COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Final Progress Report

December 18th, 2021

Keertan Krishnan (kk3446), Myles Ingram (mai2125), Rahul Agarwal (ra3097), Rahul Subramaniam (rs4128), Shaurya Malik (sm4969) *

Industry Mentors:
Rohit Cherian, Nandakumar S, Kanika Aggarwal

Faculty Advisors:
Prof. Sining Chen, Prof. Kelleher

Data Science Capstone & Ethics (ENGI E4800), Fall 2021

*In Alphabetical Order

Table of Contents

1. Introduction
    a. Problem Definition
    b. Project Goal and Deliverables
    c. Datasets
    d. Relevant Terminology

2. Proposed Framework

3. Methodology
    a. Data Pre-Processing
    b. Analytical Methods
        i. Affinity Analysis
        ii. Apriori Algorithm
        iii. Association Rule Mining
        iv. Percentage Contributions and Volumetrics

4. Results
    a. Use Case 1 - Mens Solid Colored T-Shirt Removal Test
    b. Use Case 2 - Childrens Promotion Removal Test
    c. Use Case 3 - Childrens Size Merge Test
    d. Use Case 4 - Footwear Assortment Expansion Test

5. Conclusions

6. References

Introduction

The Ralph Lauren Capstone Project is a partnership between the Data Science Institute at Columbia University and Ralph Lauren.

Problem Definition

Ralph Lauren is a leading fashion retailer that conducts a variety of tests across its stores to understand customer purchasing patterns and user experience. For the purposes of this project, our team was introduced to four such tests, i.e. use cases that modify the environments at various levels at Ralph Lauren stores.

Each use case corresponds to a certain change that was effected in a group of Ralph Lauren stores, known as the test stores, over a defined period of time. In order to analyze the differences in consumer behavior and satisfaction that this change caused, the tests have been set up in such a way that for each test store, there are a set of control stores, which for all purposes during this test are assumed to be identical to the test store except for the fact that the change we are looking to analyze was not performed in this store. The concrete use cases that have been conducted and are in the scope of this capstone project are described in Table 1.

Use Case	Description of Test	Expected Deliverable
Footwear Assortment Expansion Test	Introducing footwear to non-footwear stores under various assortment models	Analytic template for pre-post/Test-Control comparison at various time-period levels
Men's Solid Colored T-Shirt Removal Test	Mens' Solid Tees were removed and only color choices of fancy tees and fleece were given	
Children's Size Merge Test	Merged all sizes together into one cohesive area (instead of big and little) for better customer experience	
Children's Promotion Test	Removed all promotions in store in children's apparel	

Table 1: The Four Use Cases in the Ralph Lauren Capstone Project

Project Goal and Deliverables

The overall goal of the project is to construct an analytical framework to investigate and understand the differences in customer purchase behavior caused by tests that Ralph Lauren has conducted in a certain subsection of its stores. Change in customer behavior can be evaluated through various metrics such as the Average Unit Retail (AUR), Average Value per Transaction (AVT) etc.

In the future, the framework that we propose to build will be used to determine the effect of tests conducted in the future, quantifying store-level changes on customer engagement/experience and inform critical merchandising and marketing decisions by the business.

Datasets

1. **Sales Data:** The sales data provided us with granular information on every item sold in 201 Ralph Lauren stores all across the USA. Prominent features include the item department, number of units sold, net sales, facility details, date of transaction and transaction ID.
2. **Item Book Price Data:** The item book price data gives us information regarding the retail price of a given item at a certain store at a given date. The item book price data could prove to be useful in pulling out item descriptions of certain items that stand out from the rest.
3. **Customer Segmentation Data:** The customer segmentation dataset gives the gender, age group, net worth, estimated household income for each consumer ID. The dataset also includes whether children are present with the consumer and the number of children with the consumer. These characteristics are determined through store surveillance data.
4. **Store Traffic Data (Day Level):** The store traffic data gives us day-level information of the footfall in all the Ralph Lauren stores that the team is performing its analysis on. The store traffic data proves to be useful in calculation of performance metrics such as conversion ratio.
5. **Store Master Data:** This data gives us store details for each store corresponding to the different test types. Relevant features include Store name, Address, Region, Destination type, square footage area of the store.
6. **Test Description:** The test description data gives the test and control stores for each use case as well as scope of each test. Five control stores correspond to one test store. The test description also has the start dates and end dates for each use case.

Relevant Terminology

1. *Average Unit Retail (AUR, for time period t) = Total Sales / Total Number of Units Sold*
2. *Average Value Per Transaction (AVT, for time period t) = Total Sales / Total Number of Transactions*
3. Support: It gives the frequency at which at which an item appears in the given range. This is often stated as a probability count and is expressed in numerical value where:

   Support = Frequency of Item / Total Number of Transactions

   Support simply emphasizes how popular an itemset is, measured by the proportion of transactions in which an itemset appears.
4. Confidence: It is a measure of the reliability of the rule, calculated as the ratio of the number of transactions that include the consequent item (the conditional purchase - in this case, the item that is bought when the first item is purchased) to the total number of transactions with the antecedent. It is given by:

   Confidence Count = Frequency of the Consequent Item / Total Number of Transactions
5. Lift: It says how likely item Y is purchased when item X is purchased while controlling for how popular item Y is. It is defined as the ratio of the observed support to that expected if the two rules were independent. The greater the lift ratio, the stronger the association of the items. It is given by:

   Lift Ratio = (((Precedent Item + Consequent Item) / Precedent Item) / (Consequent Item / Total No. of Transactions))
6. Test Stores: A test store is the store where the test was performed.
7. Control Stores: selected stores that have similar observed sales trends as the test stores.

