Francesca Marchese, Ian Smith, Emmanuella Acheampong, Andrew Mukurazita

BUS-212A- Report 2b

BUS 212A – Advanced Data Analytics

Report 2b – Clustering

March 15, 2024

Table of Contents

Project Overview **2**

Descriptive, Exploratory Analysis **3**

Descriptive Statistics and Plots

Regression **9**

Multiple Regression, Model Features, and Interpretations

Classification **15**

Model Specifications and Logistic Regression

Reflection **34**

Citations **37**

Project Overview:

Our project focuses on leveraging data analytics to unlock valuable insights into customer behavior for MarketSphere Inc., a leading retail company in the United States. Our consultancy aims to help MarketSphere in identifying growth opportunities and refining its marketing strategies to better target its customer base.

Progress Update:

In Report 1, we improved data quality significantly, setting a solid foundation for further analysis, through meticulous preprocessing . In Report 2a, we advanced our investigation with the development of three supervised models: a regression model demonstrating substantial predictive power with an 84.28% variance explanation in wine purchases and a RMSE of 7.43; a logistic regression model showcasing a robust accuracy of 81.9% alongside balanced metrics in classifying customer responses to marketing campaigns, as reflected by an F1 score of 81.9%; and a kNN model with an accuracy of 81.06%, identifying K = 19 as the most effective parameter, evidenced by an F1 score of 80.62%. These advancements provide MarketSphere with actionable insights and underscore the logistic regression model's superior balance in precision and recall, optimizing customer response identification.

Introduction to Report 2b:

Building upon the robust groundwork laid in Report 1 and the advanced analytical applications of Report 2a for MarketSphere Inc, Report 2b ventures into the aspect of unsupervised learning and model enhancement. In this phase, our focus shifts towards uncovering latent structures within the dataset through Hierarchical and k-Means clustering techniques, and evaluating the impact of these new insights on our existing predictive models.

Descriptive, Exploratory Analysis

Regression

Justification for Regression Target Variable Selection ("Wines"):

In the context of Market Sphere's retail analytics, 'Wines' emerged as the primary product of interest due to its status as the most frequently purchased item among their diverse product range with significant contribution to the total revenue and encapsulates varied customer behaviors and preferences. This variable is an excellent candidate for regression due to its continuous numeric nature. Accurate prediction of wine purchases is crucial for MarketSphere as it aids in effective inventory management, targeted marketing strategies, financial planning, and enhances overall customer satisfaction and profitability.

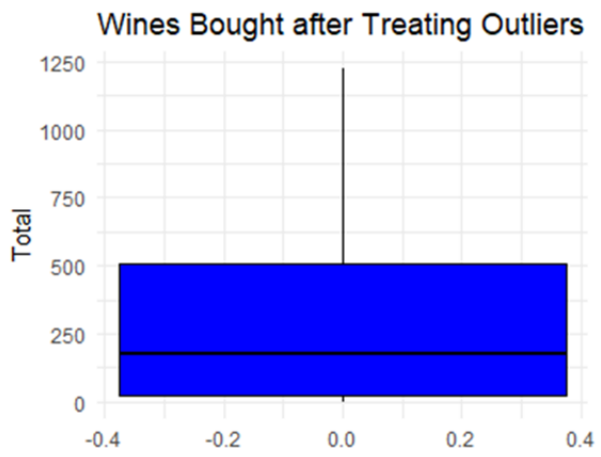
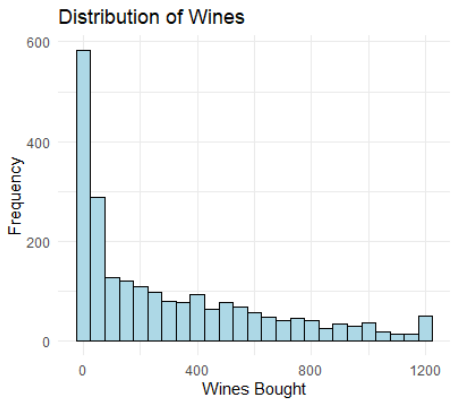


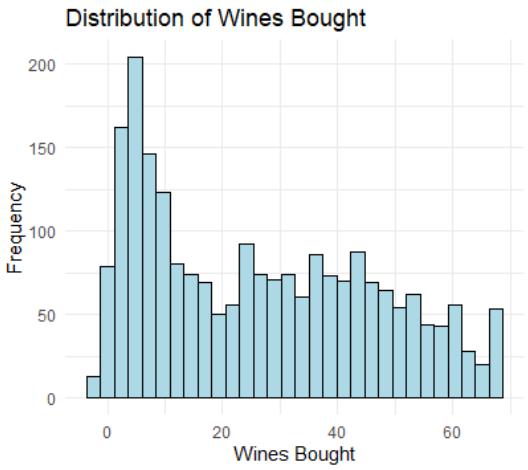
Figure 7 - Total Wines bought without outliers.

The summary statistics for 'Wines' purchases, as exhibited by the dataset encompassing 2,236 data points, demonstrates a broad spectrum of customer spending. Specifically, the data ranges from a minimum of $0 (indicating customers who did not purchase wines) to a maximum of $1,224.6, with a median value of $174. This median, in addition to the mean value of $302.3, illustrates the skewed nature of wine purchasing behavior: a significant number of customers make modest wine purchases, while a smaller, affluent segment contributes to higher sales volumes. The first and third quartiles, at 24 and 504.2 units, respectively, further highlight the diverse spending patterns among customers.

This skewed distribution is critical for regression analysis. However, linear regression models, the chosen method for this analysis, assume normality in the distribution of the dependent variable. Therefore, the initial skewness in 'Wines' distribution can lead to biased estimates, affecting the model's reliability and accuracy.



To address this, a Box-Cox Transformation was applied to the 'Wines' variable, aiming to normalize the distribution and stabilize variance. This



methodological approach is particularly effective in handling skewed data, ensuring that our regression models can produce more accurate and interpretable results.

The histograms (before and after transformation) serve as visual confirmations of the distribution's adjustment. With a focus on 'Wines' as our target variable, we aim to unravel the underlying factors influencing wine purchases at MarketSphere. Understanding these can provide actionable insights into customer preferences, enabling more targeted marketing strategies and optimized inventory management, ultimately fostering enhanced business performance and customer satisfaction.

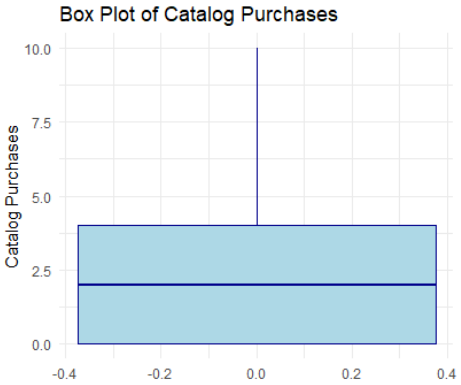
Feature Selection

Model Building and Selection:

For our initial model, an exhaustive search was conducted to identify the best subset of predictors for our linear regression model. Though about 15 variables were arrived at during the search, the final variables were chosen based on their VIF, relevance and statistical significance. The final model included 'CatalogPurchases', 'StorePurchases', 'Kidhome', and 'Complain' as predictors. These variables were identified as having a strong correlation with the target variable, while minimizing multicollinearity among themselves with VIFs (1.79, 1.81 and 1.50)

Plots for Visual Correlation and Further Statistics of Predictors:

1. CatalogPurchases:



Minimum: 0 (some customers did not make any purchases through catalogs)

Maximum: 10 (some customers made up to ten catalog purchases)

Mean: 2.625 (on average, customers made approximately three catalog purchases)

Median: 2 (half of the customers made two or fewer catalog purchases)

Standard Deviation: The range and interquartile values indicate variability in catalog purchase behavior among customers.

2. StorePurchases:

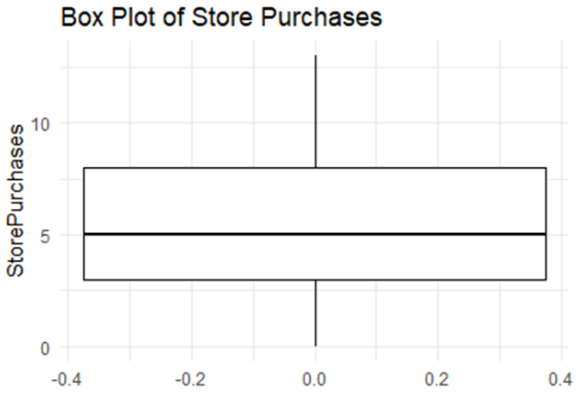


Figure 17 - Store purchases

Minimum: 0 (some customers did not make any in-store purchases)

Maximum: 13 (a few customers made up to thirteen purchases in-store)

Mean: 5.796 (on average, customers made about six in-store purchases)

Median: 5 (half of the customers made five or fewer in-store purchases)

Standard Deviation: Akin to CatalogPurchases, variability is evident given the spread between the minimum and maximum.

3. Kidhome:

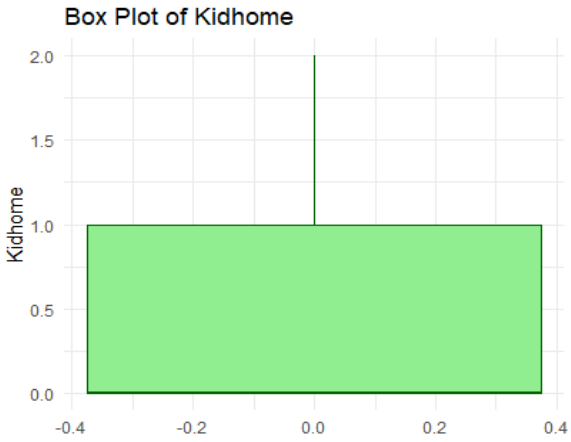
Minimum: 0 (some customers have no children at home)

Maximum: 2 (the highest number of children at home reported by customers)

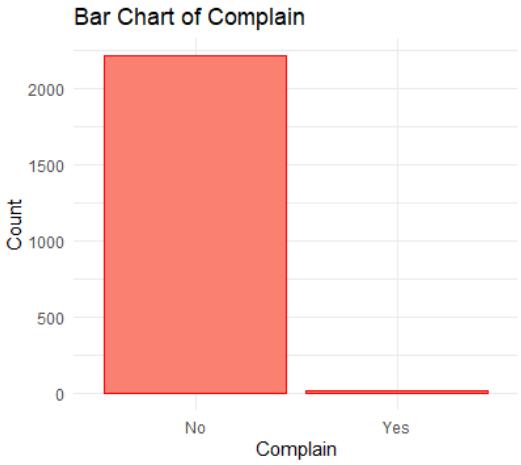
Mean: 0.4441 (indicating that, on average, less than one child is present in customers' homes)

Median: 0 (more than half of the customers reported having no children at home)

Standard Deviation: Indicate a distribution leaning towards fewer children at home.



