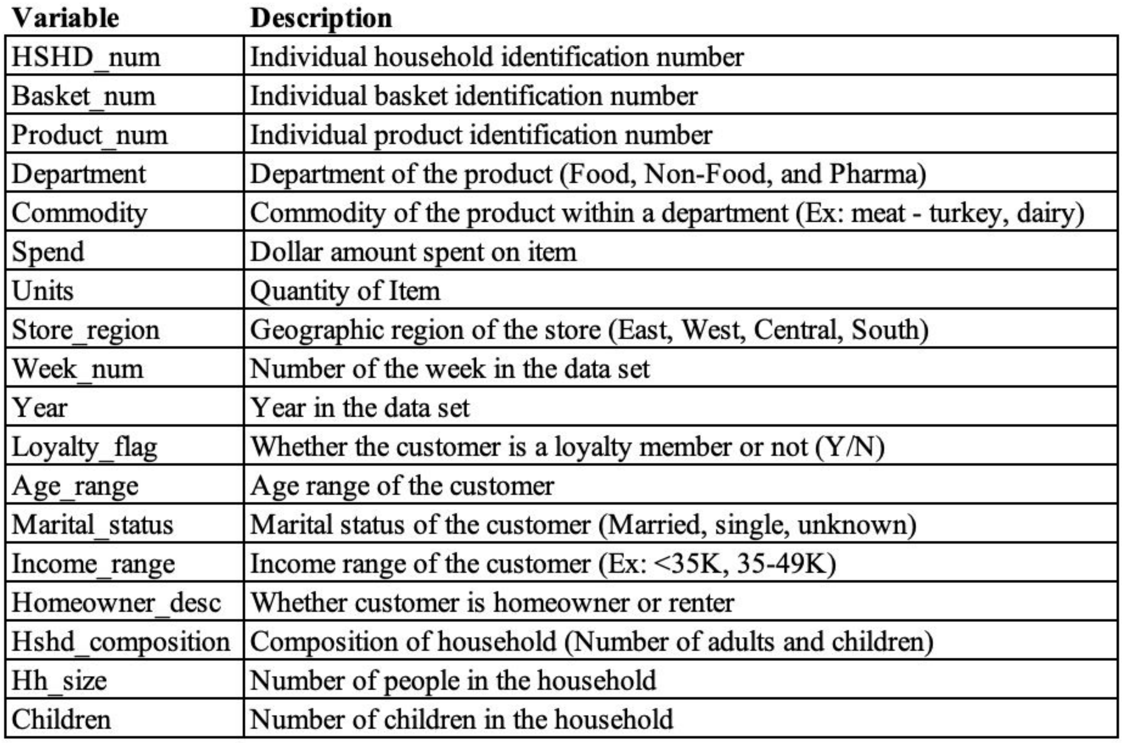
- When exploring trends, we found that it would be useful to know of current business strategies in Kroger stores. This way the analysis could also be used to measure the performance of sales, promotions, and other customer incentives.

- As discussed in our modeling conclusions, time series forecasting is not the most appropriate for this data set for several reasons, one of them being that only two years are covered. Additional years would help increase accuracy of these modeling techniques.

- Lastly, individual store data may give a more detailed picture of how customer spending trends are behaving, as well as the effectiveness of current business practices.

