**Portfolio Project**

**Business Analytics in SAS Studio- Clothing Retail Store & Distribution Company**

Joshua Humphreys

Colorado State University Global

Course Code:MIS543

Dr. Richard Morse

August 4, 2024

**Portfolio Project**

For this project I have conducted analysis of two datasets representing sales and orders from a clothing retail and distribution company. In the following pages I will describe my understanding of the business problem including the company’s strategic goals. I will then introduce four business questions (two for each dataset), which, if answered, will solve the organization’s business problem and achieve the strategic goals. Accompanying each business question are corresponding null and alternate hypotheses. I will present descriptive statistics followed by predictive statistics relevant to the datasets and questions. I will then provide an analysis of my findings with respect to the business questions and corresponding hypotheses. Finally, I will conclude by providing recommendations for further data collection and analysis.

**Business Problem and Strategic Goals**

Like most (probably all) businesses that sell products, this organization’s goal is to increase the quantity and profitability of sales orders. My understanding of their strategy is they intend to leverage data analytics to determine organizational growth opportunities and enhance the efficacy of directed marketing promotions. Armed with this information, they will be able to ensure that their application of resources to marketing promotions and growth yield greater profits.

My understanding of the business problems of the organization is that their previous marketing promotions and growth efforts have not been as effective as they expected. That is, their desire to identify those customers that respond to marketing promotions to increase the effectiveness of those promotions implies that previous efforts may have been inefficient. Furthermore, stating a desire to predict future business growth implies that they may have invested resources in growing the organization in ways that did not increase profit, thus wasting valuable resources.

**Business Questions and Hypotheses**

The organization has already identified that they want to increase the effectiveness of their marketing promotions by identifying which customers respond to marketing promotions. In the context of the sales dataset, customer clusters describe a group of customers already identified as having similarities. The question and associated null and alternate hypotheses that captures both organizational goals in context is as follows:

**BQ1**: Which customer clusters should we focus marketing promotions towards?

**H10**: There is no relationship between customer clusters and marketing promotion response.

**H1a**: A significant relationship exists between customer clusters and marketing promotion response

While directing marketing promotions at responsive customers is likely to increase marketing effectiveness, ensuring the marketing content is relevant is likely to increase effectiveness as well. Therefore, a second question with null and alternate hypotheses for the sales dataset is as follows:

**BQ2**:Which product categories should we focus marketing efforts on?

**H20**:There is no relationship between product categories and marketing promotion response.

**H2a**: A significant relationship exists between product categories and marketing promotion response.

The sales dataset seems capable of answering questions related primarily to the marketing promotion goals of the organization. However, the other objective of the organization is to identify opportunities for growth. The implication of this goal is to efficiently allocate resources to promote growth in areas likely to increase profitability. The orders dataset contains data related to business transactions across the span of the organization’s influence. Questions related to the “big picture” are more strategically relevant. That is, business strategy is more aligned with business territories and product categories than, say, individual customers and products. Therefore, the following questions of the orders dataset, if answered will assist the organization in taking advantage of growth opportunities:

**BQ3**:Which territories should we invest in?

**H30**:There is no relationship between territories and profits.

**H3a**: A significant relationship exists between territories and profits.

**BQ4**: Which product categories should we invest in?

**H40**: There is no relationship between product categories and profits.

**H4a**: A significant relationship exists between product categories and profits

**Descriptive Statistics and Feature Engineering**

My analysis of the sales dataset primarily involved the following variables: *MON*, *DAYS*, *FRE*, *CLUSTYPE*, and each of product category variables. Therefore, I started my analysis by producing summary statistics and histograms for *MON*, *DAYS* and *FRE*. I also calculated sums for each distinct value in the *CLUSTYPE* variable. The results of these initial statistics are found in figures 1 through 4 below.

**Figure 1**.

*Distribution of* MON

A graph with numbers and a number of data

Description automatically generated with medium confidence

**Figure 2**.

*Distribution of* FRE

A graph of a distribution of a number

Description automatically generated

**Figure 3**.

*Distribution of* DAYS

A graph of a number of days

Description automatically generated

**Figure 4**.

*Freq count of CLUSTYPE*

*A graph of a number of bars

Description automatically generated*

The first two business questions are related to marketing promotion response. Customers can respond to marketing promotions by buying clothing more frequently or by spending more money. Both of those actions are individually represented in the dataset. However, there isn’t a single variable in the dataset that captures both forms of promotion response. To address this problem, I created a new variable called *promo\_resp.* My original plan to create this variable was simply to average the normalized, per-customer, average daily sales and average daily visits. However (and not surprisingly given the distributions of MON and FRE), when I produced this number, the distribution was very right skewed with relatively low mean and low standard deviation. To normalize the variable, I first took the natural logarithm of the original value, divided by the minimum of the new values, and added 1. The resulting distribution is depicted in figure 5 below.

**Figure 5.**

*Distribution of promo\_resp*

**A graph with a blue line

Description automatically generated**

**Predictive Statistics**

To start my predictive statistics, I performed an analysis of variance (ANOVA) of *promo\_resp* with respect to *CLUSTYPE.* I produced an LS-mean difference plot, box plot, and means plot. Those plots are found in figures 6, 7, and 8 below.

**Figure 6.**

*LS-mean difference plot*

*A graph with blue dots and numbers

Description automatically generated*

**Figure 7.**

*ANOVA Box Plot*

*A diagram of a distribution of numbers

Description automatically generated with medium confidence*

**Figure 8.**

*Means Plot*

*A screenshot of a computer

Description automatically generated*

I then calculated the dollar amount for each clothing type by multiplying the current value by *MON.* After that, I performed correlation analysis between each clothing category and *promo\_resp*. I produced Pearson Correlation Coefficients and scatter plots for each correlation test. The correlation coefficients and scatter plots are in figures 9 and 10 below.