4. Complain:

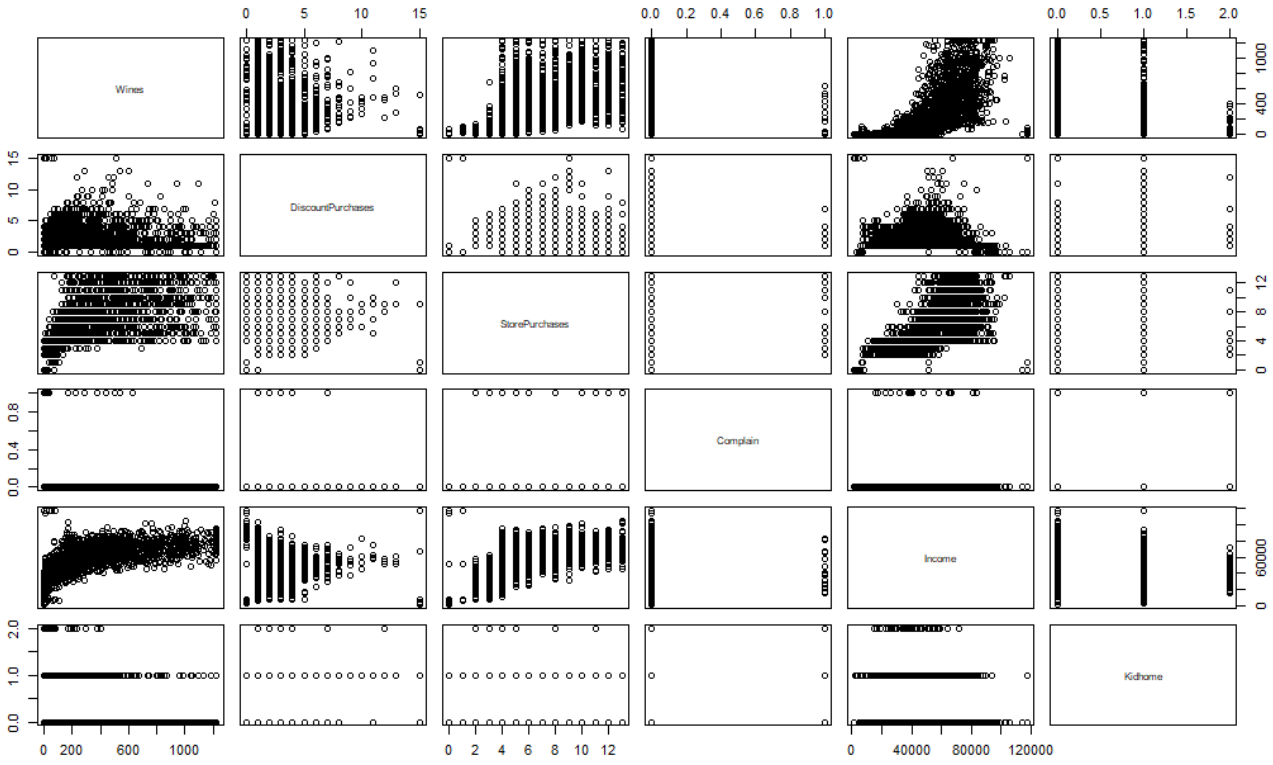


Minimum: 0 (indicating that most customers did not register complaints)

Maximum: 1 (20 complains were made in total the over 2 years)

Mean: 0.008944 (very few customers have lodged complaints)

Scatter Plot of Initially Considered Predictor Variables:



In examining the relationships between 'Wines' and the predictors, we observed that 'Wines' spending positively correlates with 'CatalogPurchases' and 'StorePurchases', indicating that customers who spend more through catalogs or in stores tend to also spend more on wines. Conversely, there is a negative correlation between 'Wines' and 'Kidhome', suggesting that having more children at home may lead to reduced wine expenditures. The variable

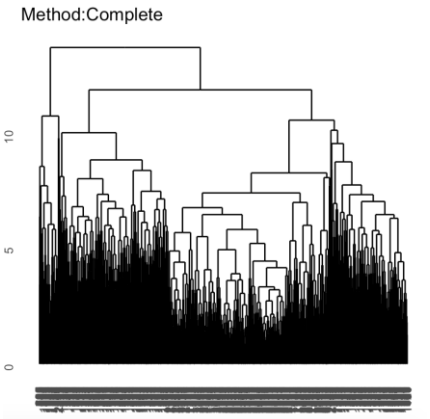
'Complain' does not exhibit a clear relationship with 'Wines', implying that customer complaints may not significantly influence wine purchasing habits. Additionally, there doesn't appear to be a relationship between the predictors, indicating that the predictors are identical and independent (i.i.d linear regression assumptions) of each other in terms of customer preference. 'Complain' was however not significant to the prediction of Wines bought, hence it was excluded from the model.

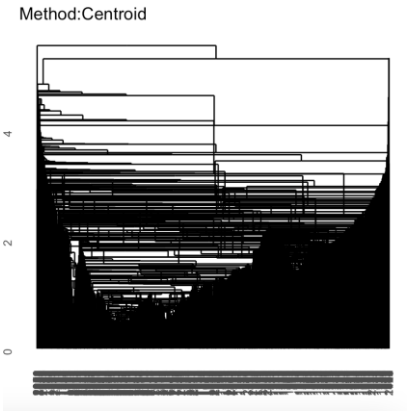
As we expanded our original model to improve its effectiveness at making accurate and meaningful predictions, we included two cluster variables - Hcluster and Kmcluster - that were both statistically significant, as their p-values were practically 0 and VIF as follows:

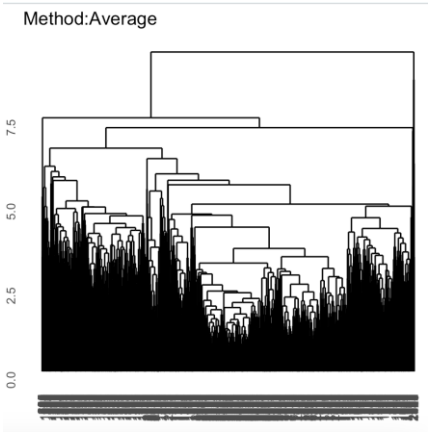
```
vif(lm_model)
CatalogPurchases StorePurchases     Kidhome       Hclusters      Kmcluster
2.48             2.19                  1.49        2.35             1.34
```

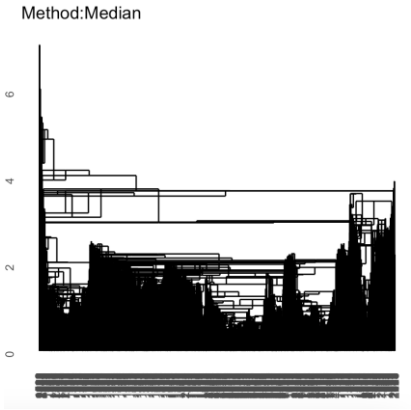
Clustering Process:

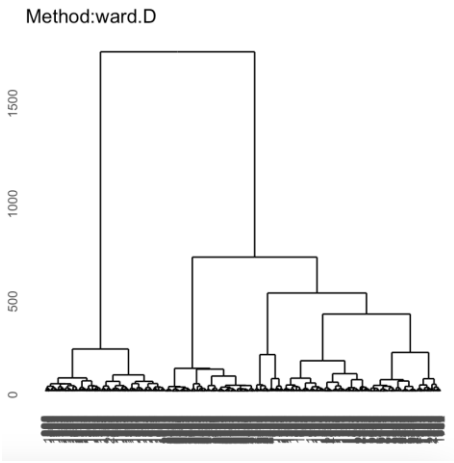
In the updated linear regression analysis, various hierarchical clustering methods, including complete, average, median, and centroid. Each method was evaluated based on its ability to produce coherent and distinct clusters. While the complete, average, median, and centroid methods did show different clustering behaviors, with differing degrees of balance and internal cluster homogeneity that provided valuable insights into consumer preferences and behaviors, the Ward.D2 method was selected as the optimal approach. Its ability to generate balanced, meaningful clusters allowed for the identification of different customer groupings that aids in MarketSphere Inc.'s ability to select ideal personas for selective target marketing that will increase customer engagement.

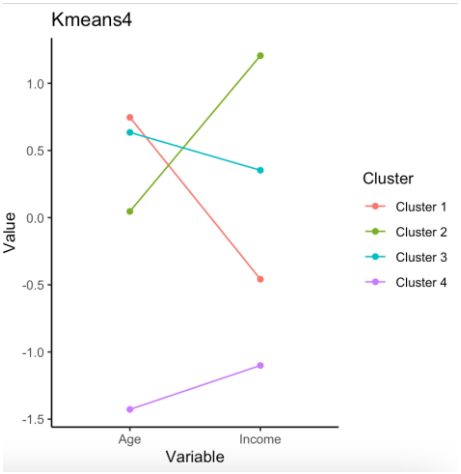








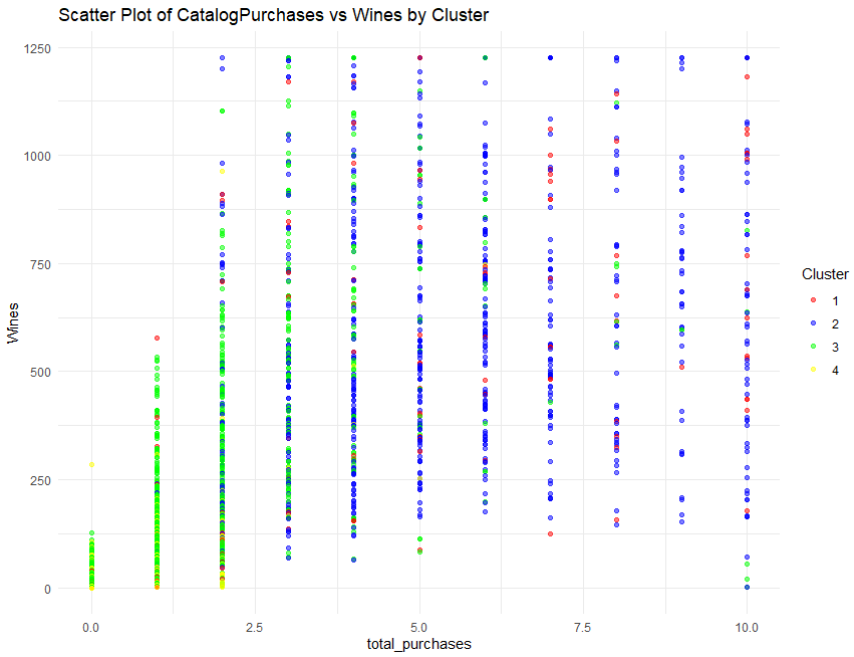




Ultimately, hierarchical clustering with the Ward.D method and k-Means clustering were explored to identify groups within our data that share commonalities. This affirms that there are meaningful differences between the clusters, highlighting distinctive customer subgroupings that can aid in MarketSphere Inc.'s ability to predict purchasing behaviors concerning consumer goods.

Clustering Justification:

The scatter plot below shows the clusters that were generated from k-Means Clustering which we observ to be similar to that generated by Hierarchical clustering.



Based on their characteristics from the scatterplot (left), we could name these clusters as follows:

Cluster 1 (Red): "Wine Enthusiasts" - This group purchases a significant number of wines compared to their total purchases, indicating a strong preference for or interest in wine.

Cluster 2 (Green): "Occasional Shoppers" - Individuals in this group have low total purchases and low wine purchases, suggesting they may shop infrequently or purchase items sparingly.

Cluster 3 (Blue): "General Consumers" - These customers have a moderate level of total purchases and include wines as a part of their broader shopping habits but not as prominently as the Wine Enthusiasts.

Cluster 4 (Purple): "Heavy Shoppers" - Characterized by very high total purchases, this group buys a variety of products, including wines, indicating they are frequent or bulk shoppers.

Linear Regression Model Performance:

The original linear regression model yielded an adjusted R-squared value of 0.8428, indicating that approximately 84.28% of the variance in wine purchases can be explained by the predictors in the model. The RMSE (Root Mean Square Error) for the training data was approximately 7.05, indicating the typical deviation from the observed values. The addition of the hierarchical cluster variable 'Hclusters' and the K-Means cluster variable 'Kmcluster', our model's adjusted R-squared increased to 0.85, while our training data's RMSE decreased just shy of 33% to 6.77. A lower RMSE is preferred, as it suggests that the model's predictions are

closer to the actual values, which can enhance MarkeShpere's ability to advise clients on particular customer groupings and characteristics.