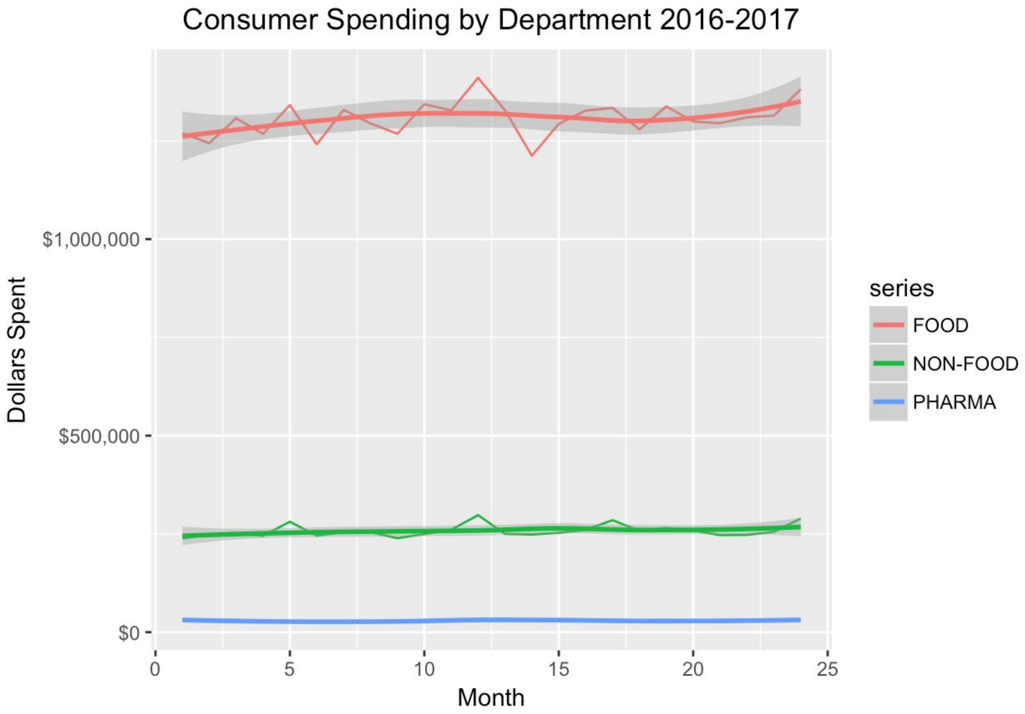
5. Appendices

5.1 Variables and Descriptions



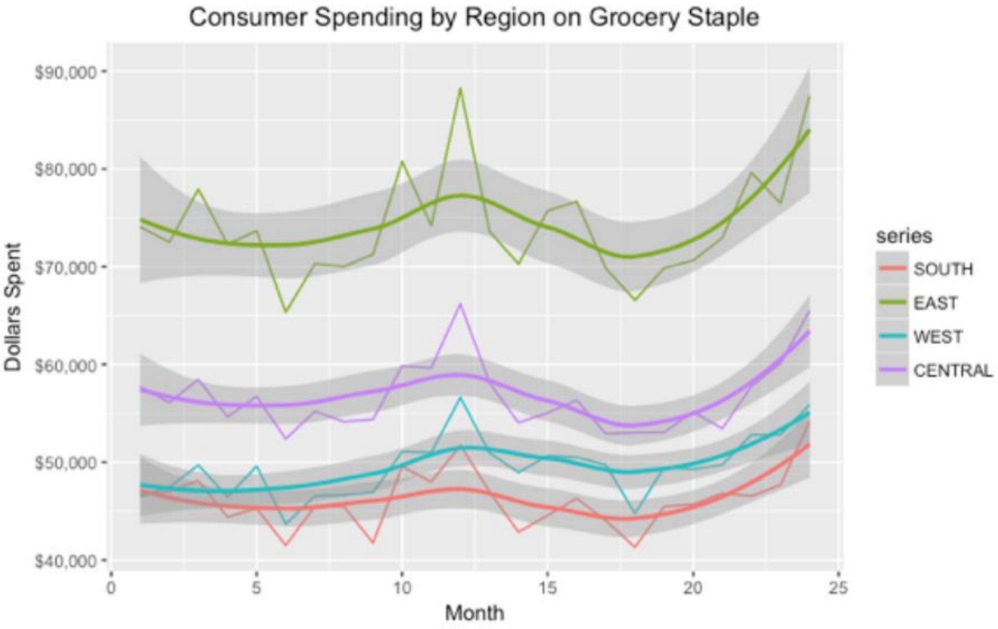
Variable	Description
HSHD_num	Individual household identification number
Basket_num	Individual basket identification number
Product_num	Individual product identification number
Department	Department of the product (Food, Non-Food, and Pharma)
Commodity	Commodity of the product within a department (Ex: meat - turkey, dairy)
Spend	Dollar amount spent on item
Units	Quantity of Item
Store_region	Geographic region of the store (East, West, Central, South)
Week_num	Number of the week in the data set
Year	Year in the data set
Loyalty_flag	Whether the customer is a loyalty member or not (Y/N)
Age_range	Age range of the customer
Marital_status	Marital status of the customer (Married, single, unknown)
Income_range	Income range of the customer (Ex: <35K, 35-49K)
Homeowner_desc	Whether customer is homeowner or renter
Hshd_composition	Composition of household (Number of adults and children)
Hh_size	Number of people in the household
Children	Number of children in the household