Proposed Framework

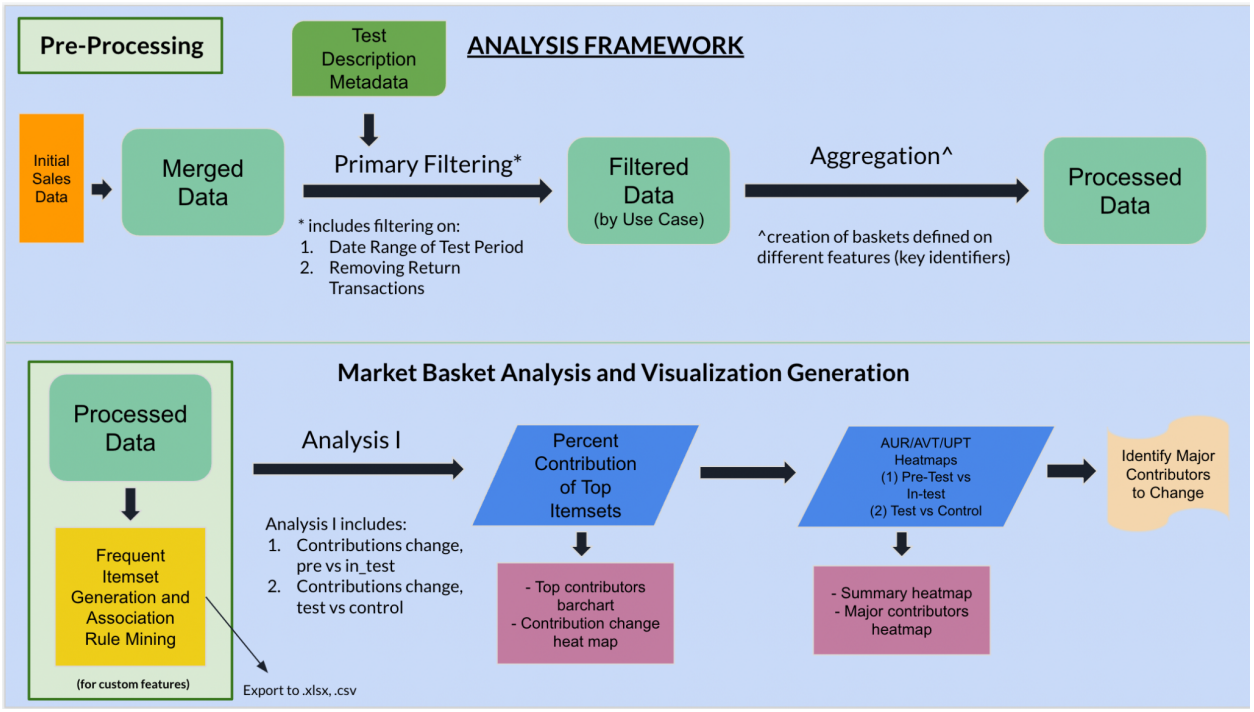


Figure 1: Analysis Framework

Our framework has been divided into two sections primarily: Pre-Processing and Market Basket Analysis (MBA) + Visualization Generation.

The first section, Pre-Processing handles all the raw sales data (which amounts to very large volumes divided into several part files) which needs to be merged per use case for further analysis. This involves an exhaustive data ingestion process and filtering mechanism. We define a DataReader class for the data ingestion process. Each DataReader object is supposed to iterate through a different kind of dataset. For each operation a separate function that performs grouping is written. The *iterate* function inside the class applies this function to each file and concatenates the output from all the parts to give the merged dataset. This merged dataset can now be filtered using a lot of test description metadata such as the stores the test was applied to, the date time period the test was performed, and other flags such as removal of returned transactions or those that have missing units or sales values. At the end of this we have filtered data per use case which is then fed as an input to an object of the DataPreProcessing class. Here we generate baskets based on an array of key identifiers for which it will be useful to conduct further analysis.

The processed data is an input to the second section, i.e. MBA and visualization. The basket level data is key to generation of sparse matrices needed for frequent itemset generation using the Apriori algorithm, which is then used for Association Rule Mining. We then individually conduct analyses through extensive visualization generation, both summarizing and detailing on finer granularities. We generate summary bar graphs to identify percent contributions of relevant items and how their contributions change across test stores during the test period. We further generate item-wise heatmaps for metrics such as AUR, AVT, UPT etc. to get a better understanding of the major contributors of change in user experience.

These outputs are relevant to the business teams to make informed decisions about the success of these tests and other recommendations they can make from the insights generated through this framework.

Methodology

Data Pre-Processing

One of our first hurdles in this project was dealing with the large amount of sales and item inventory data, amounting to 9 GB and 3 GB respectively. The size of these files meant that all the data could not be read in at once into our AWS SageMaker instance's memory.

Thus, we have created a comprehensive DataReader, implemented and encapsulated as a Python Class. This DataReader is not only capable of reading in part files for sales, item inventory and customer information, but is also able to perform all it's computation in parts as well, resulting in efficient memory usage. The initialized DataReader class is then capable of performing any function on the part files, including filtering, initializing in memory, storing intermediate results and combining results in a MapReduce fashion.

For the purposes of the use cases we have at hand, one function that has found a great deal of utility is the filter function. Any testing set-up of this kind would most likely have a set of stores where the tests were conducted, a corresponding set of stores which were held as control, and the date ranges for which the tests were conducted. In addition, these tests might affect certain departments or sub-departments within the store only. The filter function is able to account for all these possible scenarios, only returning data points that are of relevance to the test at hand.

This DataReader has been immensely useful in the ingestion, filtering and loading of the data pertinent to all the use cases we have made progress on, since these operations are a natural foundation from which downstream processing can then be carried out.

Analytical Methods

1. **Exploratory Data Analysis and Visualization:** In order to understand the data, trends, patterns, clusters and veracity of the data, we have performed comprehensive Exploratory Data Analysis for each of the use cases we are interested in. A selection of these results, which are highly pertinent to our individual use-cases have been presented here.

2. **Apriori Algorithm:** The Apriori algorithm is used for frequent itemset mining and learning association rules over transactional data. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database, using a 'bottom-up' approach. Each transaction is seen as a set of items (an itemset). Given a threshold *C*, the Apriori algorithm identifies the item sets which are subsets of at least *C* transactions in the database.

3. **Affinity analysis and Association Rule Mining:** Affinity analysis falls under the umbrella term of data mining which uncovers meaningful correlations between different entities according to their co-occurrence in a dataset. In almost all systems and processes, the application of affinity analysis can extract significant knowledge about the unexpected trends. In fact, affinity analysis takes advantage of studying attributes that go together which helps uncover the hidden patterns in big data through generating association rules.

4. **Percentage Contributions and Volumetrics:** For each frequent itemset, we also found the percentage of occurrence of other items and supplemented our overall analysis with volumetric information in terms of actual numbers in order to understand test effects fully.

Results

Use Case 1 - Mens Solid Colored T-Shirt Removal Test:

In this use case, we found that the test for removing the S/S solid colored T shirt and replacing it with additional colors Fancy and Fleece tops was not as helpful. We come to this conclusion based on the following analysis:

1. In terms of the percent of baskets which contain at least one S/S Solid tees were completely vanished in the test period, say this was x % of baskets. We would ideally expect to see these to be replaced with the additional choices provided, however, we see that the % of baskets containing these replaced items were slightly higher than the pre-period but not enough. Say it increased by y% but y is very less than x.

