Regional Analysis of a Supermarket

Ethan Ericson, Nate Moore, Jacob Romine, Jason Van Wieren

University of Arkansas

DASC 1223/H – Data Science in Today's World

Dr. Schubert

2/28/2024

Introduction

Executive Overview

Goal

The following study analyzes a collection of categorical factors to predict the product performance, measured by total profit, of goods sold at a fictitious supermarket. By establishing a connection between these factors and product performance, prescriptive action can be taken by the supermarket to improve the performance of its products; thus, it is important for these correlations to be explored. To begin our analysis, we declared two hypotheses:

1. There exists a correlation between product performance (measured by profit), the consumer segment, and product category within the superstore.
2. Product returns are correlated with the Region, State, Segment, Category and can be used to predict product performance in a specific location by pinpointing areas with high or low returns.

Methods

To perform our study, we first sourced a fictitious superstore dataset, "(US) Sample – Superstore", from Tableau. This data set contained product information concerning the profit of the product sale, the location the product was sold in, the segment of the purchaser of the product, the category of the product, and the returned status of the product from 2019 to 2022 (Martin 2023). The returned status was listed on a separate sheet of the dataset, so we cleaned and left-joined this sheet to the rest of the dataset using R. The resulting dataset was determined to be clean by our team, so we began our feature engineering. The first feature we constructed was Total Profit, which described the total profit of all orders of a certain category, grouped by consumer segment. Furthermore, to aid in the

implementation of a machine learning model, we utilized target guided ordinal encoding to factorize our categorical predictor variables, adding the factored version of these variables onto the dataset as Segment Rank, Category Rank, and Returned Rank. With our dataset now cleaned and our relevant features engineered, we created a summary of our dataset on a national level, followed by four regional analyses divided between the four members of our team. Within our regional analysis, we utilized bar charts and mosaic plots to graph our results, as well as performing Chi Squared analysis on the categorical variables we analyzed. As the final facet of our data exploration, we trained a machine learning model on our dataset to predict product performance through profit by analyzing the different combinations of categorical variables we defined in our hypotheses.

Results

Our study found that there did exist some level of correlation between Segment, Category, and Product Performance, as our first hypothesis had stipulated. Additionally, returns were also shown to be influenced by a combination of Region, State, and Segment but not Category, signifying that our second hypothesis was only partially justified. In respect to Hypothesis 1, we ascertained that the furniture category of product was consistently poor preforming across all segments, within all regions, while office supplies and technology categories were usually better preforming than furniture at a minimum. This demonstrates that a consistent correlation exists between the category of a product, its segment, and its total profit/product performance. The business can leverage this association by shifting resources away from categories and segments that do not preform well to ones that do, or by dedicating more resources towards improving areas of weakness identified by the interaction of segment, category, and total profit. In our endeavors to validate Hypothesis 2, we discovered that returns were correlated with region, state, and city. This finding connects an indicator of product performance - whether an item was returned - to locational data. This association is important to the supermarket as it will give them the ability to pinpoint areas where there are excessive amounts of

returns and allow them to work to understand and resolve the higher-than-expected level of returns in that area. Despite the success of our exploratory data analysis in validating our hypotheses, our machine learning model failed to predict total profit accurately based on predictor variables of Segment, Category, and Returned. The model, if given a combination of these variables, would output a nearly constant profit for that specific combination, leading to a highly inaccurate model. Because of the inaccuracy of this model, its predictions were assumed to be invalid in our analysis, but we included them in our report to document our processes.

Literature Review

Retail Industry

The fictitious supermarket we performed our predictive sales analytics on is a part of the retail industry. The retail industry is characterized by "all companies that sell goods and services to consumers" (Assoia, 2022), and therefore includes our supermarket, which sells furniture, office supply, and technological goods. Year by year, retail outlets must reconsider their sales practices to stay afloat in a highly competitive industry, as a company that sticks to long standing practices is more likely to be beaten out by a more flexible competitor that has taken advantage of trends in the market and has a high degree of awareness surrounding its own operations. The rise of technology in facilitating consumer spending at retail outlets has been a major driver of the need for retail businesses to reevaluate their sales practices. Omnichannel retailing addresses consumer's demand for "first-rate customer service and an integrated shopping experience" by expressing the importance of offering online shopping in addition to in-store activities, allowing a retail outlet to reach their consumers along multiple channels. (Assoia, 2022).

A Similar Study

A similar study regarding supermarket sales was conducted in 2023 prior to our analysis. This study, titled “Predictive Analytics for Grocery Sales Forecasting: A Case Study of Favorita Stores” by Rekia Iddrisu Ouedraogo, used data from La Favorita Supermarket to predict future sales using regression analysis. The study also analyzed the effect of an earthquake on the sales of this supermarket, testing the null hypothesis that the earthquake had a significant impact on sales against the alternative hypothesis that the earthquake had no significant impact on sales. The results of the study found that the Favorita Corporation targeted specific cities with its supermarkets to encourage purchases, experienced variable spikes in sales, sometimes correlating with paychecks, sometimes with holidays, and sometimes with promotions. Reasonable evidence to suggest an earthquake disrupted sales was also discovered. The study concluded with the construction of a machine learning model to predict sales using three different methods. The first method, Random Forest, had an R squared value of 0.391, the second method, Linear Regression, had an R squared value of 0.22, and the third method, ARIMA (Moving Averages), had a R squared value of 0.381 (Ouedraogo, 2023). Our team found it interesting that this study also failed to construct a machine learning model that preformed with high accuracy, potentially indicating to us that supermarket sales across a variety of products are not well predicted by the methods used in both of our studies.

