MARKETING CAMPAIGN ANALYSIS

Capstone Project Report

Contents

MARKETING CAMPAIGN ANALYSIS

Capstone Project Report

Executive Summary

Overview

Problem Statement

Marketing-to-sales conversion rates are impacting revenue and growth.

Context

Marketing-to-sales conversion rates are impacting revenue and growth. This project proposes applying a marketing strategy to better reach and be accepted by prospective consumers and generate sales. According to reports, marketing campaigns receive 100% more clicks from campaigns targeting customers grouped by similar traits than campaigning customers not grouped. Companies have experienced six to seven times growth in overall revenue.

Objective

This project proposes applying a marketing strategy to better reach and be accepted by prospective consumers and generate sales. The goal is to understand the problem by analyzing existing marketing data to identify critical insights around customer characteristics and habits, spending and purchase patterns, and then segment customers into groups with similar traits. Insights could potentially lead to data-driven recommendations for marketing activities designed to improve customer response and acceptance.

Key Questions

- Do any groups with similar characteristics exist?
- Can new variables be created to enhance insight?
- Do the new variables create new insights?
- Can customers be segmented into groups with like traits?
- What do these groups tell us about spending habits?
- Is one group more likely to respond to a specific product?
- What insights can customer demographics suggest for campaign messaging?

Problem Formulation

Problem Formulation Data Analytics uses analytical methods and techniques to understand the data and hidden insights, such as:

- Identify groups of customers with similar traits.
- Determine how groups with different characteristics respond to previous marketing activities.

- Identify correlations between customer acceptance of products and channels.
- Use the new insights to create or update current marketing activities

The problem formulation will begin with exploring and understanding the data, resulting in a data description and better preparing the data for analytics. Data analytics will provide a means for understanding, finding clusters, and identifying patterns in existing marketing data to create new marketing activities.

Proposed Methods

Proposed data analytics methods to better understand and draw insights into the data are:

1. Data Exploration and Observational Analysis
    a. Discover the shape and content of dataset
    b. Look for missing and unique values

2. Exploratory Data Analysis
    a. Separate data into numeric and categorical data
    b. Use univariate analysis on the numeric data understanding
        i. The mean, standard deviation, mean, min, and max values
        ii. Determine where most of the data lie with a quantile analysis
        iii. Draw observations
    c. Data Preparation
        i. Impute missing values
        ii. Drop unnecessary columns

3. Dataset Refining
    a. Determine which columns are best for clustering
    b. Drop unnecessary columns
    c. Correlate the data and draw observations

4. Clustering
    a. Begin looking for similarities in customers, products, channels
    b. Perform analytics using:
        i. Principal Component Analysis (PCA)
        ii. T-distributed Stochastic Neighbor Embedding (T-SNE)
        iii. K-Means

Measures of Success

What does success look like? Measure success will occurs as follows:

1. The data exploration, analysis, and analytics produced insights for marketing campaigns that include information such as:
    a. What made campaign 4 the most successful
    b. How do customers of varying incomes spend their money and on what products
    c. Are there any insight into customer characteristics and spending habits
2. From these insights, targeted marketing campaigns

   a. Segmented campaigns are created that have historically 100% more clicks creating 6 to 7 times revenue growth.
3. Market metrics from the new campaign are compared to the previous campaigns. The expectation is:
   a. Increase in customer response
   b. Increase in customer acceptance
   c. Increase in sales

Key Takeaways

What are the lessons learned? Key takeaways are:

1. A lack of understand the data impacts business
2. Understanding customer characteristics and behaviors is critical
3. Grouping customers according to similarities improves campaign effectiveness
4. Data Science should be incorporated into marketing campaign analysis
5. Optimizing Marketing activities will have a positive impact on revenue and growth

Data Exploration

Data Description

The dataset was collected in 2016 and contains 2240 marketing observations with 27 different categories. The categories consist of 24 numerical and three categorical data types. The following table breaks down the category columns:

Customer	**Products**	**Channels**	**Marketing**	**Responses**
Birth Year	Wines	Deal Purchases	Web Visits	Complaint
Education	Fruits	Web Purchases	Campaign 1	Response
Marital Status	Meats	Catalog Purchases	Campaign 2	
Income	Fish	Store Purchases	Campaign 3	
Kid Home	Sweets		Campaign 4	
Teen Home	Gold		Campaign 5	
Enrollment Date				
Days Since Purchase				
ID				

Key Observations

Missing Values

The category, Income, has 1.07 percent of its values missing and will need to be addressed as we go through data analysis. There are no other missing values in the dataset. There are various ways of handling missing values depending on how the values impact data analytics.

Unique Values

Any category, or column, with a high number of unique values does not generally offer any insights into data, and the columns dropped before analytics. The column “ID” is 100 percent unique values and will be dropped.

Exploratory Data Analysis

Numerical Column Summary Statistics

Insights are beginning to be visible. Here are some interesting observations from looking at the numerical column summary statistics.

- There are 26 numerical columns, the ID column was dropped
- Wine and Meat are purchased most often, while Fruits and Sweets are purchased least often.
- Customers tend to purchase more products in stores and purchase the least using catalogs
- Customers respond to marketing campaigns ~14.9 percent of the time but accept campaigns at a rate lower than eight percent.
- For all customers that visit the website, 50% are likely to purchase a product
- Customers tend to accept campaign four higher than all other campaigns at 7.5 percent

Categorical Column Summary Statistics

There are two categorical columns. Each column has five to eight variables. The columns “Kidhome”, “Teenhome”, and “Complain” were defined as zero and one represents Boolean variables where zero is false and one is true.

Key observations include:

- “Graduation” is the most common status in the category “Education”
- “Married” is the most common status in the category “Marital_Status”
- Most customers do not have children
- The number of complaints is very low and might be dropped later in the analysis, there have been only ~21 complaints in 2240 observations.

The categorical column “Marital_Status”

Marital_Status has the following variables eight variables:

1. Married
2. Together
3. Single
4. Divorced
5. Widow
6. Alone
7. Absurd
8. YOLO

Some of these variables will be combined into new features. “Married” and “Together” will be combined into a new column name “Relationship.” The marital status “Alone” will be replaced with “Single” combining both categories into "Single.” Finally, the categories “YOLO” and “Absurb” will be combined into a new feature (column) named “Other.” The remaining feature set will be:

- Relationship
- Single
- Divorced

- Widow
- Other

Reducing the number of variables from 8 to 5.