Without Clustering

```
Residual standard error: 7.66 on 1599 degrees of freedom
Multiple R-squared:  0.843,     Adjusted R-squared:  0.84
F-statistic:  295 on 29 and 1599 DF,  p-value: <0.0000000000000002
```

	Training	Holdout
RMSE	7.05	14.6
MAE	5.58	10.9

With Clustering

```
Residual standard error: 6.79 on 1210 degrees of freedom
Multiple R-squared:  0.85,      Adjusted R-squared:  0.85
F-statistic: 1.37e+03 on 5 and 1210 DF,  p-value: <0.0000000000000002
```

	Training	Holdout
RMSE	6.77	14.0
MAE	5.36	10.8

We observed that the RMSE and MAE of the train data appear lower than that of the holdout data. This may be a result of the differences in the distribution of both the training and the holdout set; ultimately, this is due to the skewness in the dataset. Hence, using the cook's distance metric, we were able to identify the outliers in the training dataset efficiently and address them, leading to the reduction in the training RMSE and MAE. This, however, decreased further for the model with the clustering variables. It could be an indication of the original model underfitting due to the limited variables. Hence, the addition of new variables from clustering resulted in not just an increase in adjusted R squared, but also a reduction in the RMSE and MAE. It is worth noting that our attempts to integrate higher-order terms, such as squared and interaction terms for catalog and store purchases (such as catalog_purchases_squared, store_purchases_squared and catalog_purchases*store_purchases), were thwarted by high variance inflation factors (VIF), indicating problematic multicollinearity. Consequently, these terms were excluded from the model.

Ultimately, an increase in adjusted R-squared and the decrease in RMSE suggests that the addition of cluster variables has enhanced the model's explanatory power and predictive accuracy and, therefore, is the preferred linear regression model. There are clearly underlying groupings in the data that influence wine purchases; incorporating these cluster variables aids in capturing and accounting for these patterns/shared behaviors.

Regression Coefficients:

```
Residuals:
    Min      1Q  Median      3Q     Max
-24.358  -4.541  -0.558   3.995  22.616

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)       17.9670     1.7537   10.25 < 0.0000000000000002 ***
CatalogPurchases   3.0734     0.1320   23.29 < 0.0000000000000002 ***
StorePurchases     2.4077     0.0943   25.53 < 0.0000000000000002 ***
Kidhome           -1.9457     0.4387   -4.44     0.00001003054702 ***
Hclusters         -3.3887     0.3576   -9.48 < 0.0000000000000002 ***
Kmcluster         -1.5041     0.2043   -7.36     0.00000000000033 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.79 on 1210 degrees of freedom
Multiple R-squared:  0.85,      Adjusted R-squared:  0.85
F-statistic: 1.37e+03 on 5 and 1210 DF,  p-value: <0.0000000000000002
```

Regression Coefficients Interpretation:

Intercept (17.9670): The estimated number of wine purchases when all other predictors are zero. This is the baseline against which the effects of other variables are compared.

CatalogPurchases (3.0734): For each additional catalog purchase, wine purchases increase by approximately 3.07 units. This significant positive relationship suggests that customers who buy more through catalogs tend to also buy more wine.

StorePurchases (2.4077): Similarly, for each additional store purchase, wine purchases increase by approximately 2.41 units. This indicates that in-store shopping is also positively associated with wine purchases, but slightly less so than catalog purchases.

Kidhome (-1.9457): Having children at home is associated with a decrease in wine purchases by approximately 1.95 units. This could reflect budgetary constraints or lifestyle differences of households with children.

Hclusters (-3.3887): This coefficient suggests that being part of a certain household cluster reduces wine purchases by about 3.39 units. This indicates distinct purchasing patterns or preferences among different customer clusters.

Kmcluster (-1.5041): Similar to Hclusters, membership in a certain k-means cluster reduces wine purchases by approximately 1.50 units. Again, this points to varying behavior or preferences across different market segments.

Model Fit and Statistical Significance:

Residual Standard Error (6.79): The typical deviation of the observed wine purchases from the regression line. Lower values indicate better model fit.

Multiple R-squared (0.85) and Adjusted R-squared (0.85): These indicate that approximately 85% of the variability in wine purchases is explained by the model, which is quite high, suggesting a good fit.

F-statistic: The overall significance of the regression model is very high (practically zero p-value), indicating that there is a less than 0.01% chance that these results could occur if none of these variables actually affected wine purchases.

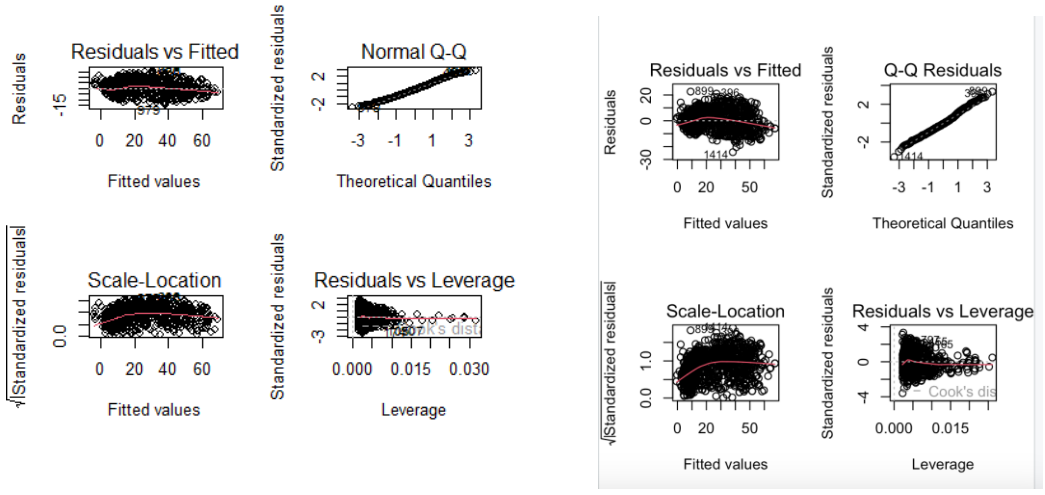
Implications for MarketSphere:

For MarketSphere, this analysis suggests that strategies aimed at increasing catalog and in-store purchases could boost wine sales. However, marketing approaches may need to be adjusted for households with children and across different customer segments identified by clustering. The distinct impacts of 'Hclusters' and 'Kmcluster' suggest that personalized marketing, tailored to the characteristics of each cluster, could be particularly effective.

Diagnostics:

The below diagnostic plots were reviewed to assess the validity of the linear regression assumptions for both the original model (left) and the linear regression featuring the clustering variables (right).

Without Clustering *With Clustering*

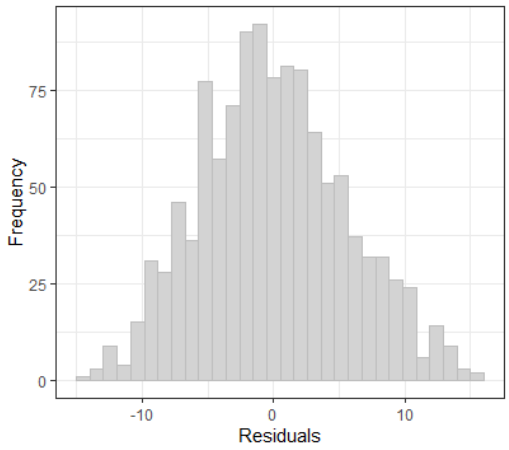


a) *Residuals vs. Fitted Values plot*: for the original model (please refer to plot on the right above) the residuals appear randomly scattered around the horizontal line at zero, without any apparent pattern. This randomness suggests that

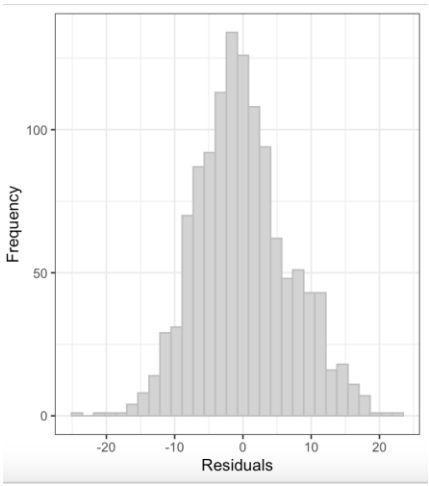
the linear model fits the data well without obvious violations of homoscedasticity or indications of non-linear relationships. However, with the introduction of the cluster variables, we observed a more random dispersion of points around the zero line, which indicates that including cluster variables might have helped capture some of the non-linear aspects the model without clustering missed.

b) The histogram of residuals of the original linear regression model (left) indicates a Gaussian distribution, showing reasonable adherence to the normality assumption but have their peculiarities: the "Without Clustering" model shows a tighter, possibly more normal distribution of residuals, whereas the "With Clustering" model shows expanded variance and potential skewness.The wider spread in the "With Clustering" residuals might reflect a more complex model capturing a broader range of behaviors but also may indicate increased model variance.

Without Clustering



With Clustering



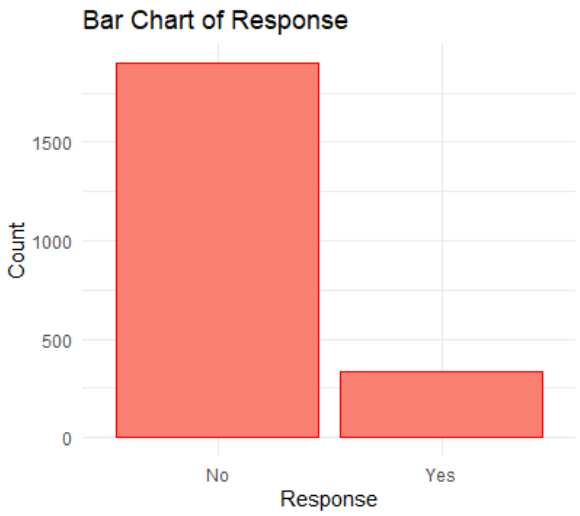
c) The *Normal Probability Plot of the Residual* confirms that the residuals are approximately normally distributed, in both the original and clustered regression models, satisfying one of the key assumptions of linear regression.

Diagnostic Residual Plots and Outliers:

The diagnostic plots of the original linear regression model provided evidence that the assumptions of linear regression were largely met. About 100 more outliers were removed from the new model using the cook's distance metric. This is however lower than that of the outliers removed from the initial model of 190 observations, representing 8.5% of the dataset. This percentage seems acceptable considering the aim was to improve model accuracy without significantly reducing the dataset's size.

Classification

Justification for Classification Target Variable Selection ("Response"):



The selection of "Response" as the target variable for classification analysis is driven by several strategic considerations. The Response variable is a categorical variable which indicates whether a customer responded to the final campaign or not. 1 indicates acceptance, while 0 indicates otherwise. Our exploratory analysis revealed significant correlations between "Response" and various predictor variables, indicating its importance in understanding customer behavior and engagement levels. In a typical business setting, knowing the response rate to marketing campaigns is essential for evaluating campaign effectiveness and optimizing marketing strategies. By accurately classifying customer responses, MarketSphere Inc. can identify target segments that are more likely to engage with marketing initiatives, tailor communication strategies to their preferences, and ultimately enhance campaign ROI and customer satisfaction. The distribution of Response was imbalanced with only 334 out of 2236 customers (representing 14.95%) responding 1(yes) to the last campaign.

