**Module 8 - Presenting Research Findings and Project Closure**

**Portfolio Project Option #1:**

**Capstone Project—Final Report and Slide Presentation: U.S. Organization**

Juliet Lachman

Colorado State University - Global Campus

MIS581 – Business Intelligence and Data Analytics

Due Date: August 1, 2021

Submitted: August 1, 2021

**Module 8 - Presenting Research Findings and Project Closure**

**Portfolio Project Option #1:**

**Capstone Project—Final Report and Slide Presentation: U.S. Organization**

**Introduction to Capstone**

The purpose of this final assignment, the capstone, is to present the culmination and interdependency of topics covered in the Master of Science in Data Analytics. Through a carefully considered scenario, the written portion of the capstone will cover the following topics:

* organization industry and history
* current business presented
* dataset
* security challenges
* results of analysis
* recommendations

The written portion will also contain References, Appendices of code, and an active link to GitHub, where the author’s work has been uploaded.

Also included in the capstone is a presentation of 10-15 slides in length. The presentation will be based on the written portion of the capstone and speaker notes. The speaker notes will be prepared for an audience of upper management of the organization. As with all presentations in the Master of Science in Data Analytics, there will also be slides to list references of 5-10 scholarly or peer-reviewed sources.

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 - 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.

**Introduction**

**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.

**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. Those states are (with the first two- or three-digits of their respective zip codes):

Alaska (AK): 995-999

Arizona (AZ): 85-86

California (CA): 900-961

Colorado (CO): 80-81

Idaho (ID): 832-839

Montana (MT): 59

Nevada (NV): 889-899

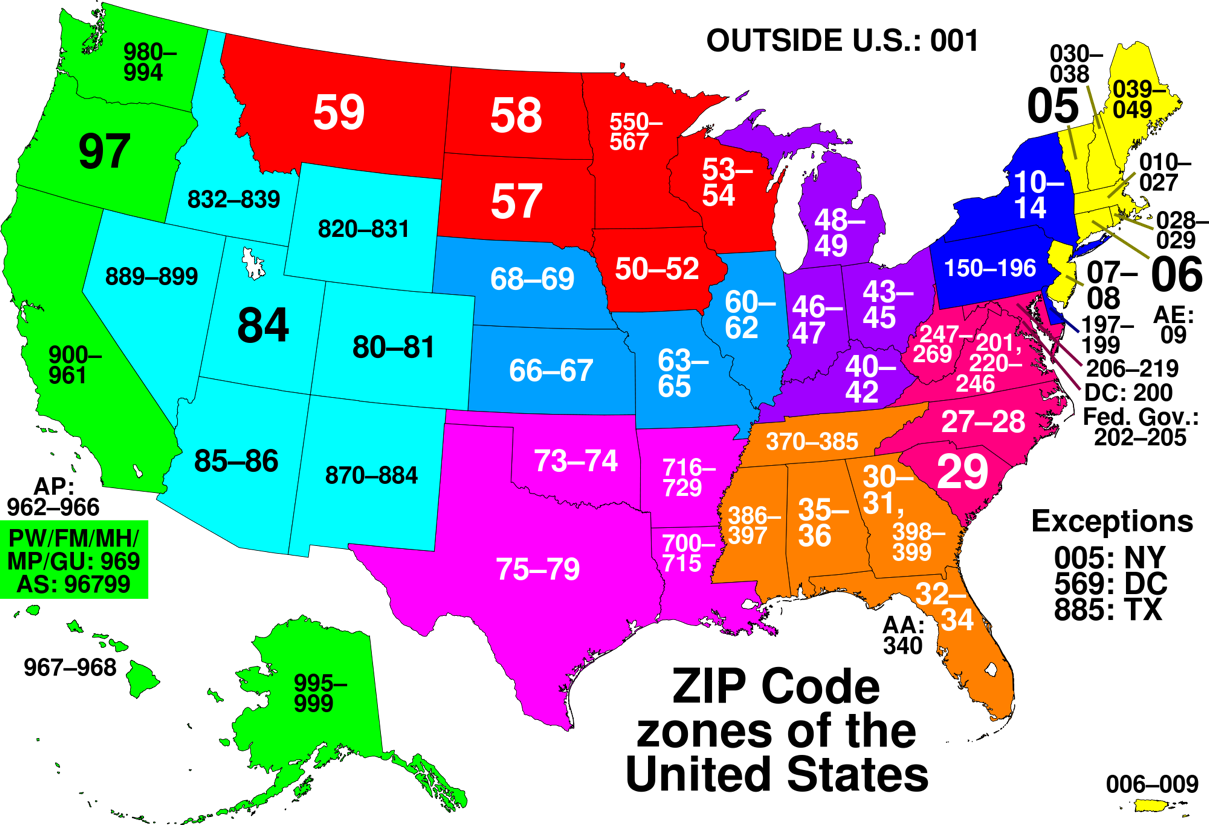
New Mexico (NM): 870-884

Oregon (OR): 97

Utah (UT): 84

Washington, (WA): 980-994

Wyoming (WY): 820-831



*Figure 1. Zip Codes of the United States from http://geometrx.com/2012/11/zip-codes-then-and-now/*

The data set contains 29 columns, all with numerical variables. To better understand the data set’s headers, they have been given synonyms, listed below:

| **Variable** | **Label** |
| --- | --- |
| customer\_id  zip\_code  num\_purch\_visits  total\_net\_sales  cc\_card  avg\_spent\_visit  p\_sweaters  p\_knit\_tops  p\_knit\_dres  p\_blouses  p\_jackets  p\_car\_pnts  p\_cas\_pnts  p\_shirts  p\_dresses  p\_suits  p\_outerwear  p\_jewelry  p\_fashion  p\_legwear  p\_collectibles  gmp  num\_mkt\_promos  num\_days\_cust\_file  mrkdn\_pntg  lifestyle\_clustype  pcnt\_rtns  days\_btwn\_purch  lifetime\_avg\_btwn\_visits | Customer ID  Zip Code  Number of Purchases in Visit  Total Net Sales  Credit Card  Average Amount Spent per Visit  Sweaters  Tops  Knit Dresses  Blouses  Jackets  Career Pants  Casual Pants  Shirts  Dresses  Suits  Outerwear  Jewelry  Fashionable Wear  Leg Wear  Collectibles  Gross Margin Percentage  Number of Marketing Promotions on File  Number of Days the Customer has been on File  Markdown Percentage on Customer Purchases  Lifestyle Cluster Type  Percent of Returns  Number of Days Between Purchases  Lifetime Average Days between Visits |