The categorical column "Education"

The category "Education" has the following five variables:

1. Graduation
2. PhD
3. Master
4. 2n Cycle
5. Basic

The variables "2n Cycle" and "Master" are the same and observations with the status "2n Cycle" will have "2n Cycle" renamed to "Master" reducing the variables from five to four. The remain features will be:

- Graduation
- PhD
- Master
- Basic

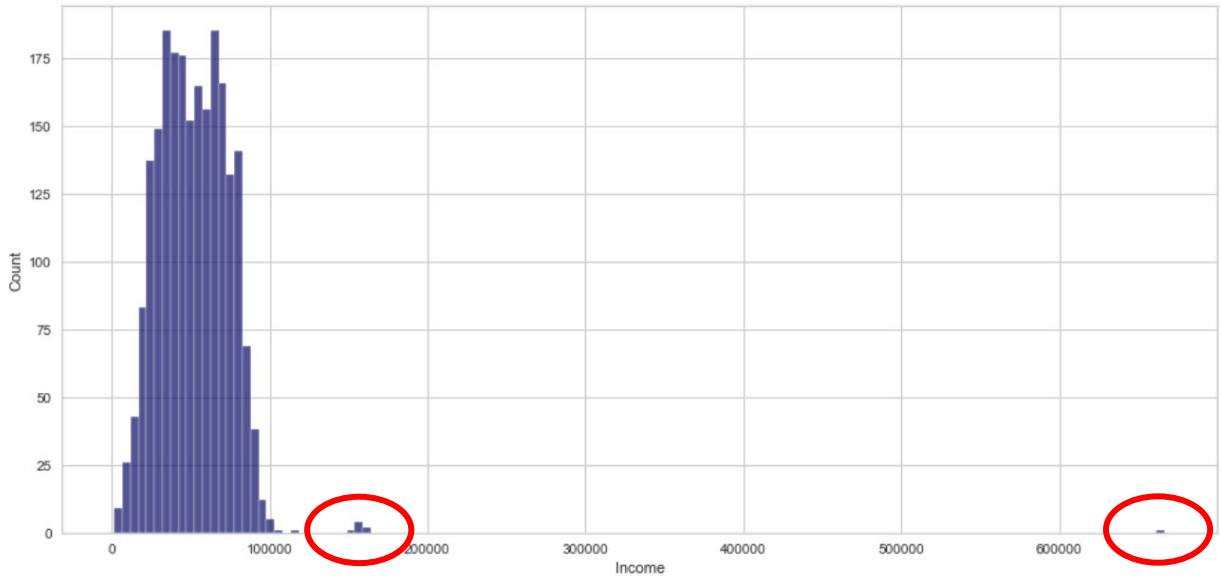
Creating New Features

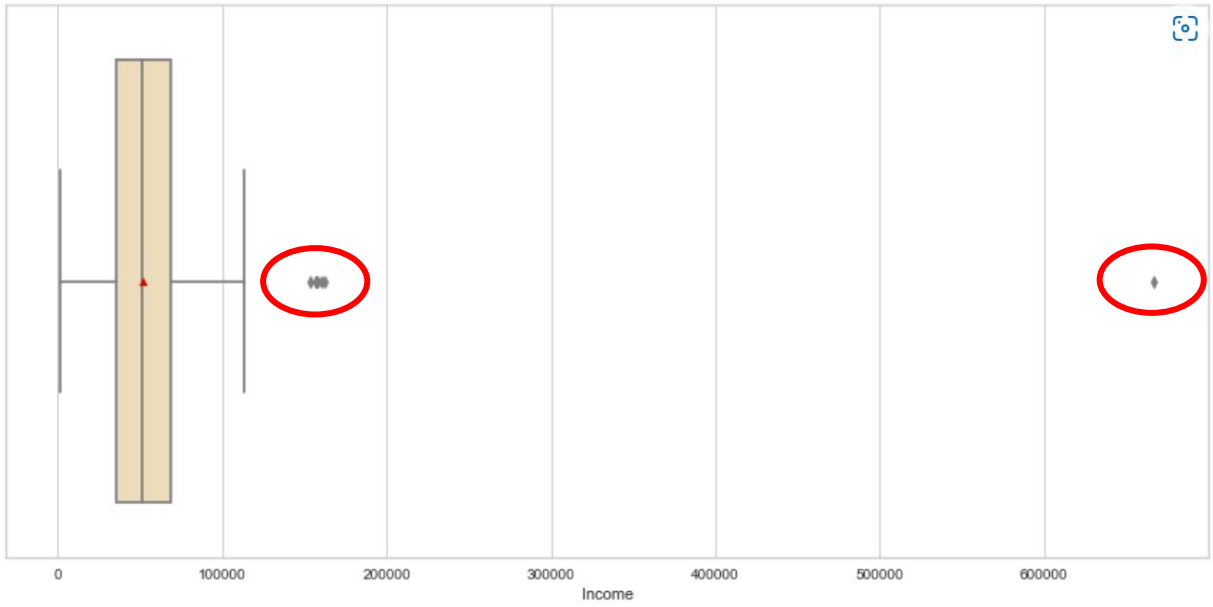
There are some variables we can create to further enhance insight into the data. A new feature named "Age" will be created adding the calculated value of Age for each observation and dropping the category "Year_Birth." "Total_Spending" will be created providing insight into customer spending on which products and/or channels. Finally, "Total_Purchases" will reflect the total number of purchases a customer has made.