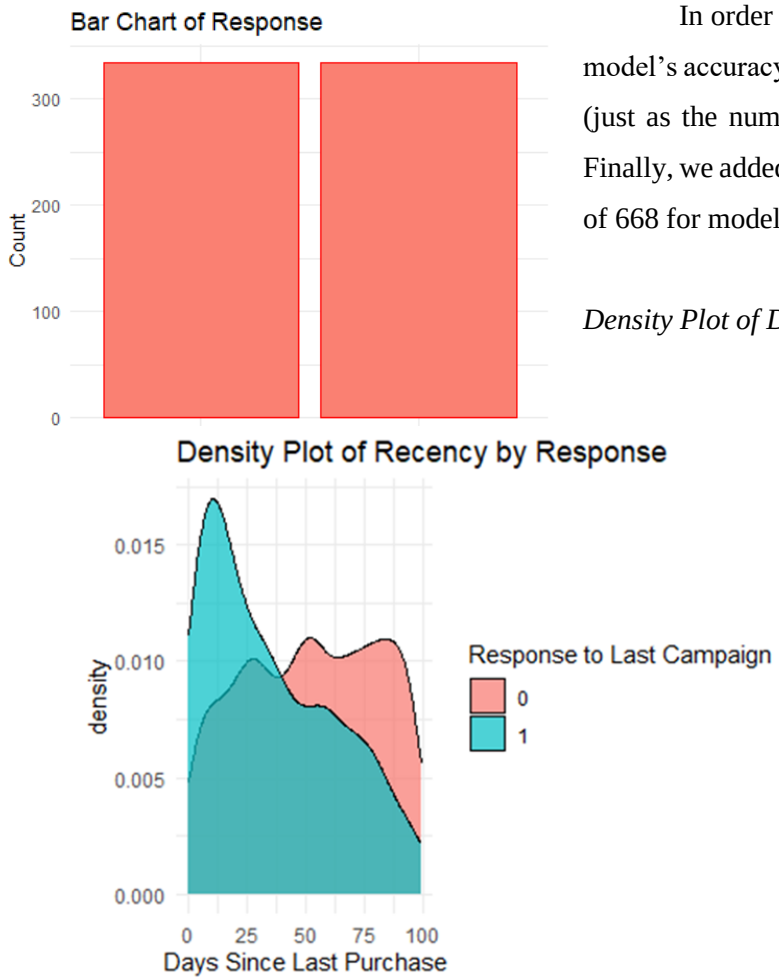


Figure 33 - Days of last purchase vs. Response

In order to prevent this imbalanced distribution from affecting the model's accuracy, we undersampled the majority class (0 responses) to 334 (just as the number of 1 responses) using the RAND function in excel. Finally, we added the sampled points to the minority class and had a dataset of 668 for modeling.

Density Plot of Days Since Last Purchase Vs. Campaign Response:

The "Density Plot of Recency by Response" illustrates the relationship between the number of days since the last purchase (recency) and the response to the last campaign. Customers with a lower number of days since their last purchase (more recent purchases) have a higher positive response (response '1') to the campaign. Conversely, the density of non-responses (response '0') increases as the number of days since the last purchase grows, implying that those who haven't made a purchase recently are less likely to respond to the campaign. This trend aligns with the idea that more recently engaged customers are more receptive to marketing efforts. For MarketSphere, focusing on customers with recent interactions may yield better campaign response rates.

CDF of Income Vs. Campaign Response

The blue line (response '0') and the red line (response '1') indicate the proportion of individuals at or below certain income levels who did not respond and who responded positively to the campaign, respectively. The red line (response '1') is consistently below the blue line (response '0'), suggesting that individuals with higher incomes are less likely to respond positively to the campaign. Both lines rise steeply in the lower income range and then gradually level off, indicating that most of the population falls within a middle-income bracket. The gap between the two lines

narrows as income increases and widens in the middle, showing that the difference in response rates between lower and higher-income individuals vary at different levels of income.

This pattern might imply that campaign messaging or offers are more appealing or relevant to those in the lower income brackets, or that those with higher incomes are less influenced by this particular campaign. This insight could be used to adjust campaign targeting or to tailor messages to be more effective for higher-income brackets.

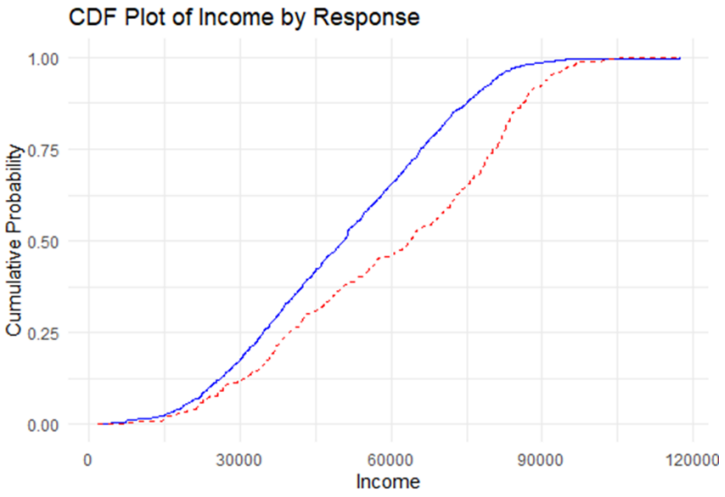


Figure 34 - Income vs. Response

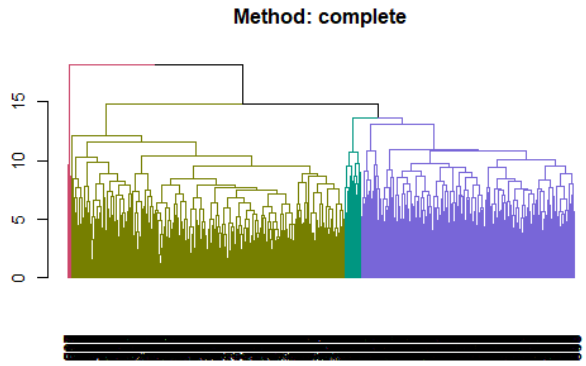
Impact of Clustering Variables on Customer Response Classification

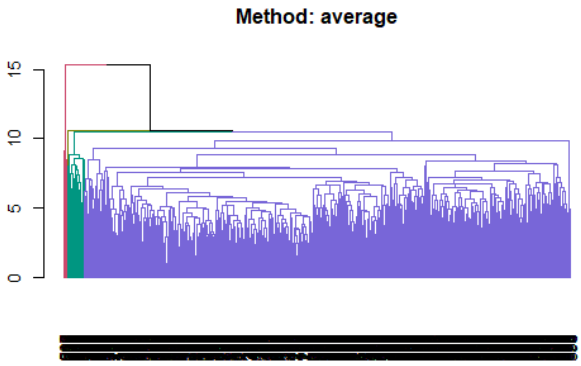
Clustering Process:

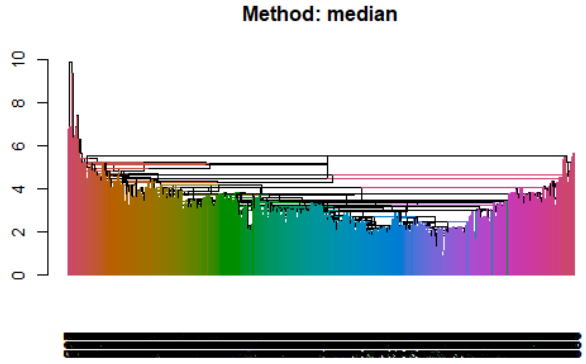
In the updated analysis, a thorough comparison was conducted among various hierarchical clustering methods, including complete, average, median, and centroid. Each method was evaluated based on its ability to produce coherent and distinct clusters:

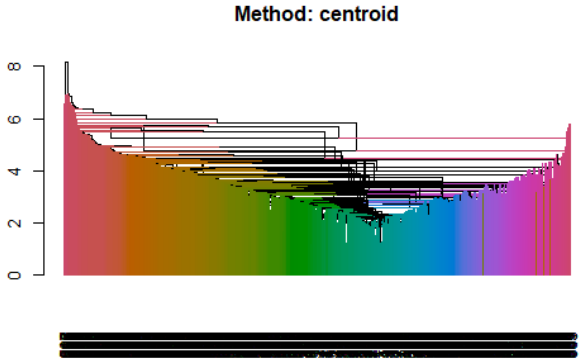
The complete, average, median, and centroid methods showed diverse clustering behaviors, with varying degrees of balance and internal cluster homogeneity. While these methods provided valuable insights, they demonstrated different levels of effectiveness in capturing the underlying structure of the data. The Ward.D2 method was selected as the optimal approach due to its proficiency in generating balanced, meaningful clusters that enables the identification of inherent customer groupings, crucial for actionable segmentation.

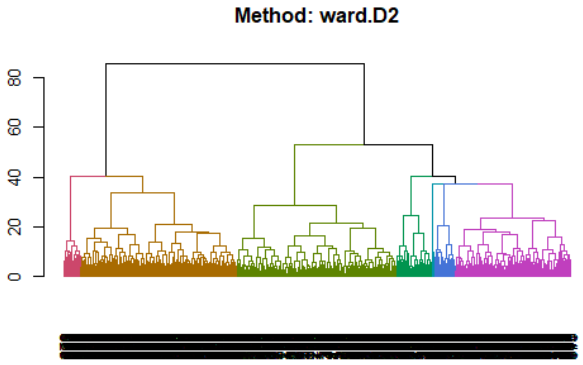
Its hierarchical structure, superior to those produced by other linkage criteria, offers a clear justification for each cluster division, aligning seamlessly with MarketSphere Inc.'s strategic objectives for customer understanding and engagement.

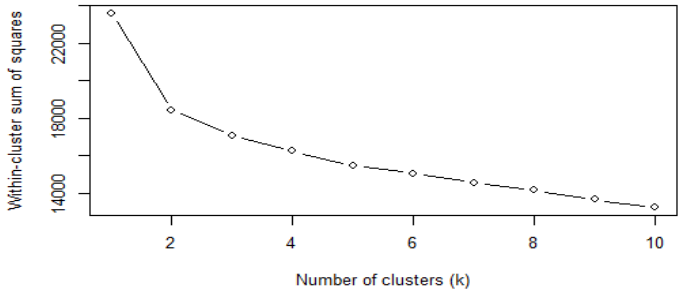










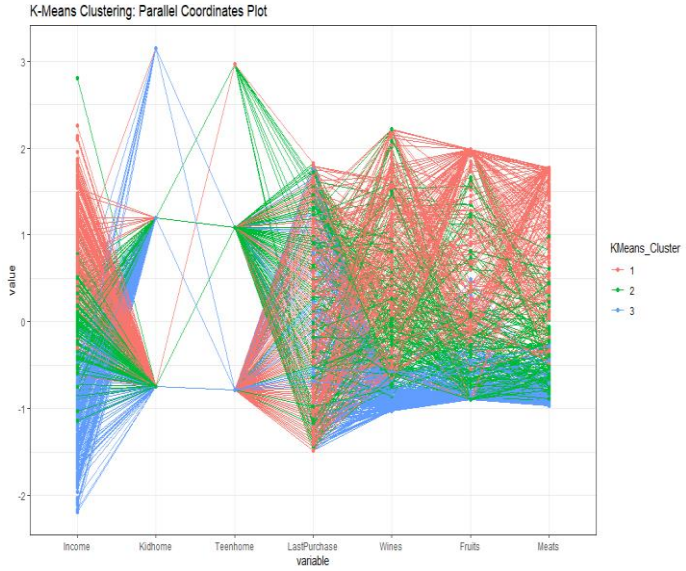


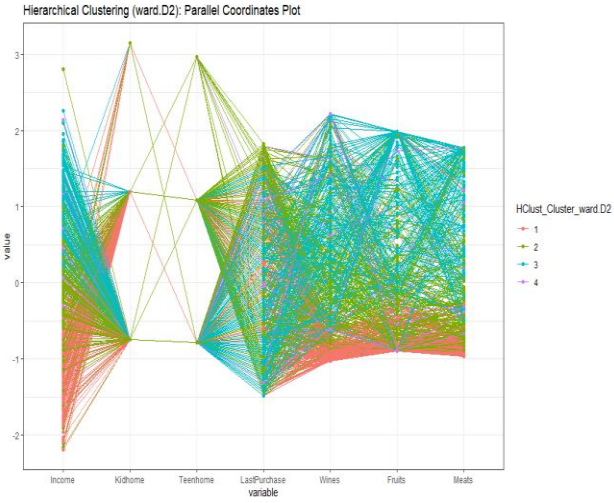
Initially, hierarchical clustering with Ward.D2 method and K-Means clustering were utilized to identify natural groupings within the entire customer dataset. We started by selecting four clusters (K = 4) for hierarchical clustering to maintain a manageable number of segments for marketing strategies, based on natural grouping observed and this resulted in the creation of model 2. However, due to unclear demarcations (please see top left and bottom left charts below) and the desire to capture more granular customer behaviors, the number of hierarchical clusters was

increased to seven (K = 7), resulting in yet another model called model 3. Both showed varying results discussed later in the report.