*Table 1. Data set variables and synonyms.*

Lifestyle Cluster Types (lifestyle\_clustype) are further expanded into 50 categories, but the top six are the following:

1. Cluster 1: Upper Crust: metropolitan families, very high income and education, homeowners, managers/professionals
2. Cluster 4: Mid-life Success: families, very high education, high income, managers/professionals, technical/ sales
3. Cluster 8: Movers and Shakers: singles, couples, students and recent graduates, high education and income, managers/professionals, technical/sales
4. Cluster 10: Home Sweet Home: families, medium-high income and education, manager/professionals, technical/sales
5. Cluster 15: Great Beginnings: young, singles and couples, medium-high education, medium income, some renters, managers/professionals, technical/sale
6. Cluster 16: Country Home Families: large families, rural areas, medium education, medium income, precision/crafts

**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.

**Summary Statistics**

| **Variable** | **Label** | **Mean** | **Median** | **Mode** | **Std Dev** | **Minimum** | **Maximum** | **Range** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| num\_purch\_visits  total\_net\_sales  avg\_spent\_visit  p\_sweaters  p\_knit\_tops  p\_knit\_dres  p\_blouses  p\_jackets  p\_car\_pnts  p\_cas\_pnts  p\_shirts  p\_dresses  p\_suits  p\_outerwear  p\_jewelry  p\_fashion  p\_legwear  p\_collectibles  gmp  num\_mkt\_promos  num\_days\_cust\_file  mrkdn\_pntg  lifestyle\_clustype  pcnt\_rtns  days\_btwn\_purch  lifetime\_avg\_btwn\_visits | Number of Purchases in Visit  Total Net Sales  Average Amount Spent per Visit  Sweaters (Total Sales %)  Tops (Total Sales %)  Knit Dresses (Total Sales %)  Blouses (Total Sales %)  Jackets (Total Sales %)  Career Pants (Total Sales %)  Casual Pants (Total Sales %)  Shirts (Total Sales %)  Dresses (Total Sales %)  Suits (Total Sales %)  Outerwear (Total Sales %)  Jewelry (Total Sales %)  Fashionable Wear (Total Sales %)  Leg Wear (Total Sales %)  Collectibles (Total Sales %)  Gross Margin Percentage  Number of Marketing Promotions on File  Number of Days the Customer has been on File  Markdown Percentage on Customer Purchases  Lifestyle Cluster Type  Percent of Returns  Number of Days Between Purchases  Lifetime Average Days between Visits | 5.0390291  473.2124633  113.5883176  0.2139460  0.0272138  0.0411240  0.0930296  0.1356939  0.0851193  0.0686125  0.0657492  0.0683635  0.0333671  0.0181944  0.0097621  0.0300010  0.0127216  0.0735088  0.5179412  11.5391159  436.9161776  0.1871020  15.1638599  0.1291021  4.7932341  3.9237425 | 3.00  261.00  92.00  0.16  0  0  0.05  0.04  0  0  0  0  0  0  0  0  0  0  0.55  12.00  445.00  0.18  11.00  0  4.82  3.95 | 1.00  98.00  98.00  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0.5900  4.00  677.00  0  10.00  0  4.9100  3.88 | 6.3491216  659.3274137  86.9808026  0.2311677  0.0680677  0.1109860  0.1355609  0.1841386  0.1411686  0.1327096  0.1167469  0.1579639  0.1300946  0.1000981  0.0364999  0.0796572  0.0500886  0.1765617  0.1722468  7.1393560  192.9708984  0.1292032  12.2464390  0.5431292  0.8727006  1.0204171 | 1.00  0.99  0.49  -0.97  -0.31  -0.71  -0.66  -0.36  -0.77  -0.50  -0.75  -0.42  -0.59  -0.73  -0.11  -0.67  -0.10  -0.44  -6.46  0  1.00  0  0  0  0  -2.41 | 115.00  24140.33  1919.88  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  0.9900  38.00  717.00  0.9500  50.00  40.9200  6.5800  5.90 | 114.00  24139.34  1919.39  1.9700  1.31  1.71  1.66  1.36  1.77  1.50  1.75  1.42  1.59  1.73  1.11  1.67  1.10  1.44  7.45  38.00  716.00  0.95  50.00  40.92  6.58  8.31 |

*Table 2. Summary Statistics for sales and percentages.*

Reviewing the data for Total Net Sales (total\_net\_sales) and Average Amount Spent per Visit (avg\_spent\_visit), the averages are $473.21 and $113.58, respectively. The average Lifestyle Cluster Type is 15, which notes medium income and not high income. The mode, however, for Lifestyle Cluster Type is 10, which does note high income.

Running a secondary Summary Statistics for frequency of certain variables (Appendix B) can help us determine additional variables to use for analysis.

**Summary Statistics**

| **Variable** | **Label** | **Mode** | **Minimum** | **Maximum** | **Range** |
| --- | --- | --- | --- | --- | --- |
| zip\_code  num\_purch\_visits  cc\_card  num\_mkt\_promos  num\_days\_cust\_file  lifestyle\_clustype  days\_btwn\_purch  lifetime\_avg\_btwn\_visits | Zip Code  Number of Purchases in Visit  Credit Card  Number of Marketing Promotions on File  Number of Days the Customer has been on File  Lifestyle Cluster Type  Number of Days Between Purchases  Lifetime Average Days between Visits | 55125.00  1.00  0  4.00  677.00  10.00  4.9100  3.8800000 | 0  1.00  0  0  1.00  0  0  -2.4100000 | 99687.00  115.00  1.00  38.00  717.00  50.00  6.5800  5.9000000 | 99687.00  114.00  1.00  38.00  716.00  50.00  6.5800  8.3100000 |

*Table 3. Summary Statistics for frequency.*

The Zip Code that appears the most at 64 times is 55125. The highest number of purchases in one visit is 115 from a Customer ID11614 in Zip Code 10314. Customer ID 32790, from Zip Code 72455 has been a Customer on File (num\_days\_cust\_file) for 677 days. Most frequent Lifestyle Cluster Type is 10, which notes high income. The most frequent Number of Days Between Purchases noted is 4.91, with only the largest number of days between purchases being 6.58. The item that yields the highest percentage in sales at 13% is Jackets (p\_jackets). Jacket sales might indicate preparation for cooler weather. Zip Codes 55125, 10314, 72455 are Woodbury, Minnesota, Staten Island, New York, and Pocahontas, Arkansas. Woodbury and Staten Island have colder winters, than Pocahontas, but all three do have winters cold enough to require jackets.