Imputing the Income Variable

Below are a bar and box blot for the income variable:

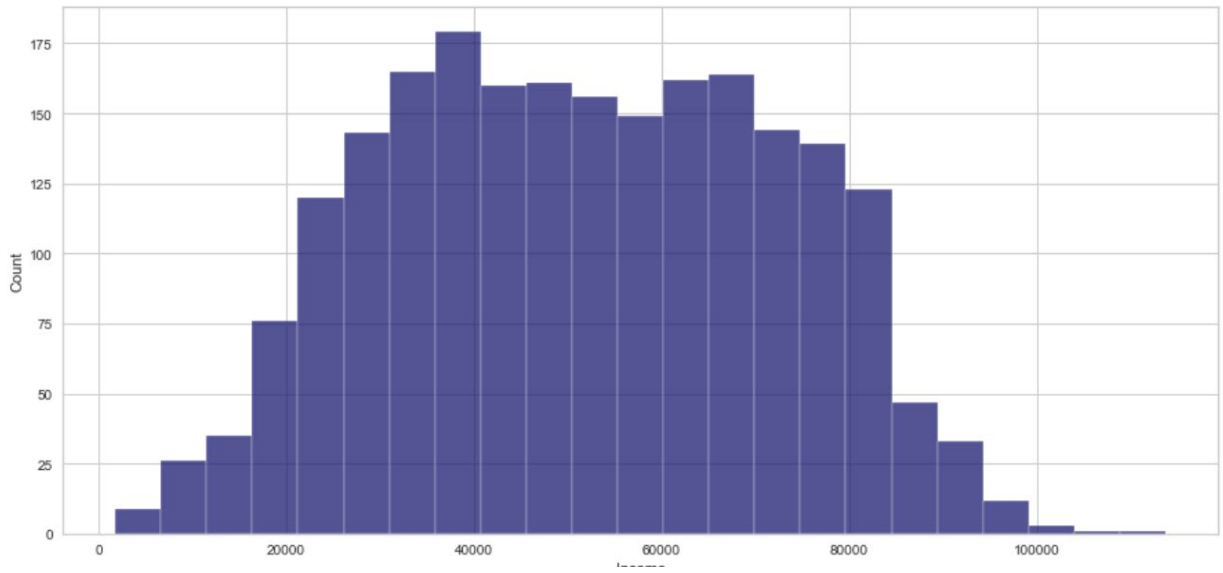
```
<function matplotlib.pyplot.show(close=None, block=None)>
```

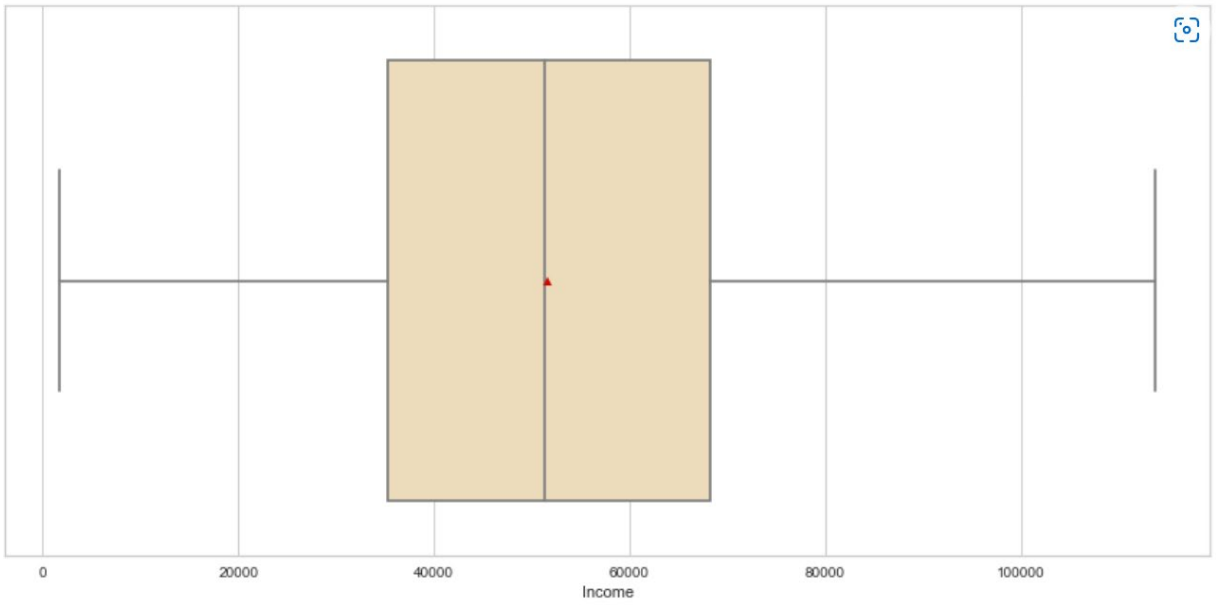




The plots clearly show there are outliers in the data shown in red above. The outliers to the far-right are extreme. We can calculate the value for the 99.5 percent quartile and determine which values occur beyond 99.5 percent. There are eight values beyond the 99.5 percent quantile and will be removed. The resulting plots after dropping the eight values are below.

```
<function matplotlib.pyplot.show(close=None, block=None)>
```





The plots now demonstrate a normal distribution, and dropping the eight values is not going to impact refined insights, however, there are still 24 missing values. These values will be imputed by calculating the median income and replacing the 24 missing values with the median. This will not impact the data analytics and will preserve the other data in those 24 observations.

Concluding Data Exploration and Analysis

The plotting of the categories “Age”, “Income”, “NumStorePurchase”, “Total_Purchases”, and “NumWebVisitsMonth” are or close to normal distributions. Several categories remain that are highly right-skewed.

