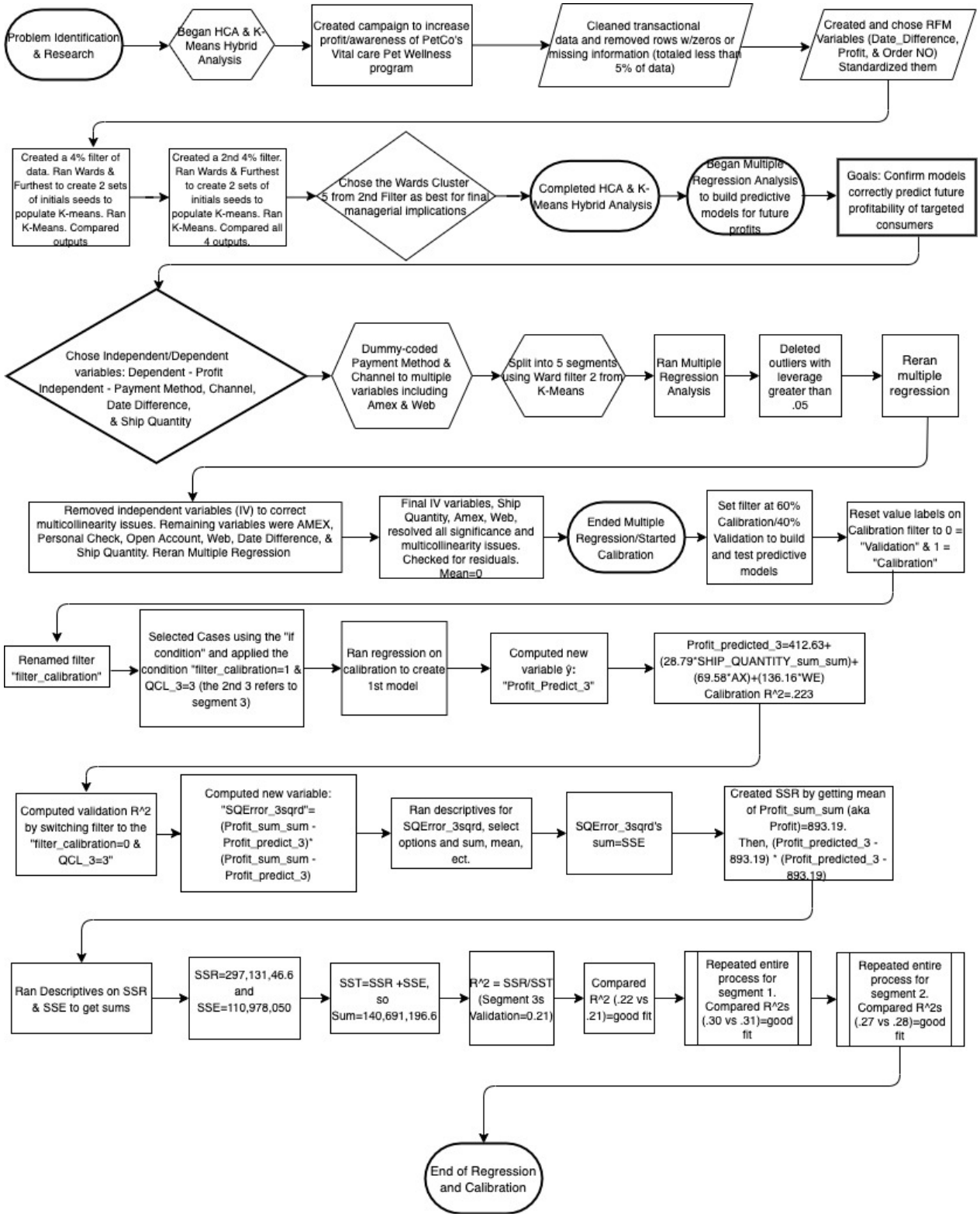
PREDICTING PROFITABILITY BASED ON PETCO'S CUSTOMER SEGMENT BEHAVIORS

METHOD USED: MULTIPLE REGRESSION ANALYSIS

TEAM 2: AISHETU AMARA, CARLA JACKSON, DELPHINE MASON, CAROL YU, LUCA ZHANG

Table of Contents

Flow Chart


Problem Identification & Research
Began HCA & K-Means Hybrid Analysis
Created campaign to increase profit/awareness of PetCo's Vital care Pet Wellness program
Cleaned transactional data and removed rows w/zeros or missing information (totaled less than 5% of data)
Created and chose RFM Variables (Date_Difference, Profit, & Order NO) Standardized them
Created a 4% filter of data. Ran Wards & Furthest to create 2 sets of initials seeds to populate K-means. Ran K-Means. Compared outputs
Created a 2nd 4% filter. Ran Wards & Furthest to create 2 sets of initials seeds to populate K-means. Ran K-Means. Compared all 4 outputs.
Chose the Wards Cluster 5 from 2nd Filter as best for final managerial implications
Completed HCA & K-Means Hybrid Analysis
Began Multiple Regression Analysis to build predictive models for future profits
Goals: Confirm models correctly predict future profitability of targeted consumers
Chose Independent/Dependent variables: Dependent - Profit Independent - Payment Method, Channel, Date Difference, & Ship Quantity
Dummy-coded Payment Method & Channel to multiple variables including Amex & Web
Split into 5 segments using Ward filter 2 from K-Means
Ran Multiple Regression Analysis
Deleted outliers with leverage greater than .05
Reran multiple regression
Removed independent variables (IV) to correct multicollinearity issues. Remaining variables were AMEX, Personal Check, Open Account, Web, Date Difference, & Ship Quantity. Reran Multiple Regression
Final IV variables, Ship Quantity, Amex, Web, resolved all significance and multicollinearity issues. Checked for residuals. Mean=0
Ended Multiple Regression/Started Calibration
Set filter at 60% Calibration/40% Validation to build and test predictive models
Reset value labels on Calibration filter to 0 = "Validation" & 1 = "Calibration"
Renamed filter "filter_calibration"
Selected Cases using the "if condition" and applied the condition "filter_calibration=1 & QCL_3=3 (the 2nd 3 refers to segment 3)
Ran regression on calibration to create 1st model
Computed new variable ŷ: "Profit_Predict_3"
Profit_predicted_3=412.63+ (28.79*SHIP_QUANTITY_sum_sum)+ (69.58*AX)+(136.16*WE) Calibration R^2=.223
Computed validation R^2 by switching filter to the "filter_calibration=0 & QCL_3=3"
Computed new variable: "SQError_3sqrd"= (Profit_sum_sum - Profit_predict_3)* (Profit_sum_sum - Profit_predict_3)
Ran descriptives for SQError_3sqrd, select options and sum, mean, ect.