While this preliminary step of running Summary Statistics on entire data set is helpful, the only Lifestyle Cluster Type values that have been requested are 1, 4, 8, 10, 15, and 16 (Appendix C). After eliminating all other Lifestyle Cluster Types, the number of records drops from 27,999 to 13,137.

**Summary Statistics**

| **Variable** | **Label** | **Mean** | **Median** | **Mode** | **Std Dev** | **Minimum** | **Maximum** | **Range** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| num\_purch\_visits  total\_net\_sales  avg\_spent\_visit  p\_sweaters  p\_knit\_tops  p\_knit\_dres  p\_blouses  p\_jackets  p\_car\_pnts  p\_cas\_pnts  p\_shirts  p\_dresses  p\_suits  p\_outerwear  p\_jewelry  p\_fashion  p\_legwear  p\_collectibles  gmp  num\_mkt\_promos  num\_days\_cust\_file  mrkdn\_pntg  lifestyle\_clustype  pcnt\_rtns  days\_btwn\_purch  lifetime\_avg\_btwn\_visits | Number of Purchases in Visit  Total Net Sales  Average Amount Spent per Visit  Sweaters (Total Sales %)  Tops (Total Sales %)  Knit Dresses (Total Sales %)  Blouses (Total Sales %)  Jackets (Total Sales %)  Career Pants (Total Sales %)  Casual Pants (Total Sales %)  Shirts (Total Sales %)  Dresses (Total Sales %)  Suits (Total Sales %)  Outerwear (Total Sales %)  Jewelry (Total Sales %)  Fashionable Wear (Total Sales %)  Leg Wear (Total Sales %)  Collectibles (Total Sales %)  Gross Margin Percentage  Number of Marketing Promotions on File  Number of Days the Customer has been on File  Markdown Percentage on Customer Purchases  Lifestyle Cluster Type  Percent of Returns  Number of Days Between Purchases  Lifetime Average Days between Visits | 5.1400624  486.5984753  114.5055979  0.2205587  0.0285164  0.0404301  0.0899665  0.1339248  0.0835495  0.0707072  0.0637086  0.0659976  0.0320773  0.0180353  0.0096506  0.0302976  0.0124595  0.0720454  0.5252463  11.8761513  443.2795920  0.1849973  8.2487630  0.1320126  4.7955682  3.9186260 | 3.00  271.38  92.73  0.17  0  0  0.04  0.04  0  0  0  0  0  0  0  0  0  0  0.56  12.00  460.00  0.18  10.00  0  4.82  3.95 | 1.00  98.00  98.00  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0.5900  4.00  677.00  0  10.00  0  4.9100  3.9500000 | 6.5942195  683.6360654  86.6382523  0.2337445  0.0721493  0.1084050  0.1328505  0.1820275  0.1392237  0.1341096  0.1140821  0.1539088  0.1232499  0.0986734  0.0366740  0.0807757  0.0475451  0.1741729  0.1570706  7.1534932  192.0417557  0.1275559  5.2993019  0.4225542  0.8680614  1.0127007 | 1.00  4.00  2.500  -0.47  -0.24  -0.71  -0.66  -0.26  -0.51  -0.5  -0.75  -0.42  -0.5  -0.73  -0.11  -0.18  -0.1  -0.31  -3.36  0  3.00  0  1.00  0  1.10  -2.4100000 | 115.00  24140.33  1725.85  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  0.83  38.00  717.00  0.91  16.00  11.70  6.580  5.90 | 114.00  24136.33  1723.35  1.47  1.24  1.71  1.66  1.26  1.51  1.50  1.75  1.42  1.50  1.73  1.11  1.18  1.10  1.31  4.19  38.00  714.00  0.91  15.00  11.70  5.48  8.31 |

*Table 4. Summary Statistics for sales and percentages on Lifestyle Cluster Types 1, 4, 8, 10, 15, and 16.*

Reviewing the data for Total Net Sales (total\_net\_sales) and Average Amount Spent per Visit (avg\_spent\_visit), the respective averages are $486.59 - which is $13.38 higher than average for entire data set - and $114.51 – which is $0.93 higher than average for entire data set. The average Lifestyle Cluster Type is 8, which notes medium income and not high income. The mode, however, for Lifestyle Cluster Type is 10.

Running a secondary Summary Statistics for frequency of certain variables (Appendix D) can help us determine additional variables to use for analysis.

**Summary Statistics**

| **Variable** | **Label** | **Mode** | **Minimum** | **Maximum** | **Range** |
| --- | --- | --- | --- | --- | --- |
| zip\_code  num\_purch\_visits  cc\_card  num\_mkt\_promos  num\_days\_cust\_file  lifestyle\_clustype  days\_btwn\_purch  lifetime\_avg\_btwn\_visits | Zip Code  Number of Purchases in Visit  Credit Card  Number of Marketing Promotions on File  Number of Days the Customer has been on File  Lifestyle Cluster Type  Number of Days Between Purchases  Lifetime Average Days between Visits | 60089.00  1.00  0  4.00  677.00  10.00  4.9100  3.9500000 | 1001.00  1.00  0  0  3.00  1.00  1.100  -2.4100000 | 99687.00  115.00  1.00  38.00  717.00  16.00  6.5800  5.9000000 | 98686.00  114.00  1.00  38.00  714.00  15.00  5.4800  8.3100000 |

*Table 5. Summary Statistics for frequency.*

The Zip Code that shows up the most at 40 times is 60089. The highest number of purchases in one visit is 115 from a Customer ID 11614 in Zip Code 10314. Customer ID 10773, from Zip Code 6482 has been a Customer on File (num\_days\_cust\_file) for 717 days. Most frequent Lifestyle Cluster Type is 10, which notes high income. The most frequent Number of Days Between Purchases noted is 4.91, with only the largest number of days between purchases being 6.58. The item that yields the highest percentage in sales at 13.39% is Jackets (p\_jackets). Jacket sales might indicate preparation for cooler weather. Zip Codes 60089, 10314, 10773 are Buffalo Grove, Illinois, Staten Island, New York, and Bronx, New York. All three have winters cold enough to require jackets.

1. **Lifestyle Cluster Type 10 and Jackets**

Combining observations between both Summary Statistics, we can begin formulating a Hypotheses. 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.

H0: Lifestyle Cluster Type who rank 10 are more likely to purchase jackets.

Ha: Lifestyle Cluster Type who rank 10 are not more likely to purchase jackets.

1. **Lifestyle Cluster Type 10 and Total Net Sales**

Since Lifestyle Cluster Type 10 has the highest percentage in sales on Jackets, the next question might be to ask to determine if Lifestyle Cluster Type 10 also has the highest Total Net Sales.