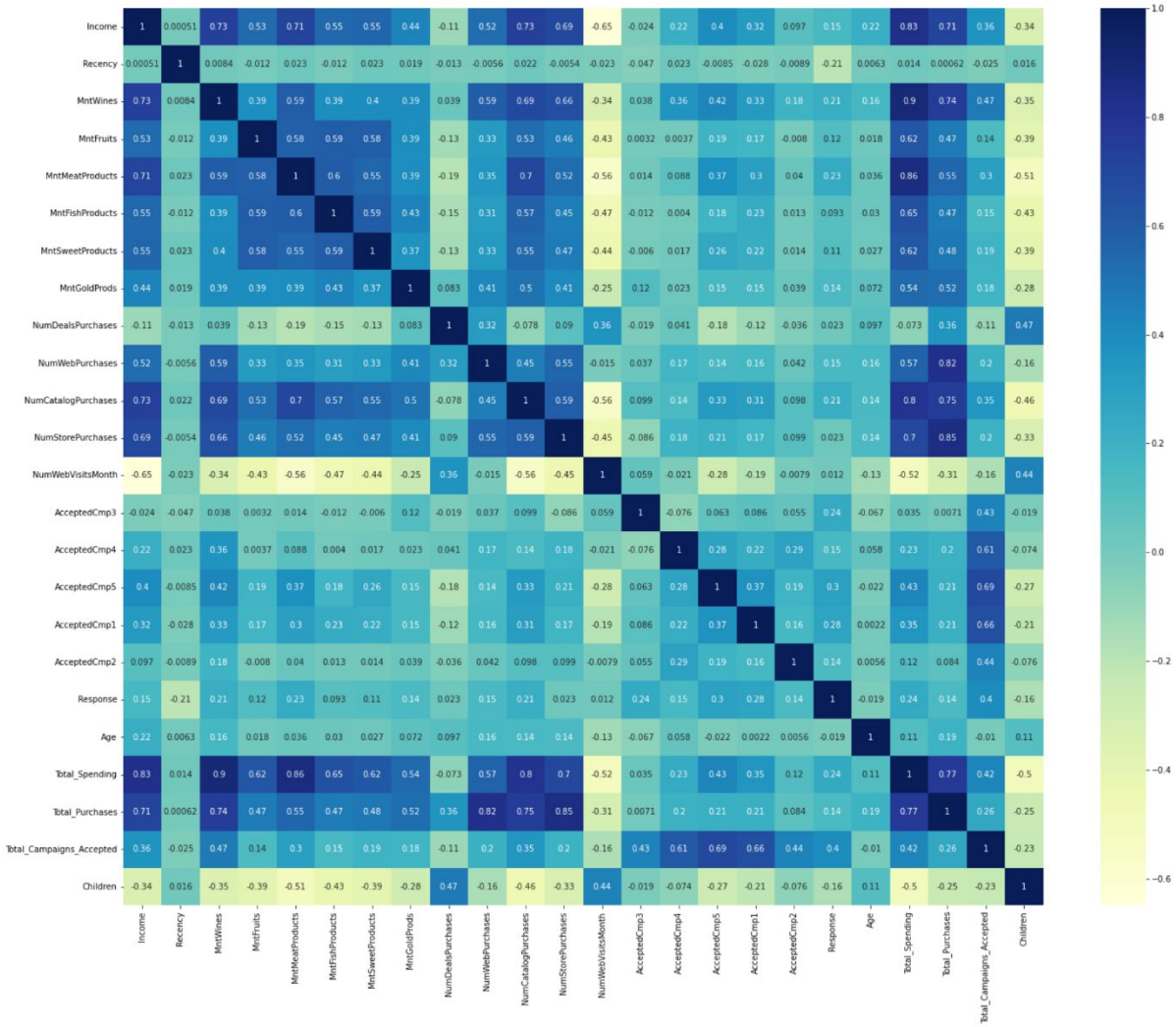
Data Analytics and Refined Insights

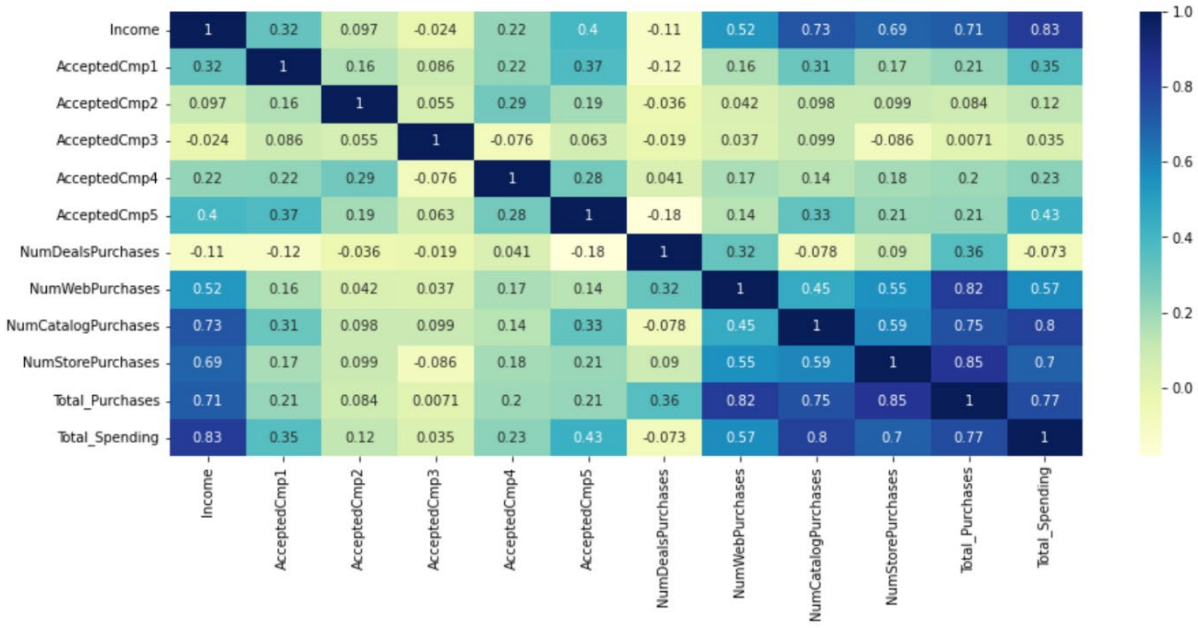
Data Correlation

For our data analytics we can drop the following columns:

Complaints	About 95% of the customers didn't complain and have the same value for this column. This variable will not have a major impact on segmentation.
Kidhome, Teenhome, Marital_Status,and Education	Distance-based algorithms cannot use the default distance like Euclidean to find the distance between categorical and numerical variables.

