Regional Analysis of a Supermarket

By Team 3: Ethan Ericson, Nate Moore, Jacob Romine, and Jason Van Wieren

Project Overview

- Dataset: US Sample Data (Fictitious)
- The goal of this project was to analyze the factors that impact product performance to predict future performance and make prescriptions for better sales practices.

Hypotheses

- Hypothesis 1: There exists a correlation between product performance (measured by profit), consumer segment, and product category within the superstore.
- Hypothesis 2: Product returns are correlated with the factors listed below and can be used to predict product performance in a specific location by pinpointing areas with high or low returns.
- Our Hypotheses predict that the factors influencing product performance are:
  - Region
  - State
  - Segment
  - Category
  - Returned Status

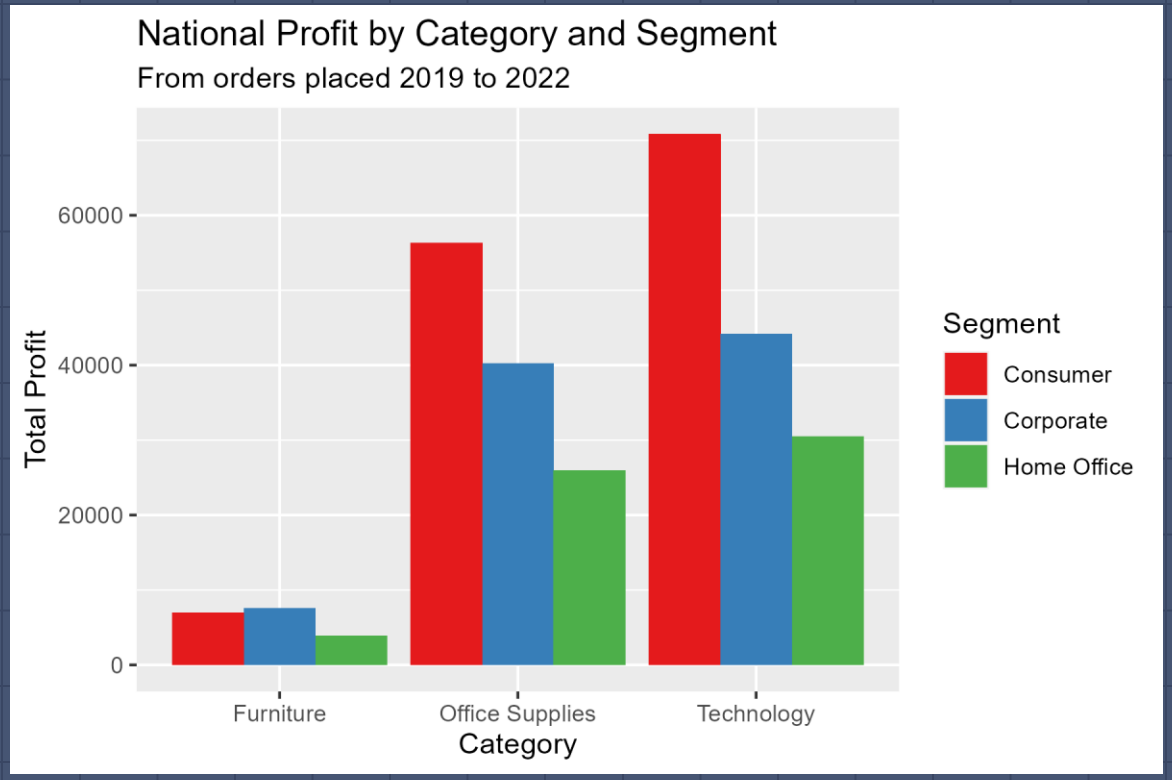
Methods

- We split our analysis up by region.
- Within each region, we explored the total profits of categories of products, separated by customer segment. We also investigated the levels of returns within each of our region.
- We constructed a machine learning model to predict product profit based on customer segment, category, and returns.

National Analysis

National Profit

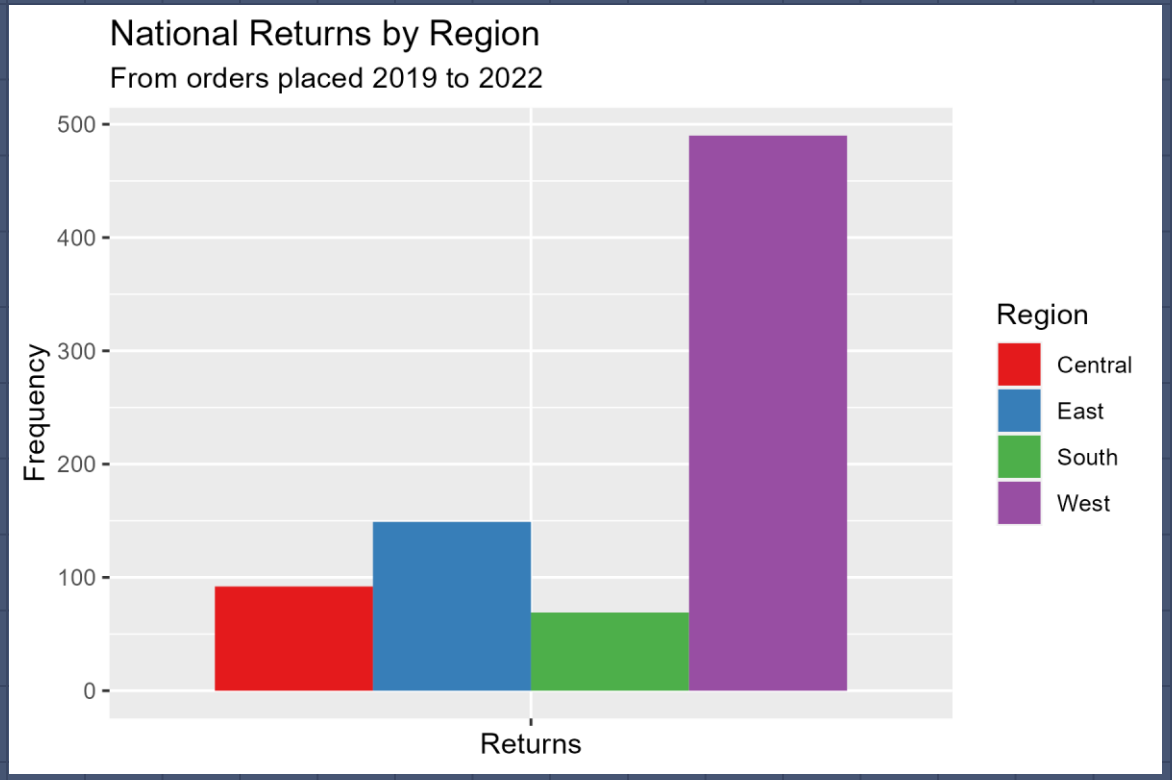
- Graph shows the total profits of a certain category, split by buyer segment.
- Furniture has the lowest profits altogether, while Office Supplies and Technology are similar to each other.
- On the surface, it appears that furniture is not preforming well across all customer segments



Source: US Superstore - Sample

National Returns

- Graph shows the total amount of returns in each of the regions.
- West region has high number of returns.
- South and Central regions have low number of returns.