5.2 Dropping "Pharma"



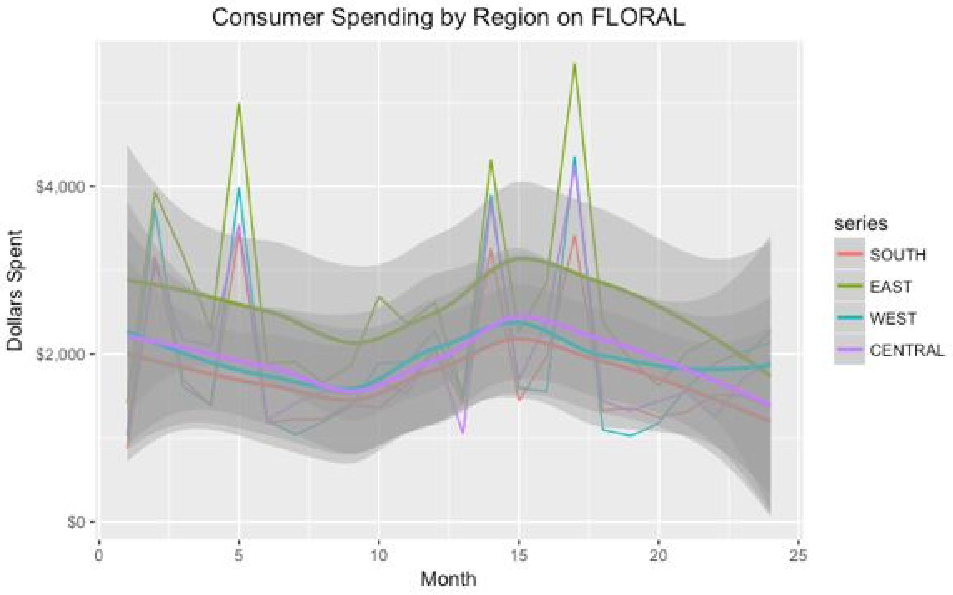
5.3 Grocery Staple

Grocery staple, because it has so many sales, was then explored by region, seen below. The East region of course is the largest buyer, and Christmas and Thanksgiving spikes are very evident. Also, we see here a spike in spending in March of both years that is of dubious origin, followed by a uniform dip around June. We attribute the June dip to consumer activity decreasing following arrangement changes following the academic year or cool-weather season. People also may be more likely to eat out than buy groceries to prepare when the weather is nicer after May.