Creating a correlation plot will reveal any relations between the different data categories shown below:



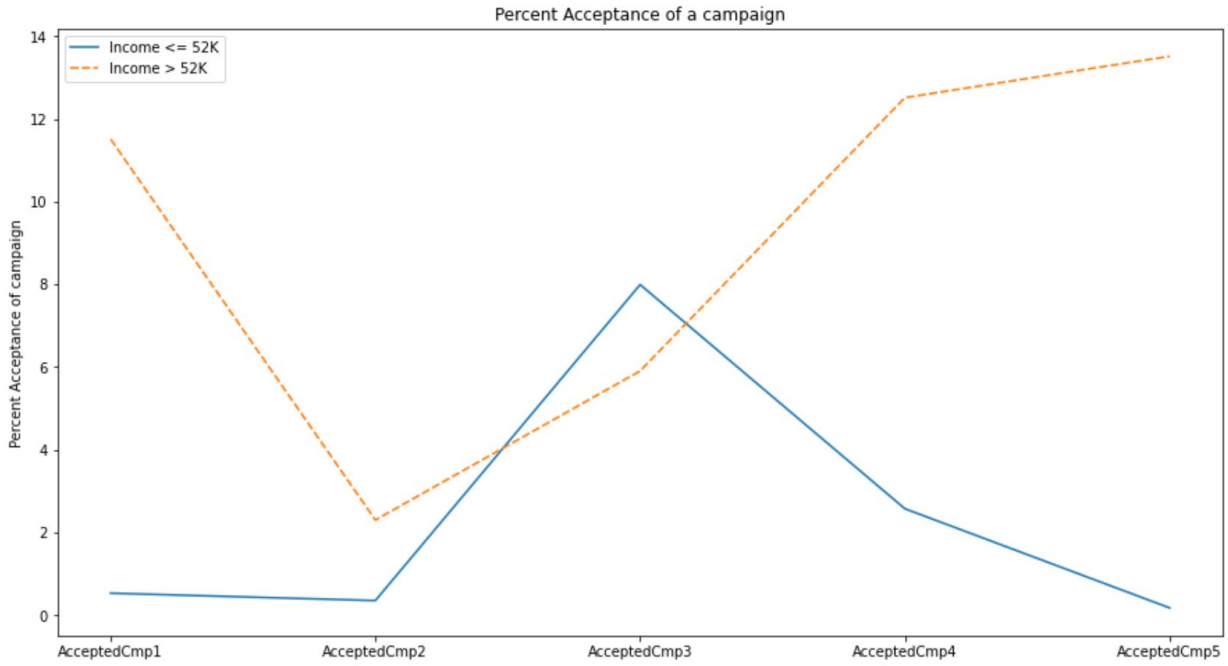


The plot reveals the following correlations:

- Income has a positive correlation to all categories except Deals and Web Visits
- Customer spending has a high positive correlation with wine and meat sales
- Wine is purchased most often using catalogs
- Total spending is highly correlated with catalog use
- Deals only have a positive correlation with web purchases regarding channels
- Meat is most often purchased from a catalog
- Customers spend more using a catalog
- Gold products are purchased most often using a catalog

Marketing Campaign Performance

The mean income for customer in the data set is $52,000. Plotting the percent each campaign is accepted by customer is valuable in understanding how the campaigns are performing illustrated in the following line plot:



Some observations can be drawn from this chart the number of customer that have purchased more than one products. For example:

- The average number of purchases is greater than 1 indicating customer loyalty across all accepted campaigns.
- **Campaign 1**
  - Inferred to be a catalog-based campaign targeting wine, meat, fish, and sweets
  - 121 customers earning greater than 52,000 have responded to this campaign
- **Campaign 2**
  - This is poorest performing campaign
  - Only 27 customers accepted it
- **Campaign 3**
  - Inferred to be a website-based campaign targeting gold
  - 149 customers have responded to Campaign 3
  - The only campaign with a higher acceptance race of customers with incomes 52,000 and less
  - Has the 2nd highest overall total percent acceptance
- **Campaign 4**
  - Inferred to be a website, store, and catalog-based campaign targeting wine
  - 154 customers have responded to Campaign 4
  - Has the highest overall acceptance
  - Near equal acceptance across customers below, equal to, and above incomes of 52k
- **Campaign 5**
  - Inferred to be a catalog-based campaign targeting wine and meat
  - 137 customers earning >> 52k have responded to this campaign

Non-Linear and Linear Dimensionality Reduction

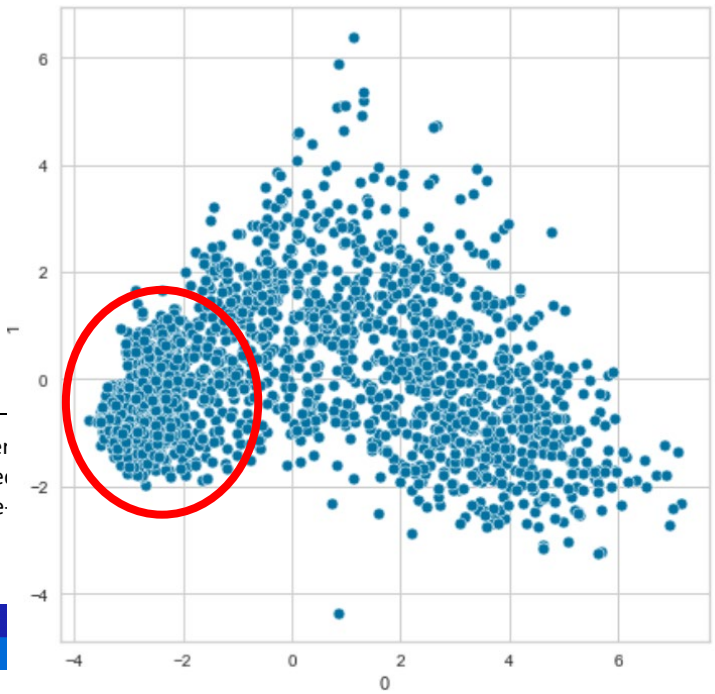
Principal Components Analysis (PCA) – Linear

Principal Components Analysis (PCA) is a well-known unsupervised dimensionality reduction technique that constructs relevant features/variables through linear (linear PCA) or non-linear (kernel PCA) combinations of the original variables (features). In this post, we will only focus on the famous and widely used linear PCA method.[1]