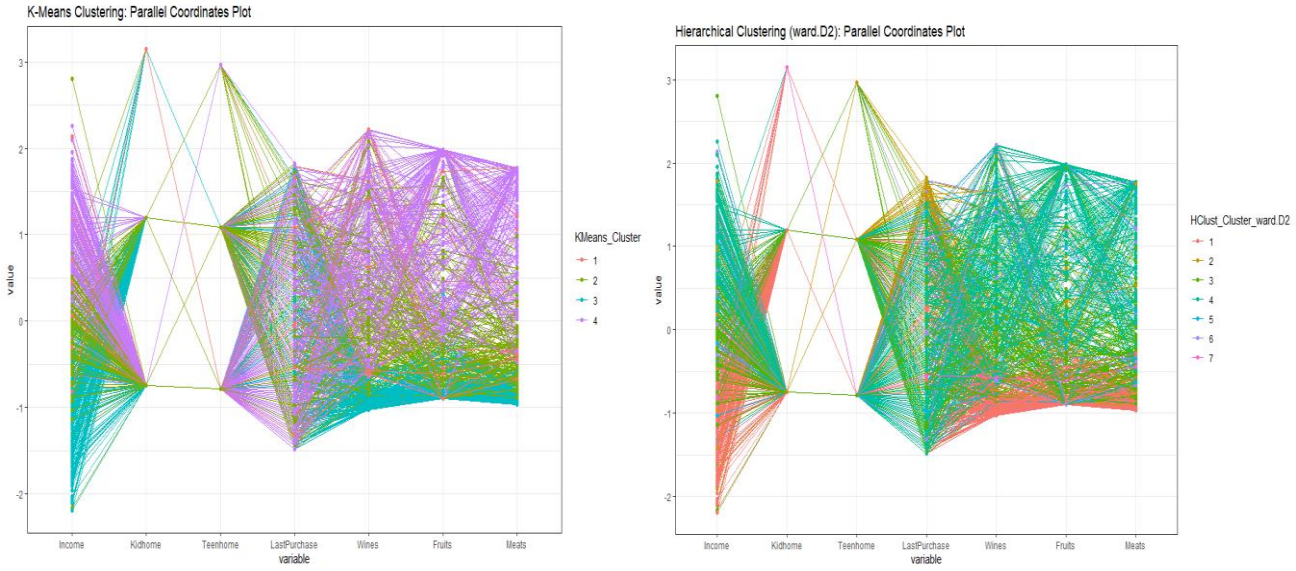
Similarly, for K-Means (please see top right and bottom right charts below) , initially, three clusters were selected based on the elbow method; however, to refine customer segmentation further, the number of clusters was increased to four. The changes were motivated by the need to explore more defined customer groups and enhance targeted marketing strategies. These clusters despite increasing the number of clusters were still not clear to explain due to too many entanglements which we believe is due to the many variations in the dataset.

Model 2 Clusters:

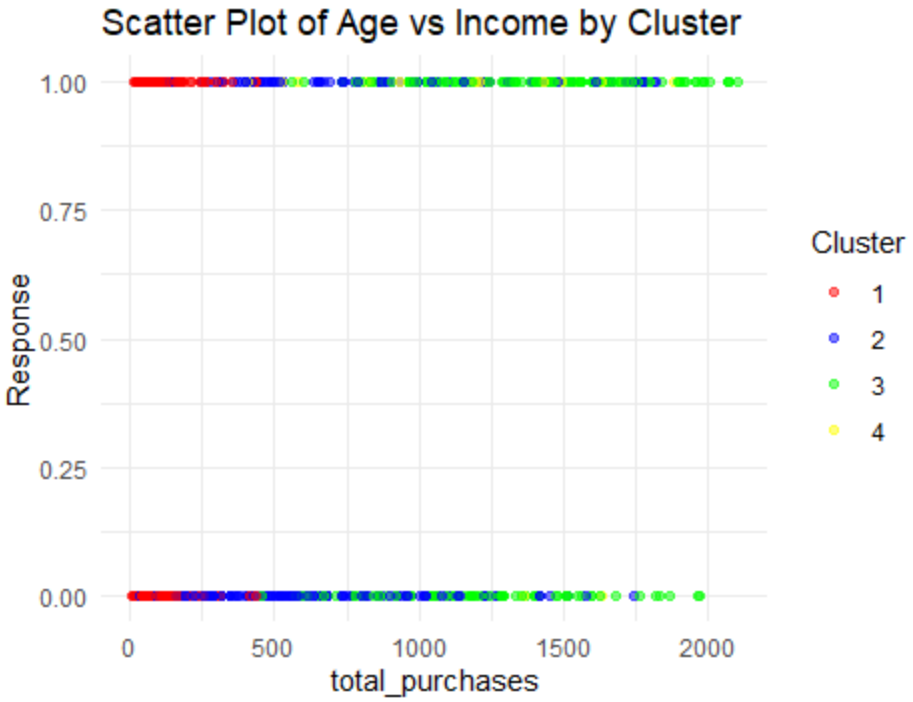




Model 3 Clusters:



To gain a much better understanding of these clusters in relation to our target variable, which is Response to marketing campaigns (0 and 1), we decided to visualize how Responses vary based on total number of purchases. We believe the relationship between customer responses to marketing campaigns and their total number of purchases will help MarketSpare tailor marketing strategies effectively, allocate resources efficiently, segment customers accurately, predict future trends, and measure campaign performance. This approach leads to more personalized marketing, better customer engagement, and improved business outcomes.



Based on their characteristics from the scatterplot (to the left), we could name these clusters as follows:

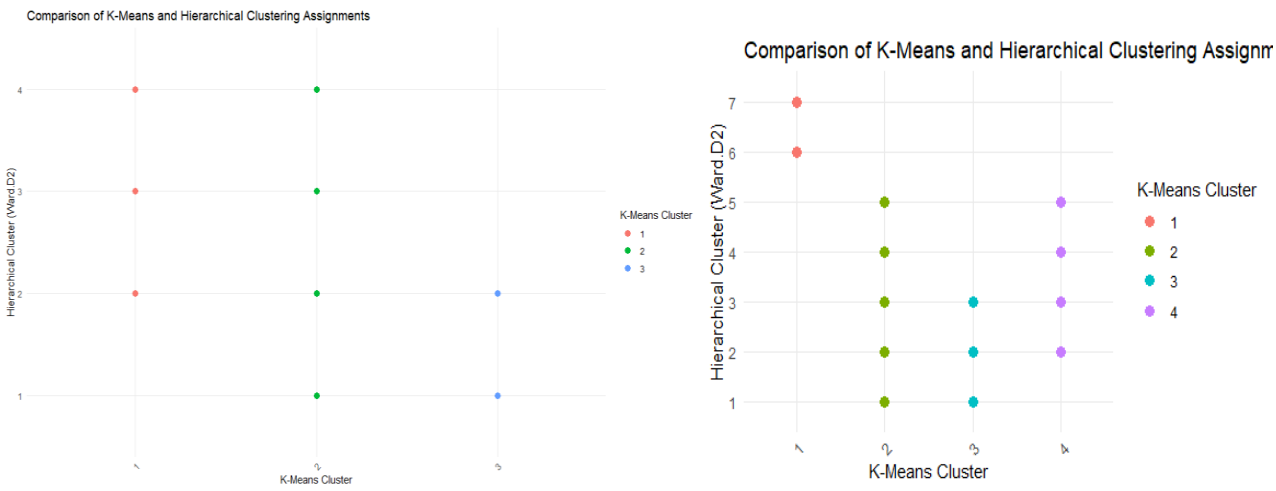
Cluster 1 - Enthusiastic Responders: They show high responsiveness to marketing efforts but have lower total purchases. This suggests they are eager or more influenced by marketing but might need additional incentives or products to increase their purchasing frequency.

Cluster 2 - Disengaged Shoppers: These customers exhibit low responsiveness to marketing campaigns and have a varying range of total purchases. They might be making necessary purchases without being influenced by marketing efforts, indicating potential disengagement.

Cluster 3 - Silent Buyers: This group also shows low responsiveness to marketing but contrasts with Cluster 2 by making more frequent purchases. They may be loyal or habitual customers who don't need marketing to make their purchase decisions.

Cluster 4 - Uncertain Group: This cluster's characteristics are less clear, possibly due to fewer data points or overlapping attributes with other clusters. They might represent a mix of behaviors or an area that requires further investigation to understand properly. To investigate a potential overlap in clusters, we decided to visualize a comparison of clusters from Hierarchical and K-Means Clustering for model 2 and 3

Model 2 Comparison *Model 3 Comparison*



The visualization comparing k-Means and Hierarchical Clustering methods shows a high degree of consistency between the cluster assignments from both methods, with no significant overlaps. Each k-Means cluster aligns well with a corresponding Hierarchical cluster, indicating that both methods are recognizing similar patterns in the data. There are clear, distinct levels for each cluster, suggesting that the two methods agree on the segmentation of the data. The only exception is the additional cluster identified by k-Means, which may suggest that k-Means is detecting a unique pattern not captured by Hierarchical Clustering. Overall, the agreement between the two methods suggests that the clustering is robust and the identified segments are reliable.

Modeling with the New variables

This section focuses on exploring the enhancement or otherwise of the models built in Report2A using kNN and Logistic Regression

Impact of Clustering Variables on Customer Response Classification using Logistic Regression:

Having added new columns (cluster labels) to enhance our supervised model, the model's complexity reflects advanced methods to encapsulate interactions and non-linear relationships in the new dataset, corresponding with cluster discovery and characterization.

Model Complexity:

```
> ##adding some interaction terms
> data$DiscStore = data$DiscountPurchases * data$StorePurchases
> data$WebCatalog = data$WebPurchases * data$CatalogPurchases  # Another interaction term
> #data$IncomeSqrd = data$Income^2  # A polynomial term (square of Income)
> data$sqtMeats = sqrt(data$Meats)  #  polynomial term (square root of Wines)
> str(data)
'data.frame':   668 obs. of  21 variables:
 $ Teenhome               : int  1 1 0 1 1 0 1 0 0 0 ...
 $ Meats                  : int  1 11 168 17 13 24 10 556 403 403 ...
 $ LastPurchase           : int  1 93 47 24 15 87 43 2 9 9 ...
 $ DiscountPurchases      : int  1 2 1 2 2 1 2 1 1 1 ...
 $ CatalogPurchases       : int  0 0 4 0 1 0 1 4 6 6 ...
 $ WebVisitsMonth         : int  6 4 1 7 4 8 5 8 3 3 ...
 $ WebPurchases           : int  1 2 3 2 1 2 2 10 7 7 ...
 $ StorePurchases         : int  2 3 7 3 4 3 4 8 6 6 ...
 $ Gold                   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp3           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp5           : int  0 0 0 0 0 0 0 0 1 1 ...
 $ AcceptedCmp1           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp2           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Education_non_graduate : int  1 0 1 1 1 1 1 1 1 1 ...
 $ Marital_Status_Married : int  0 0 1 0 1 0 1 0 1 1 ...
 $ Marital_Status_Together: int  1 1 0 0 0 1 0 1 0 0 ...
 $ Kmeans3clusters        : int  3 3 2 3 3 3 3 1 1 1 ...
 $ Response               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 2 2 ...
 $ DiscStore              : int  2 6 7 6 8 3 8 8 6 6 ...
 $ WebCatalog             : int  0 0 12 0 1 0 2 40 42 42 ...
 $ sqtMeats               : num  1 3.32 12.96 4.12 3.61 ...
```

Interaction terms like 'DiscStore' and 'WebCatalog' allow the model to assess combined influences of different purchasing behaviors on customer responses. Polynomial terms like 'IncomeSqrd' and 'sqtMeats' are used to model non-linear relationships, recognizing that consumer behavior is intricate and the effects of various factors can be multiplicative or nonlinear. These complex terms are meant to enhance the model's predictive ability, addressing the multi-faceted nature of consumer behavior. Cluster labels we determined from methods like K-Means and hierarchical clustering we used serve as new predictors, potentially improving the model's insights into consumer patterns.

