Title Slide

Bullseye: Targeting customers through analytics

Slide 2

**Abstract**

Before the internet and the convenience of shopping without leaving your seat at home, in the office, on public transportation, or in the middle of a national forest, brick and mortar establishments could rely on mostly local residents to either buy products offered or supply products for sale. The occasional customer, passing through or visiting a local resident, could be formally added to the list of customers, by some formal registry kept by the owner of the store - especially if they needed supplies delivered on a fixed scheduled - or committed to the owner’s memory. Local shop owners wanting to increase business might offer sales advertised as special overstock or special discontinued sales events. Shop owners could also strategically market seasonal products as special sales. Now, with the ability for almost any consumer to buy from almost anywhere in the world, shops can also use the internet to target specific customers. Using a dataset of customers, an organization wants to target a population of their customers who respond to direct marketing campaigns to increase sales. The dataset contains the zip codes of customers as well as the top 15 most frequently sold clothing items. Households with higher expendable incomes might seem a likely target audience for a marketing campaign on the highest selling clothing items. In this analysis, proving or disproving an initial assumption that customers who live in more affluent zip codes should be automatically included in the marketing campaign, regardless of most frequently sold clothing items, better informs marketing of their target audience based on location and clothing. SAS (Statistical Analysis Software) is the tool employed to run analyses. Summary Statistics is run to produce indisputable facts based on data pertaining to which zip codes and which lifestyle clusters spend the most and which items sell the most. One-Way ANOVAs (analysis of variance) is used to compare the means of two or more independent groups in order to determine whether there is a statistical significance in the means. A predictive analysis is run to determine which lifestyle cluster type is most likely to respond to a direct marketing campaign.

**Scenario and Business Problem**

An organization is determined to increase sales while maximizing profits through more effective direct marketing efforts. The organization needs to identify customers who respond to marketing promotions. In addition, the organization wants to discover and gain insights on how to better predict future business growth.

The business analyst is tasked to analyze the data from the data set clothing\_store\_pp\_opt1\_lc.csv. The analysis will include initial descriptive analytics tests to review what the data is initially telling us. The analysis will also should include the predictive analytics tests to assist decision makers in achieving their business goals.

**Overview of Organization**

Due to a Non-Disclosure Agreement (NDA), the author is not permitted to disclose the details of the organization’s business. For this portfolio project, the organization will be known as Big Sell. Big Sell is a global consumer products supplier that opened its doors for business in 1935. It is a publicly traded, global organization, that did over $25B in revenue in 2019, has over 16,000 employees worldwide, and is a Fortune 500 company. The term “Fortune 500” refers to a list compiled by Forbes Magazine of 500 of the largest companies in the United States. Companies can be either private or public and are ranked exclusively on annual revenues for their respective fiscal years and not on any other details of the company (e. g. benefits, location, community service, best place to work, or charitable giving.) (Hayes, 2021).

Big Sell is determined to increase sales while maximizing profits through more effective direct marketing efforts. The organization would like to cater the direct marketing effort to those customers who are likely to respond to marketing promotions. In addition, the organization wants to discover and gain insights on how to better predict future business growth.

Slide 3

**Objective**

The objective of the research is to determine which customers from the dataset are the best group to target for a direct marketing campaign.

**Overview Of Study**

The study will begin with a summary analysis of which zip codes the customers who spend the most at our stores are from. From the top zip codes with the biggest spenders, the analysis will determine which items are most often purchased and when. The business analyst is tasked to analyze the data from the data set clothing\_store\_pp\_opt1\_lc.csv. The analysis will include initial descriptive analytics tests to review what the data is initially telling us. The analysis will also include the predictive analytics tests and respective results to assist decision makers in achieving their business goals.

**Data Set**

The data set used for this analysis is clothing\_store\_pp\_opt1\_lc.csv. It has 28,799 unique records with Customer ID (cust\_id) as the unique, or primary, key. Customer ID range is from 9961 to 38759. No explanation is given for why customer list does not begin with Customer ID 1. Customer ID 9961 show s Zip Code of 0, which is not possible, so the first legitimate Zip Code is noted. The Zip Code range is from 1001 to 99387, which indicates that the organization has locations across the United States. 25,970 of the 27,999 records are from Zip Codes in the eastern half of the country, which means that only 7.24% of the customers in this data set are in the 12 states of the western half of the United States.