H0: Lifestyle Cluster Type 10 will have the highest Total Net Sales.

Ha: Lifestyle Cluster Type 10 will not have the highest Total Net Sales.

1. **Lifestyle Cluster Types and Average Amount Spent per Visit.**

Related to Total Net Sales is Average Spent per Visit (avg\_spent\_visit). While Jackets may have a limited range in MSRP (manufacture suggested retail price), average amount spent per visit could include additions items. If the same Lifestyle Cluster Types for Total Nets Sales align with Average Amount Spent per Visit, Marketing can add additional verbiage to appeal to specific Lifestyle Cluster Type.

H0: Lifestyle Cluster Types for Total Nets Sales will align with Average Amount Spent per Visit.

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

1. **Credit Card Users and Total Net Sales**

Some credit cards offer rewards or points for travel or products. Analyzing credit cards sales would help Marketing determine if targeting credit card users would help increase sales.

H0: Credit Card users increase Total Net Sales.

Ha: Credit Card user do not increase Total Net Sales.

1. **Credit Card Users and Percentage of Returns**

True profit is a result of goods sold and not returned. Another variable to analyze is percentage of returns. Using the One-Way ANOVA (Appendix I) results of credit card usage, one hypothesis can be that since credit cards are easy to use for purchase, they might be easy to use for returns.

H0: Customers who use credit cards are more likely to have a higher percentage of returns.

Ha: Customers who use credit cards are more likely to have a lower percentage of returns.

1. **Customer who have high Spent per Visit and Total Net Sales**

To help with which Lifestyle Cluster Type will be the focus of marketing’s direct campaign, information on which Lifestyle Cluster Type has a high percentage of return will be needed (Appendix J).

H0: Customer who have high Spent per Visit and Total Net Sales may likely have a higher Percent of Returns.

Ha: Customer who have high Spent per Visit and Total Net Sales may likely not have a higher Percent of Returns.

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

For an effective direct marketing campaign to be impactful, customers who are more likely to respond to the effort will be a higher priority target audience. Using analysis on Total Net Sales, we can hypothesize that Lifestyle Cluster Types that spend the most are likely to be informed of events. A One-Way ANOVA (Appendix K) can be used to analyze Lifestyle Cluster Type and Number of Marketing Promos on File (num\_mkt\_promos).

H0: Lifestyle Cluster Type who has a high Total Net Sales are likely to respond to Number of Marketing Promos on File (num\_mkt\_promos).

Ha: Lifestyle Cluster Type who has a high Total Net Sales are not likely to respond to Number of Marketing Promos on File (num\_mkt\_promos)..

1. **Predictive Analysis**

To better inform marketing on which Lifestyle Cluster Type might be best to focus on for likelihood of responding to direct marking promotions (Appendix M), a predictive analysis can use Number of Marketing Promos in File and Lifestyle Cluster Type.

Chart

Description automatically generated

*Figure 9. Predictive Analysis of Lifestyle Cluster Type and Number of Marketing Promos in File*

While all Lifestyle Cluster Types have historically responded to Promos, it appears that Types 1 and 10 are more likely than Types 4, 8, 15, and 16 to respond.

**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 are shopping for winterwear every year:

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.

**Literature Review**

Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons*, 60(3), 285-292. doi: 10.1016/j.bushor.2017.01.002

The authors of this article discuss their recommendations to address big data barriers. They suggest changing technology infrastructure, focusing on privacy, developing big data analytic skills, and creating a clear, organizational vision of the use of big data.

Eik-Andresen, P., Johansen, A., Landmark, A. D., & Sørensen, A. Ø. (2016). Controlling a Multibillion Project Portfolio - Milestones as Key Performance Indicator for Project Portfolio Management. *Procedia - Social and Behavioral Sciences*, *226*, 294–301. https://doi.org/10.1016/j.sbspro.2016.06.191

A single data analysis is a single project. A project portfolio is a collection of multiple projects managed by a portfolio manager. The authors present a case study of 2000 milestones inside a single portfolio. The case study spans six years, covers multi-billion dollar business that runs 200-300 large projects every years. The paper presents strategies that can minimize delays and reveals patterns of delays in five different business areas.

Henrion, M. (2019, Dec 4). *Why most big data analytics projects fail.*INFORMS. https://pubsonline.informs.org/do/10.1287/orms.2019.06.08/full/#:~:text=According%20to%20the%20Gartner%20survey,management%20adoption%20and%20understanding%20and

The author suggests that analysts spend too much time working with poorly structured data, debugging worksheets, decoding poorly written code, and engaging too little time with clients. Interacting with clients will better clarify objectives and decision; brainstorming will help work through decision options, explore the data, and visualize the results. Without interacting with clients, the more important issues may not be discovered.

Lane, S., O’Raghallaigh, P., Sammon, D. (2016). Requirements gathering: The journey. *Journal of Decision Systems*. Jun2016 Supplement, *25,* 302-312. 11p. DOI: 10.1080/12460125.2016.1187390.

Requirements gathering is an important, critical part of any start of a project. The authors suggest that uncovering and addressing issues early is a human-centered journey. Early conversations enforce better understanding of the requirements gathering process experience of participants and the eventual requirements gathering effectiveness. The framework the authors propose is a scientific approach and the objective is to build a template for requirements gathering.

Lee, W. W., Zankl, W., Chang, H. (2016). An ethical approach to data privacy protection. *ISACA Journal., 6,* 1-9. 9p. https://www.isaca.org/resources/isaca-journal/issues/2016/volume-6/an-ethical-approach-to-data-privacy-protection.

The authors start their article with the premise that “privacy, trust and security are closely intertwined, as are law and ethics.” Trust is necessary and violation of privacy leads to threat to security. Legal standards may provide resolutions, but ethical practice from the beginning provides frameworks for expected conduct around privacy, security, respect, and ethics

Liu, Y., Han, H., DeBello J. E. (2018). The challenges of business analytics: successes and failures. Proceedings of the 51st Hawaii International Conference on system sciences. pp. 840-849. Retrieved on September 12, 2020 from https://pdfs.semanticscholar.org/e8a6/a70f184564bf79804681ad2ee3e43c5eeae8.pdf

The authors of this paper discuss both successful and unsuccessful implementations of analytics programs. Based on lessons that had been learned, the authors offer strategies to implement a successful business analytics program.