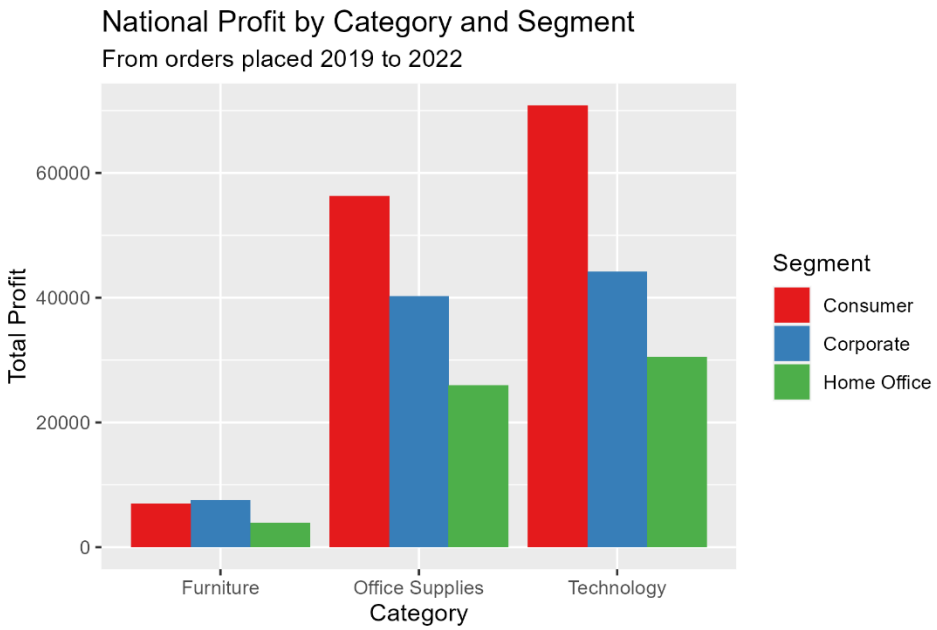
Data Analysis

National Analysis

To achieve an overall summary of our hypotheses’ related data, we graphed the national total profit of product categories, grouped by segment, as well as the total level of returns by region.

Figure 1

National Profit by Category and Segment



Source: US Superstore - Sample

Note. This figure depicts the relationship between product Category, Segment, and Total Profit.

Figure 1 illustrates the large disparity between the total profit of furniture sales and the total profits of office supplies and technology. Furniture does not perform well in comparison to these two categories and is also poor preforming across all consumer segments. Within the office supplies and technology categories, consumers are the most profitable segment to sell to, corporations are the second most profitable, and home offices are the least profitable. On a national level, total profit appears to be correlated with category as well as segment, as furniture has a much poorer performance level than office supplies and technology, and the certain segments sell much better than others in the office supplies and technology categories.

Following our national analysis of total profit, we graphed the total number of returns national in order to pinpoint areas of focus for our analysis.

Figure 2

National Returns by Region

Source: US Superstore - Sample

Note. This bar chart illustrates the total frequency of returns in the four regions of the U.S.

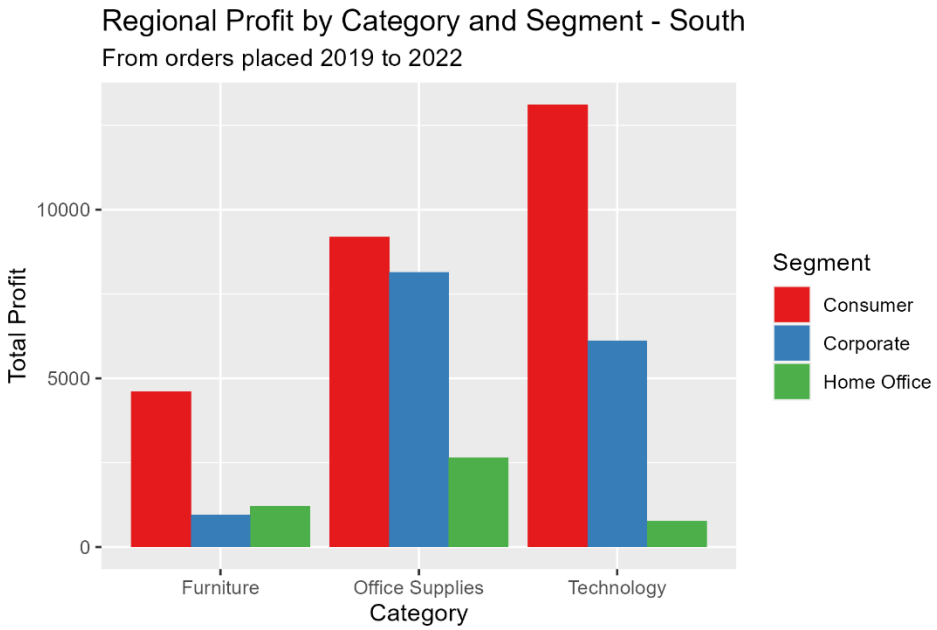
The results of Figure 2 highlight a few major areas of focus. First, the western region of the United States has a high number of returns. This high level of returns is a problem for the supermarket, as it means a sizable number of their products are not meeting consumer expectations in this area. Second, the South and Central regions of the U.S. have a low number of returns in comparison to the rest of the regions. This observation is significant to the company, as it potentially indicates products in these regions meet consumer expectations better than in the East and West regions of the U.S.

Southern Analysis

Ethan Ericson conducted his own individual analysis of the southern region of the United States, analyzing the relationship between Total Profit, Category, and Segment as well as investigating areas with higher or lower than expected levels of returns.

Figure 3

Regional Profit by Category and Segment – South



Source: US Superstore - Sample

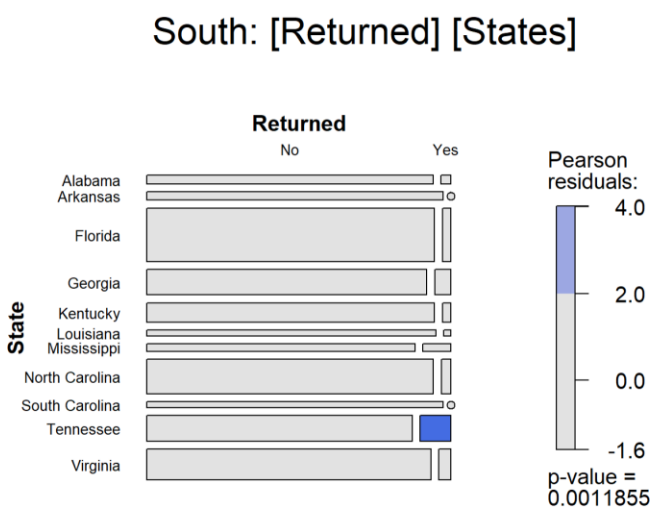
Note. This bar chart represents the relationship between southern Category, Segment, and Total Profit.