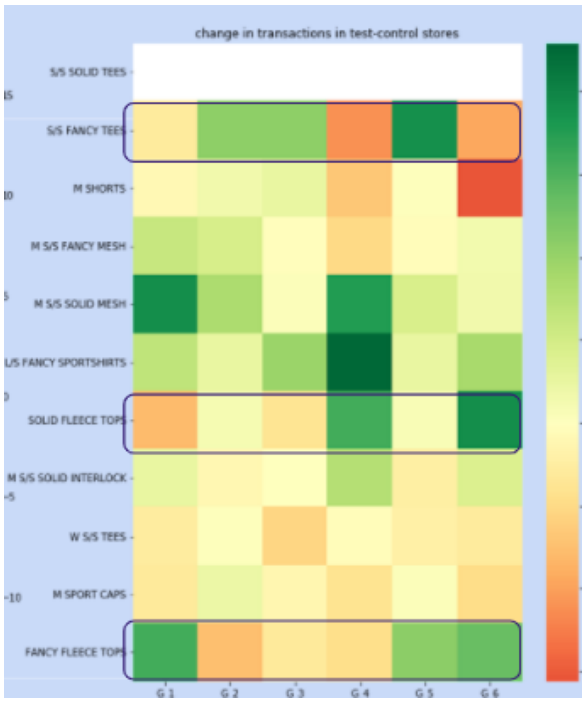


Figure 2: Top 11 Items performance in % of baskets

2. Also, when compared between test and control stores, the difference was again negligible. Implying the addition of new did not certainly help a lot to drive more sales, nor did the customers replace S/S solid tees with these items.
3. We expanded our analysis to the top 11 items and considered items till the top 5% of all baskets. We again saw similar results for all.
4. Then, to compare the impact of these individual items we verified using the AUR values of each. We again saw that there was a decrease in the value for AUR in the test period than in the pre period when compared to the control stores. We suspect this is due to the decreased number of transactions.

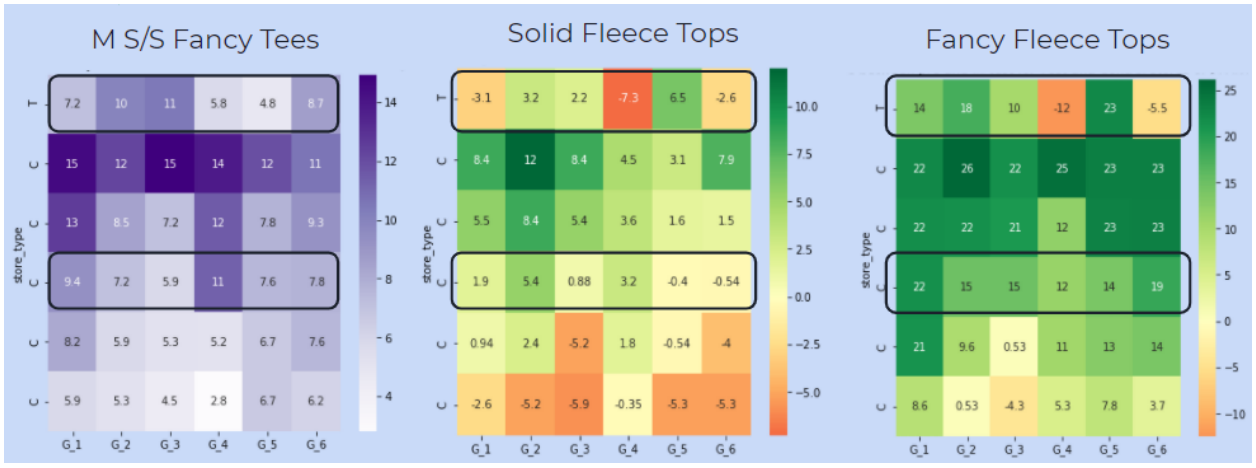


Figure 3: Top 11 Items performance in relative AUR

5. Next, We performed affinity analysis. Here we did the same comparison as we did previously with the top 5 items affiliated with S/S Solid Tees. 2 among these top 5 items performed better in terms of percentage of baskets and in AUR.

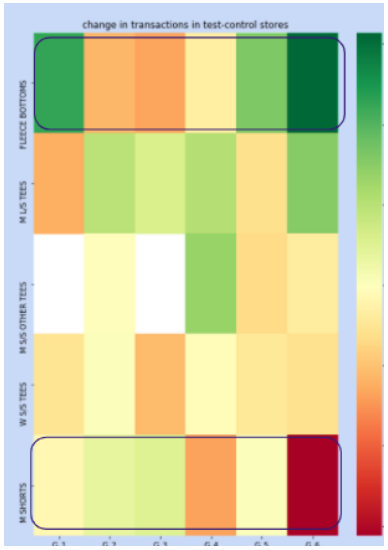


Figure 4: Top 5 affiniated Items performance in % of baskets

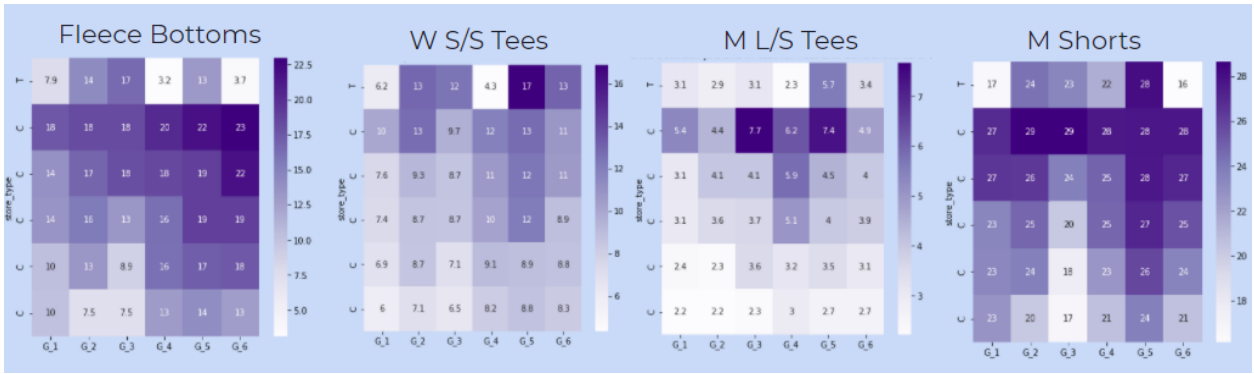


Figure 5: Top 5 affiniated Items performance in relative AUR

6. However, these 2 items were M L/S Tees and M S/S Other Tees which did not have a strong quantity of baskets. Hence, did not lead to a strong change in AVT.

Figure 6: Relative AVT for all baskets with M Clothing Items

Use Case 2 - Childrens Promotion Removal Test

In this test, we wanted to explore how the sales of children's items were affected by the removal of promotional items from the children's department. Therefore, a major portion of the analysis dealt with analysing the baskets which contained children items only. We did so by filtering based on the group division ID. The following are the insights obtained from the data provided:

Summary Insights

1. Overall net sales across all products in the children's division saw a reduction of a staggering 68% in the control stores and 50% in the test stores.