Applying this linear technique to our data we can reduce the number of principal components from 14 to 7 which is a 50 percent reduction explaining more than 88 percent of the variance. Using the median of PC1 as a threshold, we can see correlations between categories. In the illustration below, blue indicates a positive correlation and pink indicates a negative correlation:

	PC1	PC2	PC3	PC4	PC5
Income	0.320000	-0.050000	0.100000	-0.230000	0.080000
MntWines	0.290000	0.150000	0.360000	-0.070000	-0.010000
MntFruits	0.240000	-0.170000	-0.310000	0.200000	-0.300000
MntMeatProducts	0.290000	-0.180000	0.040000	-0.010000	-0.130000
MntFishProducts	0.250000	-0.190000	-0.270000	0.240000	-0.130000
MntSweetProducts	0.250000	-0.170000	-0.250000	0.190000	-0.380000
MntGoldProds	0.210000	0.090000	-0.210000	0.550000	0.730000
NumDealsPurchases	-0.020000	0.610000	-0.200000	0.020000	-0.040000
NumWebPurchases	0.230000	0.420000	-0.050000	0.030000	-0.190000
NumCatalogPurchases	0.310000	-0.030000	0.060000	-0.020000	0.120000
NumStorePurchases	0.280000	0.160000	-0.080000	-0.390000	0.090000
NumWebVisitsMonth	-0.220000	0.380000	0.040000	0.410000	-0.340000
Total_Spending	0.340000	-0.020000	0.140000	0.050000	-0.050000
Total_Purchases	0.300000	0.350000	-0.080000	-0.160000	0.010000
Total_Campaigns_Accepted	0.140000	0.000000	0.710000	0.390000	-0.100000

From the illustration above we can infer that income has a positive correlation with consumer spending and purchases. Popular products are meats and wines purchase using the store and catalog channels. There is also a strong correlation between deals and visits to the website where earlier in this report it was concluded that 50 percent of customers visiting the website make a purchase.

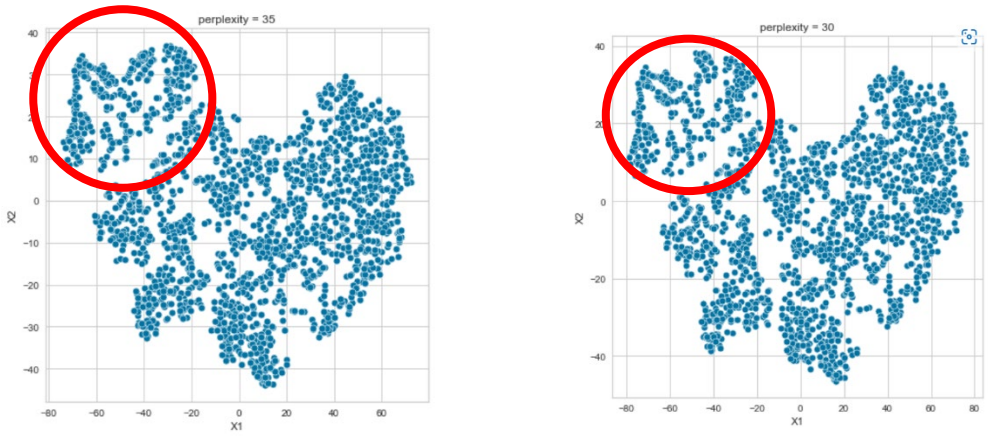


[1] Loukas, S. (2021, October
python. Medium. Retrieve
it-and-feature-importance
portance: A guide in
d-how-when-why-to-use-

PCA shows a tighter cluster on the left side shown in red, but there is no clear picture at this point identifying this cluster.

T-distributed Stochastic Neighbor Embedding (T-SNE)

T-SNE is a nonlinear dimensionality reduction algorithm allowing for the separation of data that cannot be separated by a straight line.[2] The algorithm is used to understand and cluster data. For the analysis, we set the perplexity to both 30 and 35, which is the target number of neighbors from a central point.



You can see the data beginning to cluster indicated in red, but there isn't a real clear separation into groups.

Clustering

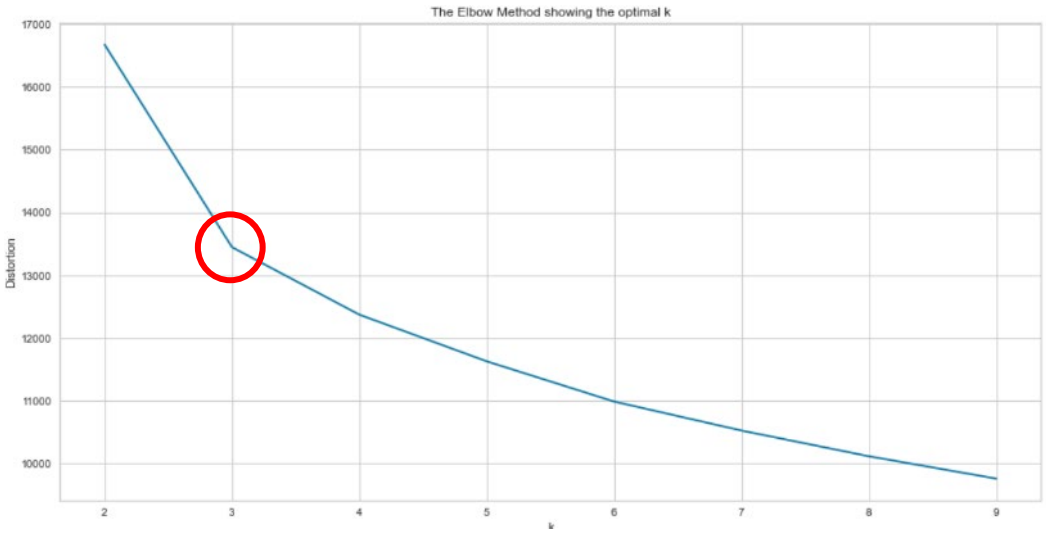
K-Means Clustering

K-means clustering is one of the most widely used methods, unlike PCA and T-SNE, K-Means can be used for both linear and non-linear use cases. It uses randomly selected data points provided as the k value. Think of these data points as the center of a cluster. If the k value is three, then three randomly selected data points will represent the center of three distinct groups. The distance between all the remaining data points and the center points is measured, with each data point getting assigned to the closest cluster. The distance can be measured (linear) or calculated (non-linear). The Euclidian distance is used to measure