As in our national analysis, furniture is once again a low performer in respect to the other categories. This observation reinforces our first hypothesis that there exists a correlation between Total Profit, Category, and Segment within the supermarket. Additionally, the consumer segment continues to be the dominating force in product profits across office supplies and technology. Contrary to our national analysis, the consumer segment in the furniture category accounts for the majority of total profit within the furniture category. This indicates that the consumer segment within the South has a higher demand for furniture than the other segments, meaning that the supermarket should focus marketing of furniture to this segment.

To investigate the returns within each region and the relationship between returned statues and location, Ethan constructed mosaic plots and used Chi Squared statistical testing.

Figure 4

South: Returned by States



Source: US Superstore - Sample

Note. This mosaic plot represents the residuals of returns within each southern state.

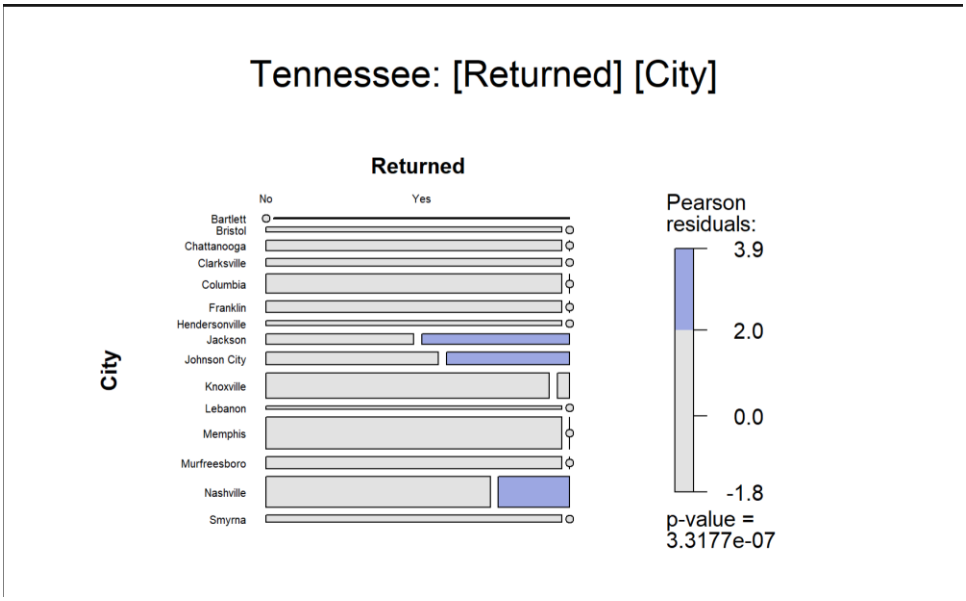
The mosaic in Figure 4 displays several relevant pieces of information. Looking at the mosaic bars, we can see that they are different widths. These different widths represent the total number of products ordered within those states. Next, the vertical dividing lines on the bars indicate the proportion of kept products to returned products. Residuals were calculated for the collection of states based on expected values drawn from a random distribution. Blue shading represents a high positive residual, and red represents a high negative, while grey represents a residual within two units of zero. Finally, a p-value is depicted in the mosaic graph, symbolizing the result of the Chi Squared test of returns and States. A p-value under 0.05 represents a statistically significant relationship between the two variables, meaning

that Returned and State are associated on a level not accounted for by random chance. Figure 4 indicates with a small p-value that States and Returns are correlated, supporting our second hypothesis, and that Tennessee, the state that possessed the largest residuals, has a higher-than-expected level of returns. These results can convey important information regarding product performance to the supermarket, specifically that people in Tennessee are not satisfied with their products and are returning them.

Ethan dove further into the relationship between location and returns by creating mosaic plots for the cities within Tennessee, and the segments within Nashville.

Figure 5

Tennessee: Returned by City



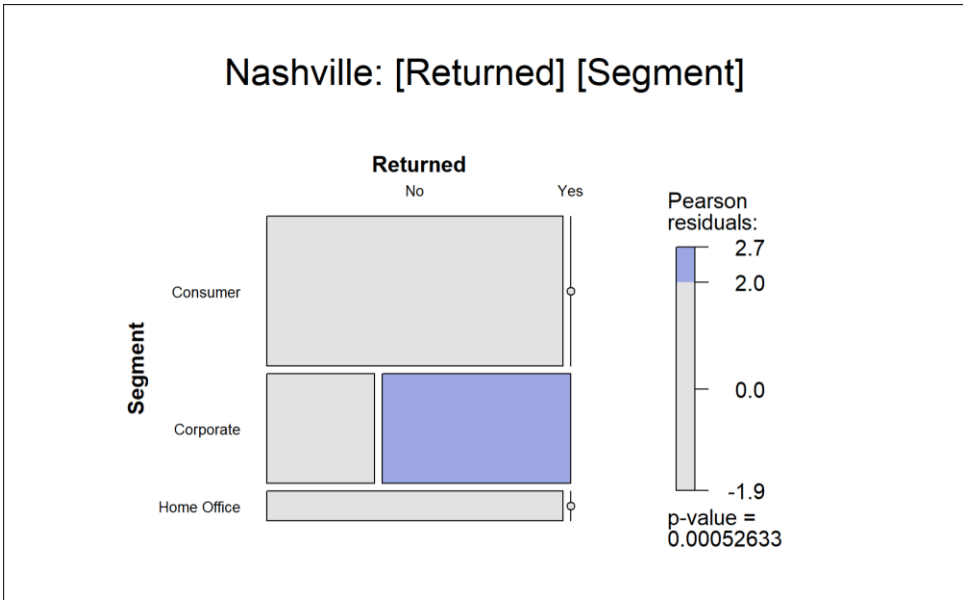
Source: US Superstore - Sample

Note. This mosaic plot represents the residuals of returns within each city in Tennessee.