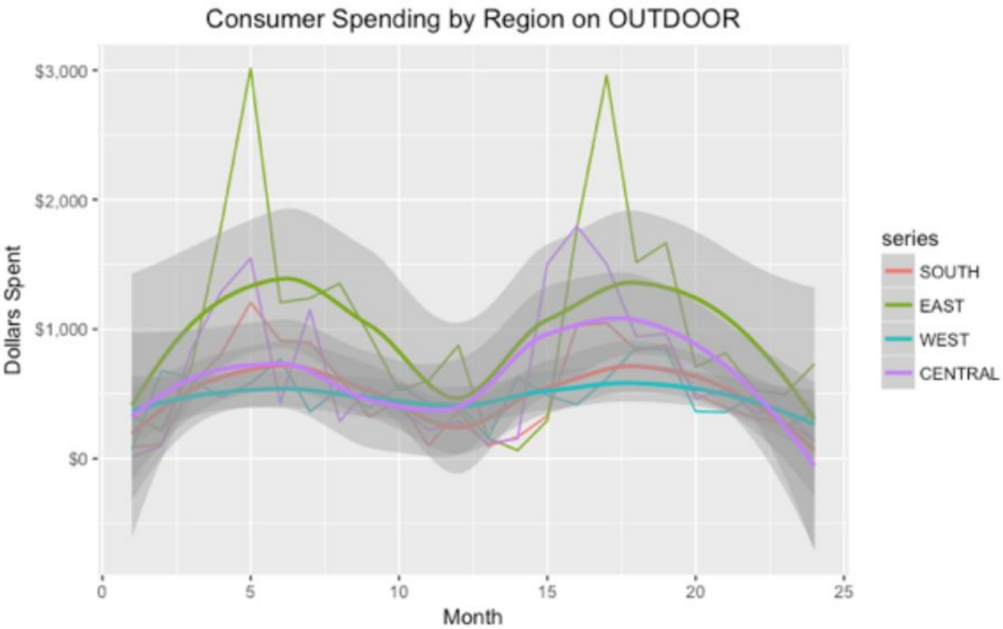


5.4 Niche Trends

Diving into some more niche spending categories, we explored spending on floral, outdoor, and international food commodities over time. Floral, as expected, rises in February (Valentine's Day) and May (Mother's Day), but the May spike is actually noticeably higher than the February jump for both years in the data set. It is important to note the how small the floral sales are compared to sales across all commodities.

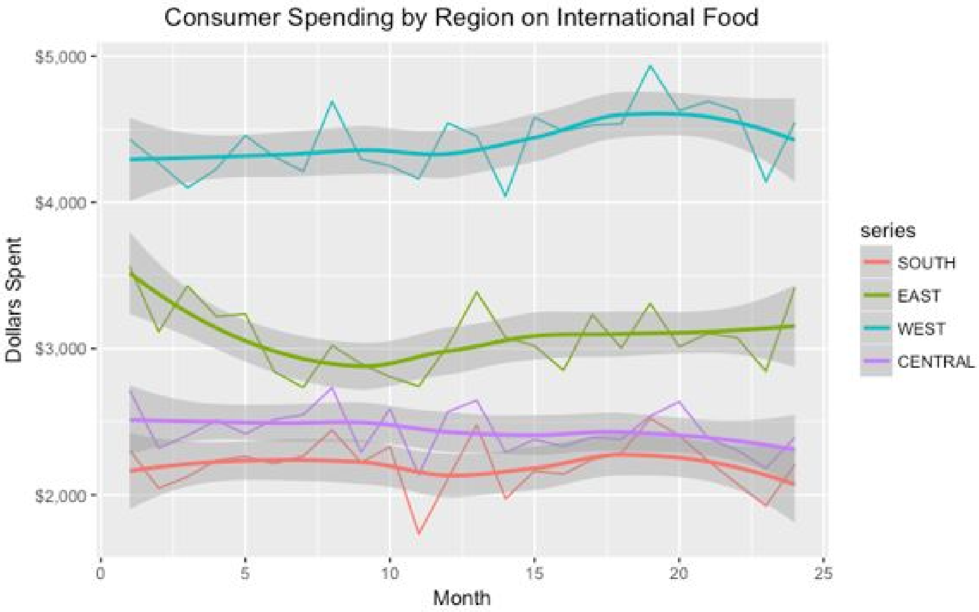


An even smaller niche market, the outdoor commodity, was also explored. It's maximum monthly sales for any region were only $3,000 per month, but it's patterns are highly predictable as they fluctuate with the warm seasons. It is notable that the Central region bought noticeably more outdoor products in the second year of the data set for unclear reasons.