SQError_3sqrd's sum=SSE
Created SSR by getting mean of Profit_sum_sum (aka Profit)=893.19. Then, (Profit_predicted_3 - 893.19) * (Profit_predicted_3 - 893.19)
Ran Descriptives on SSR & SSE to get sums
SSR=297,131,46.6 and SSE=110,978,050
SST=SSR +SSE, so Sum=140,691,196.6
R^2 = SSR/SST (Segment 3s Validation=0.21)
Compared R^2 (.22 vs .21)=good fit
Repeated entire process for segment 1. Compared R^2s (.30 vs .31)=good fit
Repeated entire process for segment 2. Compared R^2s (.27 vs .28)=good fit
End of Regression and Calibration

Technical

In Project Two, we successfully identified five target customer segments using a hybrid method of HCA and K-Mean analysis. However, in order to predict how certain behaviors within our customer segments will affect future profitability, and thus our overall managerial decisions, we applied multilinear regression analyses using two distinct customer groups.

Data Preparation & Selection of Independent/Dependent Variables:

First, we selected our independent and dependent variables. Our objective was to determine which predictor variables would affect profits. We chose PROFIT (Profit) as our dependent variable, SHIP_QUANTITY (Ship_Quantity), DATE_DIFFERENCE (Date_Difference), PAY_METHOD (Payment Method), and CHANNEL (Channel) as our independent variables. However, since Payment Method and Channel are Nominal variables, we converted their data into multiple dummy-coded variables that included Amex and Web respectively. Then, in SPSS, we split the data file using the K-Means filter where our chosen segments were housed (QCL_3) and selected "Organize output by groups." This step was necessary, so that when we ran the regressions, we could easily compare each segment's output simultaneously.

Multiple Regression Analysis: Testing Regression Assumptions

We ran the regression and tested for all assumptions. Our results revealed that the p-values for all of the segments were statistically significant (0.00). Therefore, there was overwhelming evidence to reject the Null Hypothesis (H_o: $\beta_1 = \beta_2 = \beta_3 = 0$) and accept the alternative hypothesis (H_1: not all B_i's = 0) because there was a relationship between at least one of the chosen independent variables and Profit.

Influential Observations – Outliers

After reviewing the data from our Leverage and Mahalanobis tests for outliers, we saw that we needed to delete all of the cases with Leverage numbers greater than .05 (*Total of 27 rows*). We reviewed all cases between .05 - .025 and decided to keep the small numbers of cases that were there. We re-ran the regression without outliers and found that the last segment of our five-segment cluster was eliminated (*11 cases*).