Within Figure 5, several cities contain high residuals, the most prominent of these being Nashville. The p-value for the association of City and Returned is also very small, indicating a significant association between these two variables, again supporting our second hypothesis.

Figure 6

Nashville: Returned by Segment



Source: US Superstore - Sample

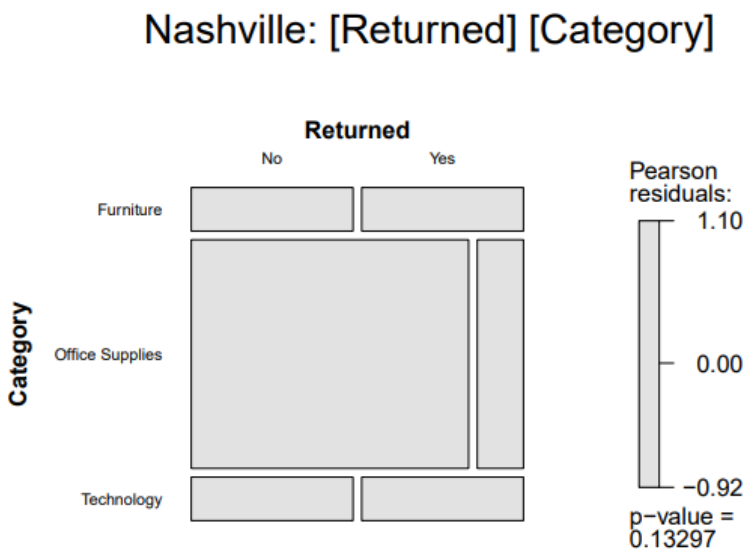
Note. This mosaic represents the residuals of returns within the consumer segments of Nashville.

Figure 6 demonstrates that the consumer segment is responsible for most purchases within Nashville, and that there is a significant association between Segment and Returned, as indicated by the small p-value. The consumer segment does not have a high residual for positive returns, indicating that the supermarket is providing products to consumers that are meeting their expectations. On the other hand, the corporate segment is returning products more often than expected. Seeing as they make up the second highest number of purchases within Nashville, this is a worrying indicator of product performance within Nashville and potentially explains why Tennessee has higher than expected returns.

To complete testing of our second hypothesis, Ethan examined the relationship between returns and category within Nashville.

Figure 7

Nashville: Returned by Category



Source: US Superstore – Sample

Note. This mosaic depicts the relationship between Returned and Category.

Figure 7, contrary to our expectations, does not display a statistically significant p-value, indicating that Category and Returned do not possess a strong correlation. This observation runs counter to our second hypothesis, which predicted that category would be correlated with returns.

Through Ethan's analysis of the Southern region, we observed that Total Profit is correlated with Category and Segment, and that returns have a significant association with location. The correlation between returns and segment is strong, but there is not a significant relationship between returns and

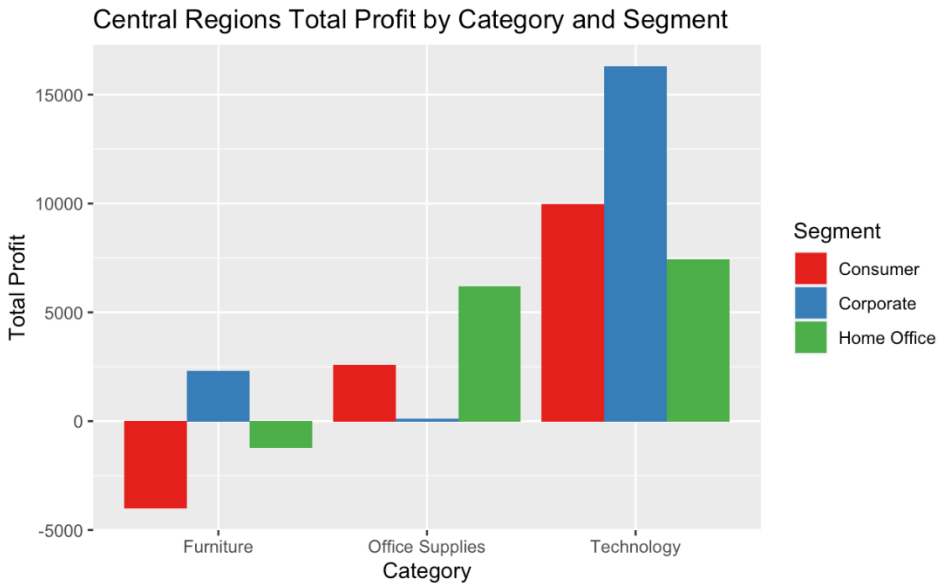
Category. These results are important as they tell the company what they can change to affect product performance, and where they can deploy their efforts, excluding Category.

Central Analysis

Jason Van Wieren was responsible for the analysis of the central region of the U.S., analyzing the relationship between Total Profit, Category, and Segment as well as investigating areas with higher or lower than expected levels of returns. Additionally, Jason compared the total profits of furniture across central states and investigated outliers within these observations.

Figure 8

Central Regions Total Profit by Category and Segment



Source: US Superstore – Sample

Note. This bar chart depicts the relationship between central Total Profit, Category, and Segment.