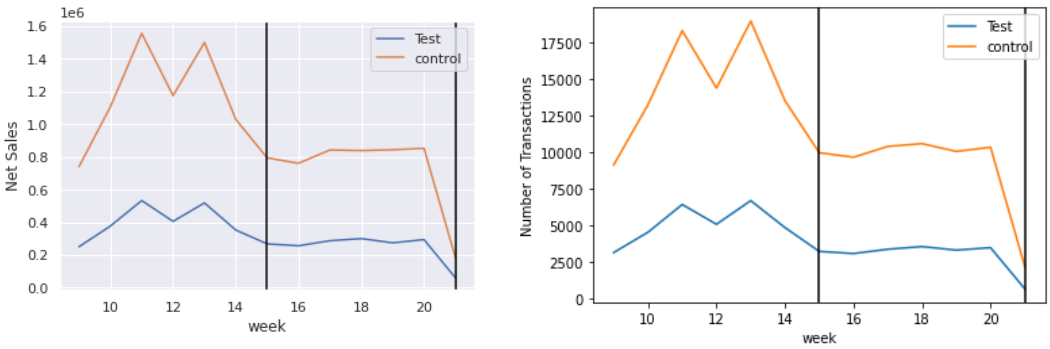


Figure 7: Diagram showing the weekly net sales (left) and Weekly # of transactions (right)

2. 11 out of the 17 test stores underperformed the corresponding control stores by approximately 0.5 Units per transaction while the remaining stores overperformed by 0.2 UPT.

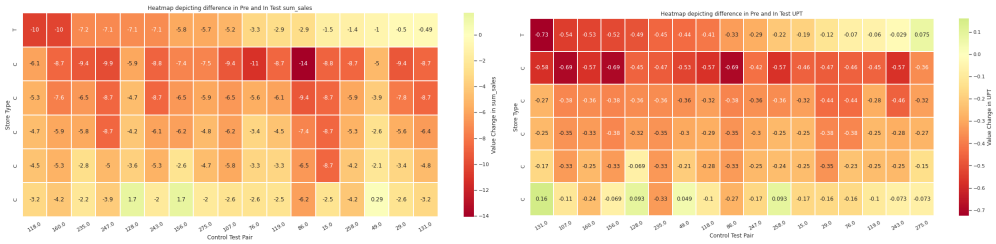


Figure 8 : Heatmap diagrams showing the Net sales(left) and UPT difference for all transactions

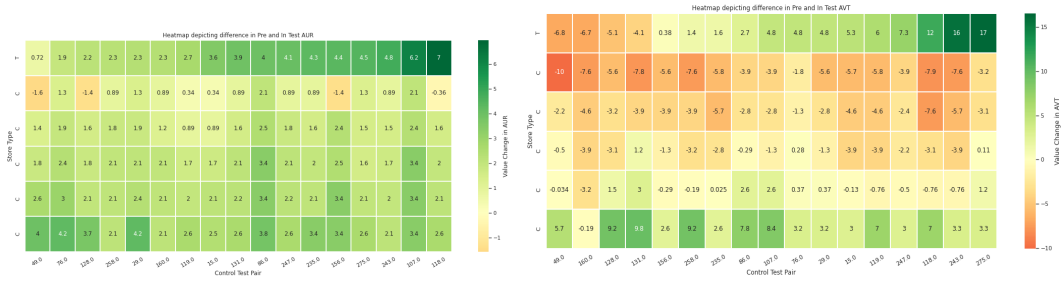


Figure 9 : Heatmap diagrams showing the AUR(left) and AVT difference for all transactions

Formula used for each of them is: (Pre_value of Metric for store)-(Post_value of metric for store)

Affiliated Departments Insights

1. Analysing affiliated departments helps remove seasonality differences to some extent. This is so because departments are grouped on the basis of age groups and not the clothes. The most affiliated departments identified as a result of the test were

    - Boys 2-7, Boys 8-20, Girls 2-6X, Girls 7-16 among the children departments
    - Mens tees/Fleece, Womens Tees/Fleece, Mens Knits, Mens Bottoms under non childrens divisions.

Based on the feedback obtained, we went ahead and analysed the children's departments only.

2. For Boys 2-7, 12 Stores underperformed in comparison to their corresponding control stores while the remaining outperformed. For Boys 8-20, 11 of them underperformed by an and the remaining outperformed the corresponding control store.

Formula : (Post UPT - Pre UPT)_TEST - (Post UPT - Pre UPT)_CONTROL

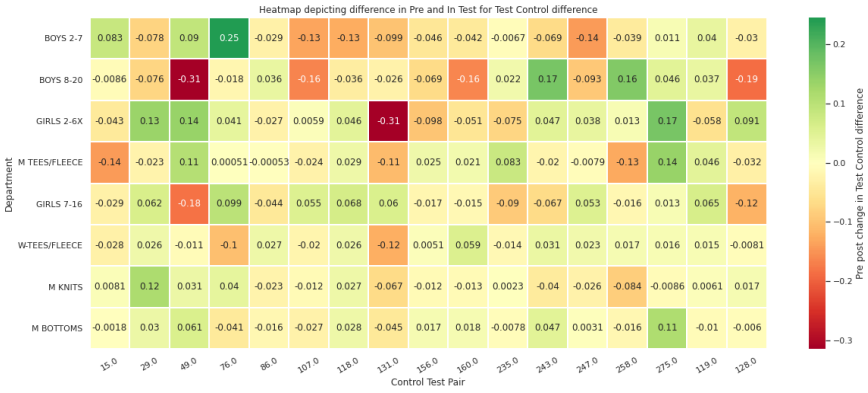


Figure 10: Diagram showing the UPT pre post analysis for the top affiliated departments

Here pre implies values before the test and post implies values after the test.

3. In terms of Average Unit Retail, the general trend observed was that the Test Stores outperformed the corresponding control stores.
4. Coming to the departments, In Boys 2-7 14 out of the 17 test stores outperformed the corresponding control stores, while in Boys 8-20, 12 of the Test stores outperformed corresponding control stores.
5. Stores of the West corridor comprehensively outperformed the stores of the eastern corridor by a relative AUR of $2.1.

Formula : (Post AUR - Pre AUR)_TEST - (Post AUR - Pre AUR)_CONTROL

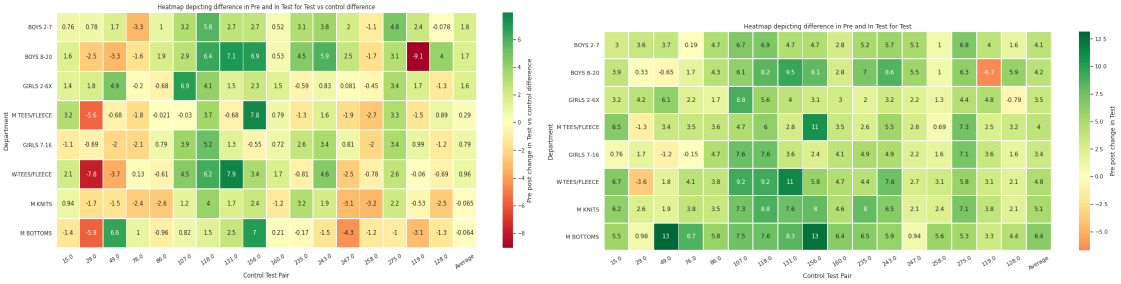


Figure 11: Diagram showing the AUR pre post analysis for the top affiliated departments and the corresponding control store

6. In terms of Average Value per Transaction, the general trend observed was that the Test Stores outperformed the corresponding control stores.
7. Anomaly was observed in Boys 8-20, wherein a mere 9 out of Test stores outperformed the corresponding control stores while 8 underperformed. Boys 2-7 shows results as expected with 13 test stores outperforming control.
8. Girls 2-6X and Mens tees/fleece were two other departments that showed significant increase in AVT.