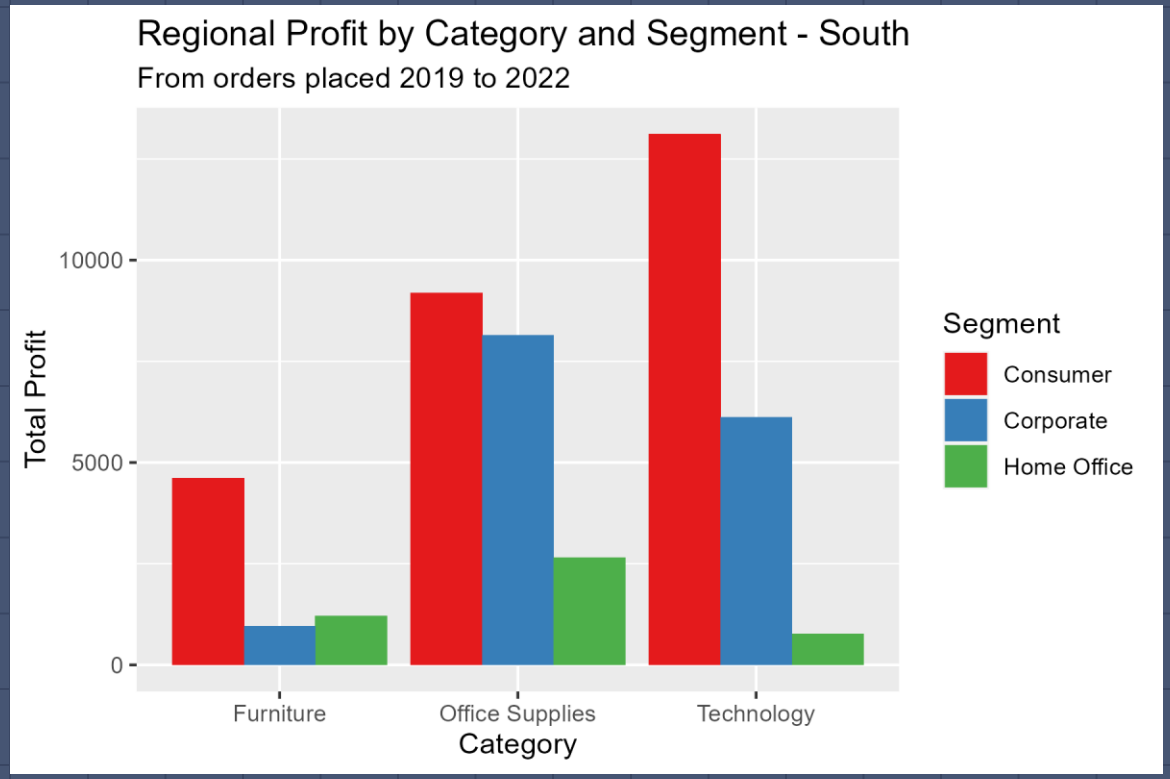


Source: US Superstore - Sample

Southern Analysis

Southern Profit Analysis – Category and Segment

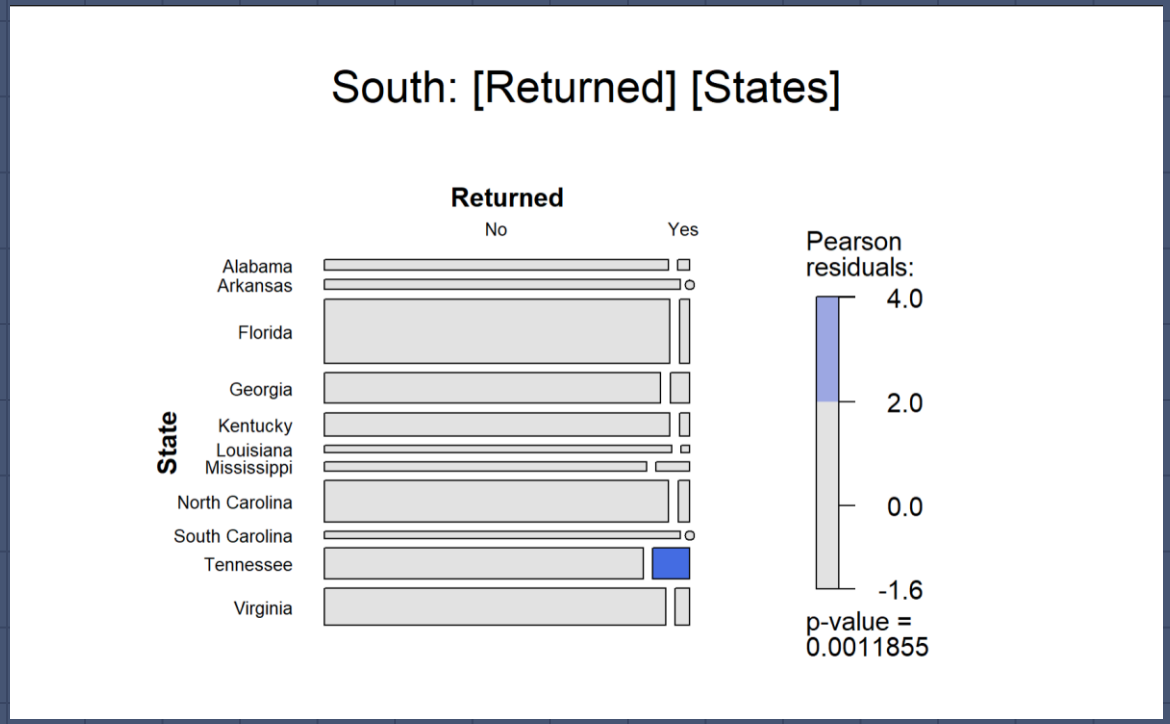
- Consumers represent the most profitable segment to sell products to.
- The furniture category is the worst preforming out of the three the company sells.
- So, category and segment affect the profit the business earns.