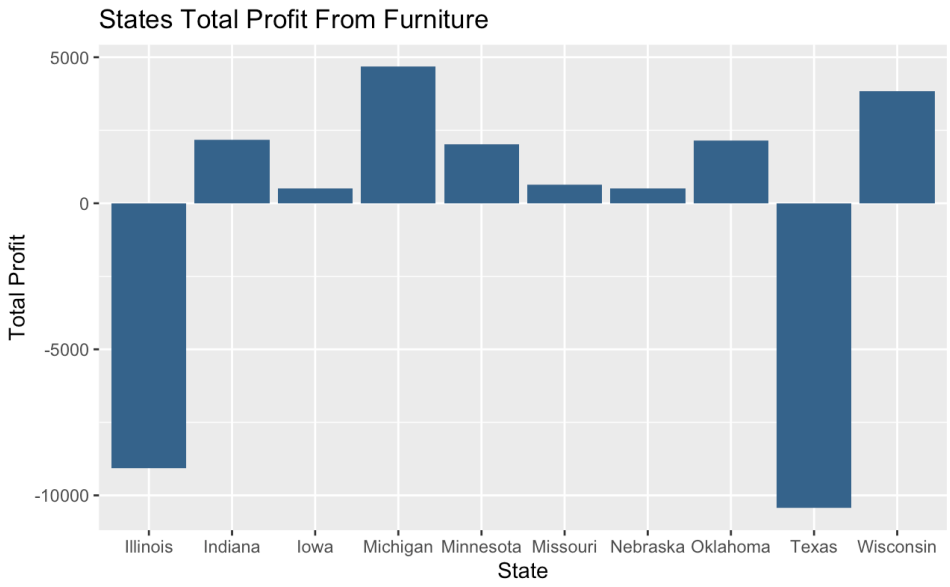
Contrary to the national and southern total profits of furniture, total profit is negative for both the consumer and home office segments, indicating very poor performance of furniture in the central

region, specifically in relation to consumers and home office segments. This regional difference in furniture profits demonstrates that our observations of total profit can differ by region, establishing a correlation between total profit and region. Furthermore, furniture's negative total profits in the consumer and home office segments but positive profit in the corporate sector imply that there exists a correlation between segment and total profit within the central region of the United States. Comparing the technology segment with office supplies and furniture, we observed that technology has a much higher total profit across all segments than every other category, demonstrating that technology in the central region preforms much better than the other categories. These insights support our first hypothesis that there exists a correlation between Total Profit, Category, and Segment.

Delving deeper into the make up of furniture's negative total profit in the central region, Jason graphed the total profit of furniture by state to identify potential outliers that may be influencing the results found in Figure 8.

Figure 9

States Total Profit From Furniture



Source: US Superstore – Sample

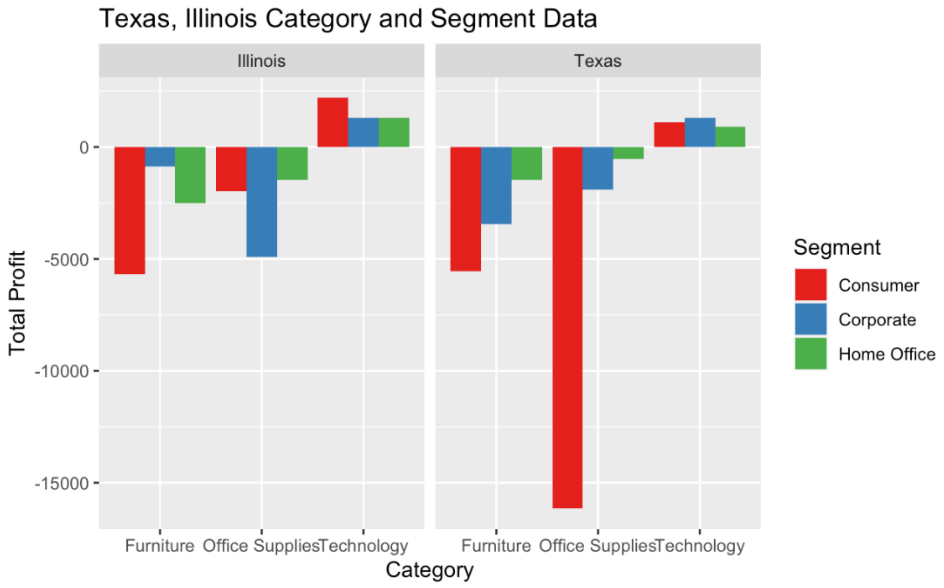
Note. This bar chart depicts the total profit of furniture per central state, minus Kansas, and South Dakota.

When creating this visualization, Jason removed Kansas and South Dakota to decrease the number of states cluttering the bar chart. He chose these states as they had total profits close to 0 and did not represent a strong negative outlier. From Figure 9, Jason surmised most central states do not contain a total profit that exceeds $2,500, and that Illinois and Texas account for much of the negative Total Profit seen in Figure 8.

Jason next compared the total profits for all categories, grouped by segment, within Illinois and Texas.

Figure 10

Texas, Illinois Category and Segment Data



Source: US Superstore – Sample

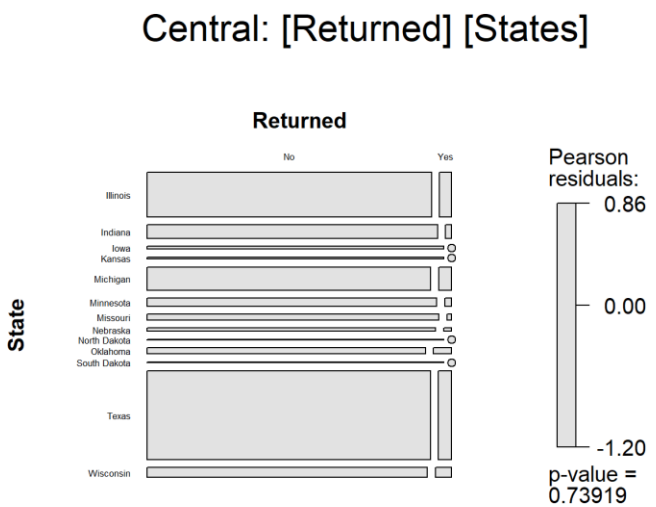
Note. This bar chart depicts the relationship between Total Profit, Category, and Segment within Illinois and Texas.