**Descriptive Analysis**

Having been made aware that the goal of the analysis is to create more effective direct marketing efforts, the first variable for consideration is to review where, and by whom, the highest dollars are found in Total Net Sales (total\_net\_sales), Average Amount Spent per Visit (avg\_spent\_visit), and Lifestyle Cluster Types 1, 4, 8, and 10. It would seem that initial focus would be on these. Running Summary Statistics (Appendix A) would be a better analysis to determine the more likely variables that will be used for further analysis.

Slide 4

**Analysis**

One-Way ANOVA (analysis of variance) - used to compare the means of two or more independent groups in order to determine whether there is a statistical significance in the means.

Lifestyle Cluster Type 10 and Jackets

Lifestyle Cluster Type 10 and Total Net Sales

Lifestyle Cluster Types and Average Amount Spent per Visit

Credit Card Users and Percentage of Returns

Credit Card Users and Percentage of Returns

Customers who have high Spent per Visit and Total Net Sales

Lifestyle Cluster Type who has a high Total Net Sales are likely to respond to Number of Marketing Promos on File

Predictive Analytics

Slide 5

**Research Hypothesis**

When determining a hypothesis, researchers may begin with details that are relatable. For instance, during summer months, there might be less desire to stay warm, so there is not as much a need for more expensive products like winterwear. During the winter months, however, the desire to stay warm is not a want, but a need. The winter season is a good time to push winterwear products like jackets, seaters, fleece-lined pants, snow booth, gloves, and scarves.

Combining observations between both Summary Statistics, we can begin formulating a Hypothesis. Since the highest percentage in sales of 13.39% has been on Jackets the marketing campaign can use this one item to appeal to a target demographic. After zip codes have been analyzed and Lifestyle Cluster Types have been determined, hypothesizing can begin. Since, winterwear can be a necessary expense, an assumption can be made that the biggest spenders, Lifestyle Cluster Type 10, will be the customers most likely to purchase jackets.

H0: The biggest spenders, Lifestyle Cluster Type 10, are more likely to purchase jackets.

Ha: The biggest spenders, Lifestyle Cluster Type 10, are not more likely to purchase jackets.

Slide 6

**Methodology**

According to O’Leary (2017), there is either a quantitative approach or a qualitative approach to projects. Quantitative researchers explain “phenomena by collecting numerical data that are analyzed using mathematically based methods (in particular statistics)”. Qualitative researchers are interested in understanding the meaning people have constructed, that is, how people make sense of their world and the experiences they have in the world. (Merriam, 2009, p. 13). The project will produce better results when researchers place no meanings on the data (i.e. there personal associations on products of the organization or the city or states where they do business); the data will better serve the project if there are no biases attached (Kord, 2012). A quantitative approach will used on this project.

**Methods**

SAS (Statistical Analysis Software) is the tool employed to run analyses. Summary Statistics is run to produce indisputable facts based on data pertaining to which zip codes and which lifestyle clusters spend the most and which items sell the most. One-Way ANOVAs (analysis of variance) is used to compare the means of two or more independent groups in order to determine whether there is a statistical significance in the means. A predictive analysis is run to determine which lifestyle cluster type is most likely to respond to a direct marketing campaign.

**Limitations**

While there are many records in the dataset used for this analysis, we are not sure if the data is a complete list of all customers, purchases, or zip codes. For customer information, not all customers who paid in cash may have been logged. In addition, any customer who has explicitly requested that no data be kept on them may not be included in this dataset. Another limitation to consider is that the method for collecting data may not be known as well as if there is a previous dataset available.

Slide 7

**Common Security, Privacy, and Ethical Challenges**

The issues of privacy in data security might elicit a knee-jerk reaction that might be verbalized as “All data should be private.” The door closes on the issue of data privacy, including an adherence to withholding ethical handling of now-private data. Without exposure, there are no ethical issues. Done. But without data exposure, how do we track and establish the patterns of a pedophile, how do we monitor and decode the chatter between terrorists?