Source: US Superstore - Sample

Southern Return Analysis - States

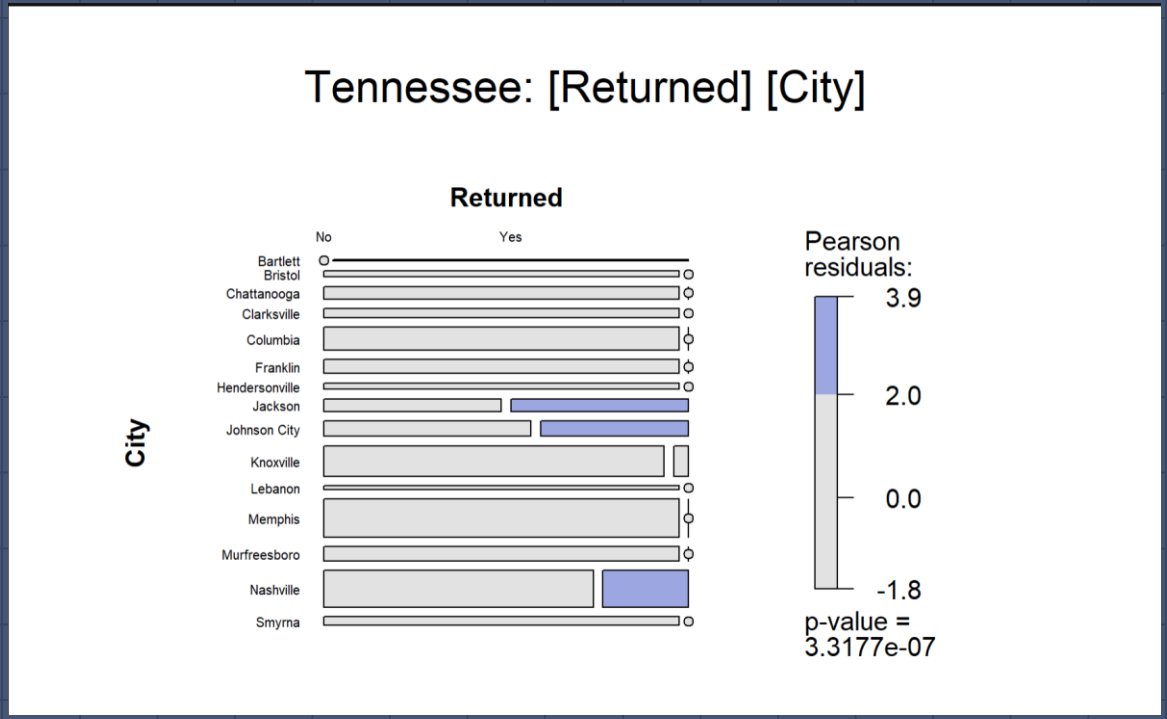
- P values measure probability of attaining the observed results given the null hypothesis is true.
- A residual measures how far away an observation is from an expected value give the null hypothesis is true.
- Tennessee has a high positive residual for returns, which indicates higher than expected returns.



Source: US Superstore - Sample

Southern Return Analysis – Tennessee Cities

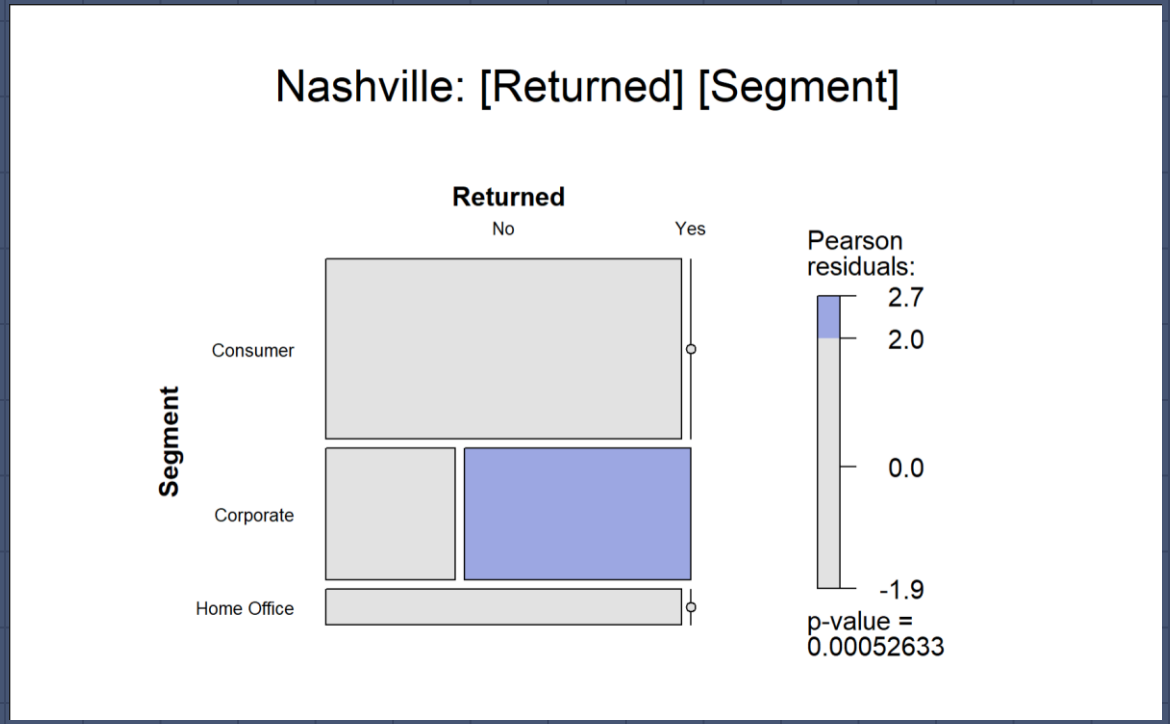
- Note the p-value of 0
- Nashville, along with Jackson and Johnson City have high residuals, so there must exist a strong relationship between returns and cities within Tennessee.



Source: US Superstore - Sample

Southern Return Analysis – Segments within Nashville

- Note the p-value is < 0.05.
- Large residuals in the Corporate segment indicate higher than expected returns from corporate buyers.
- Nashville corporations are likely are driving the higher than expected returns in Tennessee.