**Figure 9.**

*Pearson Correlation Coefficients*

*A screenshot of a computer

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

**Figure 10.**

*Scatter Plot Matrices*

*A diagram of a scatter plot matrix

Description automatically generated*

*A screenshot of a graph

Description automatically generated*

*A diagram of a scatter plot matrix

Description automatically generated*

Lastly, I simply calculated the sum of *net\_sale* for each territory then each category. I produced bar graphs to visualize those sums. The bar graphs can be found in figures 11 and 12 below.

**Figure 11.**

*Net Sales by Territory*

*A graph of numbers and columns

Description automatically generated*

**Figure 12.**

*Net Sales by Category*

*A graph of a number of bars

Description automatically generated with medium confidence*

**Analysis**

Prior to conducting my analysis, I went to great lengths to ensure that the promotion response variable reasonably closely followed a normal distribution. The justification for this is that as we compare promotion response to other variables, we can be sure that a distribution that deviates from a normal distribution can be attributed to the predictor rather than deviations in promotion response.

Examination of the ANOVA test reveals notable variances among clusters relative to promotion response. For instance, cluster 19 had a very high LS-mean. Clusters 18, 22, 36, 41, 42, and 44 also show relatively higher LS means. This indicates that certain clusters have higher promotion responses. In reference to the first business question: “which customer clusters should we focus marketing promotions on?” We clearly see that we can reject the null hypothesis and conclude that certain customer clusters are likely candidates for focusing marketing efforts. With that said, if we refer to figure 4, frequency count of *CLUSTYPE,* we see that clusters 36, 41, 42, and 44 all possessed relatively low overall frequencies. In other words, in terms of cluster types where we’ve collected more data, it seems that clusters 18 and 22 are likely to respond to promotions.

The next business question is about which product categories we should focus marketing efforts on. The null hypothesis is that there is no relationship between product categories and promotion response. However, figure 9, which depicts the correlation analysis between categories and promotion response, clearly demonstrates significant correlations among product categories and promotion response. That is, we can reject the null hypothesis. Notably, we have high correlations with dresses, blouses, jackets, and pants, while legwear, outerwear, and jewelry have low correlations.

On examination of the scatter plot matrices, we see clusters with high promotion responses in the previously identified categories. Though the differences are subtle, categories like legwear show obvious lower promotion responses. Thus, we can conclude that dresses, blouses, jackets, and pants would be good candidates for marketing investments. While investments in marketing of legwear, outerwear, and jewelry are unlikely to be as efficient.

The next business questions pertain to which territories and product categories to invest in. The null hypothesis for each is that there is no relationship between territories or categories and profits. The analysis clearly indicates that territories 20852, 27403, and 27511 have significantly higher sales than others. Similarly, category ID 4 shows higher sales than other categories. Therefore, we can reject both null hypotheses.

While it’s certainly true that sales vary by territory and category, it’s difficult to decide as to which territory and category represent the greatest opportunity. While it makes sense to invest in historically high-performing categories and territories, it’s essential to consider trends and other related questions (more on this later). Nevertheless, we can feel confident that investing in high-performing categories and territories would likely continue to yield positive results.

**Recommendations**

The first recommendation I would make is based on the data missing from both datasets. I highly recommend collecting dates associated with orders and sales. By collecting dates, we can analyze trends, investigate seasonality, and understand the direction in which sales are trending. For example, certain articles of clothing are likely to sell better during specific times of the year. Collecting time data will allow us to investigate these patterns more clearly.

Another recommendation is to gather more granular data on the nature of promotions. Instead of looking at the aggregate number of promotions delivered to a customer, we should record when each promotion was given, what it included, when the customer responded, and what they purchased. This detailed information would help us understand whether customers are responding to a promotion or if their purchase and the promotion are coincidental. Furthermore, this more granular data would help to understand the way in which customers respond to promotions.

More granular promotion data coupled with dates would help identify time lags. For instance, it’s likely that customers buy winter coats before they need them. Understanding the time lag between weather (for example) and when people buy weathe-specific clothing would help determine the optimal time to start marketing winter coats, shorts, skirts, and other season-specific clothing.

For the orders dataset, one unanswered question from this analysis is: which territories represent opportunities? To address this, we need to tie the datasets together and link more data points. We want to understand where investing will make a difference. In the same way we measure promotion response, we should analyze territories categories where we have historically invested and determine which ones responded positively.

In terms of external datasets, there's a wealth of information available related to marketing, clothing categories, customer types, regions, and more. For example, large online retail stores like Amazon have extensive data on customer behavior, the kinds of clothing they purchase, and where these customers live. This data could be invaluable in informing our marketing strategies.

Additionally, clothing tends to be very cyclical or trendy, changing with time based on the sentiment of the masses. Social media data, web scraping, and similar tools could be very useful in aggregating information on current and future fashion trends. Fashion trends tend to cycle over time. For instance, I remember bell bottoms being popular when I was young (I wasn’t alive in the 1970s). I also see young people wearing many of the things that were popular when I was young. Investigating these cyclical trends in fashion could help us predict and prepare for future trends.

**References**

Anishnama (2023, March 31). *Understanding ANOVA Test: Definition, Types, and Applications.* Medium. <https://medium.com/@anishnama20/understanding-anova-test-definition-types-and-applications-d9781b641854>

Geeksforgeeks. (2024, March 19). *What is correlation analysis?* https://www.geeksforgeeks.org/what-is-correlation-analysis/

Miles, M. (2024, May 13). *3 Problem Statement Examples and Steps to Write your own.* Betterup. <https://www.betterup.com/blog/problem-statement>

Navidi, W. & Monk, B.(2022). *Essential Statistics.* McGraw Hill LLC.