Formula : (Post AVT - Pre AVT)_TEST - (Post AVT - Pre AVT)_CONTROL

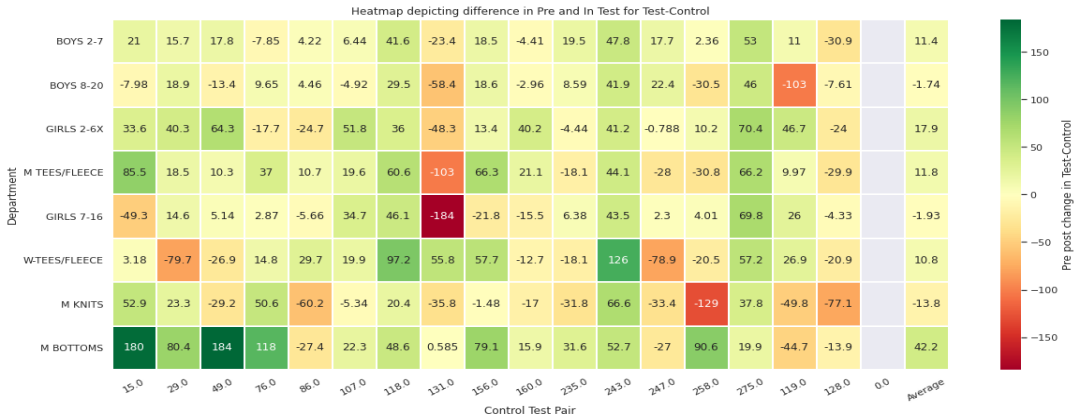


Figure 12: Among affiliated baskets, Metro Stores were favored by AVT while domestic, beach and rural stores saw a reduction in AVT.

Use Case 3 - Children's Size Merge Test

For this use case, the following actions were taken during the test period in the children's department:

- Boys: Sizes 2-7 (Little) and 8-20 (Big) merged into one region in the test store. Control stores' configuration remained the same
- Girls: Sizes 2-6X (Little) and 7-16 (Big) merged into one region in the test store. Control stores' configuration remained the same

This was done with the expectation that this would lead to a cohesive shopping experience, with new assortments with family and corresponding increases in UPT, thus leading to a greater AVT as a result.
The approach for this use case was to conduct a thorough subdept level analysis during the pre and during test period for both test and control stores, so that we have full comprehension of all trends. We also looked at frequent itemsets in both the boys and girls category and chose to focus on this category. For these frequent itemsets, we also found their correspondingly most affiniated categories and trends and looked at the changes in those categories as well. As before, we only considered baskets having at least one children's section item in them. All other baskets in the dataset were discarded since they are irrelevant to our analysis. In order to supplement this basket percentage-wise analysis, we also looked at overall volumetric in terms of total units sold in order to have the complete picture in terms of both percentages as well as actual numbers.

Some of these frequent itemsets are shown in the following figure:

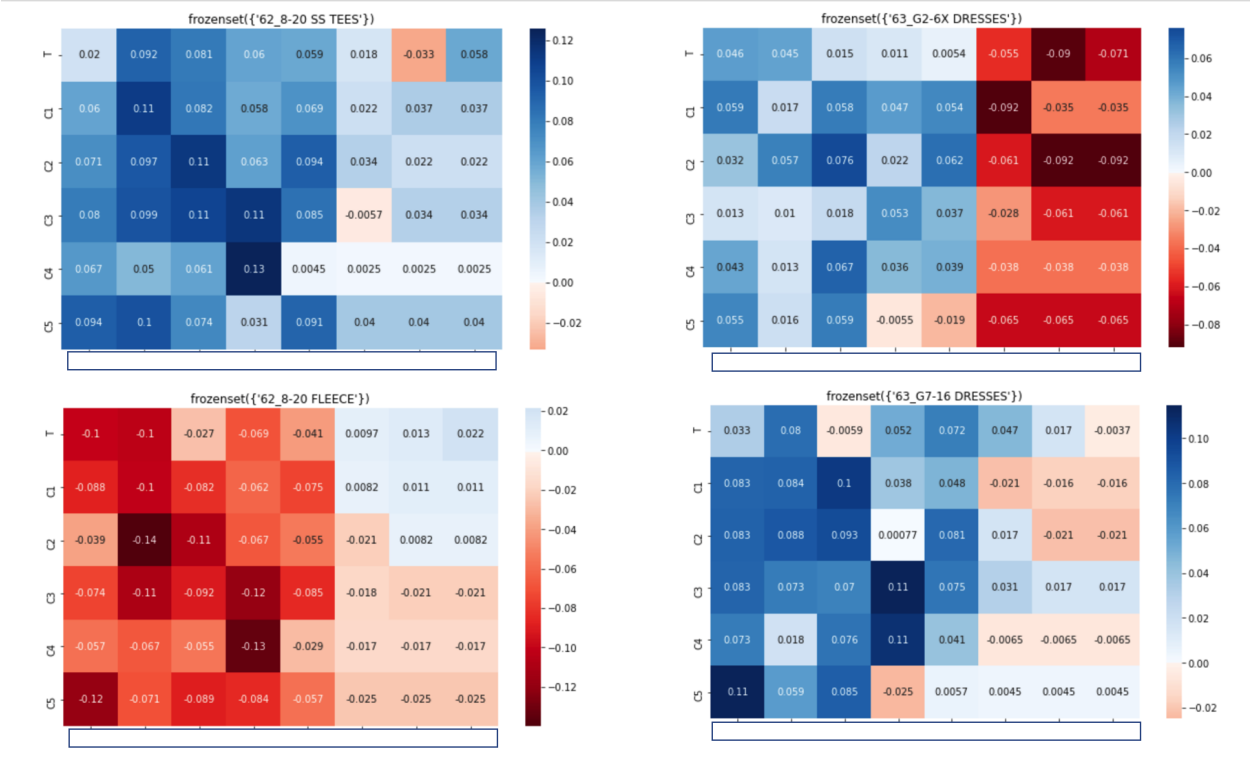


Figure 13: Difference in support values for each test-control group from test to pre-test period

These heatmaps on the left column are for the boys department and the ones on the right are for the girls department. Each column corresponds to one Test-Control Group (defined as test store + it's

corresponding stores). The rows correspond to each one of these stores, with the test store being marked as T in the top row. The value in the boxes corresponds to the difference percentage between the period before the test and the period during the test. For instance, we see that 8-20 SS Tees were found in 2% more baskets during the test period compared to the period before the test for the test store. We can see that there are no marked differences between the test and control stores on an overall basis for these sub departments. However, we see that the last 3 columns often have very different trends in comparison to the rest of the stores. This is due to the fact that these stores are all of a different type and contain only children's items, whereas the other stores also contain items from the mens' department for example. Having completed this exercise and found the frequent itemsets along with the fact that there seems to be no real discernable difference between the test and control stores, we now find important association rules, or what we call "driving factors" for these frequent itemsets.