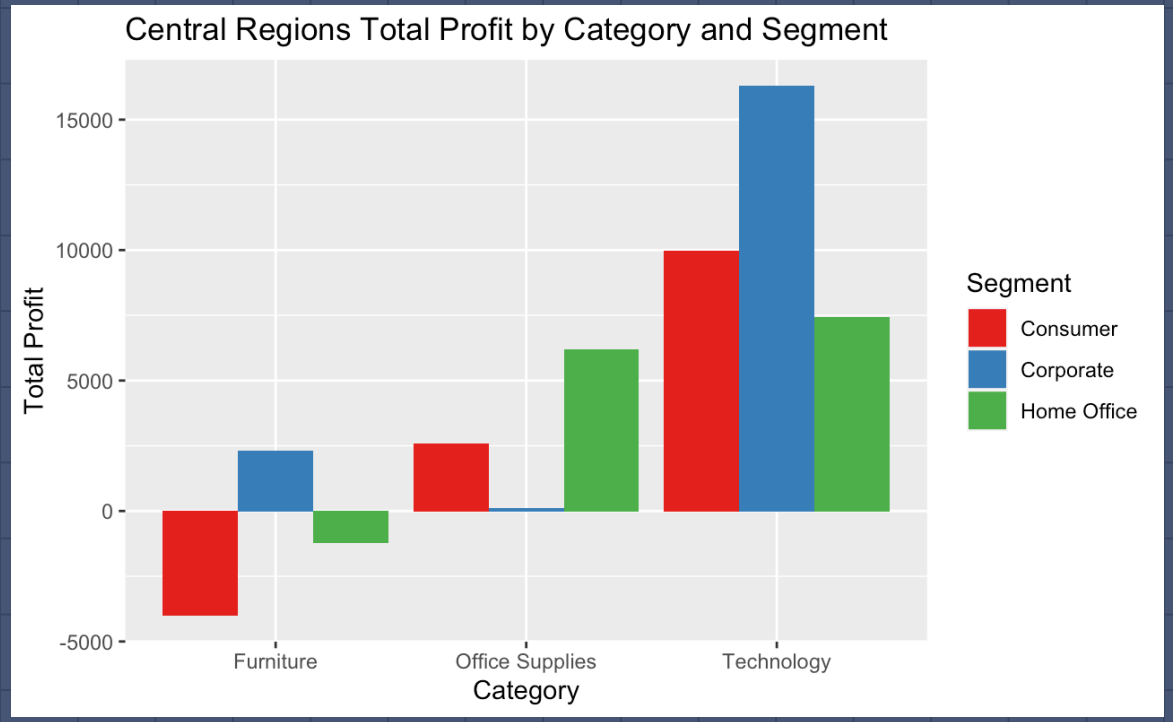


Source: US Superstore - Sample

Central Analysis

Central Profit Analysis – Category and Segment

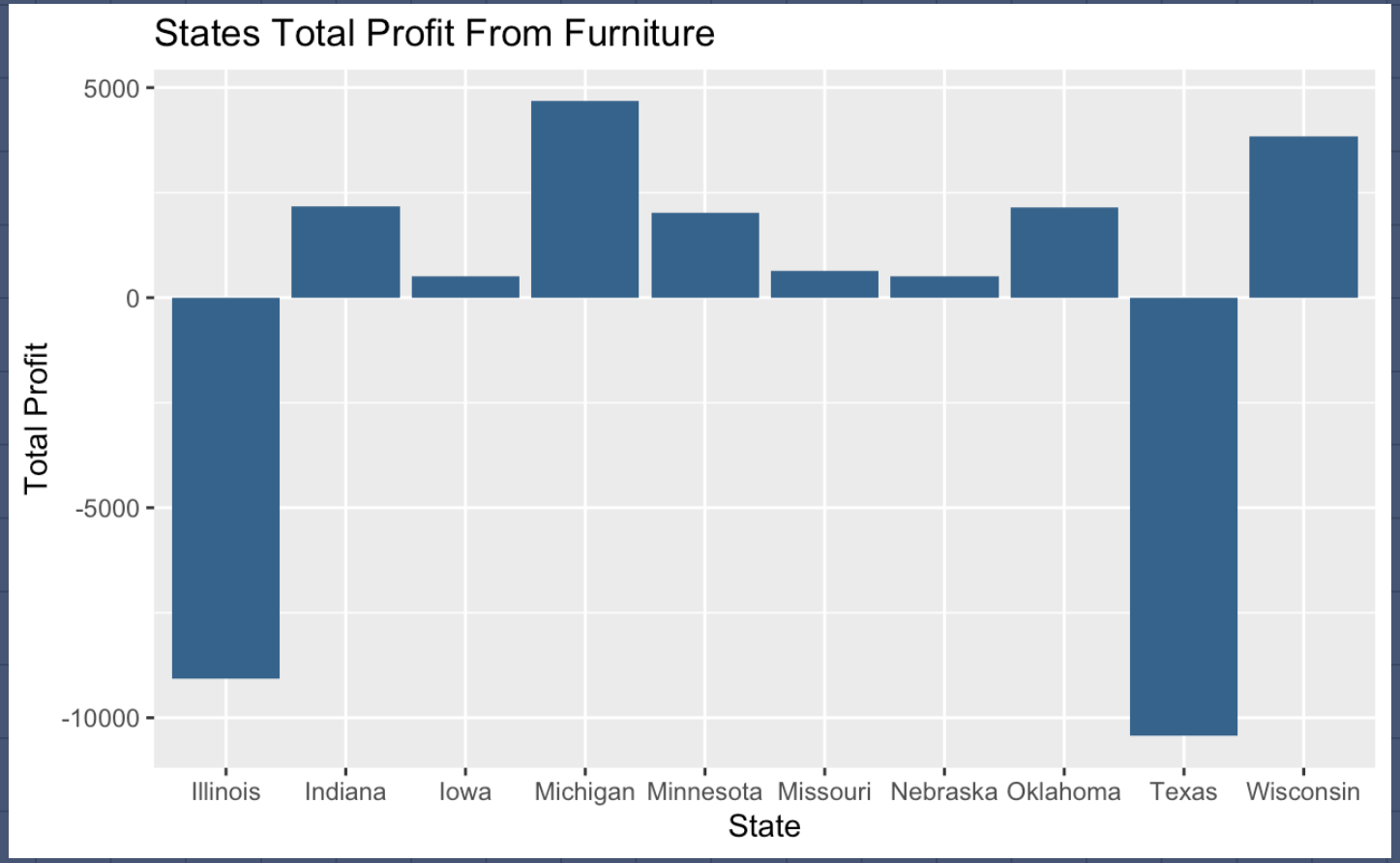
- Furniture is being sold at a loss on average.
- Technology by far the most profitable category.
- Corporate is the most profitable segment.
- This shows both category and segment play a role in profits in the central region.



Source: US Superstore - Sample

Central State Profit Analysis – Furniture

- Illinois and Texas stand out as outliers in the data.
- Most other states making profit of less than $2,500.
- Texas and Illinois alone make profits from furniture in the central region a deficit.