From Figure 10, Jason discovered that the consumer segments of Illinois and Texas experienced the greatest negative total profit for the furniture category, illustrating a correlation between Segment, Category, and Total Profit. However, Texas and Illinois had different relative levels of profit loss in their corporate segments for furniture, demonstrating that state also plays a role in product performance through its interaction with category and segment. This observation was not initially accounted for by our hypothesis but is understandable after viewing the visualizations. Another significant observation gleamed from Figure 10 is the very large negative total profit that exists within the office supplies category of Texas. This extreme value is a significant issue for the company, as it means their sales of office supplies in Texas are losing them a lot of money. Figure 10 also illustrates that technology still experiences profit, even in these states that do not perform well in other categories.

Jason concluded his analysis with a mosaic plot of returns to test our second hypothesis within the central region.

Figure 11

Central regions total Profit by Category and Segment



Source: US Superstore – Sample

Note. This mosaic illustrates the correlation between State and Returned Status.

The absence of large residuals in central states indicates that the observed values of returns within these states were close to what was expected by randomly assigned values. This means that central supermarkets are not witnessing high levels of returns. The p value of this graph is very high, which signifies a lack of correlation between State and Returned with central U.S. The lack of strong residuals in central states indicated to us that returns could not be reliably used to pinpoint areas of focus within the central region. This conclusion decreases support for our second hypothesis.

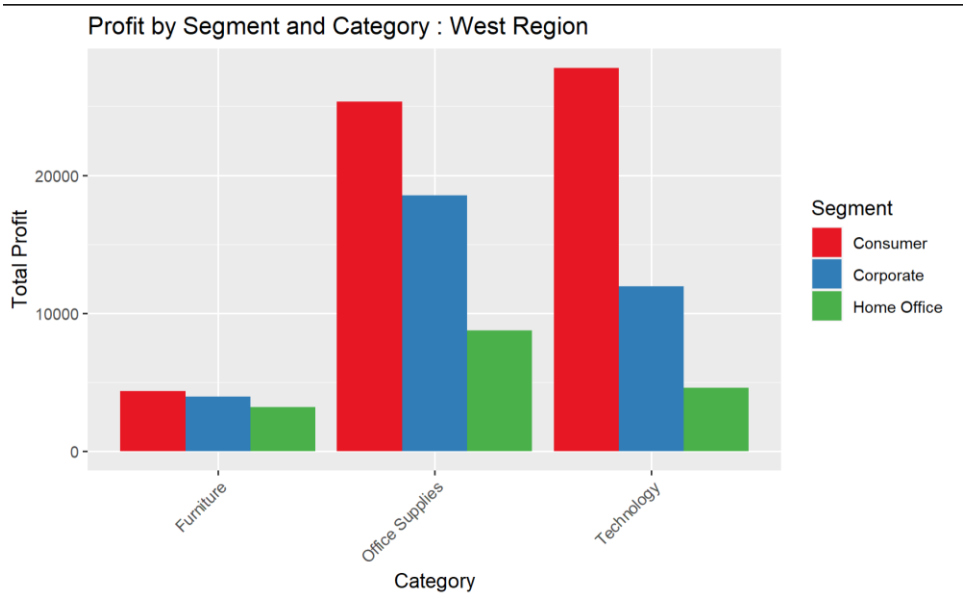
Overall, Jason's investigation into the central region yielded several important observations. First, furniture is still consistently performing poorly, and performs worse than in the southern region, illustrating a correlation between Total Profit and Region that was not accounted for by our first hypothesis. Additionally, state impacts Total Profit in a similar way that was also not explained by our first hypothesis. State and Returned were shown to have little correlation, decreasing the utility of using returns to determine areas of focus within the central region, and going against our hypothesized expectations.

Western Analysis

Nate investigated the western region of the United States to test our hypotheses. He did through the creation of a bar chart and a mosaic plot.

Figure 12

Profit by Segment and Category: West Region



Source: US Superstore – Sample

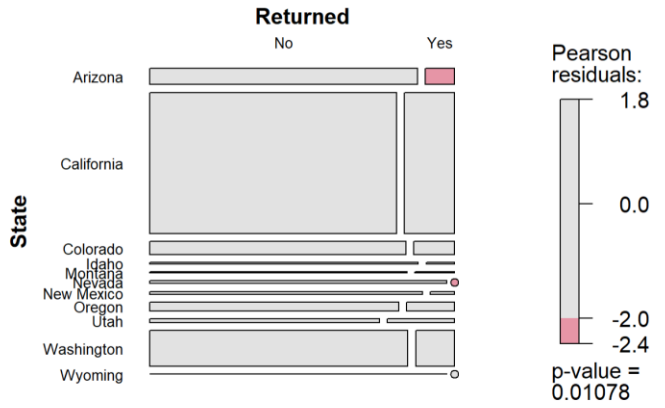
Note. Figure 12 is a bar chart that compares western Total Profit, Category, and Segment.

Figure 12 mirrors the southern distribution of Total Profit by Segment and Category, depicting low total profit for the furniture category, and high total profit for office supplies and technology. Also similar to the southern analysis, the consumer segment makes up most of the total profit within office supplies and technology, indicating that the difference in region has not affected those results significantly. The correlations between Total Profit, Category, and Segment as expected by our first hypothesis exist, and the regional differences we had seen in the central region are not present within the West, indicating that our first hypothesis is validated by Figure 12.

Nate also created a mosaic plot to test the relationship between returns and states within the western region.

Figure 13

West: Returned by States



Source: US Superstore – Sample

Note. This mosaic displays the relationship between returns and states within the western region.

Within Figure 13, Arizona and Nevada both contain negative residuals, symbolizing lower than expected returns. The presence of lower-than-expected returns is a good thing for the supermarket, as it means customers are returning their products less than expected, so the supermarket could gain knowledge from studying the operations of its stores within Arizona and Nevada to understand what they do well that leaves customers satisfied with their products. The low p-value from our Chi Squared test indicates that State and Returned are correlated; thus, our second hypothesis is supported by Figure 13.