To judge the effectiveness of these additions, we would need to compare key performance metrics before and after their inclusion. If there's no improvement, it may indicate that the existing variables already adequately explain the response, or the model might be overfitting. It's vital to test the model's enhanced complexity on new data to ensure it generalizes well beyond the training dataset.

Model Accuracy:

Without Clustering *With Clustering*

```
>  cm
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 220  49
         1  48 219

               Accuracy : 0.819
                 95% CI : (0.7838, 0.8507)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6381

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8209
            Specificity : 0.8172
         Pos Pred Value : 0.8178
         Neg Pred Value : 0.8202
             Prevalence : 0.5000
         Detection Rate : 0.4104
   Detection Prevalence : 0.5019
      Balanced Accuracy : 0.8190

       'Positive' Class : 0

> sensitivity = cm$byClass[1] # also known as recall
> specificity = cm$byClass[2] # also known as precision
> # Let's calculate F1, which combines sensitivity and specificity
> F1 = (2 * sensitivity * specificity) / (sensitivity + specificity)
> cat( "F1 statistic = ", round(F1,3))
F1 statistic =  0.819
```

```
> cm
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 219  47
         1  49 221

               Accuracy : 0.8209
                 95% CI : (0.7858, 0.8524)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6418

 Mcnemar's Test P-Value : 0.9187

            Sensitivity : 0.8172
            Specificity : 0.8246
         Pos Pred Value : 0.8233
         Neg Pred Value : 0.8185
             Prevalence : 0.5000
         Detection Rate : 0.4086
   Detection Prevalence : 0.4963
      Balanced Accuracy : 0.8209

       'Positive' Class : 0

> sensitivity = cm$byClass[1] # also known as recall
> specificity = cm$byClass[2] # also known as precision
> # Let's calculate F1, which combines sensitivity and specificity
> F1 = (2 * sensitivity * specificity) / (sensitivity + specificity)
> cat( "F1 statistic = ", round(F1,3))
F1 statistic =  0.821
```

The comparison between the logistic regression models without clustering and with clustering indicates that both models have very similar accuracies.

For the model without clustering, the accuracy is 0.819. This represents a substantial improvement over the "No Information Rate" of 0.5, showing the model's predictive power is significantly better than random guessing. The model has a Kappa statistic of 0.6381, which implies a good level of agreement beyond what would be expected by chance alone. This model shows a well-balanced sensitivity and specificity with values of 0.8209 and 0.8172, respectively, suggesting an equitable predictive performance for both positive and negative classes.

The model with clustering, on the other hand, has a slightly higher accuracy of 0.8209. It also presents an improved Kappa statistic of 0.6418, suggesting a very slight increase in the level of agreement beyond chance. The sensitivity of this model is slightly lower at 0.8172, while the specificity is marginally higher at 0.8246. The F1 statistic for this model is 0.821, marginally higher than the model without clustering.

These metrics suggest that the inclusion of clustering as an additional predictor in the logistic regression model has not significantly changed the accuracy or the balance between sensitivity and specificity. Both models perform well above the No Information Rate, indicating they are both effective at making predictions beyond what would be expected by chance. The very slight increase in overall accuracy, Kappa, and F1 statistic in the model with clustering suggests a minimal improvement. The choice between the two models might come down to considerations such as simplicity, interpretability, and computational efficiency, given the marginal difference in predictive performance.

Without Clustering *With Clustering*

```
                                          Pr(>|z|)
(Intercept)                               0.072622 .
Teenhome                                  0.006997 **
Meats                                     0.900204
LastPurchase                              7.36e-08 ***
DiscountPurchases                         0.645818
CatalogPurchases                          0.085159 .
WebVisitsMonth                            2.20e-06 ***
WebPurchases                              0.462638
StorePurchases                            0.001353 **
AcceptedCmp4                              0.408290
AcceptedCmp3                              0.000131 ***
AcceptedCmp5                              5.55e-05 ***
AcceptedCmp1                              0.003729 **
AcceptedCmp2                              0.163109
Education_non_graduate                    0.002254 **
Marital_Status_Married                    1.23e-05 ***
Marital_Status_Together                   8.29e-07 ***
`\\`Recency_Intervals_More than 90 days\\`` 0.378494
Income                                    0.455037
DiscStore                                 0.701426
WebCatalog                                0.981636
IncomeSqrd                                0.443797
sqtMeats                                  0.289311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

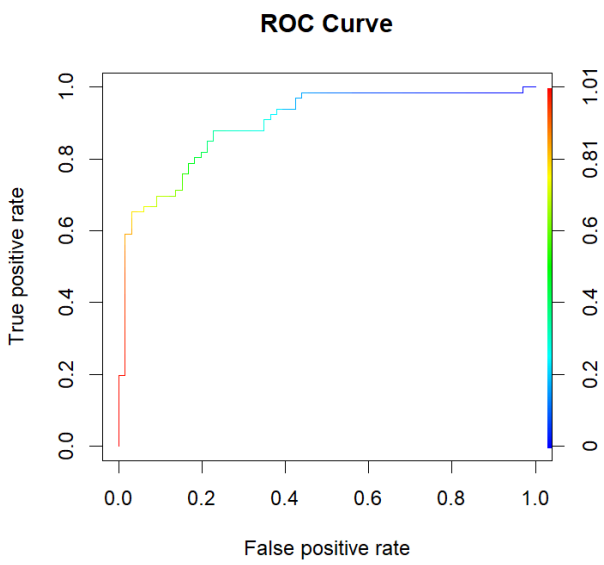
```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.2681591  1.5558802  -1.458 0.144896
Teenhome                 -1.0249148  0.3078894  -3.329 0.000872 ***
Meats                     0.0005429  0.0041571   0.131 0.896101
LastPurchase             -0.0273644  0.0044489  -6.151 7.71e-10 ***
DiscountPurchases         0.1320667  0.2044720   0.646 0.518350
CatalogPurchases          0.2702879  0.1483368   1.822 0.068436 .
WebVisitsMonth            0.4640447  0.0949750   4.886 1.03e-06 ***
WebPurchases              0.0966814  0.1085748   0.890 0.373219
StorePurchases           -0.2694692  0.0870710  -3.095 0.001969 **
Gold                     -0.0001133  0.0041100  -0.028 0.978008
AcceptedCmp3              1.4571558  0.3965818   3.674 0.000239 ***
AcceptedCmp5              2.4578253  0.5294737   4.642 3.45e-06 ***
AcceptedCmp1              2.0025099  0.6099333   3.283 0.001026 **
AcceptedCmp2              1.9443916  1.2391474   1.569 0.116616
Education_non_graduate    0.7997598  0.2559706   3.124 0.001782 **
Marital_Status_Married   -1.3291724  0.2971368  -4.473 7.70e-06 ***
Marital_Status_Together  -1.5990526  0.3230626  -4.950 7.43e-07 ***
Kmeans3clusters          -0.0571114  0.4397818  -0.130 0.896675
DiscStore                 0.0063951  0.0251845   0.254 0.799549
WebCatalog               -0.0034871  0.0211358  -0.165 0.868954
sqtMeats                  0.0959796  0.1263430   0.760 0.447449
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

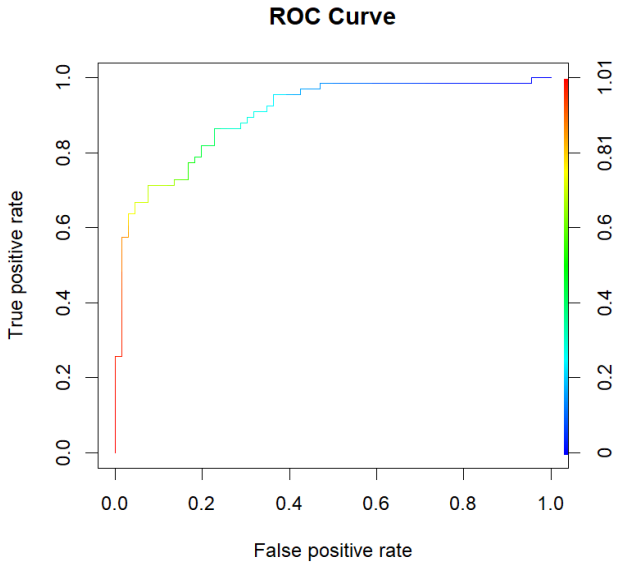
When comparing the p-values between the two logistic regression models – one without clustering and one with clustering, it's evident that the inclusion of clustering has an impact on the statistical significance of some predictor variables. In both models, predictors like *'Teenhome'*, *'LastPurchase'*, *'WebVisitsMonth'*, and *'StorePurchases'* retain their statistical significance, with p-values well below 0.05, confirming their influential roles in the models. Campaign acceptance variables *'AcceptedCmp3', 'AcceptedCmp5', and 'AcceptedCmp1'* continue to show very strong statistical significance in both models, emphasizing the importance of these features in predicting the outcome.

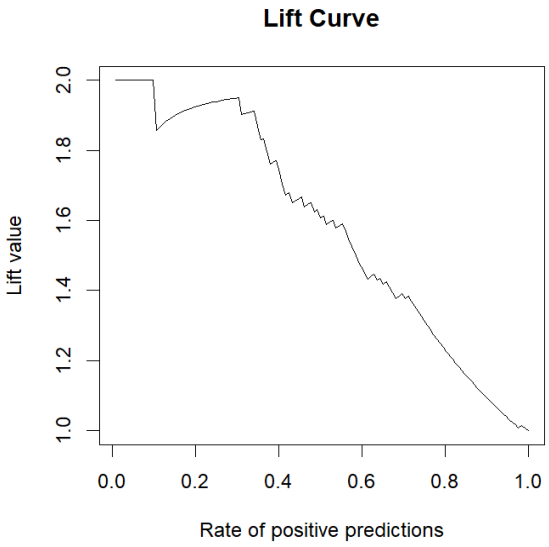
For the model with clustering, *'Marital_Status_Married'* and *'Marital_Status_Together'* also remain highly significant, with p-values less than 0.001, suggesting that marital status is a strong predictor of the response variable in the context of these data. Variables like *'Meats'*, *'DiscountPurchases'*, and *'Gold'*, which have p-values greater than 0.1 in the clustered model, suggest that these variables may not be as significant. This can indicate that these features do not have a strong linear relationship with the log odds of the outcome, or their effects are possibly being captured by other variables or the clusters themselves. The clustering variable *'Kmeans3clusters'* has a p-value greater than 0.1, indicating that the clusters as a whole may not be providing additional predictive power beyond what is already captured by the other variables in the model. This could suggest that while clustering helped in understanding the structure of the data, it might not have introduced additional predictive power to this logistic regression model.