[2] (burnpiro), K. E. (2022, July 21). T-SNE clearly explained. Medium. Retrieved July 28, 2022, from https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

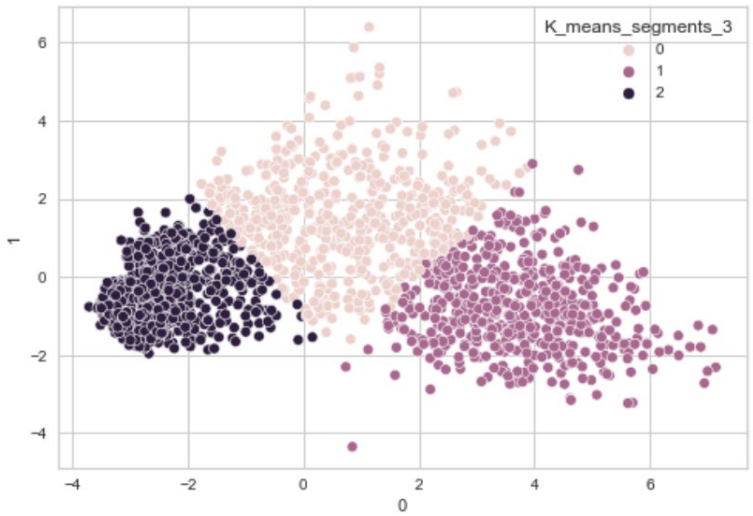
the distance between the data points and the selected center of each cluster in non-linear use cases. After assigning each data point to a cluster, there will be k clusters or groups of data points. The K-means algorithm checks each cluster's quality by calculating the mean of all the values and measuring the variance of the mean value to the previously selected cluster centers. This process repeats if the variance is high until the clusters no longer change. At the end of this process, the K-means algorithm uses the clusters with the least variance.

One problem with K-means clustering is determining the optimal value of k. Variance can be reduced by increasing the k value until the variation equals zero, meaning each data point is its own cluster and thus not a desirable result. The best way to determine the value of k is by plotting the reduction variance per value for k. Larger reductions can be identified to provide the optimal number for k. Plotting the reduction



variance is named an “elbow plot.” The marketing data shows a sharper turn (more significant reduction) at value three, as shown in red in the plot below.

Using three as the value of k, the K-means cluster algorithm has identified three distinct clusters of data points shown below.



K-Means Customer Grouping Results

Group 0 Characteristics and Behaviors

- Income is higher than all the segments
- Spend the most money of the three segments
- Visit the website the least of all three segments
- Most items are purchased using the catalog channel
- Both catalogs and stores are used most frequently\
- Deals and the website are channels used the least in this group

Group 1 Characteristics and Behaviors

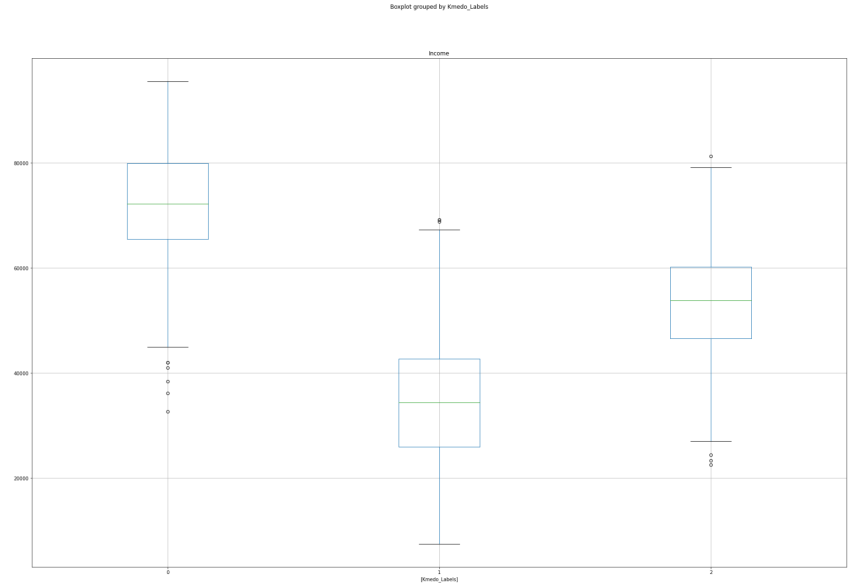
- They purchase more items than any other group
- Deals are the most used channel which may explain why this group purchases more, but spends less that group 0
- The website is the preferred channel of making purchases which is driven by deals
- Purchase wine and gold the most often

Group 2 Characteristics and Behaviors

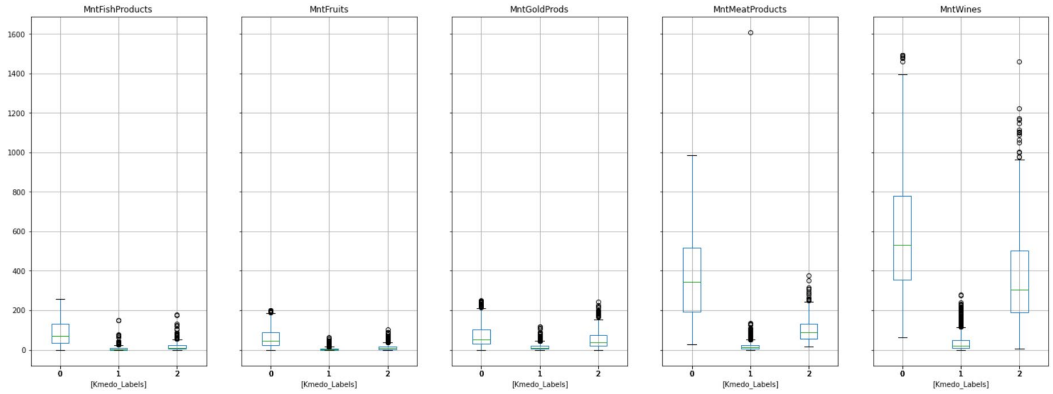
- Spend and purchase the least of the groups
- They used deals the most of any group
- They use the website to make purchases
- Purchase gold and fruit the most, but also purchase fish and sweets
- Wine is purchased the least of the groups