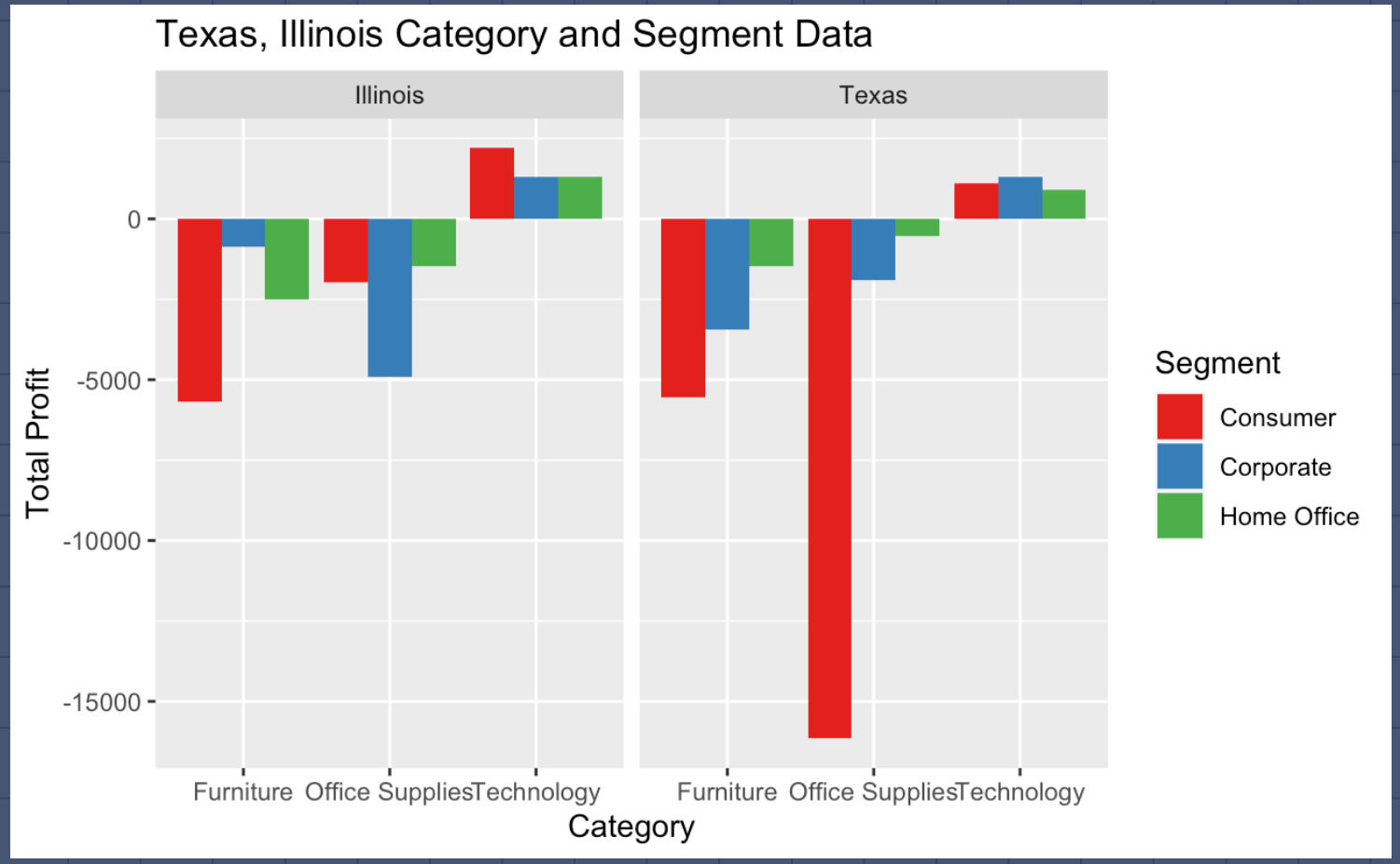


(Kansas and South Dakota not included)

Source: US Superstore - Sample

Texas & Illinois Profit Analysis – Category & Segment

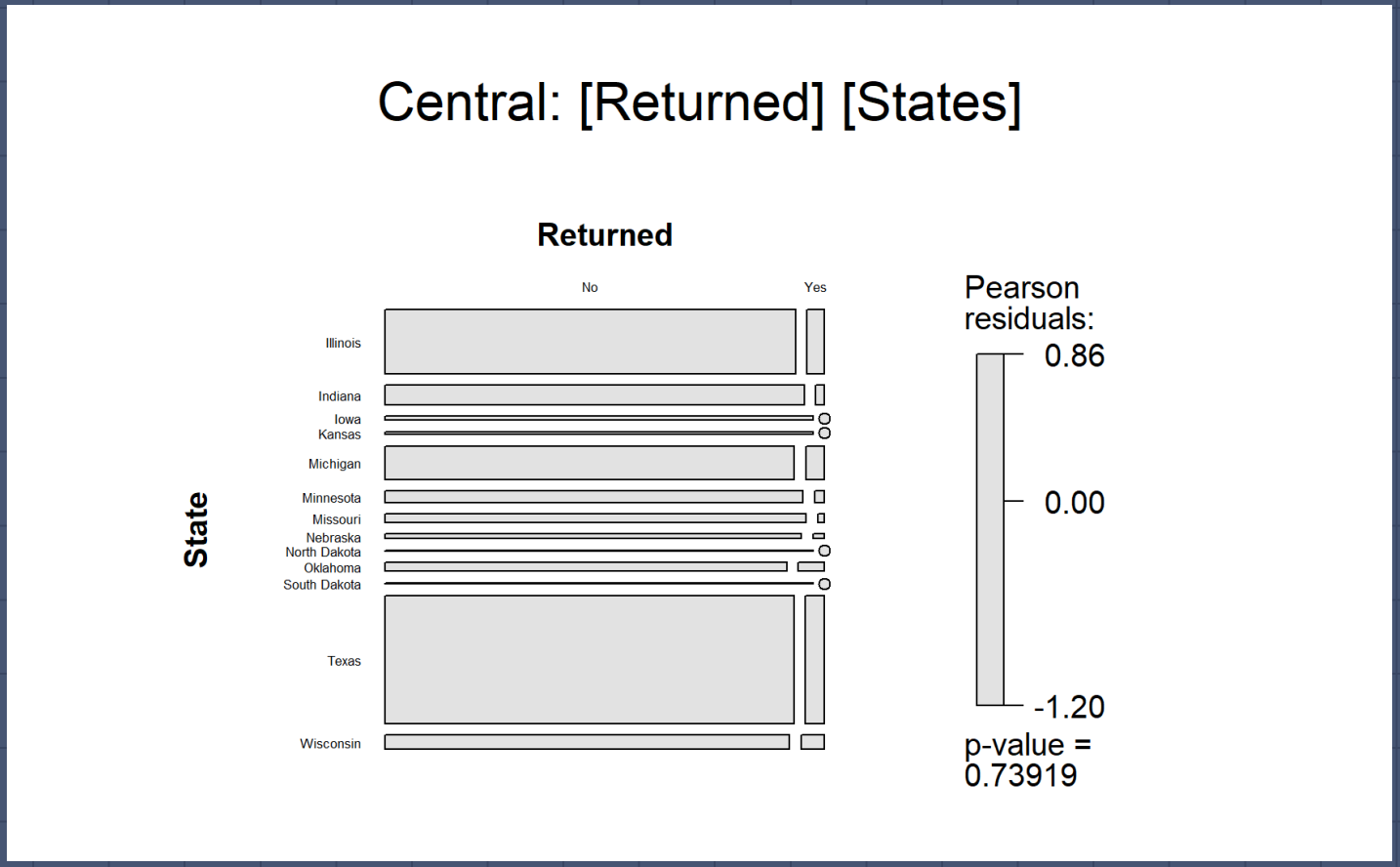
- Both furniture and office supplies being sold at a loss in all segments in both states
- Technology still the most profitable category



