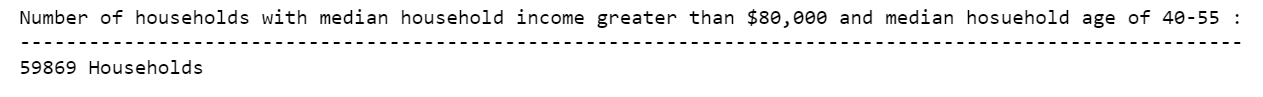
**SLM Business Analyst Assignment: Mitchell Keomoungkhoune**

**SECTION 1:**

**Question a) How many households are in the postal routes that have a median household income greater than $80,000 and a median household age of 40-55?**

The number of households that have a median household income greater than $80,000 and a median household age between 40 and 55 years old is 59,869 households.



**Question b) Let's say you want to target 100K households in Atlanta for our SLM shared mail program. Which Geocodes do you choose and why?**

To meet the 100,000-household target in Atlanta for SLM’s shared mail program the following line of code was used to effectively filter the data frame that would generate a mailing population (high income families) that would meet the 100,000-household target:

Geocodes = data1A[(data1A['MedianIncome'] > 55000) & (data1A['CityName'] == 'ATLANTA') & (data1A['Median Age'] <= 44) & (data1A['PercentHouseholdWithChildren'] <= 28)]

To adequately meet this target the median income was set to above $55,000, the city was constrained to the Atlanta area, the median age was set to less than or equal to 44 years of age, and the percent household with children was constrained to less than or equal to 28%. This filtered population generated a household target of 104,548 households while using 143 Geocodes.

A screenshot of a computer

Description automatically generated with medium confidence

**Question c) Let's say that a client gives you a zip ranking for Atlanta (Dataset 1B). How would you use this client info to improve targeting?**

To improve overall targeting for the client by including the “Rank” of zip codes. This was done by merging Dataset 1A (Postal Route Data for Atlanta) with Dataset 1B (Client Zip Ranking) to populate a dataset that would incorporate postal route data with zip rank. This merge could improve targeting for client that seeks to survey certain zip code and ranks with customers more likely to purchase their product as a result of receiving direct mailing campaigns.

Geocodes\_rank = data1A\_1B[( data1A\_1B['MedianIncome'] >= 45000) & ( data1A\_1B['CityName'] == 'ATLANTA') & ( data1A\_1B['Median Age'] <= 65) & ( data1A\_1B['PercentHouseholdWithChildren'] <= 45 ) & (**data1A\_1B['Rank'] == 3**)]

The above code shows an example of using rank to enhance customer targeting. In this example, ‘Rank’ = 3 and is linked with other variable parameters such as median income and median age to filter the data to generate a population preferable to the client. This particular example generated a targeting household population of 16,759 households and 23 geocodes.

Graphical user interface, text

Description automatically generated with medium confidence

**Question d) Dataset 1C is a list of Geocodes that have been previously mailed. Using your answer from question 2, what % of this market has been mailed before? What % has been mailed 2 times?**

Graphical user interface, text, application

Description automatically generated

The above figure shows the breakdown of percentages of geocodes that have been mailed before and geocodes that have been mailed twice before. It was determined that nearly 62% of the geocodes were previously mailed while nearly 40% of the geocodes have been mailed twice before.

Chart, pie chart

Description automatically generated

**Question e) What is the breakdown of household counts by zip code?**

Below figure shows a breakdown of households by zip code. Some zip codes that contained high number of households include **30309, 30339, 30328, and 30324**.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**SECTION 2:**

**Question a) What steps did you take to prepare the raw client order match data for campaign analysis?**

Text

Description automatically generated

One of the first steps in preparing the raw client order match data for the campaign analysis that were taken included any necessary data preprocessing techniques such as checking for null/missing values and examining the datatype of Dataset 2A (List of Orders at line-item level attributed to mailed and holdout customers for recent campaign). From this initial investigation, it was observed that all 1016 entries for each column were non-null values and of type “object”.