In particular, we are interested in cross-size rules since these cross-size shopping customers are who are most likely to be affected by this test. What we found from this analysis was that the cross-size shopping rules were quite weak. On further analysis, we found that the reason for this is that most of the transactions occurring for the children's department only contained one size. The percentage of transactions containing more than one size for the children's department was less than 10%. This meant that for the stores also containing the men's and womens' department, the rules for the cross-size items were extremely weak, if significant at all. Thus, we chose to focus our further analysis on only the stores which contain only children's items in order to derive strong insights.

We also conducted a full AUR, AVT, UPT analysis for both the boys' and girls' department to support this, and the results are shown in the following figure.

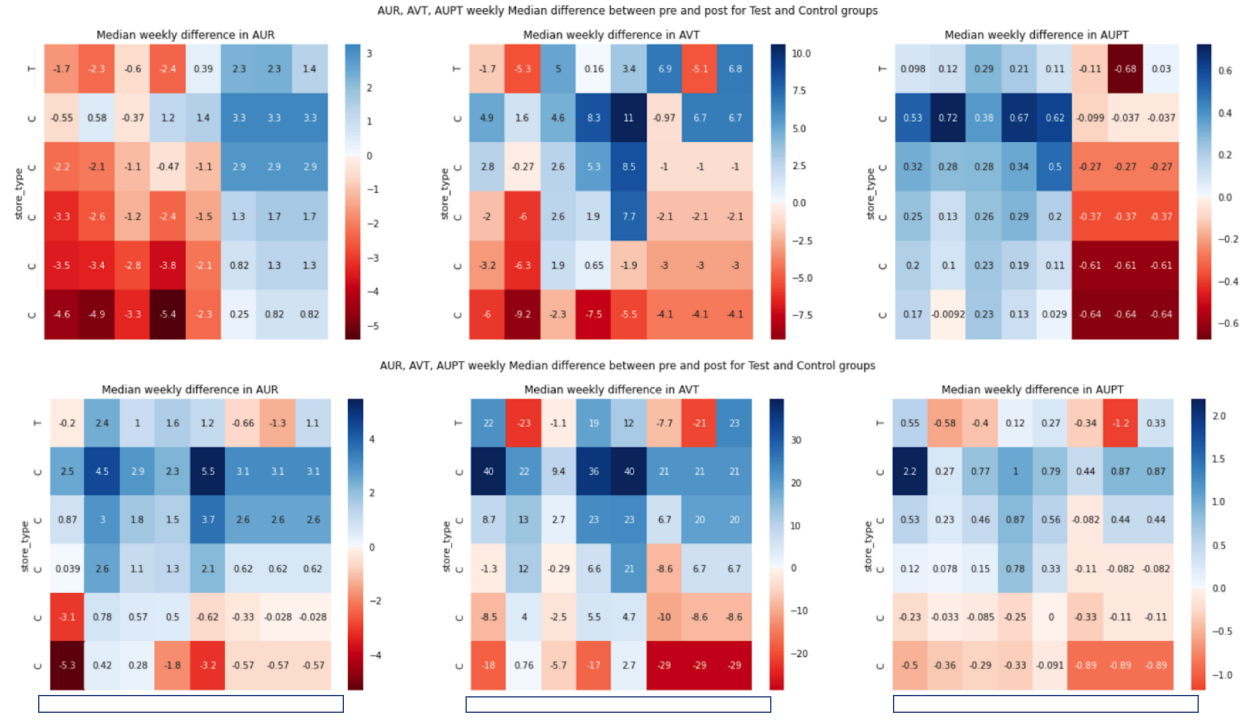


Figure 14: AUR, AVT and UPT differences for inter-size shopping for boys (top) and girls (bottom)

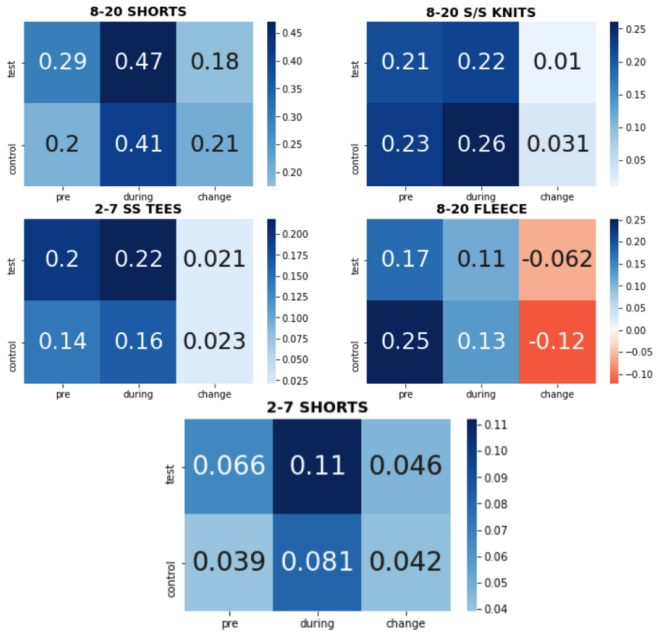
As before, each column corresponds to one test-control group, with each row corresponding to a single store, with the test store being on the first row. The value inside the heatmap depends on the heatmap. The value denotes the difference between the value of the metric (AUR/AVT/UPT) during the test and the

value of the metric before the test. Of course, the duration of the period prior to the test was adjusted to be equal to the duration of the test itself.

What we gather from this is the fact that there seems to be a split result once again in terms of AVT, AUR and UPT, and we see inconclusive results. What we do gather, however, is that the children-exclusive stores clearly follow different trends from the rest of the store groups only for the boy's department, but not the girls department.

Diving deeper, we also looked at each individual sub department and its corresponding affiliated items. We analysed what was the percentage of this affiliated item occurring in baskets having this frequent item. We also show this analysis in terms of raw numbers in a stacked bar chart as shown below.

We also see that for the 8-20 SS Tees subdepartment, for the test store, the period prior to the test saw 8-20 Shorts being included in the baskets 29% of the time. However, for the period during the test, we see that this percentage increases to 47% of baskets, an increase of 18%. Similarly, for the control store, we see a 21% increase for this particular test-control group. This is also reflected in the volumetric analysis in the stacked bar chart below. The two sets of bars correspond to the test and control stores respectively. The y-axis is the number of items sold. Within each set, the bar on the left shows the number prior to the test period and the right bar shows the number during the test period. We see a greater co-occurrence of 8-20 shorts with 8-20 SS Tees during the test period for both the test and control stores, as seen by the larger bottom bar. Thus, the volumetric analysis agrees with the basket analysis.