Loon, R. V. (2020, Jun 8). *How businesses can navigate the ethics of big data*. Simpli Learn. https://www.simplilearn.com/how-businesses-can-navigate-big-data-ethics-article

The abundance of data and the use of big data has opened up possibilities for analysis as well as access for researchers and organizations who are interested in reliable data for Artificial Intelligence (AI). The ethics of big data, however should not be overlooked and privacy and security should not be compromised. All users should still comply with ethical principles around big data.

Majeed, M., Jain, V., & Varma, S. (2017). Charting an effective big data strategy. *Pharmaceutical Executive*, 37(9), 54-55.

The authors recognize that big data is a valuable business asset. Using it to gain competitive advantages, however, required the right combination of strategy, technology and execution. Many companies might opt for the latest technologies in their excitements to solve problems in new ways. The authors caution that even if technology components are implemented, outputs might still be misaligned.

**Research Design**

**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.

The first step in analyzing the dataset is to run summary statistics (Appendix A). For this project, SAS is the tool used for analysis. When beginning to evaluate the organization’s goals to increase profit through direct marketing, one initial direction is to assess which products are likely to be targeted in achieving the organization’s goal of increased profits. The summary statistics shows that sweaters and jackets are the highest selling products. Using this information, the organization can then target the zip code in colder regions of the country where there are more likely to be more movement of these products (Appendix B). Another key element for analysis is the lifestyle cluster types into which customers are categorized. This will better inform which demographics are likely to purchase jackets and sweaters. For lifestyle cluster types, and sweater and jacket sales, one-way ANOVAs will be used (Appendices E-K). A one-way ANOVA is a one-way analysis of variance that is used to on two sample means to check for significant differences between the means. If there is a significant difference among the results of cluster lifestyle types, then the marketing group may consider targeting the more likely demographic for direct marketing on jackets and sweaters.

**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.

**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.

**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

1. Comply with national data protection or privacy law, national contract law, and other legal requirements or regulations relating to data privacy.
2. Comply with current security standards to protect stored personal data from illegitimate or unauthorized access or from accidental access, processing, erasure, loss or use.
3. Implement an easily perceptible, accessible and comprehensible privacy policy with information on who is in charge of data privacy and how this person can be individually contacted, why and which personal data are collected, how these data are used, who will receive these data, how long these data are stored, and whether and which data will be deleted or rectified upon request.
4. Instruct employees to comply with such privacy policies and avoid activities that enable or facilitate illegitimate or unauthorized access in terms of IDPPs.
5. Do not use or divulge any customer data (except for statistical analysis and when the customer’s identity remains anonymous), unless the company is obliged to do so by law or the customer agrees to such use or circulation.
6. Do not collect customer data if such collection is unnecessary or excessive.
7. Use or divulge customer data in a fair way and only for a purpose related to activities of the company.
8. Do not outsource customer data to third parties unless they also comply with standards comparable to these IDPPs.
9. Announce data breaches relating to sensitive data.
10. Do not keep personal data for longer than necessary.
11. Do not transfer personal data to countries with inadequate or unknown data protection standards unless the customer is informed about these standards being inadequate or unknown and agrees to such a transfer.
12. In the case of a contract between the company and the customer in which the customer commits to pay for services or goods:
    * Inform the costumer individually and as soon as reasonably possible in the event of a data breach.
    * Inform the customer upon request about which specific data are stored, and delete such data upon request unless applicable laws or regulations require the company to continue storing such data.
    * Do not use or divulge content-related personal data.
    * Do not use or divulge any other personal data without the customer’s explicit, separate and individual consent.
    * Do not store, use or divulge any customer data, unless applicable laws or regulations require the company to continue storing such data.
13. In the absence of a contract between the company and the customer in which the customer commits to pay for services or goods:
    * Inform the customer as soon as reasonably possible in the event of data breaches.
    * Inform the customer upon request what types of sensitive data are stored and delete such data upon request when such data are outdated, unless applicable laws or regulations require the company to continue storing such data.
    * Do not use or divulge sensitive data without the customer’s explicit, separate and individual consent.

* The hexa-dimension metric operationalization framework for executing policies, standards and guidelines.
* Identify the relevant critical factors depending on the target end users (corporatewide or a functional unit or nature of operation). For example, environmental impact is critical for a mining company or a factory but could probably be skipped for an information security unit.
* Secure the support of the board of directors with respect to corporate policy aspects and the supporting infrastructures that include the organization’s human resources (HR) management, legal, finance, and information and communications technology functional units with respect to technical support and reference. An appraisal of ethical consistency in conduct should be included during annual performance reviews (by HR).
* Determine a schedule for quantifying the elements of each factor for measuring, prioritizing and balancing the factors. The attributes/factors with help determine the steps to be taken to measure the effectiveness.

**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.

**Findings**

1. **Lifestyle Cluster Type 10 and Jackets**

Using One-Way ANOVA (Appendix E), we can compare analyze Jacket Sales and Lifestyle Cluster Types.

*Chart, scatter chart

Description automatically generated*

*Figure 2. One-Way ANOVA of Lifestyle Cluster Type and Jacket Sales.*

Figure 2. One-Way ANOVA of Lifestyle Cluster Type and Sweater Sale shows that of the six Lifestyle Cluster Types used for this analysis, Type 10 has the highest percentage sales, which supports H0: Lifestyle Cluster Type who rank 10 are more likely to purchase jackets.

1. **Lifestyle Cluster Type and Total Net Sales**

Using One-Way ANOVA (Appendix F), we can analyze Total Net Sales and Lifestyle Cluster Types.

*Chart, scatter chart

Description automatically generated*

*Figure 3. One-Way ANOVA of Total Net Sales and Lifestyle Cluster Type.*

Lifestyle Cluster Type 10 does not have the highest Total Net Sales. Lifestyle Cluster Type 4 has the highest Total Net Sales. With Lifestyle Cluster Type 10 following immediately after. For this analysis, support is for Ha: Lifestyle Cluster Type 10 will not have the highest Total Net Sales.

1. **Lifestyle Cluster Types and Average Amount Spent per Visit.**

Using One-Way ANOVA (Appendix G), we can analyze Average Amount Spent per Visit and Lifestyle Cluster Types.

Chart, scatter chart

Description automatically generated

*Figure 4. One-Way ANOVA of Average Amount Spent per Visit and Lifestyle Cluster Types.*

Average Amount Spent per Visit and Total Net Sales do not align. Results support Ha: Lifestyle Cluster Types for Total Nets Sales will not align with Average Amount Spent per Visit.

1. **Credit Card Users and Total Net Sales**

Using One-Way ANOVA (Appendix H), we can analyze Total Net Sales and Credit Card purchases.