Nate's analysis of the western United States supported our existing hypotheses by demonstrating that furniture is still poor preforming, regardless of region, implying that it is a predictor of Total Profit. Additionally, within all categories, the relative segment total profits were similar to expected results, validating our first hypothesis. In respect to returns, the presence of negative residuals and a low p-value indicate that the supermarket can pinpoint areas of focus based on returns and in the

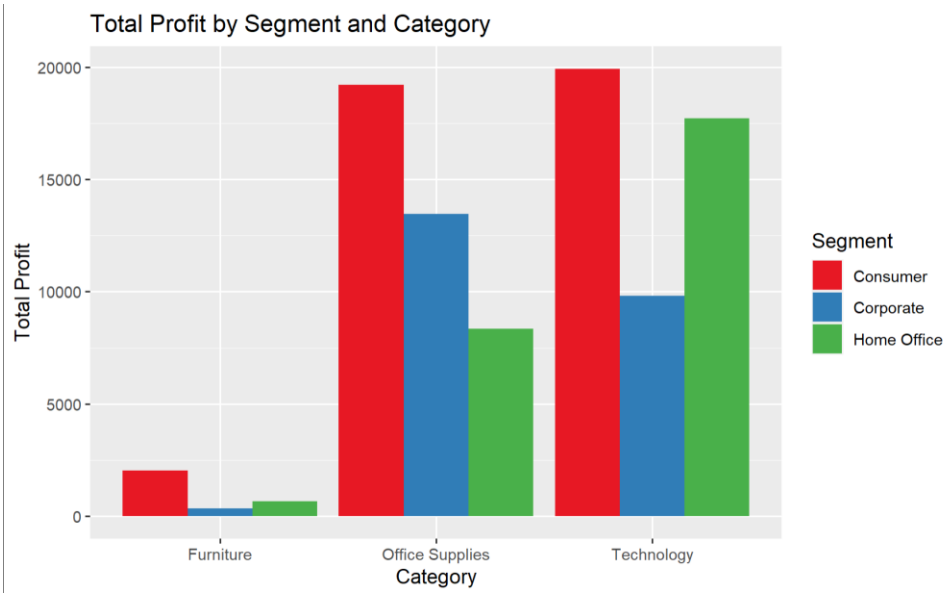
western region, and that there exists a correlation between State and Returned, as predicted by our second hypothesis.

Eastern Analysis

Through the use of bar charts and mosaics, Jacob examined Total Profit trends as wells as returns within the eastern region of the U.S.

Figure 14

Total Profit by Segment and Category – East



Source: US Superstore – Sample

Note. This is a bar chart illustrating the relationship between eastern Total Profit, Category, and Segment.

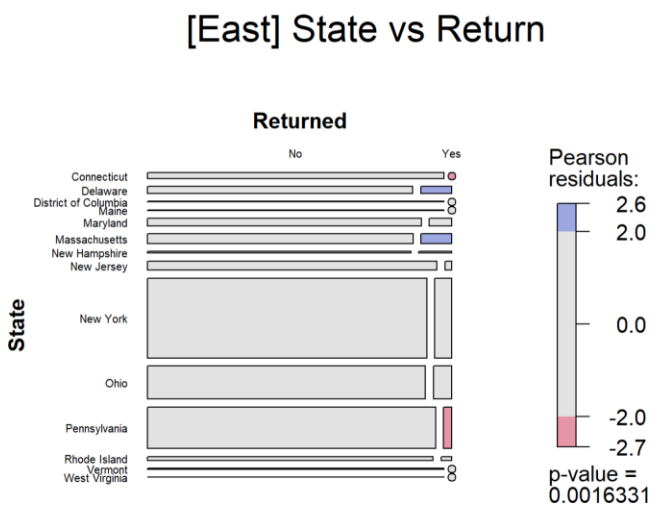
Once again, furniture Total Profit is very low across all segments, indicating that it is not affected by region, and is an all-around weak point for the company. Office supplies and technology both have high total profits, resembling southern and western levels except for the decreased total profit of corporate purchasers of technology. The disparity of this segment from other regions is the only regional

difference displayed in Figure 14, meaning that for the most part, the interactions between Total Profit, Segment, and Category follow our first hypothesis.

Jacob also graphed returns against eastern states to evaluate the utility of using returns as a method of pinpointing areas of focus within the East.

Figure 15

East: State vs Return



Source: US Superstore – Sample

Note. This mosaic demonstrates the relationship between Returned and State.

Figure 15 depicts negative residuals in Connecticut and Pennsylvania, with positive residuals in Delaware and Massachusetts, providing a range of satisfied and unsatisfied customers. The presence of residuals, in addition to the small p-value, demonstrates that returns are a good way to pinpoint areas of focus within the East, and that State and Returned are correlated within the East. These conclusions support our second hypothesis and mirror the conclusions of our southern and western analyses.

Jacob's analysis of the eastern United States supported our hypotheses through the discovery of a correlation between Total Profit, Category, and Segment, as well as a correlation between location and returns. The eastern results were similar to those in the southern and western regions, demonstrating a symmetry between these three regions.

Predictions

EDA-Based Predictions

Through our EDA, we discovered multiple trends that corroborated our hypotheses, as well as some that ran counter to what we expected. Based on our observations, we have several predictions for the direction of product performance at the supermarket:

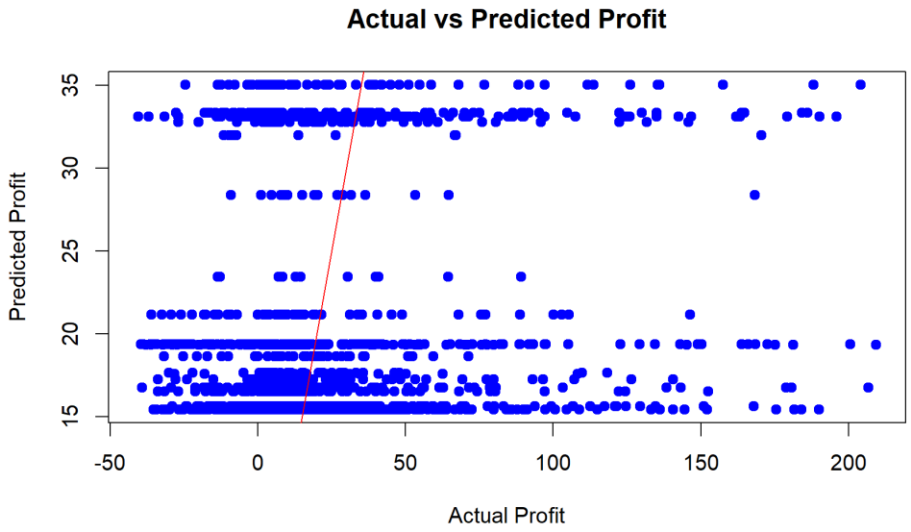
1. Furniture will continue to perform poorly across all segments nationally if no action is taken.
2. Technology will continue to be the dominant segment by Total Profit across all segments nationally.
3. The consumer segment will maintain its position as the most profitable segment nationally.
4. Within the central region, furniture will continue to lose the supermarket money, and office supplies will do the same unless corrective action is taken.