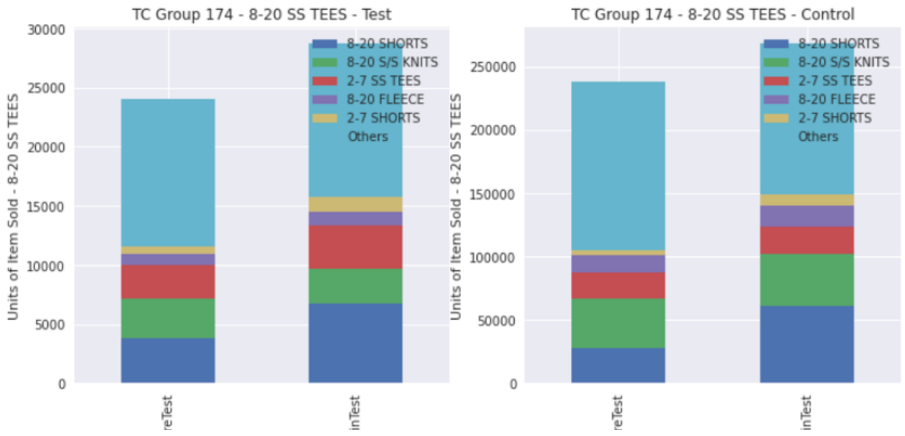


Figure 15: Frequent Itemsets Analysis and Volumetric Analysis

We conducted similar analysis for every single one of the frequent itemsets. The overarching results are presented in the conclusion section.

Use Case 4 - Footwear Assortment Expansion Test

In this use case, there were 5 types of footwear display types: boxes on the floor, boxes off the floor, from our home to yours (FOHTY), try on ship to home, and full service door. The rollout period was during Q2 (April - June) and Q3 (July - September).

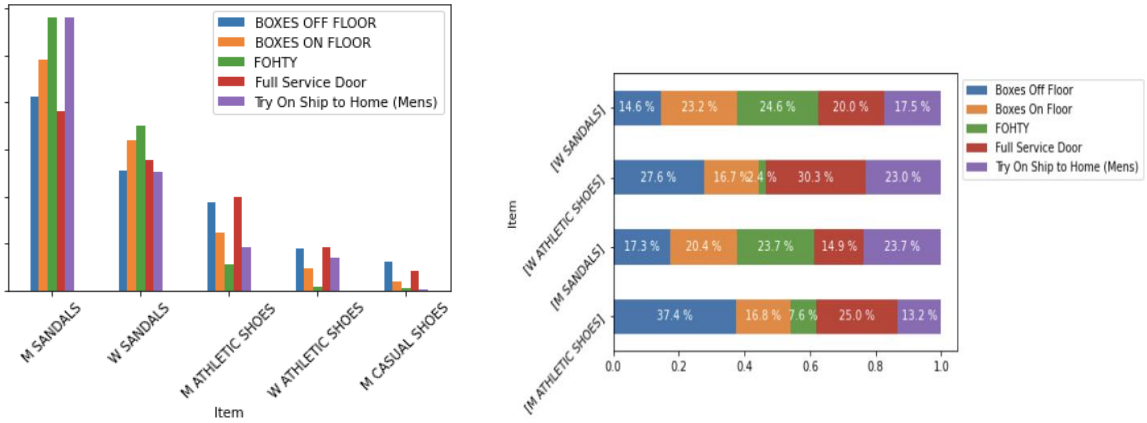


Figure 16 : distribution of footwear sales in rollout period (left) and Purchase likelihood of footwear item by footwear display type (right)

In our analysis, we found that few of all the transactions done in the rollout period (Q2-Q3) included an item of footwear. From those transactions, sandals were the most popular footwear to buy. Figure 15 (left) shows that for every display type, the most popular items sold were men's and women's sandals. Athletic shoes were below sandals in terms of popularity and casual shoes were below athletic shoes. We also noticed that most footwear baskets contained only one pair of footwear. Sandals were slightly more likely to be bought from stores with boxes on the floor display or FOHTY display, and outdoor stores near beaches, rural areas, and metro areas. Athletic shoes were more likely to be sold from stores with a full service door display or boxes off the floor display and indoor domestic/international stores. (Figure 15 (right)-16) In these figures purchase likelihood represents the chance a sold footwear item was purchased from that store type.

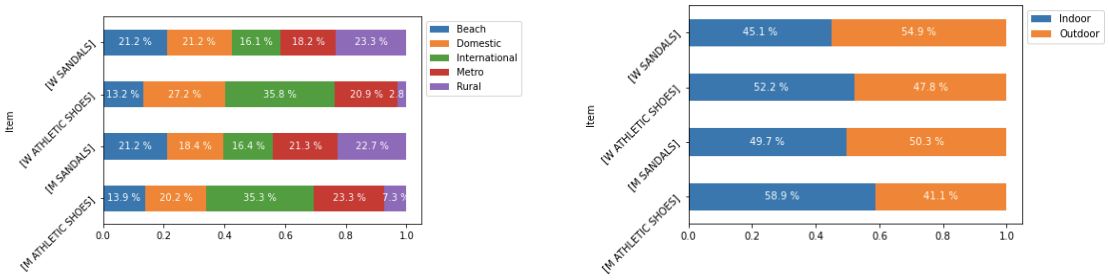


Figure 17: Purchase likelihood of footwear item by destination(left) Purchase likelihood of footwear item by indoor/outdoor (right)

In instances where footwear was bought with another item, men's short sleeve mesh tees, hats, swimwear, men's socks and underwear are being purchased with men's sandals and women's short sleeve tees, dresses, women's knit bottoms, and women's socks are being purchased with women's sandals. For these items, we saw increases in AVT for most of them except for men's underwear and hats. We saw a particularly high increase in AVT for dresses in both Q2 and Q3. (Figures 17-18)