Graphical user interface, application

Description automatically generated

*Figure 5 One-Way ANOVA of Total Net Sales and Credit Card purchases.*

Logically, each visit that results in a purchase using a credit card would increase the total spending of the customer. Figure 4, where 0 = no Credit Card used and 1 = Credit Card used, Total Net Sales is higher for Credit Card used, showing average Total Net Sales of $781.67 for Credit Card used vs $301.63 for no Credit Card used, supporting H0: Credit Card users increase Total Net Sales.

1. **Credit Card Users and Percentage of Returns**

Graphical user interface, application

Description automatically generated

*Figure 6. One-Way ANOVA of Percentage of Returns and Credit Card purchases.*

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%

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

Chart, scatter chart

Description automatically generated

*Figure 7. One-Way ANOVA of Percentage of Returns and Lifestyle Cluster Types.*

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

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

*Chart, scatter chart

Description automatically generated*

*Figure 8. One-Way ANOVA of Number of Marketing Promos on File and Lifestyle Cluster Types.*

Lifestyle Cluster Type 1 is the most likely Type to respond to promotions, but they rank 5th in Total Net Sales. This supports the Ha: Lifestyle Cluster Type who have a high Total Net Sales are not likely to respond to Number of Marketing Promos on File (num\_mkt\_promos).

1. **Predictive Analysis**

Chart

Description automatically generated

*Figure 9. Predictive Analysis of Lifestyle Cluster Type and Number of Marketing Promos in File*

While all Lifestyle Cluster Types have historically responded to Promos, it appears that Types 1 and 10 are more likely than Types 4, 8, 15, and 16 to respond to direct marketing campaign efforts.

**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. 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.

|  | **Rank** | Average Amount Spent per Visit | Total Net Sales | Percentage of Returns | Number of Marketing Promotions on File |
| --- | --- | --- | --- | --- | --- |
| Ranking of Lifestyle Cluster Groups from Highest to Lowest | 1 | 4 | 4 | 1 | 1 |
| 2 | 1 | 10 | 4 | 4 |
| 3 | 8 | 15 | 10 | 8 |
| 4 | 16 | 16 | 15 | 10 |
| 5 | 15 | 1 | 8 | 16 |
| 6 | 10 | 8 | 16 | 15 |

*Table 6. Rankings of Lifestyle Cluster Groups from Highest to Lowest.*

**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.

**References**

Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons*, 60(3), 285-292. doi: 10.1016/j.bushor.2017.01.002

Ayyagari, R., Jaejoo Lim, & Hoxha, O. (2019). Why do not we use password managers? A study on the intention to use password managers. *Contemporary Management Research*, 15(4), 227–244. https://doi.org/10.7903/cmr.19394

Davis, K. (2012). Ethics of big data. O'Reilly Media, Inc. ISBN: 9781449311797

Dimon, R. (2013). Enterprise performance management done right: An operating system for your organization. Hoboken, NJ: Wiley.

GitHub reviews (2021). From https://www.trustradius.com/products/github/reviews?qs=pros-and-cons

Hayes, A. (2021). Fortune 500. From https://www.investopedia.com/terms/f/fortune500.asp

Henrion, M. (2019, Dec 4). *Why most big data analytics projects fail.*INFORMS. https://pubsonline.informs.org/do/10.1287/orms.2019.06.08/full/#:~:text=According%20to%20the%20Gartner%20survey,management%20adoption%20and%20understanding%20and

I.T. Toolkit Magazine. (n.d.). Using project milestones to track accomplishments and status. https://www.ittoolkit.com/articles/project-milestones

Introduction to quantitative research. (n. d.). Retrieved from https://www.sagepub.com/sites/default/files/upm-binaries/36869\_muijs.pdf (Links to an external site.)

Jovancic, N. (2020). 5 research design types + key elements and characteristics*.* Lead Quizzes. https://www.leadquizzes.com/blog/research-design-types/

Lane, S., O’Raghallaigh, P., Sammon, D. (2016). Requirements gathering: The journey. *Journal of Decision Systems*. Jun2016 Supplement, *25,* 302-312. 11p. DOI: 10.1080/12460125.2016.1187390.

Lee, W. W., Zankl, W., Chang, H. (2016). An ethical approach to data privacy protection. *ISACA Journal., 6,* 1-9. 9p. https://www.isaca.org/resources/isaca-journal/issues/2016/volume-6/an-ethical-approach-to-data-privacy-protection.

Lund, T. (2005). The qualitative- quantitative distinction: Some comments. *Scandinavian Journal of Educational Research, 49*(2), 115–132.

Majeed, M., Jain, V., & Varma, S. (2017). Charting an effective big data strategy. *Pharmaceutical Executive*, 37(9), 54-55.

Majeed, M., Jain, V., & Varma, S. (2017). Charting an effective big data strategy. *Pharmaceutical Executive, 37*(9), 54-55.

Marr, B. (2015). Big data using SMART big data, analytics and metrics to make better decisions and improve performance. West Sussex, UK, John Wiley and Sons, Ltd.

Marr, B. (2016). Big data in practice: How 45 successful companies used big data, analytics, and metrics to make better decisions and improve performance. Chichester, West Sussex, UK: John Wiley and Co.

Merriam, S. (2009). Qualitative research: A guide to design and implementation. San Francisco, CA: Jossey-Bass.

O’Leary, Z. (2017). The essential guide to doing your research project (3rd edition). Thousand Oaks, CA: Sage Publishing.

Qualitative research: Defining and designing. (n. d.). Retrieved from http://www.sagepub.com/sites/default/files/upm-binaries/48453\_ch\_1.pdf

Sorensen, D. (2016). Beyond S&OP and IBP to enterprise planning and performance management. *Foresight: The International Journal of Applied Forecasting*, (40), 27-36.

Sunmola, H. O. (2020). Evaluation of motivating and requiring factors for milestones in IT projects. *Procedia Manufacturing*, *51*, 1469–1477. https://doi.org/10.1016/j.promfg.2020.10.204

Swinhoe, D. (2021). The 15 biggest data breaches of the 21st century. Retrieved from https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html

Trochim, W. M. K. (2020, Mar 10). *Research methods knowledge base*. Conjoint.ly. http://socialresearchmethods.net/kb/desintro.php

Williams, S. & Williams, N. (2010). How do we achieve strategic alignment. In the Profit Impact of Business Intelligence. San Francisco, CA: Morgan Kaufman.