Machine Learning Predictions

In addition to the predictions we made based on our EDA, we also trained a machine learning model to predict total profit. We used a Random Tree Regression method, and target guided ordinal encoding to factorize our predictor variables of Segment, Category, and Return Status. Our model was not very accurate, with an MSE of 1211.12.

Figure 16

Actual vs Predicted Profit



Source: US Superstore – Sample

Note. This is a scatterplot illustrating the actual profit for a collection of predictor variables vs the profit the machine leaning model predicted.

The red line in Figure 16 represents a correct prediction of profit by the machine learning model. Our machine learning model rarely achieved a correct prediction, resulting in a highly inaccurate model. Because of the model's inaccuracy, we chose not to take into consideration its predictions.

Conclusion

Overall Findings

Through this study, we found that both of our hypotheses were partially validated. Our first hypothesis predicted a correlation between product performance measured by Total Profit, Category, and Segment. Because of the repeated inferior performance of furniture across all regions, and because of the prevalence of technology as a profitable category, we validated that Category was correlated with Total Profit. The consistent prevalence of the customer segment within each category for most regions, except central, demonstrated that segment was usually correlated with Total Profit and Category, but

not always. Within the central region, state was correlated with Total Profit, which was also not accounted for by our first hypothesis. Our second hypothesis, which predicted that product returns were correlated with Region, State, Segment, and Category, and could be used to pinpoint areas of interest to the supermarket, proved to be true in respect to the correlation between Returned Status, Region, State, and Segment; however, no correlation was found between returns and category, indicating that our second hypothesis was partially incorrect. Returns were effective at pinpointing locations of interest to the supermarket using residuals in all regions except central. Overall, our first hypothesis successfully described the correlation between Total Profit, Segment, and Category that we witnessed in the majority of regions, even though it failed to account for deviations in the central regions results. Our second hypothesis also accurately represented the utility of using returns to pinpoint areas of interest to a company, provided Region, State, and Segment have some form of residuals in their returns.

Prescriptions

Using our hypotheses as a guide, our team prescribes several actions to the supermarket in order to increase product performance:

1. Use residuals on returns to pinpoint areas with higher-than-expected returns, then investigate the segment(s) that is driving the high returns. Based on feedback from members of the dissatisfied segment(s), reform supermarkets in the area to increase that segment's product satisfaction.
2. Interview buyers of supermarket's goods nationally, across all segments, to ascertain if furniture is a sufficiently demanded good from the supermarket. If it is, work to decrease the costs associated with bringing the furniture to the store to increase furniture profits. If it is not demanded, stop selling furniture.

3. Within the central region, specifically Illinois and Texas, focus special attention on raising profits of furniture and office supplies through methods outlined in the second prescription. In Texas, stop selling office supplies immediately while re-tooling is underway.
4. For product categories that are discontinued, reinvest those resources into the technology category.

Reflections

Following the completion of this study, we reflected on several challenges we faced during our analysis. First, because of our fictitious dataset, we had a challenging time creating time series plots that demonstrated anything conclusive. Second, it was difficult to find a dataset that could accommodate four people working on it at the same time, as well as contain enough numerical data to be useful in a machine learning model. Due to the abundance of categorical variables and the lack of numeric continues variables, we were unable to create a machine learning model that accurately predicted the profit of a product.

We felt that our team did a good job of communicating and coordinating efforts, however. We were able to consistently meet and share findings, which led to an exploratory analysis of our data that was very in-depth and informative. Overall, we felt that our team did well with the dataset we chose, despite its limitations, and produced a quality final result.

Bibliography

Assosia. (2022). *Retail Sector | Retail Market Analysis & Industry Insights*. Assosia. https://www.assosia.com/sectors-channels/retail#:~:text=The%20retail%20industry%20consists%20of

Bhalla, D. (n.d.). *A complete guide to Random Forest in R*. ListenData. https://www.listendata.com/2014/11/random-forest-with-r.html

Friendly, M. (2023, August 21). *Permuting variable levels*. R-Packages. https://cran.r-project.org/web/packages/vcdExtra/vignettes/mosaics.html

Helsloot, R. (2020, November 8). *Change the size of labels in mosaic function, R*. Stack Overflow. https://stackoverflow.com/questions/61579713/change-the-size-of-labels-in-mosaic-function-r

Martin, S. (2023, January 18). *Where Can I Find Superstore Sales? - On a Tableau Quest... - Confluence*. Datawonders.atlassian.net. https://datawonders.atlassian.net/wiki/spaces/TABLEAU/blog/2022/10/26/1953431553/Where+Can+I+Find+Superstore+Sales#Workbooks-and-Data-Sources

Mudadla, S. (2023, April 27). *Target guided ordinal Encoding with Example*. Medium. https://medium.com/@sujathamudadla1213/target-guided-ordinal-encoding-with-example-450323fea78e

Ouedraogo, R. I. (2023, April 30). *Predictive Analytics for Grocery Sales Forecasting: A Case Study of Favorita Stores*. Www.linkedin.com. https://www.linkedin.com/pulse/predictive-analytics-grocery-sales-forecasting-case-rekia/

STHDA. (n.d.). *Chi-Square Test of Independence in R - Easy Guides - Wiki - STHDA*. Www.sthda.com. http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r