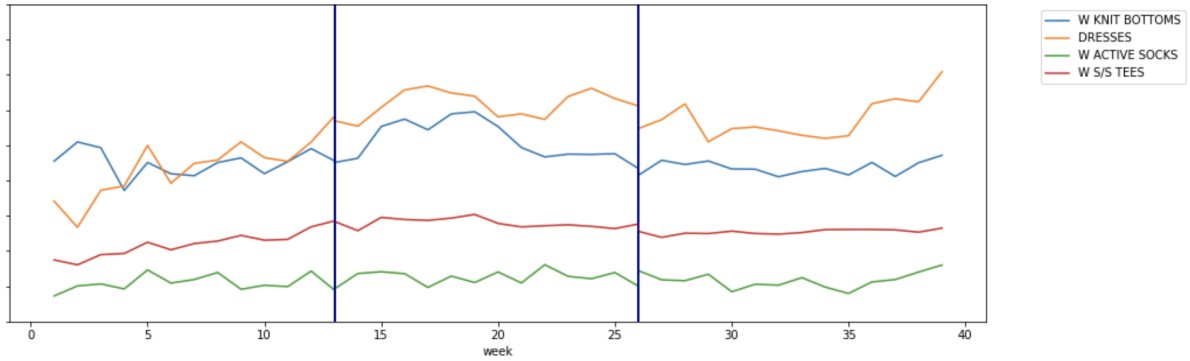


Figure 18: AVT of items affiliated with women's sandals

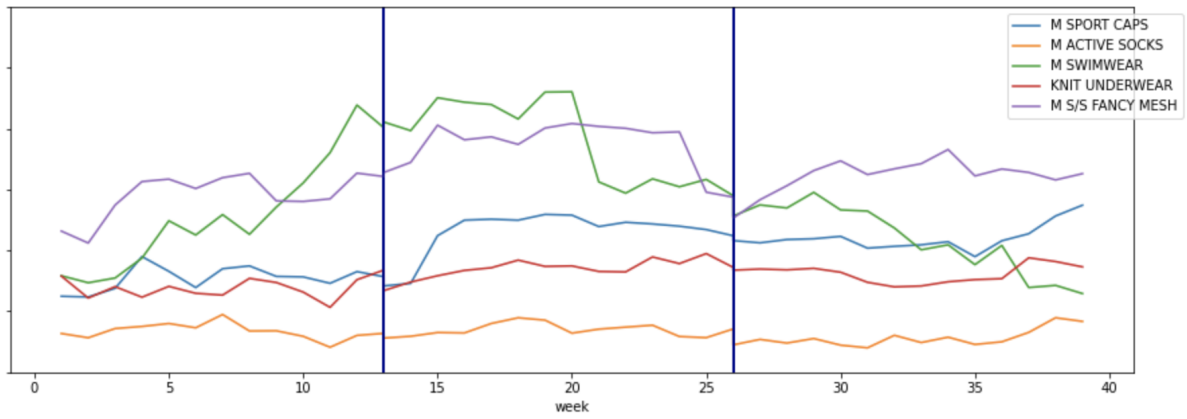


Figure 19: AVT of items affiliated with men's sandal

Conclusions

Use Case 1 - Mens Solid Colored T-Shirt Removal Test:

Our Analysis as explained previously inclines with not removing the S/S Solid Tees and replacement with Fleece and Fancy options. Additionally, we also observe an overall decrease in Fleece Tops sales during the test period. Based on the time of the year the test was conducted we hypothesize this might be due to Seasonality change. However, M S/S Solid Mesh is one category which saw a better performance than the rest, despite no change in their composition. Also, we observe them being highly affinitated with other Tees used in the test. Hence, we recommend using M S/S Solid Mesh as a better option to replace S/S Solid Tees for future such use cases.

Use Case 2 - Childrens Promotion Removal Test

The conclusions boil down to two components: Did the promo removal test have an overall impact on the sales, and What department/items were affected as a result of the analysis. Coming to the first component, The expected results of the Children promotion test was that the number of transactions would go down, but the question of whether net sales would be affected or not was in question. The analysis reveals that the number of items did see a reduction, but the overall net sales weren't really affected as a result of the test. Coming to the second component of the analysis, we observed that the Boys 8-20 department saw a great amount of movement across all Test control pairs. Furthermore, Girls 2-6X saw a marginal increase in the units per transaction (when compared to corresponding control stores). The AVT, AUR increase was higher in test than control stores while the UPT increase was more in control than test stores.

Use Case 3 - Children's Size Merge Test

What we found after the entire analysis is as follows. For inter-size shopping (Ex. 2-7 and 8-20 being bought together), test store and control store patterns are largely identical, meaning that people are not picking up more across sections just because the items were placed together. We know found that the antecedent-consequent pairs ie. driving factors/rules for inter-size shopping are quite weak, showing minimal change from pre to during test. Thus, we can conclude that 2-7 size items are not driving 8-20 size items sales and vice versa for the boys' section. They seem to be independent of each other, but co-occurring often. Thus, placing them together had no effect on affinities/sales. A similar trend is found in the case of the girls' section for the 2-6X and the 7-16 sizes. For intra-size shopping, overall rules are stronger. We also see that driving factors are stronger. However, other than seasonality, there are no overall affinity changes with both test stores and control stores following very similar patterns. Thus, we can conclude that intra-size shopping was largely unaffected by the test. This is as expected, since placing another section close, had no effect on shoppers only looking to pick up from one section only.

Use Case 4 - Footwear Assortment Expansion Test

Through our analysis we recommend that sandals, the most popular footwear type, be introduced in all stores, especially at outdoor, rural, beach stores with boxes on floor and FOTHY display. We also strongly recommend athletic shoes, the second most popular footwear type, be introduced in international, domestic, and indoor stores with full service doors and boxes off floor display Other types of shoes, on the other hand, were not very popular and should be considered for only select stores. Additionally, we also found that most customers opt to buy only footwear. The items that had the most affinity with footwear sales were socks, dresses, underwear, caps, and men's short sleeve mesh tees, and women's tees.

All items except for M active socks and M swimwear saw increases in AVT, particularly in Q2 of the rollout period.

Future work

To generate additional insights into our data and to confirm some of our insights, it would be helpful to conduct certain statistical analyses and causal inference methods. These analyses could help us further understand the impact of changes in the total sales of a store.

Furthermore, in addition to sales data, we would also like to incorporate other datasets into our analysis to understand other aspects of customer purchase behavior. This could involve customer level data such as age and demographics, whether they are a returning customer, if they have kids, etc. We could also include store level data such as geographic information, footfall and other key features that could drive another dimension of insights.

Ethical considerations

Given that the data we were given is real-world data collected from Ralph Lauren stores, we need to make sure that the personal identifiers such as contact information, names, addresses and emails are removed. The customer data we received did not contain any of these personal identifiers. That being said, we still had information in the customer data regarding income level and gender information. With this, we needed to ensure that our analyses and recommendations made to Ralph Lauren did not include any inadvertent information which could lead to the alienation of one or groups of people, having unintended and serious implications for Ralph Lauren as a brand. Our decision to use the customer data very sparingly meant that we did not have this possible bias enter the data we were working with.

References

1. https://pbpython.com/market-basket-analysis.html
2. https://towardsdatascience.com/affinity-analysis-market-basket-analysis-c8e7fcc61a21
3. https://en.wikipedia.org/wiki/Affinity_analysis
4. http://www.cse.msu.edu/~cse960/Papers/MiningAssoc-AgrawalAS-VLDB94.pdf [Apriori Algorithm]
5. https://www.kaggle.com/akhilram7/affinity-analysis-of-market-basket
6. http://www09.sigmod.org/disc/disc99/disc/sigmod_papers/slisp_efficiently_mining_l/slides.pdf