Since the datatype for Line-Item Value was of type “object” the below code was used to convert object values in this column to float values that can be more easily manipulated (indicated by red box):

data2A['Line Item Value'] = pd.to\_numeric(data2A['Line Item Value'].replace('[^0-9\.-]','', regex = True))

Table

Description automatically generated

It can also be observed that the data type for Line-item value changed to “float64” as opposed to object after the conversion is made using pd.to\_numeric code implementation.

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

A boxplot was also used to investigate and visually identify any outliers that might be associated with line-item values. Once initial data preprocessing was conducted raw data was consolidated by grouping Order IDs and summing line-item values. Dates for 90-day attribution window were also accounted for by filtering the dates outside the attribution window in excel and rereading the csv file back into the IDE.

**Question b) Build a report to show the client's results. What is the cost per customer order of each Test Cell? Which Test Cell was more successful on this metric?**

Graphical user interface

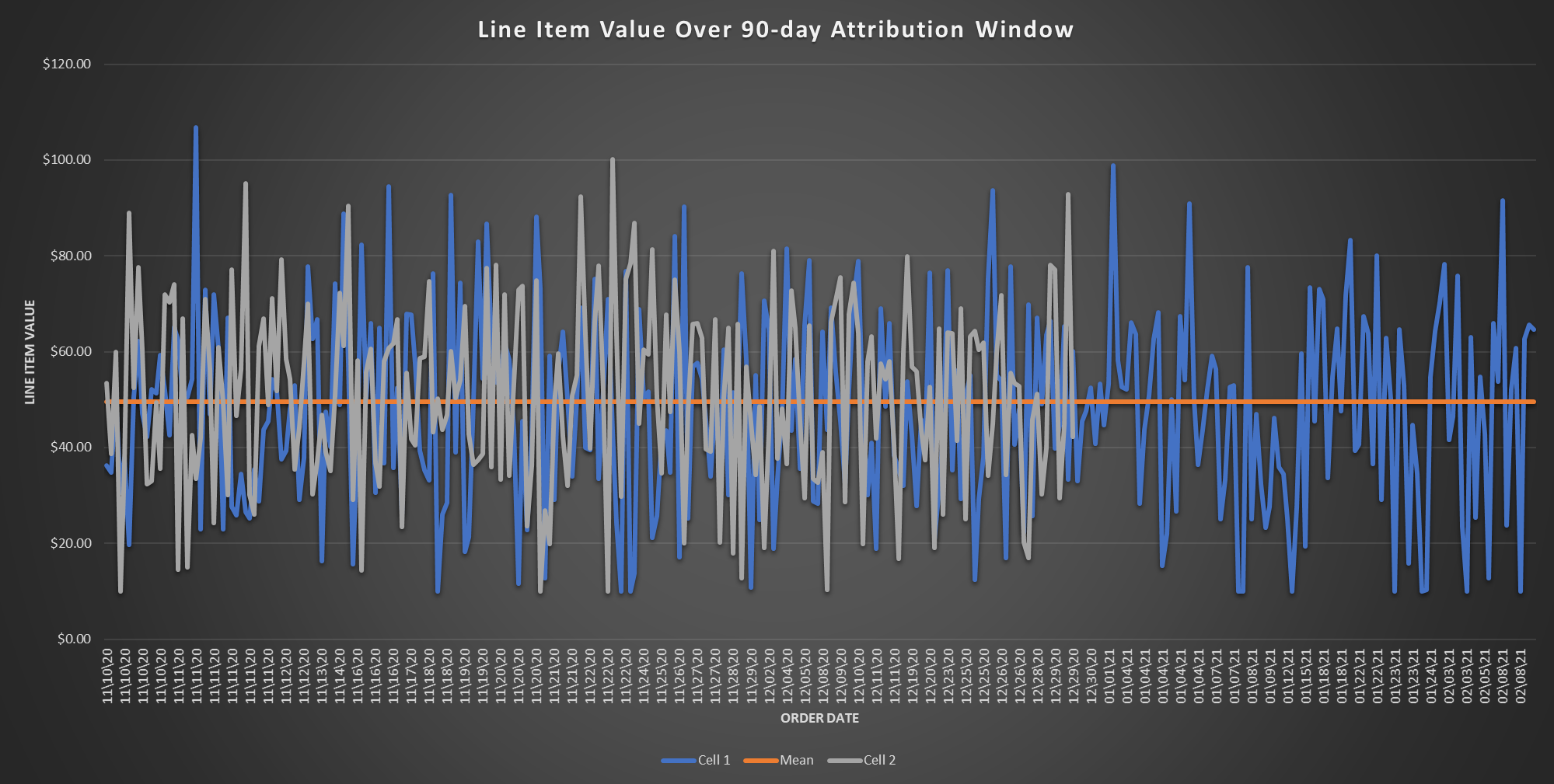
Description automatically generated with medium confidence