Big-Data ethics, according to Davis, are Identity, Privacy, Ownership, and Reputation. While multiple identities, such as profiles, usernames, or accounts, might lead to more anonymity, Davis notes that Mark Zuckerberg asserts that multiplicity demonstrates a “lack of identity (Davis, 2012).” Big data, however, allows other organizations and individuals the ability to create stories about people based on what is shared, with or without permission.

“Data privacy (or information privacy or data protection) is about access, use and collection of data, and the data subject’s legal right to the data. (Lee, et al, 2016). There are two perspectives to privacy (Davis, 2016): online and offline. Online, sites might offer an opportunity to “opt-out” of receiving emails or other notifications, but this does not mean you are private. Opting out might simply mean your information has been captured, but you will not receive communication. A more appropriate offer would be the refusal of allowing information to be stored. Also under data privacy is not just keeping information free from unauthorized access, it is the cost of mitigation if privacy is breached.

Ownership of data is complicated. While one might opt-out of data gathered during an online session, the information that was used for the online session may not be exclusive to that session: one’s name, address, and phone number can be found at one’s employer, with the Department of Motor Vehicles, or on a saved list of gift recipients on an online shopping store account that belongs to someone else.

Reputation used to be created by the observations made by others, which could include where one lived, what kind of car did one own, where did one work, who were one’s friends, and possibly if one was religious. Now, even if privacy is intact, a reputation can be constructed by gathering information on what online forums are visited, where shopping is done, and even what email provider is used (outlook, gmail, yahoo, etc). A reputation can also further evolve based on language used in reviews, tweets, and comments.

According to Davis (2016), big data ethics does not start with questions about doing the right thing. Discussions over identity, privacy, reputation, or ownership are not starting points. Big data ethics are not about one or two issues. “Ethical practices are an outcome of ethical enquiry (Davis, 2016).” Davis suggests starting with values. Values common to an organization - formal or informal, business or social - are already implicit in decisions of that organization. In analyzing data, however, identity, privacy, reputation, and ownership are challenges.

**Privacy**

“Data privacy (or information privacy or data protection) is about access, use and collection of data, and the data subject’s legal right to the data. (Lee, et al, 2016). There are two perspectives to privacy (Davis, 2016): online and offline. Online, sites might offer an opportunity to “opt-out” of receiving emails or other notifications, but this does not mean you are private. Opting out might simply mean your information has been captured, but you will not receive communication. A more appropriate offer would be the refusal of allowing information to be stored. Also under data privacy is not just keeping information free from unauthorized access, it is the cost of mitigation if privacy is breached.

**Ownership**

Ownership of data is complicated. While one might opt-out of data gathered during an online session, the information that was used for the online session may not be exclusive to that session: one’s name, address, and phone number can be found at one’s employer, with the Department of Motor Vehicles, or on a saved list of gift recipients on an online shopping store account that belongs to someone else.

**Reputation**

Reputation used to be created by the observations made by others, which could include where one lived, what kind of car did one own, where did one work, who were one’s friends, and possibly if one was religious. Now, even if privacy is intact, a reputation can be constructed by gathering information on what online forums are visited, where shopping is done, and even what email provider is used (outlook, gmail, yahoo, etc). A reputation can also further evolve based on language used in reviews, tweets, and comments.

According to Davis (2016), big data ethics does not start with questions about doing the right thing. Discussions over identity, privacy, reputation, or ownership are not starting points. Big data ethics are not about one or two issues. “Ethical practices are an outcome of ethical enquiry (Davis, 2016).” Davis suggests starting with values. Values common to an organization - formal or informal, business or social - are already implicit in decisions of that organization. In analyzing data, however, identity, privacy, reputation, and ownership are challenges.

Slide 8

**Plans, Tools, and/or Techniques to Address Challenges**

Lee, et al (2016), from their article in the Information Systems Audit and Control Association (ISACA), suggests a three-fold approach to addressing the challenges of working with big data:

* The International Data Privacy Principles (IDPPs) for establishing and maintaining data privacy policies, operating standards and mitigation measures
* The hexa-dimension metric operationalization framework for executing policies, standards and guidelines.