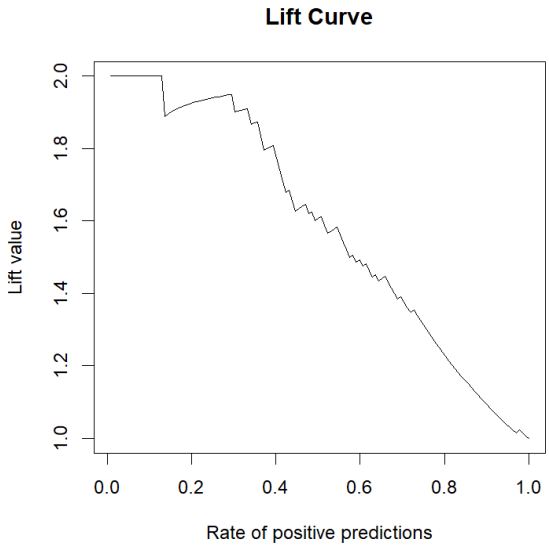
Plots:

Without Clustering *With Clustering*







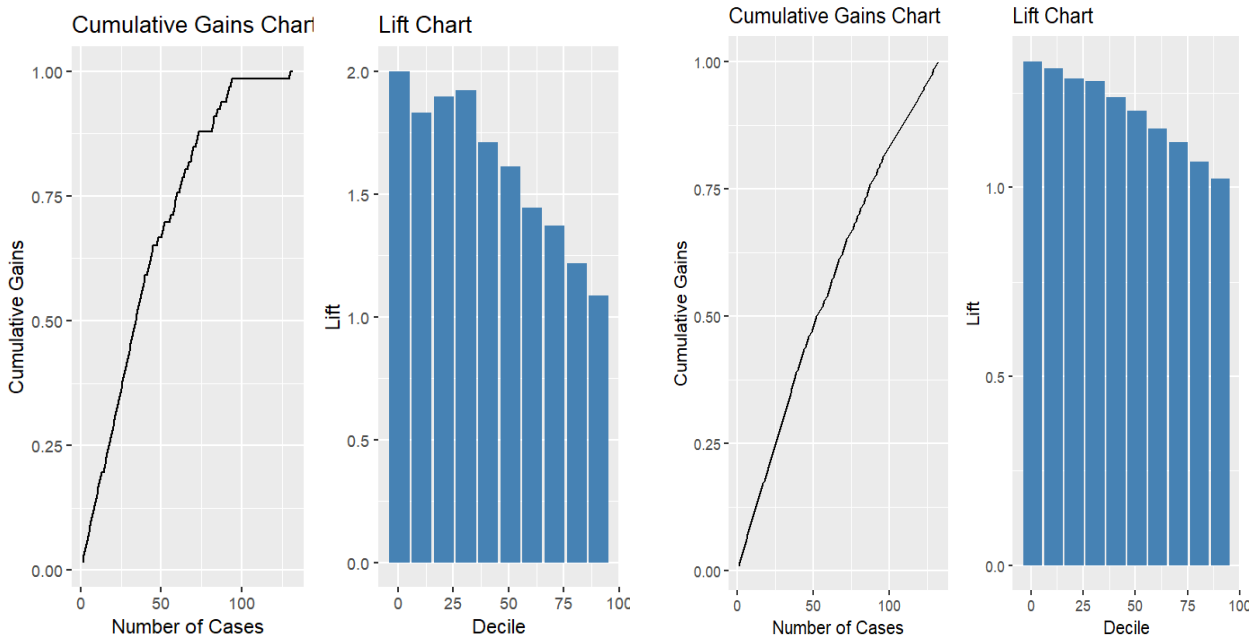


ROC Curve: The ROC curves for the logistic regression with and without clustering exhibit similar patterns, indicating comparable discriminative abilities between the two models. Both curves approach the top-left corner, which suggests a strong ability to correctly classify true positives while minimizing false positives. The visual similarity of the curves implies that adding clustering as a predictor does not significantly impact the model's overall classification performance.

Lift Curve: The lift curves for both models illustrate the effectiveness of the logistic regression in ranking positive cases above random chance. Initially, both models provide a lift significantly better than random, indicating strong

predictive performance in prioritizing cases most likely to be positive. As more cases are considered, the lift for both models gradually declines, suggesting that targeting becomes less efficient beyond a certain point. These curves can guide the optimal threshold for targeting in marketing campaigns, emphasizing the value of strong early predictions.

Without Clustering *With Clustering*



Cumulative Gains Chart: Both charts show a curve that rises sharply at the beginning, which indicates that both models are effective at identifying positive cases early on when targeting a smaller percentage of the total cases. This is beneficial because it means the model can capture a high proportion of the positive outcomes with a lower number of cases. Both curves appear very similar, indicating that the introduction of clustering doesn't drastically change the model's ability to identify positive cases early in the selection process.

Lift Chart: The Lift Charts display the lift value across different population deciles. Both charts show a descending pattern, where the lift is highest at the first decile and decreases as we move to the higher deciles. This decrease indicates that as we target more of the customers, the less likely we are to correctly identify as many positive cases compared to the best cases. In these images, both with and without clustering, the pattern is similar, suggesting that clustering doesn't have a major impact on how the model prioritizes and identifies positive outcomes across different segments of the population.

In both types of charts, the models without and with clustering seem to perform similarly, suggesting that clustering hasn't provided a significant advantage in these particular measures of model performance.

Interpretation of Odds Ratios:

Without Clustering

	Pr...z..	odds
(Intercept)	0.07262	0.13142
Teenhome	0.00700	0.39431
Meats	0.90020	0.99945
LastPurchase	0.00000	0.97432
DiscountPurchases	0.64582	1.10218
CatalogPurchases	0.08516	1.28337
WebVisitsMonth	0.00000	1.60078
WebPurchases	0.46264	1.08301
StorePurchases	0.00135	0.75492
AcceptedCmp4	0.40829	1.51438
AcceptedCmp3	0.00013	4.59338
AcceptedCmp5	0.00006	9.38602
AcceptedCmp1	0.00373	6.38006
AcceptedCmp2	0.16311	5.66205
Education_non_graduate	0.00225	2.18837
Marital_Status_Married	0.00001	0.27090
Marital_Status_Together	0.00000	0.20274
`\\`Recency_Intervals_More than 90 days\\``	0.37849	0.53306
Income	0.45504	0.99997
DiscStore	0.70143	1.01006
WebCatalog	0.98164	0.99950
IncomeSqrd	0.44380	1.00000
sqtMeats	0.28931	1.15220

With Clustering

	Estimate	Std..Error	z.value	Pr...z..	odds
(Intercept)	-2.26816	1.55588	-1.45780	0.14490	0.10350
Teenhome	-1.02491	0.30789	-3.32884	0.00087	0.35883
Meats	0.00054	0.00416	0.13059	0.89610	1.00054
LastPurchase	-0.02736	0.00445	-6.15076	0.00000	0.97301
DiscountPurchases	0.13207	0.20447	0.64589	0.51835	1.14118
CatalogPurchases	0.27029	0.14834	1.82212	0.06844	1.31034
WebVisitsMonth	0.46404	0.09498	4.88596	0.00000	1.59049
WebPurchases	0.09668	0.10857	0.89046	0.37322	1.10151
StorePurchases	-0.26947	0.08707	-3.09482	0.00197	0.76378
Gold	-0.00011	0.00411	-0.02757	0.97801	0.99989
AcceptedCmp3	1.45716	0.39658	3.67429	0.00024	4.29373
AcceptedCmp5	2.45783	0.52947	4.64202	0.00000	11.67938
AcceptedCmp1	2.00251	0.60993	3.28316	0.00103	7.40763
AcceptedCmp2	1.94439	1.23915	1.56914	0.11662	6.98938
Education_non_graduate	0.79976	0.25597	3.12442	0.00178	2.22501
Marital_Status_Married	-1.32917	0.29714	-4.47327	0.00001	0.26470
Marital_Status_Together	-1.59905	0.32306	-4.94967	0.00000	0.20209
Kmeans3clusters	-0.05711	0.43978	-0.12986	0.89667	0.94449
DiscStore	0.00640	0.02518	0.25393	0.79955	1.00642
WebCatalog	-0.00349	0.02114	-0.16499	0.86895	0.99652
sqtMeats	0.09598	0.12634	0.75968	0.44745	1.10074

Comparing the odds ratios from the two models, with and without clustering, provides insights into the relationships between the predictor variables and the likelihood of a positive response.

For the model without clustering, certain campaign acceptance variables, like 'AcceptedCmp3' and 'AcceptedCmp5', have odds ratios of 4.59338 and 9.38602, respectively. This suggests strong positive relationships with the response—accepting these campaigns increases the likelihood of a positive response by approximately 4.6 and 9.4 times. Marital status variables like 'Marital_Status_Married' and 'Marital_Status_Together' are associated with significantly lower odds of a positive response, by 73% and 80%, respectively.

On the other hand, in the model with clustering, the 'AcceptedCmp3' odds ratio slightly decreases to 4.29373, but 'AcceptedCmp5' significantly increases to 11.67938, indicating an even stronger association with the likelihood of a response after considering clustering. The marital status variables show a similar pattern of association with lower odds of a positive response, though 'Marital_Status_Together' shows a marginally stronger negative association in the model with clustering. The 'Income' variable shows an odds ratio close to 1 in both models, suggesting no significant direct linear effect on the likelihood of a positive response. However, the presence of 'IncomeSqrd' with an odds ratio significantly less than 1 in the clustered model indicates a more complex, non-linear relationship with income. For 'sqtMeats', the non-clustered model suggests a 15.2% increase in the odds of a positive response for each unit increase in the square root of meat purchases. In contrast, the clustered model shows this variable with an odds ratio of 1.10074, indicating a similar but slightly less pronounced effect.

These differences in the odds ratios between the models with and without clustering highlight the importance of considering underlying data structures, such as clusters, when predicting outcomes. Clustering appears to enhance the model's sensitivity to the nuances of the data, as indicated by the changes in the strength of associations reflected in the odds ratios.

Interpretation of Confusion Matrix:

Without Clustering

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 220  49
         1  48 219

               Accuracy : 0.819
                 95% CI : (0.7838, 0.8507)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6381
```

With Clustering

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 219  47
         1  49 221

               Accuracy : 0.8209
                 95% CI : (0.7858, 0.8524)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6418
```

Comparing the logistic regression models, we see that the addition of clustering has a subtle yet positive effect on the predictive accuracy. The model without clustering (left) has an accuracy of 81.9% with a Kappa statistic of 0.6381, while the model with clustering (right) shows a slight improvement in accuracy to 82.09% and a Kappa of 0.6418. Both models have high true positive rates, with the non-clustering model yielding 219 and the clustering model 221, illustrating their competence in correctly identifying positive instances. The reduction of false positives from 49 to 47 in the clustering model suggests a refined ability to identify true negatives. The precision in predictions for both models far exceeds random guessing, indicated by P-Values significantly lower than 0.05. The models'

performances are quite comparable, with clustering providing a nuanced improvement in identifying and predicting the correct classes.

Impact of Clustering Variables on Customer Response Classification using kNN

Overview:

We continued with the initiative to refine customer response prediction to marketing campaigns. This integrated report compares the original k-Nearest Neighbors (k-NN) classification model's performance against 2 revised models that incorporate clustering variables derived from both K-Means and Hierarchical clustering methods, the original model without clustering variables, a second model incorporating two new clustering variables, and a third model adding two additional clustering variables, totaling four variables.

Data Preparation and Preprocessing for kNN Algorithm:

The initial steps remained consistent with previous analyses, involving data loading, dummy variable generation, and feature selection based on demographics, past purchase behavior, and engagement metrics. Notably, the updated analysis included additional preprocessing steps to integrate clustering variables obtained from Hierarchical and K-Means clustering into the dataset, aimed at capturing intrinsic groupings within the customer base.

Methodological Update:

The core methodology of using the k-NN algorithm persisted across all models with an 80-20 training-testing split. The novel approach in the subsequent models involved integrating cluster membership variables, which encapsulated additional patterns related to customer behaviors and responses. This strategy aimed at leveraging the nuanced distinctions within customer segments to predict campaign responses more accurately.

Model Evaluation

Method: k-NN algorithm.

First Model's Performance

This model showed an accuracy of 81.06% (This indicates that 81% of the model's predictions were correct), with a precision of 78.79% and a recall of 82.54%. The F1 score stood at 80.62%. A K = 19 was the best K obtained through cross validation

```
> print(confusionMatrix)
          Actual
Predicted  0  1
        0 55 14
        1 11 52
> accuracy <- sum(diag(confusionMatrix)) /
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.810606060606061"
```

```
[1] "Precision: 0.787878787878788"
> print(paste("Recall:", recall))
[1] "Recall: 0.825396825396825"
> print(paste("F1 Score:", F1))
[1] "F1 Score: 0.806201550387597"
```

First Model Confusion Matrix:

True Negatives (TN): The model correctly predicted 55 instances where the actual outcome was 0 (negative class).