After data cleaning and manipulation, the calculated cost per customer order was $47.93 and $50.61 for campaign cell 1 and campaign cell 2, respectively for the 90-day attribution window (11/10/20 – 2/8/21). Test cell 2 was more successful on this metric.

**Question c) Treating the holdout as a control, how would you assess the performance of each test mailed group in driving additional orders? Does this performance vary over time? What is the significance of any observations?**

Campaign cell 2 based on cost per customer order seemingly is the more successful campaign. When the performance is viewed of cell 1 and cell 2 over time populated by viewing the line-item value over the 90-day attribution window it is visually apparent that cell 2 during this time-period slightly exceeds cell 1 in terms of cost per customer order. However, there are still instances where test cell 1 (format: postcard, population: 10000, total cost to mail: $5000) outperforms test cell 2 shown in the line graph plotting line-value for both test cells over the 90-day window.

****

**Question d) Do you have any other observations about the data that may be of interest to the client or in assessing the campaign? What recommendations would you make?**

RR or the response rate, is the percentage of people on the mailing list who respond to the direct mailing campaigns. According to the Direct Marketing Association, the average response rate for direct mailing house lists is 9% and 5% for more prospective lists. Response rate could also be another KPI (key performance indicator) to use to assess the efficiency of both campaign cells for the client. In regard to the client data, RR can be determined by comparing number of mailings to number of orders or instances where consumer products are purchased. In this case, the client can correlate order IDs to total mailings and generate a response rate and establish which mailing campaign was more effective based on calculated RR values.

Versatility and personalization of data points and relevant variables are a significant factor for the successful targeting and response of direct mailing campaigns. Recommend that client improves customer response rate by also improving mailing list and targeting initiatives.

**SECTION 3:**

**Question a) Conceptually: What types of additional data points might you want for a model build? What types of modeling techniques come to mind as having potential, and what programs and/or software do you think might be applicable or relevant to actually run a model?**

Logistic regression is a regression analysis technique used when a dependent variable is dichotomous or binary i.e., 0 or 1 for pass/fail, win/lose, alive/dead, or healthy/sick or a scale from 0 to 5. This can be implemented for variables such Median Age, Median Income, and Total Households by classification where ranges can be designated to simple value predictors (i.e., Age 55 to 65 = 1). Implementing concepts of logistic regression could help enhance the analysis and prediction of the likelihood of consumer response to direct mailing campaigns as well. Additional data points to help choose “best” postal routes can include other demographic data points such as martial status and gender. Psychographic data can also help enhance a predictive modeling process by pinpointing certain psychographic traits such as hobbies, interests, likes/dislikes, and values.

With this predictive model the client can rank order perspective consumers based on their likelihood to respond and take into account their likelihood to respond into a final decision. A RFM Model or Recency, Frequency, and Monetary Value Model can also be considered to help improve overall KPI. RFM is marketing analysis tool used to identify or an organization’s best customers by measuring and analyzing spending habits (how recently they made purchase, how often they purchase product, and how large purchase are). This action can augment targeting and therefore improve response rates for clients aiming for populations that better respond to their products.

Some software’s/programs that can considerably enhance ML (machine learning) models are deep learning software’s such as Torch, TensorFlow, Keras, PyTorch, and Caffe.