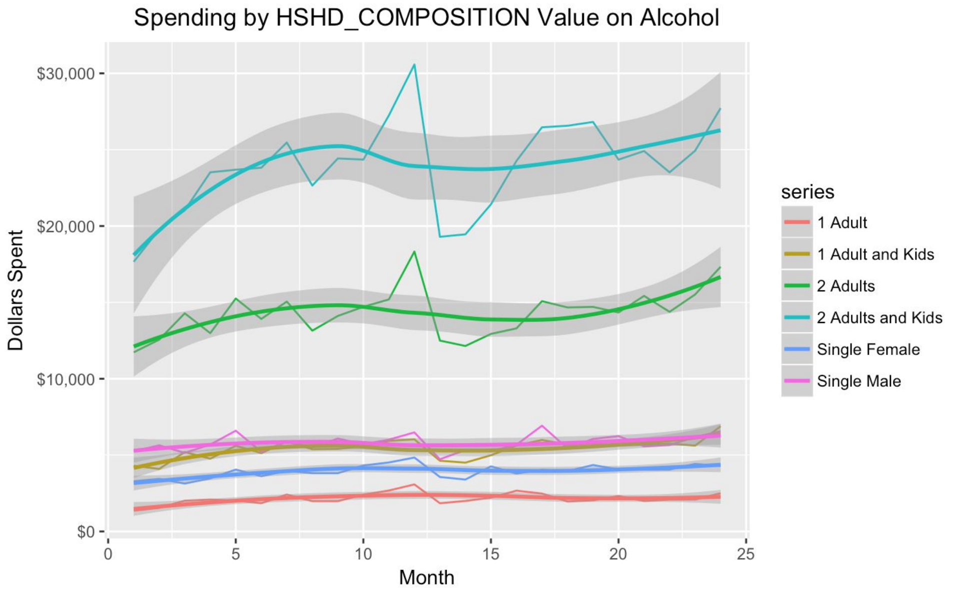


Lastly, the international food commodity patterns were additionally analyzed. Contrary to overall spending, spending on international foods was dominated by the

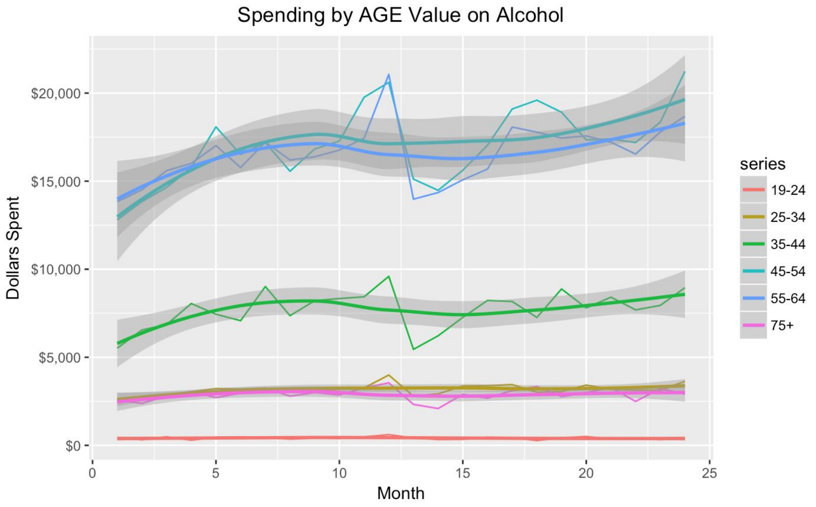
West region. It is interesting to note that the spending in the West on these food items appears to be gradually increasing, while it appears to remain constant in the other three regions.



5.5 Alcohol Trends



In the plot above, household composition in relation to spending on alcohol was explored. Notable aspects of this plot include the fact that most spending comes from the nuclear households containing 2 adults and kids. 2 adult households spend the next highest amount, and the rest are quite comparable at a lower amount. In the below plot, we see that the middle-aged customers (45-64) buy the most alcohol by sales, while counterintuitively the younger drinkers spend the very least on it.