APPENDICES

APPENDIX A

/\* import - begin \*/

libname PERM '/folders/myfolders/';

proc import

datafile='/folders/myfolders/sasuser.v94/clothing\_store\_pp\_opt1.csv'

dbms=csv out=WORK.MIS581\_M8\_PP\_OPT1 replace;

GETNAMES=YES;

run;

/\* data - begin \*/

data sales;

infile '/folders/myfolders/sasuser.v94/clothing\_store\_pp\_opt1.csv' DLM=',';

input customer\_id zip\_code num\_purch\_visits total\_net\_sales cc\_card avg\_spent\_visit p\_sweaters p\_knit\_tops p\_knit\_dres p\_blouses p\_jackets p\_car\_pnts p\_cas\_pnts p\_shirts p\_dresses p\_suits p\_outerwear p\_jewelry p\_fashion p\_legwear p\_collectibles gmp numb\_mkt\_promos num\_days\_cust\_file mrkdn\_pntg lifestyle\_clustype pcnt\_rtns days\_btwn\_purch lifetime\_avg\_btwn\_visits;

run;

proc datasets lib=WORK nolist;

modify MIS581\_M8\_PP\_OPT1;

label customer\_id='Customer ID';

label zip\_code='Zip Code';

label num\_purch\_visits='Number of Purchases in Visit' ;

label total\_net\_sales='Total Net Sales';

label cc\_card='Credit Card';

label avg\_spent\_visit='Average Amount Spent per Visit';

label p\_sweaters='Sweaters (Total Sales %)';

label p\_knit\_tops='Tops (Total Sales %)';

label p\_knit\_dres='Knit Dresses (Total Sales %)';

label p\_blouses='Blouses (Total Sales %)';

label p\_jackets='Jackets (Total Sales %)';

label p\_car\_pnts='Career Pants (Total Sales %)';

label p\_cas\_pnts='Casual Pants (Total Sales %)';

label p\_shirts='Shirts (Total Sales %)';

label p\_dresses='Dresses (Total Sales %)';

label p\_suits='Suits (Total Sales %)';

label p\_outerwear='Outerwear (Total Sales %)';

label p\_jewelry='Jewelry (Total Sales %)';

label p\_fashion='Fashionable Wear (Total Sales %)';

label p\_legwear='Leg Wear (Total Sales %)';

label p\_collectibles='Collectibles (Total Sales %)';

label gmp='Gross Margin Percentage';

label num\_mkt\_promos='Number of Marketing Promotions on File';

label num\_days\_cust\_file='Number of Days the Customer has been on File';

label mrkdn\_pntg='Markdown Percentage on Customer Purchases';

label lifestyle\_clustype='Lifestyle Cluster Type';

label pcnt\_rtns='Percent of Returns';

label days\_btwn\_purch='Number of Days Between Purchases';

label lifetime\_avg\_btwn\_visits='Lifetime Average Days between Visits';

run;

/\* data - end \*/

/\* SUMMARY STATISTICS \*/

/\* preliminiary summary statistics - begin \*/

ods noproctitle;

ods graphics / imagemap=on;

proc means data=WORK.MIS581\_M8\_PP\_OPT1 chartype mean median mode std min max range n vardef=df;

title 'Summary Statistics';

var num\_purch\_visits;

var total\_net\_sales;

var avg\_spent\_visit;

var p\_sweaters;

var p\_knit\_tops;

var p\_knit\_dres;

var p\_blouses;

var p\_jackets;

var p\_car\_pnts;

var p\_cas\_pnts;

var p\_shirts;

var p\_dresses;

var p\_suits;

var p\_outerwear;

var p\_jewelry;

var p\_fashion;

var p\_legwear;

var p\_collectibles;

var gmp;

var num\_mkt\_promos;

var num\_days\_cust\_file;

var mrkdn\_pntg;

var lifestyle\_clustype;

var pcnt\_rtns;

var days\_btwn\_purch;

var lifetime\_avg\_btwn\_visits;

run;

/\* preliminiary summary statistics - end \*/

APPENDIX B

/\* One-Way Anova on lifestyle\_clustype and p\_jackets - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=15.0in height=6.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class lifestyle\_clustype;

model p\_jackets=lifestyle\_clustype;

means lifestyle\_clustype / welch plots=none;

lsmeans lifestyle\_clustype / plots=(meanplot);

run;

quit;

/\* One-Way Anova on lifestyle\_clustype and p\_jackets - end \*/

APPENDIX C

/\* import - begin \*/

libname PERM '/folders/myfolders/';

proc import

datafile='/folders/myfolders/sasuser.v94/clothing\_store\_pp\_opt1.csv'

dbms=csv out=WORK.MIS581\_M8\_PP\_OPT1 replace;

GETNAMES=YES;

run;

/\* data - begin \*/

data sales;

infile '/folders/myfolders/sasuser.v94/clothing\_store\_pp\_opt1.csv' DLM=',';

input customer\_id zip\_code num\_purch\_visits total\_net\_sales cc\_card avg\_spent\_visit p\_sweaters p\_knit\_tops p\_knit\_dres p\_blouses p\_jackets p\_car\_pnts p\_cas\_pnts p\_shirts p\_dresses p\_suits p\_outerwear p\_jewelry p\_fashion p\_legwear p\_collectibles gmp numb\_mkt\_promos num\_days\_cust\_file mrkdn\_pntg lifestyle\_clustype pcnt\_rtns days\_btwn\_purch lifetime\_avg\_btwn\_visits;

run;

proc datasets lib=WORK nolist;

modify MIS581\_M8\_PP\_OPT1;

label customer\_id='Customer ID';

label zip\_code='Zip Code';

label num\_purch\_visits='Number of Purchases in Visit' ;

label total\_net\_sales='Total Net Sales';

label cc\_card='Credit Card';

label avg\_spent\_visit='Average Amount Spent per Visit';

label p\_sweaters='Sweaters (Total Sales %)';

label p\_knit\_tops='Tops (Total Sales %)';

label p\_knit\_dres='Knit Dresses (Total Sales %)';

label p\_blouses='Blouses (Total Sales %)';

label p\_jackets='Jackets (Total Sales %)';

label p\_car\_pnts='Career Pants (Total Sales %)';

label p\_cas\_pnts='Casual Pants (Total Sales %)';

label p\_shirts='Shirts (Total Sales %)';

label p\_dresses='Dresses (Total Sales %)';

label p\_suits='Suits (Total Sales %)';

label p\_outerwear='Outerwear (Total Sales %)';

label p\_jewelry='Jewelry (Total Sales %)';

label p\_fashion='Fashionable Wear (Total Sales %)';

label p\_legwear='Leg Wear (Total Sales %)';

label p\_collectibles='Collectibles (Total Sales %)';

label gmp='Gross Margin Percentage';

label num\_mkt\_promos='Number of Marketing Promotions on File';

label num\_days\_cust\_file='Number of Days the Customer has been on File';

label mrkdn\_pntg='Markdown Percentage on Customer Purchases';

label lifestyle\_clustype='Lifestyle Cluster Type';

label pcnt\_rtns='Percent of Returns';

label days\_btwn\_purch='Number of Days Between Purchases';

label lifetime\_avg\_btwn\_visits='Lifetime Average Days between Visits';

run;