Testing for Multicollinearity

The results displayed through VIF and Tolerance indicated that we had issues with multicollinearity and significance with a number of independent variables. Therefore, we re-ran the multiple regression, after reducing the independent variables (IV) whose p-values were greater than .05. We continued to re-run our regression until all of the significance values and multicollinearity issues had been addressed for our remaining IVs. As a result of this, our final independent variables were Ship_Quantity, Amex, and Web. (See Table 1. For multicollinearity stats)

Table 1. Multicollinearity Statistics for Independent Variables

Collinearity Stats - Tolerance	Low-End (1)	Low-End (2)	High-End (3)	Multicollinearity Test (>.10)	Average (1-2)
Ship Quantity	0.995	0.994	0.988	Acceptable	0.995
Amex	0.993	0.994	0.983	Acceptable	0.994
Web	0.991	0.991	0.991	Acceptable	0.991

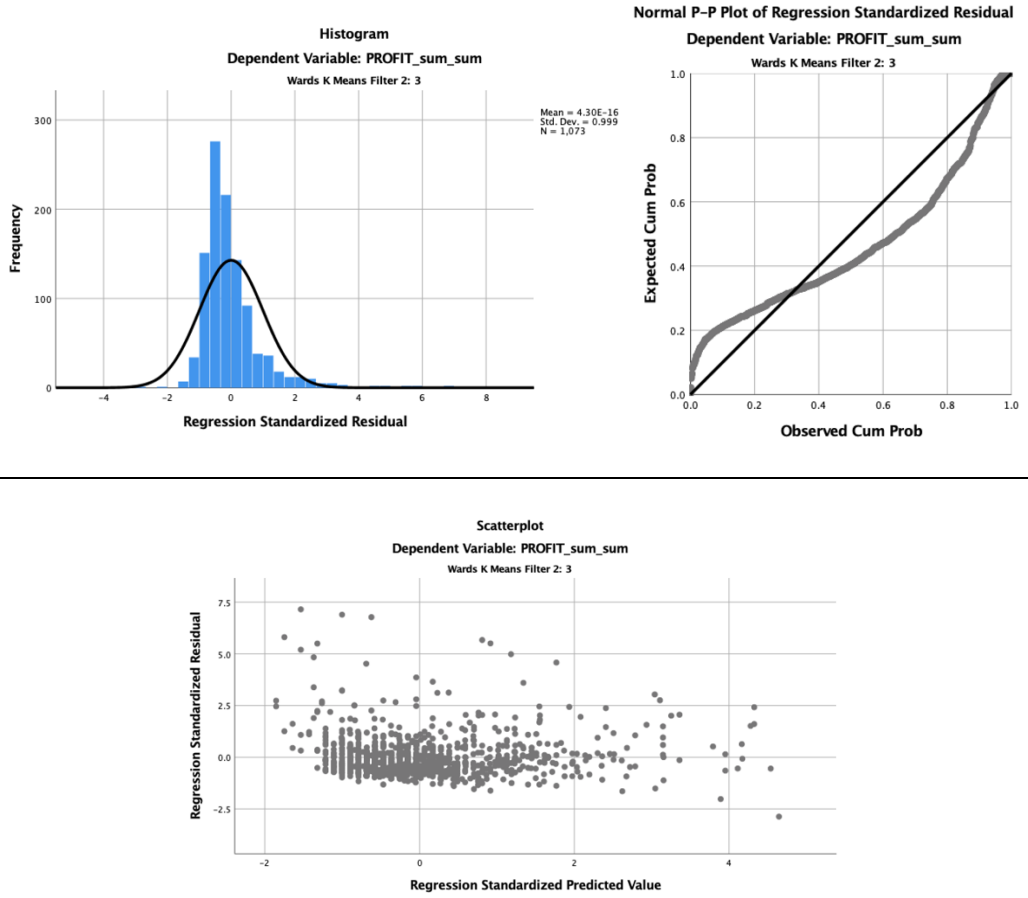
Collinearity Stats - VIF	Low-End (1)	Low-End (2)	High-End (3)	Multicollinearity Test (<4)	Average (1-2)
Ship Quantity	1.005	1.006	1.012	Acceptable	1.006
Amex	1.007	1.006	1.017	Acceptable	1.007
Web	1.009	1.009	1.009	Acceptable	1.009

Checking for Residuals

Following this, we checked for residuals through the creation of plots, as well as our outputs' value statistics for all four segments. The mean for each of the segments' residual was

equivalent to zero. A representation of all three of the segments' plots is displayed below in Figure 1. using the charts from segment 3.