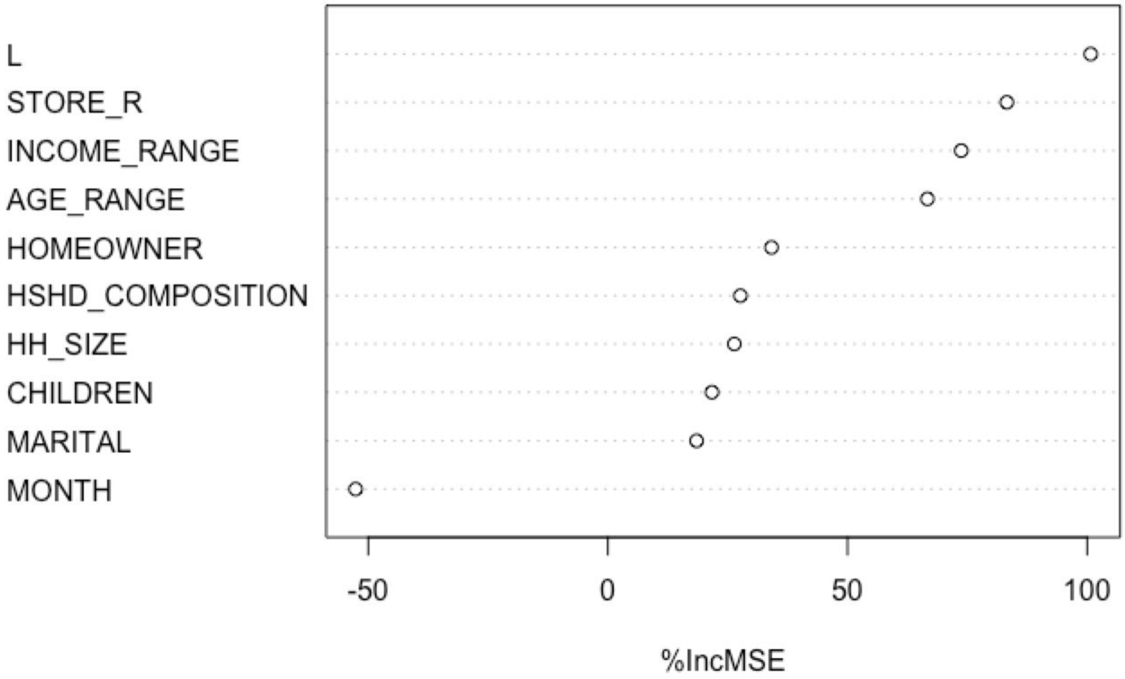
5.6 Variable Importance Lists

Elastic Net
"MONTH.Dec"
"STORE_R_CENTRAL"
"STORE_R_EAST"
"STORE_R_SOUTH"
"L_Y"
"AGE_RANGE35.44"
"AGE_RANGE45.54"
"AGE_RANGE65.74"
"AGE_RANGE75."
"INCOME_RANGE100.150K"
"INCOME_RANGE150K."
"INCOME_RANGE35.49K"
"INCOME_RANGEUNDER.35K"
"HSHD_COMPOSITION_2.Adults.and.Kids"
"HH_SIZE1"
"HH_SIZE5."
"CHILDREN2"