/\* data - end \*/

/\* SUMMARY STATISTICS \*/

/\* preliminiary summary statistics - begin \*/

ods noproctitle;

ods graphics / imagemap=on;

proc means data=WORK.MIS581\_M8\_PP\_OPT1 chartype mean median mode std min max range n vardef=df;

title 'Summary Statistics';

var num\_purch\_visits;

var total\_net\_sales;

var avg\_spent\_visit;

var p\_sweaters;

var p\_knit\_tops;

var p\_knit\_dres;

var p\_blouses;

var p\_jackets;

var p\_car\_pnts;

var p\_cas\_pnts;

var p\_shirts;

var p\_dresses;

var p\_suits;

var p\_outerwear;

var p\_jewelry;

var p\_fashion;

var p\_legwear;

var p\_collectibles;

var gmp;

var num\_mkt\_promos;

var num\_days\_cust\_file;

var mrkdn\_pntg;

var lifestyle\_clustype;

var pcnt\_rtns;

var days\_btwn\_purch;

var lifetime\_avg\_btwn\_visits;

run;

/\* preliminiary summary statistics - end \*/

APPENDIX D

/\* secondary summary statistics - begin \*/

ods noproctitle;

ods graphics / imagemap=on;

proc means data=WORK.MIS581\_M8\_PP\_OPT1 chartype mode min max range vardef=df;

title 'Summary Statistics';

var zip\_code;

var num\_purch\_visits;

var cc\_card;

var num\_mkt\_promos;

var num\_days\_cust\_file;

var lifestyle\_clustype;

var days\_btwn\_purch;

var lifetime\_avg\_btwn\_visits;

run;

/\* secondary summary statistics - end \*/

APPENDIX E

/\* One-Way Anova on lifestyle\_clustype and p\_jackets - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=15.0in height=6.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class lifestyle\_clustype;

model p\_jackets=lifestyle\_clustype;

means lifestyle\_clustype / welch plots=none;

lsmeans lifestyle\_clustype / plots=(meanplot);

run;

quit;

/\* One-Way Anova on lifestyle\_clustype and p\_jackets - end \*/

APPENDIX F

/\* One-Way Anova on lifestyle\_clustype and total\_net\_sales - begin \*/  
Title;  
ods noproctitle;  
ods graphics / reset width=8.0in height=4.0in imagemap;  
  
proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);  
 class lifestyle\_clustype;  
 model total\_net\_sales=lifestyle\_clustype;  
 means lifestyle\_clustype / welch plots=none;  
 lsmeans lifestyle\_clustype / plots=(meanplot);  
 run;  
quit;  
/\* One-Way Anova on lifestyle\_clustype and total\_net\_sales - end \*/

APPENDIX G

/\* One-Way Anova on lifestyle\_clustype and avg\_spent\_visit - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=8.0in height=4.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class lifestyle\_clustype;

model avg\_spent\_visit=lifestyle\_clustype;

means lifestyle\_clustype / welch plots=none;

lsmeans lifestyle\_clustype / plots=(meanplot);

run;

quit;

/\* One-Way Anova on lifestyle\_clustype and avg\_spent\_visit - end \*/

APPENDIX H

/\* One-Way Anova on cc\_card and total\_net\_sales - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=5.5in height=2.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class cc\_card;

model total\_net\_sales=cc\_card;

means cc\_card / welch plots=none;

lsmeans cc\_card / plots=(meanplot);

run;

quit;

/\* One-Way Anova on cc\_card and total\_net\_sales - end \*/

APPENDIX I

/\* One-Way Anova on cc\_card and pcnt\_rtns - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=5.5in height=2.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class cc\_card;

model pcnt\_rtns=cc\_card;

means cc\_card / welch plots=none;

lsmeans cc\_card / plots=(meanplot);

run;

quit;

/\* One-Way Anova on cc\_card and pcnt\_rtns - end \*/

APPENDIX J

/\* One-Way Anova on lifestyle\_clustype and pcnt\_rtns - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=5.5in height=2.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class lifestyle\_clustype;

model pcnt\_rtns=lifestyle\_clustype;

means lifestyle\_clustype / welch plots=none;

lsmeans lifestyle\_clustype / plots=(meanplot);

run;

quit;

/\* One-Way Anova on lifestyle\_clustype and pcnt\_rtns - end \*/

APPENDIX K

/\* One-Way Anova on lifestyle\_clustype and num\_mkt\_promos - begin \*/

Title;

ods noproctitle;

ods graphics / reset width=5.5in height=2.0in imagemap;

proc glm data=WORK.MIS581\_M8\_PP\_OPT1\_LC plots(only);

class lifestyle\_clustype;

model num\_mkt\_promos=lifestyle\_clustype;

means lifestyle\_clustype / welch plots=none;

lsmeans lifestyle\_clustype / plots=(meanplot);

run;

quit;

/\* One-Way Anova on lifestyle\_clustype and num\_mkt\_promos - end \*/

APPENDIX L

/\* predictive - lifestyle\_clustype - begin \*/

ods noproctitle;

ods graphics / reset width=5.5in height=5.5in imagemap;

proc reg data=WORK.MIS581\_M8\_PP\_OPT1\_LC PLOTS(MAXPOINTS=13137) alpha=0.05 plots=all;

model num\_mkt\_promos=lifestyle\_clustype;

output out=WORK.MIS581\_M8\_PP\_OPT1\_LC\_P p=p\_ lcl=lcl\_ ucl=ucl\_ rstudent = r;

run;

quit;

/\* predictive - lifestyle\_clustype - end \*/

APPENDIX M

GitHub Details

Code and Details uploaded to:

Author’s GitHub Account

Graphical user interface, text, application, email

Description automatically generated *Figure 1. Author’s GitHub account showing contents of directory.*

Graphical user interface, text, application, email

Description automatically generated *Figure 2. Author’s GitHub account showing details of README.md file.*

Graphical user interface, text, application, email

Description automatically generated *Figure 3. Author’s GitHub account showing first few lines of capstone\_appendices.md file.*