Slide 9

**Methods for Eliminating Security, Privacy, and Ethical Challenges When Presenting**

It is one focus to analyze data and adhere to privacy best practices, but it is another focus to discuss findings openly, present results, or share analytical steps without compromising the data privacy guidelines set forth at the beginning of a project. If the focus of the presentation is to show results of analysis, and the presentation is not overwhelming (O’Leary, p. 401), then identifying details will not likely be needed, only the illustration of the result. Likewise for discussions or sharing analytical steps, identifying factors can be avoided if the purpose is to discuss results. If data sets are to be shared, then, researchers will need to revisit initially agreed guidelines set at the beginning of the project. Research participants may need to be consulted as to whether or not their data can be transferred to an interested, but not initially-approved, research group.

The challenges around data privacy can be a multi-layered hurdle of checks and balances. While keeping data private might seem easily solved with strong passwords, omission of unrequired details when filling out web forms, or creating multiple profiles, the data that other organizations might keep or share through social, government, or financial institution sites could be at risk. Further, what about the data that individuals give freely but expect privacy or even anonymity? There are limits and guidelines that both researcher and subject can use to ensure when and how data is used. When the best practices of technological, social and laws and regulations are applied to data, privacy is better secured.

Slide 10

**Findings**

1. Lifestyle Cluster Type 10 and Jackets

Supports Ha: Lifestyle Cluster Type who rank 10 are more likely to purchase jackets.

2. Lifestyle Cluster Type 10 and Jackets

Supports Ha: Lifestyle Cluster Type and Total Net Sales

3. Lifestyle Cluster Types and Average Amount Spent per Visit.

Supports Ha: Lifestyle Cluster Types for Total Nets Sales will not align with Average Amount Spent per Visit

Credit Card Users and Total Net Sales

Supports H0: Credit Card users increase Total Net Sales.

Slide 11

**Findings (don’t’d)**

5. Credit Card Users and Percentage of Returns

Average percentage of returns on Credit Card purchases is higher at 17.7%, while average percentage of returns on non-Credit Card purchases is 10.3%

6. Customers who have high Spent per Visit and Total Net Sales

The Lifestyle Cluster Type 1, while the having the highest income and education, is also the Type to have the highest returns.

7. Lifestyle Cluster Type who has a high Total Net Sales are likely to respond to Number of Marketing Promos on File

Supports Ha: Lifestyle Cluster Type who have a high Total Net Sales are not likely to respond to Number of Marketing Promos on File

Slide 12

**Conclusion**

Based on the total analysis of the data set, clothing\_store\_pp\_opt1\_lc.csv, which focusses on Lifestyle Cluster Type, Types 1, 4, and 10 should be the target of the direct marketing promo. Since the prescribed which Lifestyle Cluster Types were already decided, Marketing could have begun working on a direct marketing effort for this reduced data set of 13,137 customers. The six of 50 Lifestyle Cluster Types chosen are high and medium income and education groups. Reviewing additional variables, however, reveal that certain items have a typically higher profit percentage. Rather than a direct marketing campaign for all products considered, Summary Statistics reveal that customers spend a significantly more on jackets than any other item.

Slide 13

In addition, the Lifestyle Cluster Type that spends the most on Jackets is Lifestyle Cluster Type 10, even though they have the lowest ranking for Average Amount Spent per Visit, this groups spend the least per visit and rank 3rd in Percentage of Returns. They do, however, rank 2nd in Total Net Sales.

Slide 14

Rankings of the Lifestyle Cluster Groups 1, 4, 8, 10, 15, and 16

Slide 15

**Recommendations**

Lifestyle Cluster Type 1 scores favorably higher than others in in Average Amount Spent per Visit and Number of Marketing Promotions on File; but they also rank first in Percentage of Returns. Lifestyle Cluster Type 4 ranks higher than most other Lifestyle Cluster Types in each category, but not as high as Lifestyle Cluster Group 1 in Percentage of Returns. If the focus is on Lifestyle Cluster Types and Products, then Marketing efforts for a direct mail campaign should focus on the following:

* Target Highest Net Sales by Lifestyle Cluster Type: Lifestyle Cluster Type 4 and Legwear
* Target High End Customer and largest selling item: Lifestyle Cluster Type 10 and jackets.

Additional recommendation are to explore products, total net sales, and zip codes of customer who are in Lifestyle Cluster Type 16, who are have the lowest percentage of return dollars.