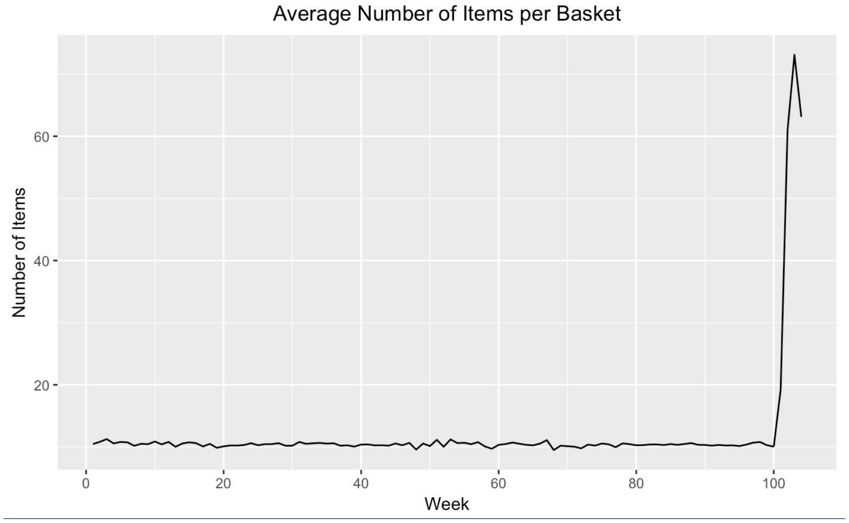
Elastic Net, Modified
"MONTH.Jan"
"MONTH.Feb"
"MONTH.Apr"
"MONTH.May"
"MONTH.Jun"
"MONTH.Jul"
"MONTH.Sep"
"MONTH.Oct"
"MONTH.Nov"
"MONTH.Dec"
"STORE_R_CENTRAL"
"STORE_R_EAST"
"STORE_R_SOUTH"
"L_Y"
"AGE_RANGE35.44"
"AGE_RANGE45.54"
"AGE_RANGE65.74"
"AGE_RANGE75."
"MARITAL_Married"
"INCOME_RANGE100.150K"
"INCOME_RANGE35.49K"
"INCOME_RANGEUNDER.35K"
"HOMEOWNER_Renter"
"HSHD_COMPOSITIONSingle.Female"
"HH_SIZE1"
"HH_SIZE4"
"HH_SIZE5."
"CHILDREN2"
"CHILDREN3."

Random Forest	IncMSE
L	100.7401
STORE_R	83.23505
INCOME_RANGE	73.7213
AGE_RANGE	66.67058
HOMEOWNER	34.14242
HSHD_COMPOSITION	27.65312
HH_SIZE	26.37984
CHILDREN	21.72236
MARITAL	18.51345

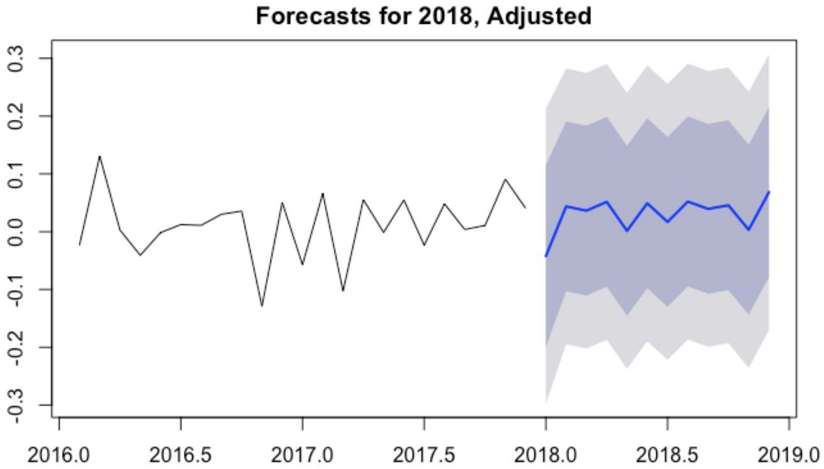
The first list corresponds to the variable importances of the Elastic Net model in descending order of importance, while the second list corresponds to the variable importances of the modified Elastic Net model in descending order of importance. The final list corresponds to the variable importances of the Random Forest model, listed with the change in mean squared error due to each variable (IncMSE). What this signifies is how much the mean squared error of the model would change if the variable in question was randomly shuffled, and the greater the change the more significant the variable. A negative value (namely “Month”) would signify that a random variable could perform better than using the variable in question for splitting the trees of the forest. We believe this may be due to the Random Forest eliminating the time significance of the model and instead attributing general trends to the demographic factors. Since the time trends seemed to be relatively similar across any demographic splits, the model most likely assumed this to be less significant than finding other appropriate splits. This data for Random Forest variable importances can also be plotted as below:



5.7 Items per Basket

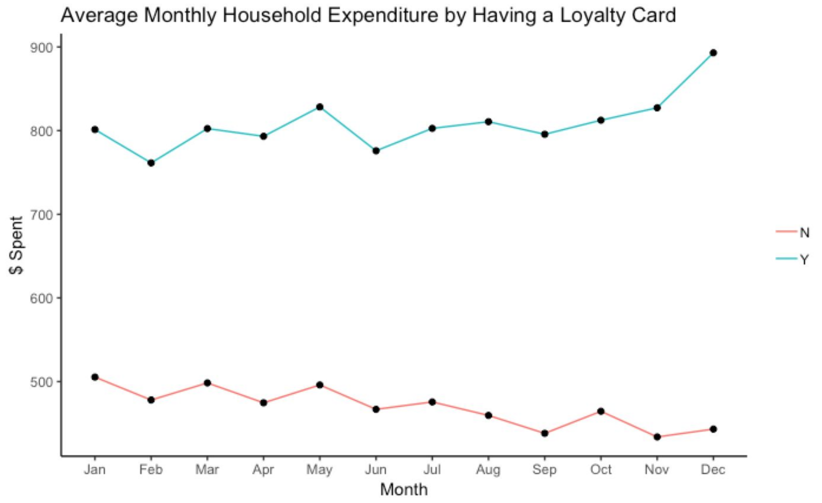


5.8 Time Series Forecasting

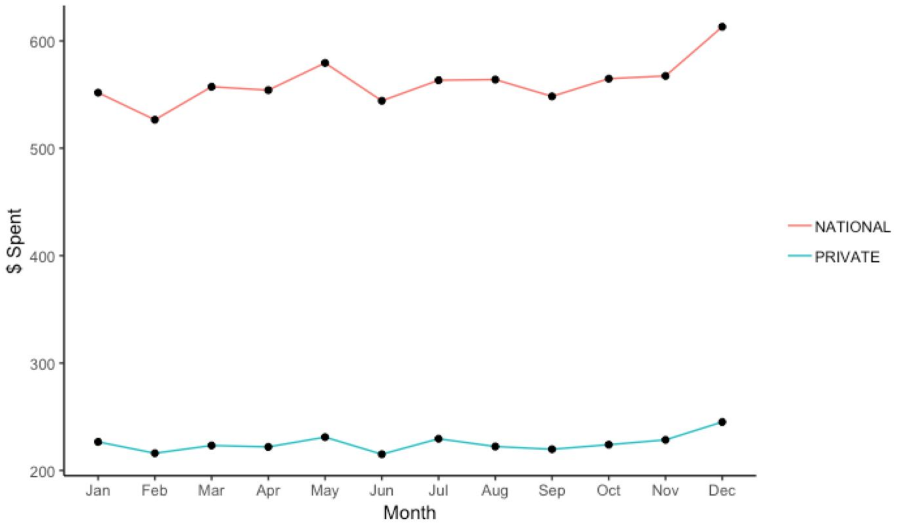


The adjusted model actually seems to be approaching a more standard time series by using the mean difference across the years to approximate a stationary series from the original time series. While this does suggest that more data may help with time series analysis, there is still not enough data to draw reasonable conclusions.

5.9 Loyalty Membership Trends



5.10 Brand Loyalty Trends



5.11 System Specifications

Our methods were obtained by using the R programming language and RStudio. These tasks were run on MacBook Pros, Early 2015 model, with a 2.7 GHz Intel Core i5 and 8GB 1867 MHz DDR3 RAM. Some limitations to our analysis could be in part related to our system, but the analyses that were performed should be reproducible on machines that match these specifications.

5.12 Contributions

Sahil Kumar

Sections 2.1, 2.2*, 2.3, 3.1*, 3.2, 4.1.1**, 4.1.2, 5.6, 5.8, 5.9, 5.10
*Basket Pairing analysis
**Brand Label conclusions

Jonah Somers

Sections 1.1,1.2, 2.2, 3.1, 4.1.1, 4.2, 5.1, 5.2, 5.3, 5.4, 5.5, 5.7

6. References

The Complete Journey 2.0. (www.8451.com/area51)

7. Code

Please see additional files:

- clean_model_trend_pairs.Rmd
- geography_and_trends.Rmd
- items_per_basket.Rmd