Source: US Superstore - Sample

Central Return Analysis - States

- Note the high p-value.
- Few states have high residual values meaning what was observed was mostly expected.

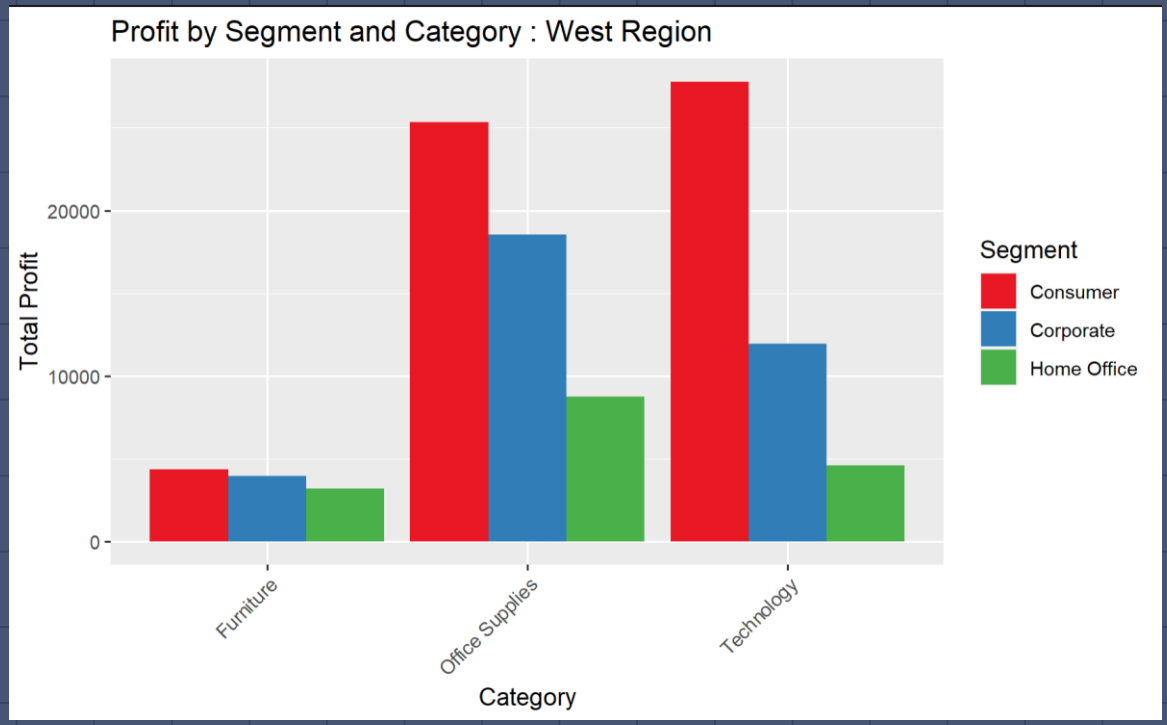


Source: US Superstore - Sample

Western Analysis

Western Profit Analysis - Category and Segment

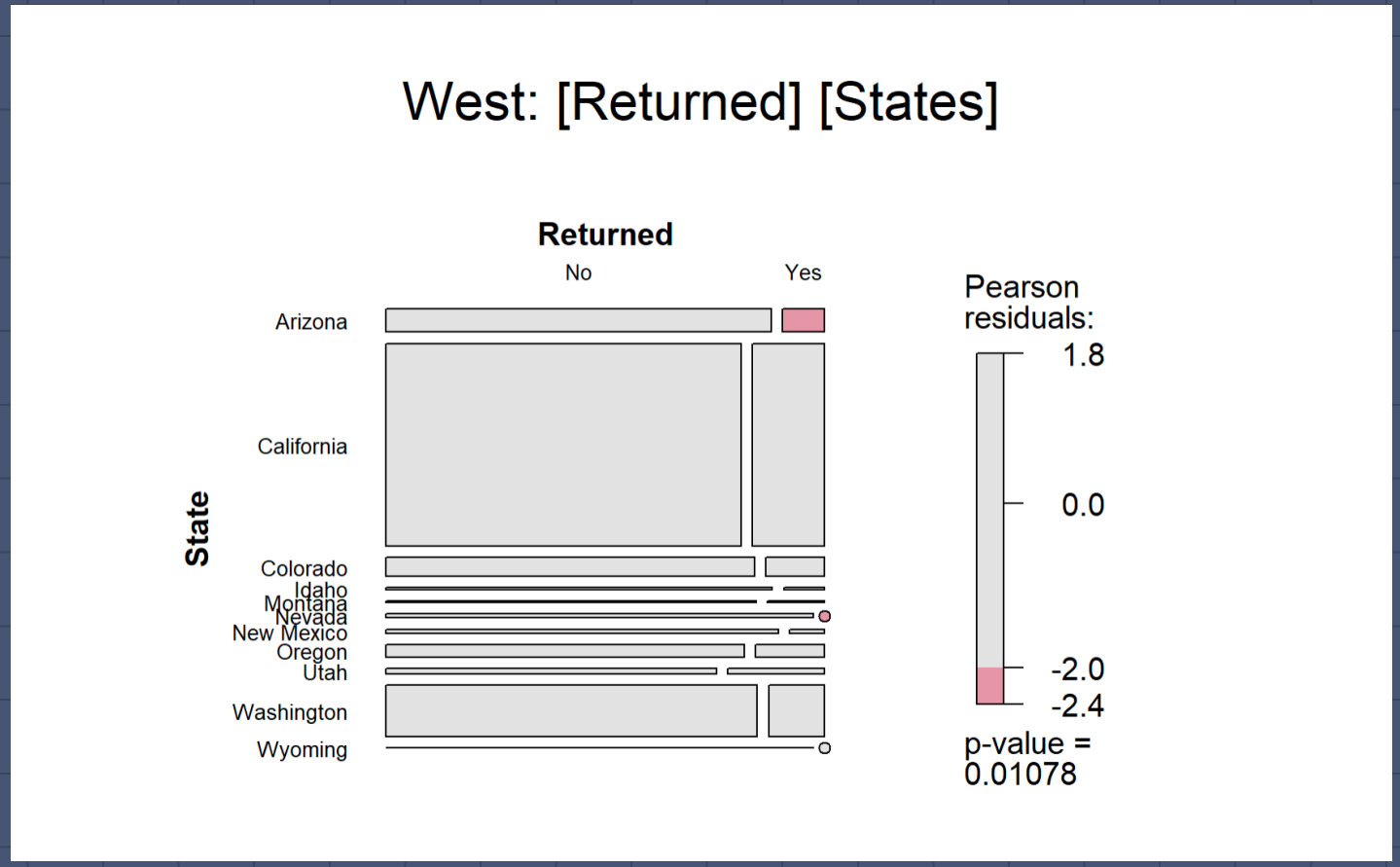
- As you can see, Furniture seems to be overall one of the least profitable
- Consumers make up majority of the volume in profit
- Corporate and home office have similar interest in products purchased



