Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

The most important decision that needs to be made is whether the company will send the catalog to the 250 new customers, based on the expected profit.

2. What data is needed to inform those decisions?

In order to calculate the expected profit, two datasets will be needed. One is ***p1_customers*** *and it's the one that will be used to calculate linear regression's coefficients and the other is* ***p1_mailinglist*** *which is the dataset where linear regression is applied. The data needed to calculate expected profit from this decision are:* ***Customer_Segment, Score_Yes, Avg_Num_Products_Purchased*** *and* ***Avg_Sale_Amount****. Finally, gross margin and cost of producing are needed.*

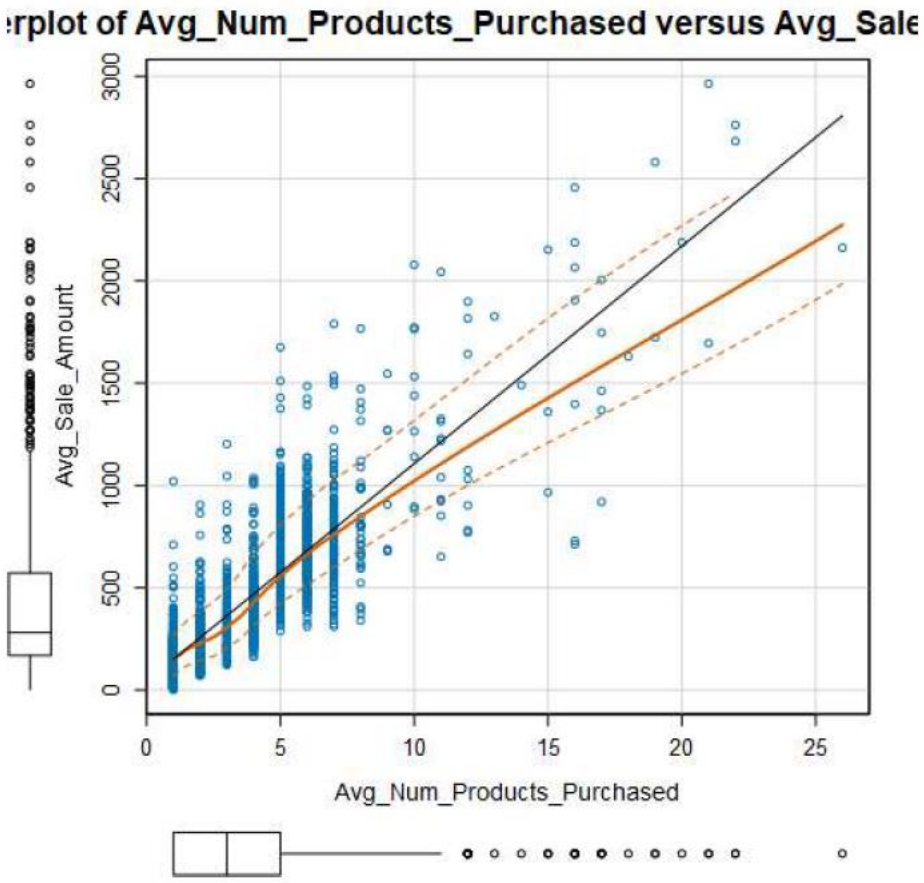
Step 2: Analysis, Modeling, and Validation

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)

*The file p1****-customers.xlsx*** *was chosen as the base data file to set up the linear regression model. Only* ***Customer_Segment*** *and* ***Avg_Num_Products_Purchased*** *were chosen as predictor variables. This was because of the following reasons:*

a) Responded to last catalog not relevant as new customers sought
b) Rest is address determiners and have no impact on new purchases

Also, target variable was Average sales. The Linear regression tool output was used to score the ***p1-mailinglist.xlsx****.*



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)

Report for Linear Model Linear_Regression_10

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = 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 degrees of freedom (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

R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. The higher the r-squared, the higher the explanatory power of the model. The linear model developed is a good model because multiple R-squared (0.8369) and adjusted R-squared (0.8366) values are high. P-value is the probability that the coefficient is zero. P-values, in this case, are below 0.05 which makes the predictor variables statistically significant.

3. What is the best linear regression equation based on the available data?

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

Step 3: Presentation/Visualization

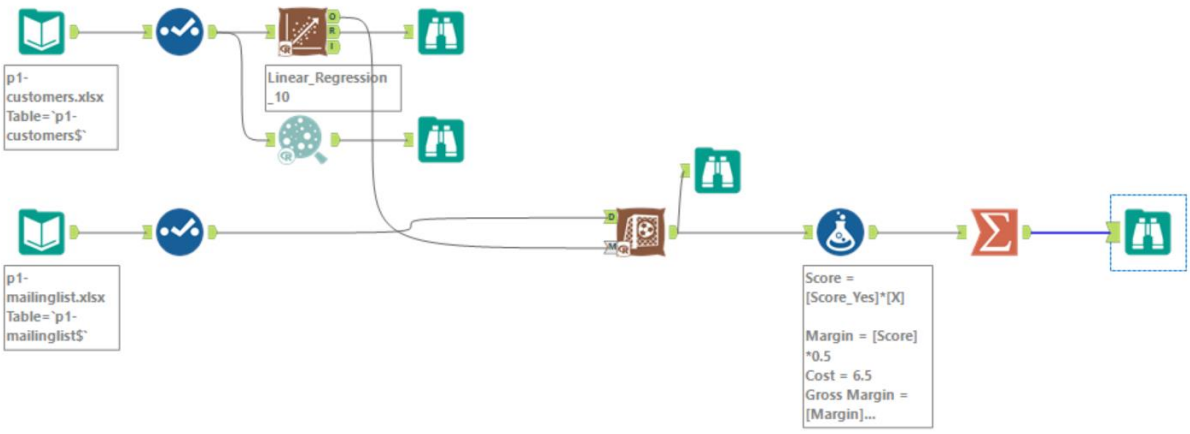
1. What is your recommendation? Should the company send the catalog to these 250 customers?

*Yes, if the profit is greater than **$10,000**.*

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

The X field gives the predicted sales amount for the 250 customers. Then it is multiplied with score_yes which is the probability of buying products. These individual values are summed. This is multiplied with gross margin of 50% on products sold via catalog. The total of (250 $6.50) is then subtracted from it. $6.50 is the cost of printing one catalog and 250 is the total number of people in it. The profit is greater than $10000 and hence profitable to send catalog to the new customers.*

Alteryx Workflow:



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

*Expected profit is **$21.987,44**.*

Record #	Sum_Margin	Sum_Cost	Sum_Gross Margin
1	23612.435687	1625	21987.435687