K-Medoids Clustering

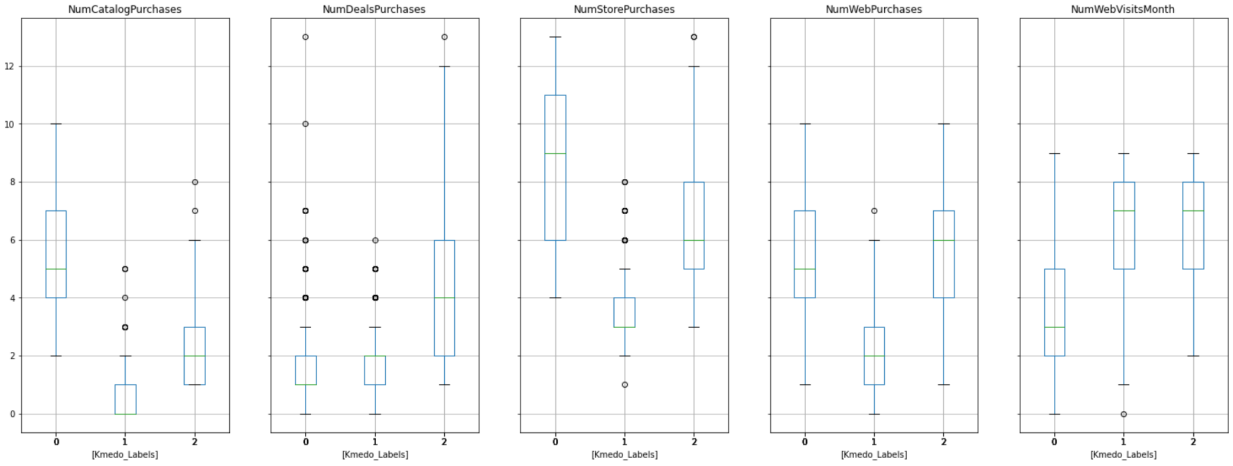
The K-Means method is sensitive to outliers because it uses the mean of a cluster as the center point of reference for the cluster. K-Mediods is another method to cluster the data and it is not sensitive to outliers because the algorithm uses a random point in the cluster and calculates the some of the square's error (SSE). The algorithm continues to enumerate the cluster and the point with the lowest SSE becomes the central point of the cluster. The difference between K-Means and K-Medoids is using a mean vs an actual data point as the center reference point of a cluster. Applying the K-Medoids model and plotting the data based on income, products, and channels results in the following plots:



Boxplot grouped by Kmedo_Labels



Boxplot grouped by Kmedo_Labels



K-Medoids Customer Grouping Results

Group 0 Profile

- Has the highest average income
- Meat and wine are the favorite products to purchase, although there is a high about of variance in both
- Purchase wine more than any other group
- This group prefers purchasing in stores, however will use both a catalog and website as well
- Likely to respond to campaign 4 most often, and campaign 1

Group 1 Profile

- Has the lowest average income
- Meat and wine are the favorite products to purchase
- The website is the preferred channel of making purchases followed by visits to stores

Group 2 Profile

- Incomes are closer to the average, the middle between Group 0 and Group 1
- Prefer wine, meat, and gold products
- They use the stores and websites equally followed by catalogs to make purchases
- Responds best to deals

Conclusion

Key Takeaways

What are the lessons learned? Key takeaways are:

1. A lack of understand the data impacts business
2. Understanding customer characteristics and behaviors is critical
3. Grouping customers according to similarities improves campaign effectiveness
4. Data Science should be incorporated into marketing campaign analysis
5. Optimizing Marketing activities will have a positive impact on revenue and growth

Proposed Model

The purposed model is the K-Medoid method for customer profiling. While the K-Means method is a widely used method, it is sensitive to outliers. The observed Marketing data has a high-amount of outliers that could impact insights due to K-Means use of the mean for calculating the center of a cluster. The K-Medoid method is not impacted by outliers.

Key Problem solved

The key problem solved is a lack of extracting insights into customer characteristics and behaviors. Insights that include identifying customers group by similarities. Segmenting customer has shown to improve campaign acceptance, generate revenue and drive growth.

Recommendations

1. The details and content of the campaigns was not included in the dataset. The data suggests that Campaign 4 performed well across all incomes, three channels, targeting wine. I recommend

reviewing that campaign content and create new campaigns for other products, such as meat, fish, and gold.

2. Capture additional data around customer health habits and reapply the model above. There may be additional insight suggesting bundles of health products, such as fish.
3. Sweets are generally purchased around holidays, there were only five campaigns in the data and there are at least five US holidays (alone) where sweets are popular. Holiday campaigns around sweets is recommended.
4. Now that we're aware wine is a popular product, market research around what wine is good with different kinds of meat and fish. It is suggested that the website incorporates a recommendation system if it hasn't already. If there is a recommendation system, it is recommended to review its configuration and performance based on the customer insights provided in this data science notebook
5. The data suggests that none of the current marketing campaigns are attracting new customers. It appears that once customers have purchased, they continue to purchase. It is recommended that a review of advertising channels to increase traffic to the website, fifty percent of users visiting the website purchase a product
6. Customer aren't seeing enough value in products, their purchase generally are less that $250. It is recommended that a review of product messaging

Further Analysis

1. More detailed data for meat. Does meat include chicken? Collect more data around customer health interests.
2. The use of recommendation systems on the website such as customer that buy this particular wine have also purchase this type of meat. Present users with data driven options.
3. How can the use of bundling products and discounts improve marketing performance and drive sales of less popular items?

Stakeholder Action Items

1. Use this report to create marketing campaigns.
2. Review the metrics from the measures of success, set goals on the expected improvement and compare to the metrics.
3. Look at incorporating Data Science permanently to study marketing and business analytics.
4. Review the data collected and include data points that may produce better insights.
5. Research the benefit of adding a Data Science program with the cost of maintaining a Data Science program.