Source: US Superstore - Sample

Western Return Analysis - States

- Note p-value < 0.05
- California accounts for most of the returns
- Negative residuals in Arizona and Nevada, indicating more products kept than expected.

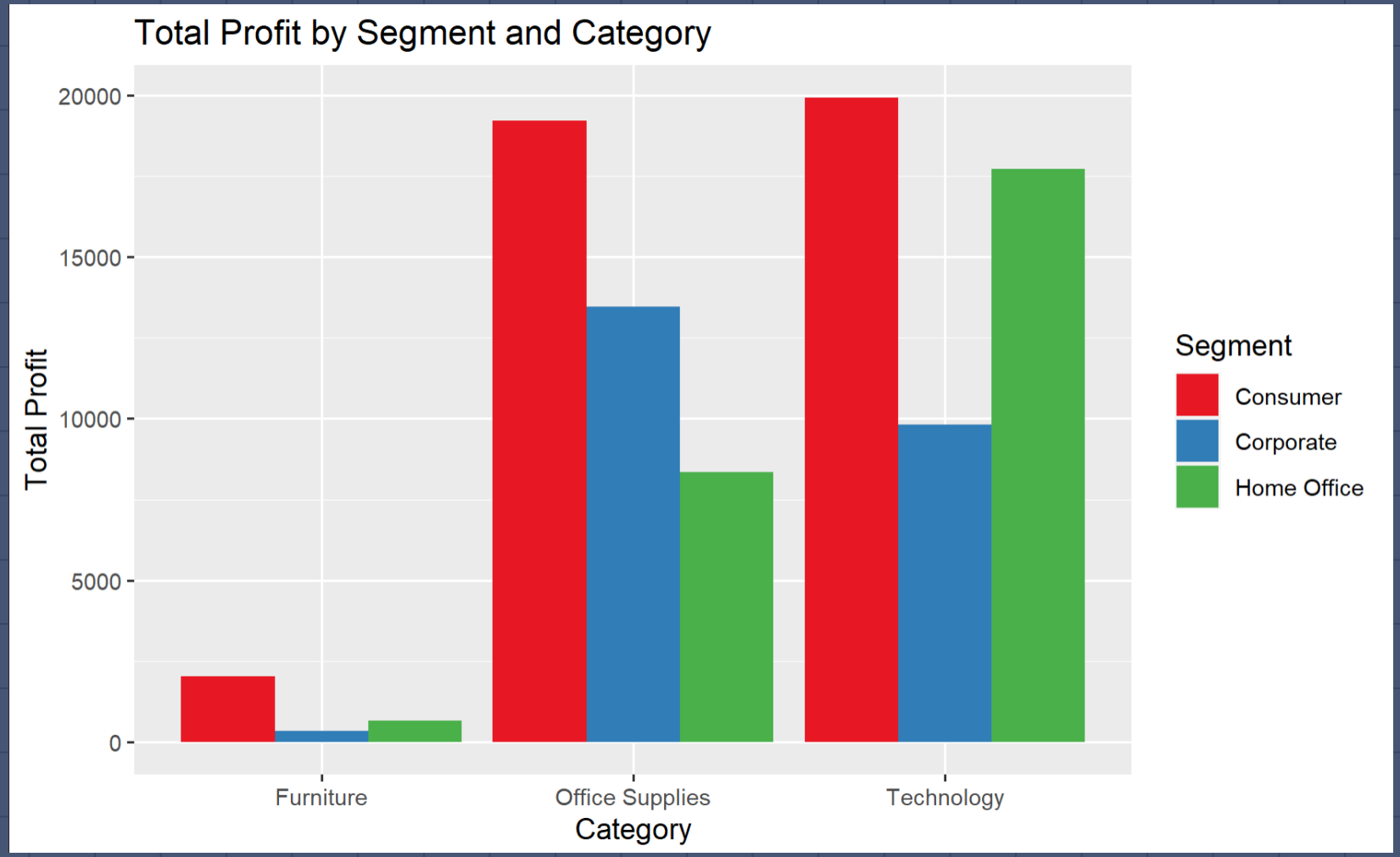


Source: US Superstore - Sample

Eastern Analysis

East Total Profit by Segment and Category

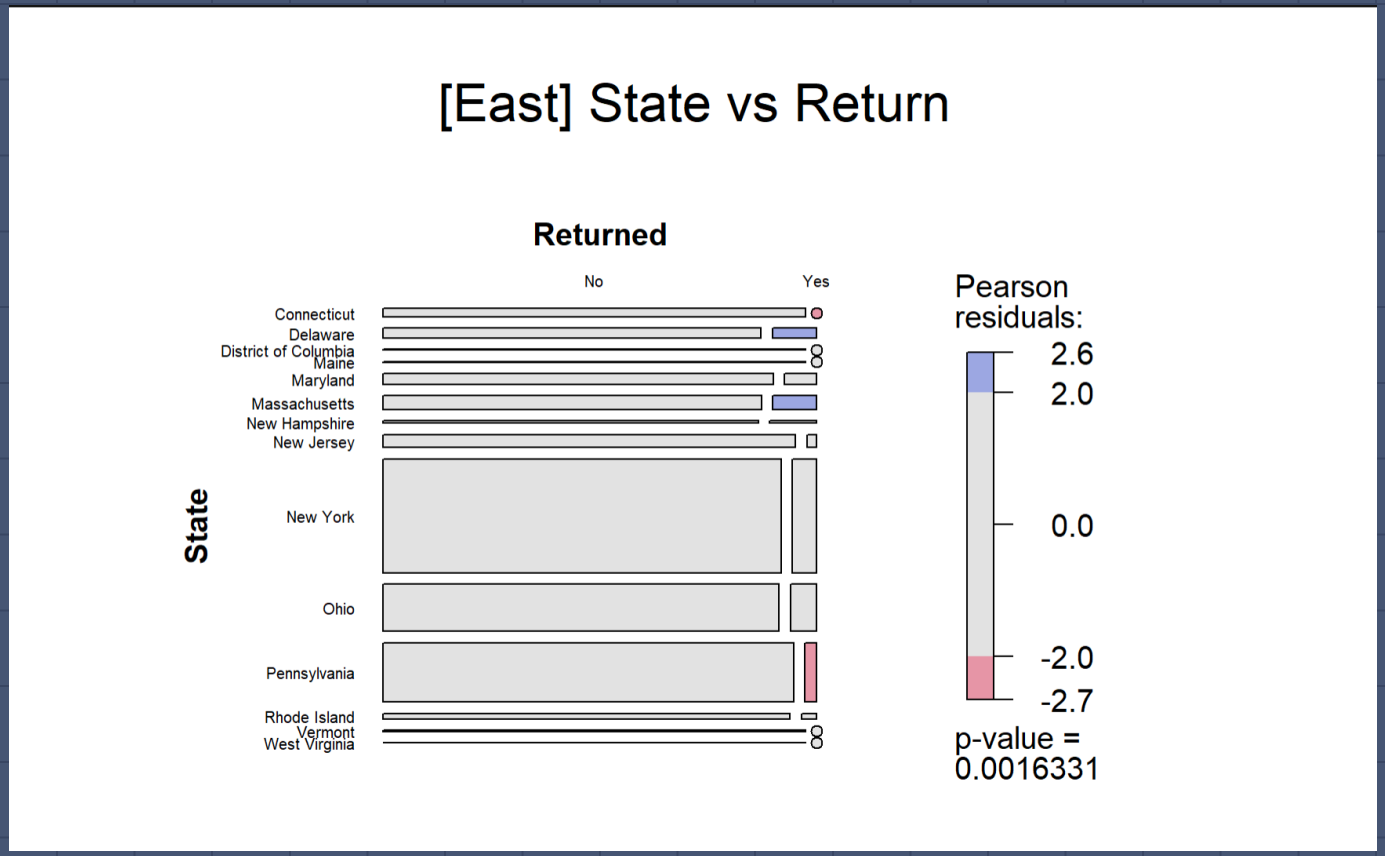
- Consumer is the most profitable by far, except in technology, where home office is profitable as well.
- Furniture is by far the worst, almost losing profit with corporate buyers.



Source: US Superstore - Sample

East State versus Return Mosaic Plot

- New York (NYC specifically) accounts for most of the returns
- Negative residuals in Connecticut and Pennsylvania.
- Positive residuals in Delaware and Massachusetts.

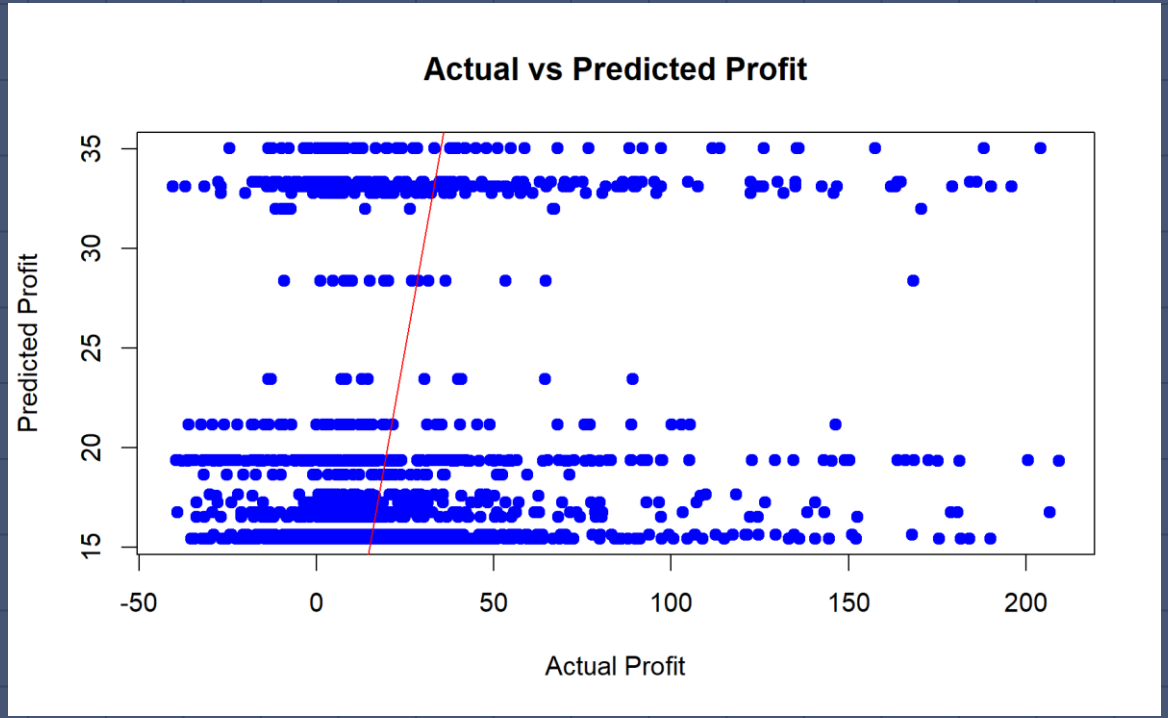


Source: US Superstore - Sample

Machine Learning

Machine Learning Model

- Correlation plot of predicted profits vs the actual profits based on Segment, Category, and Return Status
- We tried to use Target Guided Ordinal Encoding to account for the use of only categorical variables, but we still failed to accurately predict profits.
- Used Random Tree Regression model.
- MSE: 1211.12 (Very high = Inaccurate)



Source: US Superstore - Sample

Conclusion

Overall Findings

- We found that furniture under-preformed in all of our regions; thus, the company should re-evaluate whether it sells furniture or not.
- Continued sale of furniture will not produce large dividends.
- Product returns are influenced by region, state, and city overall.
- These observations have the implication that the company should divert resources away from furniture sales towards office supplies and technology, and that the company should investigate the levels of returns within each regional location.

Challenges

- We found that due to our data being fictitious it was hard to find any incredibly conclusive evidence that time and the data points correlated.
- We found very few publicly available datasets that had enough data to conduct a Exploratory Data Analysis split among 4 people. We settled for a fictitious dataset which had its own problems.

References

- 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
- 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

Questions?