Figure 1: Histogram, P-Plot, and Scatterplots for Segment Three



Final Evaluation of Customer Segments

We focused on the three segments chosen from Project Two, segments 1, 2, and 3. Since customer segments 1 and 2 had similar average profits and counts, we consolidated the two groups. On the other hand, segment 3 was our high-end segment, whose customers averaged about $900 in profits, but had little more than a thousand people count. Segment 4's average profit (almost $300) per customer fits between the two groups and will be considered as part of another promotion at a future date. It is, therefore, not included in the present analysis. Also, as

noted, segment 5's cases were outliers and were eliminated earlier. (See Table 2. for remaining segments' profit and count information)

Table 2: Cluster Results (Wards/Filter 2)

SPSS (Cluster Results)	Low-End (1)	Low-End (2)	High-End (3)	Average (1-2)
Profits	76	71	897	73
Count	42,121	48,323	1,109	45,222

Confirmation of Results: Calibration/Validation

The final stage of our process was to check how good our regression model could predict profit in the future. This was done by comparing the segments' actual R^2 calibrations with their computed R^2 from the validation tests. We computed R^2 through this formula: R^2=SSR/SST. To predict future profits without future data, we split the data of each segment into two sample groups. 60% went towards a calibration sample and the remaining 40% went to a validation (test) sample. We built our regression model in the calibration sample and tested it on the validation sample. If each of our calibration R^2 values were close in number to each of their corresponding validation R^2 values, this would prove that our models for all three segments were a good fit, and we could reliably make managerial decisions based on these predictions. Please see the steps and calculations that were performed in the SPSS process below:

We changed the name of the newly created variable to "filter_calibration," defined its label to be calibration=1 and validation=0, and used the following filter to access our three segments: (filter_calibration=1) & (QCL_3=3). QCL_3 refers to a filter created in an earlier iteration of the project (*Revitalizing Petco's VitalCare Wellness Plan*). The "=3" allows us to access segment 3 through this method.

Profit_predicted became our Y dependent variable, and was created through the Transform>>**Compute variable** function. After running the initial regression, we inserted the

coefficients of the IVs, producing the model: **Profit_predicted_3** = 412.63 + (28.79*SHIP_QUANTITY_sum_sum) + (69.58*AX) + (136.16*WE). The calibration R^2=.223. To access our validation cases for testing, we changed our filter to "If condition is satisfied"=(filter_calibration=0) & (QCL_3=3).

Knowing that SSE=Sum of square due to the regression, we calculated it through this formula using the Compute variable function: (Profit-Profit_predicted_3) * (Profit-Profit_predicted_3).

To get SSR (Sum of square due to the error), we had to know the mean of our original Profit. Therefore, we added Profit to the **Analyze>>Descriptive** function which gave us 893.19 as the mean. We created SQregression_3 through this formula: SQregression_3 = (Profit_predicted_3 - 893.19) * (Profit_predicted_3 - 893.19).

Now that we had SSR and SSE, we put both of them into the Descriptive function and selected Sum. The result showed SSR=297,131,46.6 and SSE=110,978,050. Their sum=140,691,196.6. Knowing that R^2=Sum of Squares due to Regression/Total Sum of Squares, we divided the SSR by SST. The result was the validation R^2, 0.21.

Finally, we compared the validation R^2 of 0.21 to the calibration R^2 of 0.22. Overall, we conclude that the model is a good fit since the difference between them is very small. Then, we repeated all of the steps above for Segment 1 and 2. Segment 1's calibration R^2 is 0.30 versus 0.31 for its validation R^2. Segment 2's calibration R^2 is 0.27 versus 0.28 for its validation's R^2. Again, all of the differences were minor. Therefore, we concluded that all of the models are a good fit.

Key Findings

Our multi-regression model analyzes the affect that customers using the web as a preferred purchasing channel, actual quantities shipped, and payment methods have on Petco's profits. Our findings below highlight the key insights from our model.

- Amongst our Mass or Low-end (profit) customer (segments 1 and 2), purchasing online decreased profits, while for our high-end customers (segment 3), online purchasing behavior increased profits. We have several theories that may explain these results: our mass customer groups may not be as tech-savvy when it comes to online purchases, and therefore may purchase less often as a consequence. Unfortunately, our web operational processing costs are high, which makes us lose money when small infrequent purchases are made. We will need to explore how to lower those costs. Additionally, we may also need to simplify the online checkout process, so that our customers can move seamlessly from shopping to purchasing. Finally, we will review our distribution costs to make sure that they are cost-efficient.
- We also noted that our high-end customers (segment 3) spend a lot more when they purchase using their American Express cards. Overall, profits from all three customer groups increased with the use of American Express cards (AX). Therefore, we think that a promotional campaign or partnership between Petco and American express could be lucrative for both companies and provide further rewards/savings for customers on both sides. (See Figure 3. for Coefficient information)

Figure 3: Multi-Regression Results

Evaluating Coefficients (B)	Low-End (1)	Low-End (2)	High-End (3)	Average (1-3)	Average (1-2)
Ship Quantity	25.43	22.02	25.99	24.48	23.72
Amex	6.77	7.45	102.18	38.80	7.11
Web	-3.84	-5.48	118.55	36.41	-4.66