Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   The company needs to decide whether it's worth sending out catalogs to the new customers or not.

2. What data is needed to inform those decisions?

   We need data that will help us to predict which new customers will respond to the catalog, i.e. make purchases, and how much they will spend. Also, we need to know the total cost involved in sending out the catalogs to the new customers.

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

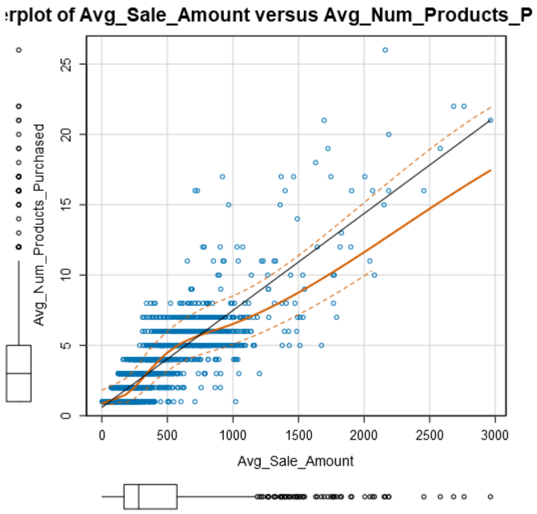
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

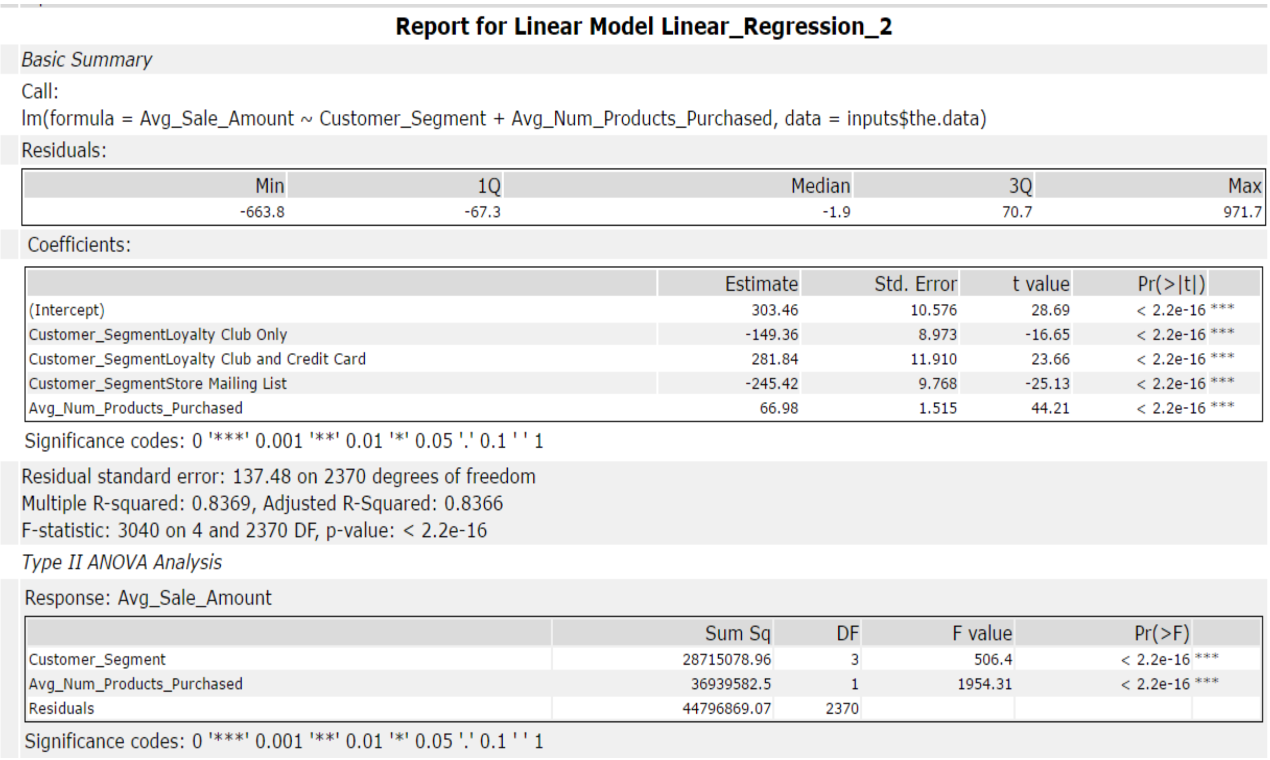
   At the outset, it looks like Customer Segment, State, Responded to Last Catalog, Avg Num Products Purchased, and Number of Years as Customer can be used as predictor variables. However, further analysis reveals that the data has only one state (CO), so it doesn't make much sense to have it as predictor variables. Likewise, Responded to Last Catalog, and Number of Years as Customer also doesn't contribute much as we are going to predict for new customers and not for existing customers. So, the predictor variables chosen would be "Customer Segment" and "Avg Num Products Purchased".

The scatter plot below shows a linear relationship between the Avg Sale Amount and Avg Num Products Purchased.