False Positives (FP): The model incorrectly predicted 14 instances as 1 (positive class) when they were actually 0 (negative class).

False Negatives (FN): The model incorrectly predicted 11 instances as 0 (negative class) when they were actually 1 (positive class).

True Positives (TP): The model correctly predicted 52 instances where the actual outcome was 1 (positive class).

Second Model's Performance (with two clustering variables: hcWardD4clusters, Kmeans3clusters)

The second model, enhanced with two new clustering variables, showed improved performance over the original. The optimal k value, determined through cross-validation, was 15, leading to an accuracy of 82.58%. This model achieved a precision of 78.79% and a recall of 85.25%, culminating in an F1 Score of 81.89%.

```
> print(confusionMatrix)
          Actual
Predicted  0  1
        0 57 14
        1  9 52
> accuracy <- sum(diag(confusionMatrix)) /
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.825757575757576"
```

```
[1] "Precision: 0.787878787878788"
> print(paste("Recall:", recall))
[1] "Recall: 0.852459016393443"
> print(paste("F1 Score:", F1))
[1] "F1 Score: 0.818897637795276"
```

The incorporation of clustering variables has evidently refined the model's predictive capabilities, enabling more accurate identification of customer responses. This improvement assists MarketSphere Inc. in tailoring its marketing efforts more effectively, ensuring resources are optimally allocated and campaigns are more personalized.

Second Model Confusion Matrix:

True Negatives (TN): The model correctly predicted 57 negative (0) cases. This means that 57 individuals who were not supposed to be in the positive (1) class were correctly identified by the model.

False Positives (FP): The model incorrectly predicted 14 positive (1) cases as negative (0). These are cases that the model incorrectly classified as positive when they were actually negative.

False Negatives (FN): The model incorrectly predicted 9 negative (0) cases as positive (1). These are instances where the model failed to identify positive cases, mistaking them for negative ones.

True Positives (TP): The model correctly predicted 52 positive (1) cases. This means that the model correctly identified 52 instances that were supposed to be classified as positive.

Third Model's Performance (with 4four clustering variables: hcWardD4clusters, Kmeans3clusters, hcWardD7clusters, Kmeans4clusters)

The third model achieved an accuracy of 79.55%. This marks a decrease from the second model's 82.58% and original model's 81.06%. Precision was 81.82%. This shows improvement from the second model's 78.79% and the original model's precision of 78.79% . Recall was also 78.26%, lower than the second model's 85.25% but comparable to the original model. F1 Score: 80%. This is slightly below the second model's 81.89% but shows improvement over the original model's 80.62%. The decrease in accuracy and recall suggests a trade-off, where the model might miss out on identifying some true positives, possibly due to over-specialization or the clustering variables introducing noise for some observations. For MarketSphere, the F1 score is especially relevant in predicting customer responses to marketing campaigns. Since both over-targeting (leading to wasted resources and potential customer annoyance) and under-targeting (missing out on potential sales) are costly, the F1 score helps evaluate how well the model identifies truly interested customers. A higher F1 score indicates that the model effectively balances catching genuine opportunities (recall) without excessively flagging uninterested customers (precision).

```
> print(confusionMatrix)
          Actual
Predicted  0  1
        0 51 12
        1 15 54
> accuracy <- sum(diag(confusionMatrix)) / 
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.795454545454545"
```

```
> print(paste("Precision:", precision))
[1] "Precision: 0.818181818181818"
> print(paste("Recall:", recall))
[1] "Recall: 0.782608695652174"
> print(paste("F1 Score:", F1))
[1] "F1 Score: 0.8"
```

Third Model Confusion Matrix:

True Negatives (TN): The model correctly predicted 51 instances where the actual class was 0 (negative). This means 51 times the model correctly predicted the non-event.

False Positives (FP): The model incorrectly predicted 12 instances as class 1 (positive) when they were actually class 0 (negative). These are cases where the model predicted an event incorrectly.

False Negatives (FN): The model incorrectly predicted 15 instances as class 0 (negative) when they were actually class 1 (positive). These are situations where the model missed the event.

True Positives (TP): The model correctly predicted 54 instances where the actual class was 1 (positive). This means the model correctly identified the event 54 times.

When to Use Which Model;

The second model, with two new clustering variables, showed the highest accuracy (82.58%) and recall (85.25%), indicating it was the most reliable in predicting customer responses overall.

The third model, which added four new clustering variables, improved in precision (81.82%) compared to the other models, indicating it's better at identifying true positive responses but had a slight decrease in accuracy (79.55%) and recall (78.26%) compared to the second model.

The original model, without clustering variables, had a balanced performance but was surpassed in precision by the third model and in accuracy and recall by the second model.

The choice between these models should be based on what the company values more: precision (favoring the third model) or a combination of accuracy and recall (favoring the second model).

Comparison between F1 Score Summary with Confidence Intervals for Logistic Regression & kNN

Logistic Regression Model

F1 Score: 0.821

95% Confidence Interval: [0.7858, 0.8524]

k-NN Model:

F1 Score: 0.8189

95% Confidence Interval: [0.678, 0.804]

Interpretation and Discussion;

Performance Comparison: Both models demonstrate high F1 scores, suggesting they are effective for the task at hand. The logistic regression model has a slightly higher F1 score than the k-NN model, but the difference is very small.

Confidence Intervals: The confidence interval for the k-NN model's F1 score, [0.678, 0.804], indicates more variability in performance compared to the logistic regression model, whose F1 score confidence interval is [0.7858, 0.8524]. This suggests that while the k-NN model can perform comparably to the logistic regression, its performance might be more sensitive to the specific data it's trained and tested on.

Statistical and Practical Considerations: The overlap in confidence intervals suggests comparable performance between the models, but the wider range for k-NN underscores greater variability. When choosing between the two models, consider additional factors like interpretability (a strength of logistic regression) and flexibility or the handling of non-linear relationships (where k-NN may have advantages).

How We Avoided Overfitting

To avoid, manage, or minimize overfitting in this project, we:

Scaling Predictors: Features were scaled to ensure uniformity; this step is crucial as it prevents variables with larger scales from disproportionately influencing the model, a common pitfall that can lead to overfitting.

Categorical Variable Transformation: We transformed categories that skewed distribution into binary variables, such as converting 'Education' into 'graduate' and 'non-graduate'. This approach reduces complexity and helps models focus on significant distinctions.

Exhaustive Search for Predictors: By using an exhaustive search, we systematically evaluated combinations of predictors to identify the most impactful ones, thereby avoiding the inclusion of irrelevant features that could lead to overfitting.

Variance Inflation Factor (VIF) and Pair Plots: We used VIF to identify and eliminate multicollinearity between variables, and pair plots to assess relationships and redundancies among predictors, further refining our feature set.

Handling Outliers: The elimination of outliers from the dataset ensured that our models were not swayed by anomalous data points, which can distort model training and lead to overfitting.

Balanced Dataset: For classification tasks, we used a balanced dataset by undersampling the majority class, which prevents the model from being biased toward the more prevalent class and helps in generalizing better to unseen data.

Additional Measures and Insights:

Cross-Validation: While not explicitly mentioned earlier, cross-validation techniques, particularly Leave-One-Out Cross-Validation (LOOCV) used in k-NN model optimization, are instrumental in evaluating the model's generalizability and thus combating overfitting.

Reflection

The exploration through three distinct models highlights the evolving understanding of customer behaviors and preferences. The integration of clustering variables has shown potential in enhancing the predictive power of the models. MarketSphere Inc. should leverage the insights gained from the most effective model to tailor marketing strategies that resonate with distinct customer segments.

Continuous refinement and validation of these models with new data are recommended to maintain their relevance and accuracy. Additionally, exploring other predictive modeling techniques and clustering methods could provide further insights into customer behavior patterns.

Upon reflection on the entire analysis process, we embarked on a comprehensive journey, examining the relationship between 'Wines' purchases and various predictors to enhance MarketSphere Inc.'s understanding and predictive capabilities regarding customer behaviors.

Initially, we discovered positive correlations between 'Wines' spending and both 'CatalogPurchases' and 'StorePurchases,' implying that higher spending in these areas typically translates into increased wine purchases. Conversely, the presence of children at home ('Kidhome') showed a negative correlation with wine spending, suggesting a potential shift in household priorities or budgets. Surprisingly, 'Complain' did not exhibit a significant relationship with 'Wines,' indicating that customer complaints might not directly impact wine purchasing behaviors. These initial findings laid a solid foundation, highlighting key factors influencing wine purchases, but also pointing towards potential areas for model enhancement.

Transitioning to an advanced analytical phase, we integrated clustering variables into our linear regression model to refine our predictions and insights. The inclusion of 'Hcluster' and 'Kmcluster' variables, derived from meticulous clustering processes, was statistically justified given their significant p-values and manageable variance inflation factors (VIFs), ensuring we did not introduce detrimental multicollinearity.

The clustering process itself, particularly using hierarchical methods and k-Means clustering, illuminated distinct customer segments within MarketSphere's data. Our selection of the Ward.D2 method, after evaluating

various clustering techniques, proved pivotal. It enabled us to identify coherent customer groupings, thereby enhancing our model's granularity and applicability. Naming these clusters - 'Wine Enthusiasts,' 'Occasional Shoppers,' 'General Consumers,' and 'Heavy Shoppers' - provided actionable insights, allowing for targeted marketing approaches tailored to distinct customer preferences and behaviors.

Enhancements to our linear regression model reflected through improved performance metrics, notably an increased adjusted R-squared and decreased RMSE in training data, underscore the added value of incorporating cluster variables. This evolution from a basic to an advanced model not only improved predictive accuracy but also deepened our understanding of the underlying dynamics affecting wine purchases.

Furthermore, diagnostic evaluations of both original and enhanced models revealed critical insights. While both models adhered largely to linear regression assumptions, the introduction of clustering variables appeared to address specific discrepancies and enhance model robustness. Particularly, the diagnostic plots post-clustering amendment provided reassurance about the refined model’s validity and reliability.

Comparison between the logistic regression and k-NN models revealed both to be effective, with logistic regression slightly outperforming k-NN in terms of F1 score and confidence interval stability. This discrepancy underscores the need to weigh interpretability against flexibility when selecting the optimal model.

Ultimately, our analysis journey has equipped MarketSphere with actionable insights, paving the way for tailored marketing strategies that cater to the nuanced preferences of diverse customer segments. The journey underscores the potent blend of data-driven analytics and strategic marketing, setting a foundation for enhanced decision-making and targeted engagement in MarketSphere’s operations.

In conclusion, the methodical expansion of our analysis to include clustering variables has undeniably fortified our regression model, providing MarketSphere Inc. with a more nuanced, powerful tool for predicting wine purchase behaviors. This journey from basic correlations to sophisticated clustering-informed regression has not only bolstered our analytical acumen but also offered MarketSphere actionable strategies to engage distinct customer segments effectively. Ultimately, this reflects a significant stride towards leveraging data-driven insights for strategic decision-making and customer-centric marketing.

Best Model Selection:

From the three models, the choice of the best model depends on the specific marketing objectives and the trade-offs between model complexity and interpretability. If the goal is to gain a deeper understanding of customer segments, the third model with four clustering variables provides a more nuanced segmentation. However, for more straightforward campaigns or when computational resources are limited, the original model or the second model might suffice.

Ultimately, the best model should align with MarketSphere Inc.'s strategic goals, balancing predictive performance with practical applicability in targeted marketing campaigns.

Citations

Boudet, J., Brodherson, M., Robinson, K., & Stein, E. (2023, June 26). *Beyond belt-tightening: How marketing can drive resiliency during uncertain times.* McKinsey & Company. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/beyond-belt-tightening-how-marketing-can-drive-resiliency-during-uncertain-times#/

Checa, A., Heller, C., Stein, E., & Wilkie, J. (2023, April 5). *Modern marketing: Six capabilities for multidisciplinary teams.* McKinsey & Company. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/modern-marketing-six-capabilities-for-multidisciplinary-teams

Customer Personality Analysis. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

Appendix Back to Top