2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

   As per the report below, all predictor variables show p-values under 0.05, which are statistically significant and the Adjusted R-squared value of 0.8366 also shows that there is a linear relationship.



Report for Linear Model Linear_Regression_2

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

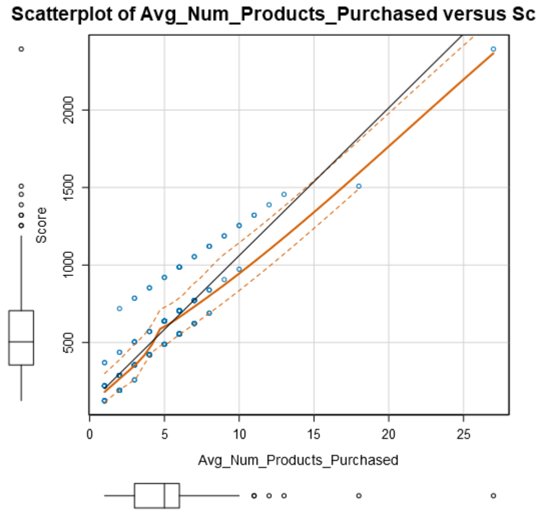
Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

This linear relationship can also be seen in the below scatterplot between the Score and the Avg Num Products Purchased.



3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Y = 303.46 + (281.84 * Customer_SegmentLoyalty Club and Credit Card) + (-149.36 * Customer_SegmentLoyalty Club Only) + (-245.42 * Customer_SegmentStore Mailing List) + (66.98 * Avg_Num_Products_Purchased) + (Credit Card * 0)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

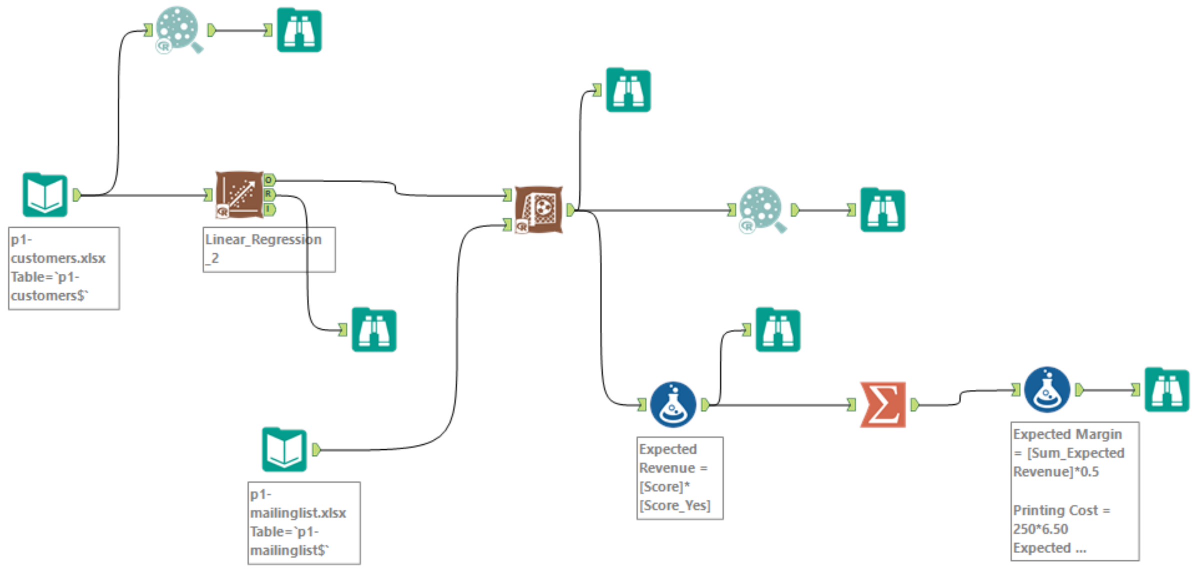
1. What is your recommendation? Should the company send the catalog to these 250 customers?

   Yes, the company should be sending the catalog to the 250 new customers.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

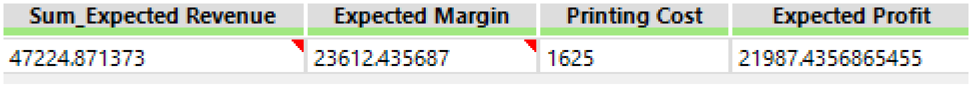
   Step 1: Generate linear regression with the predictor variables for the existing customers (p1-customers.xlsx)
   Step 2: Combine the overall mailing list (mailinglist.xlsx) with the results of the linear regression and create the scores.
   Step 3: Once the scores are generated, calculate the expected revenue, which would [Score] * [Score_Yes]
   Step 4: Once the expected revenue is calculated, summarize them.
   Step 5: Calculate the Margin and Profit as below.
   Expected Margin = Sum of the Expected Revenue * 0.5
   Printing cost = 6.50 * 250 (new customers)
   Expected Profit = Expected margin – Printing cost.

   The Alteryx workflow for this is shown below.



3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

   The expected profit is **$21987.436**



Sum_Expected Revenue	Expected Margin	Printing Cost	Expected Profit
47224.871373	23612.435687	1625	21987.4356865455