Project 1: Predicting Catalog Demand

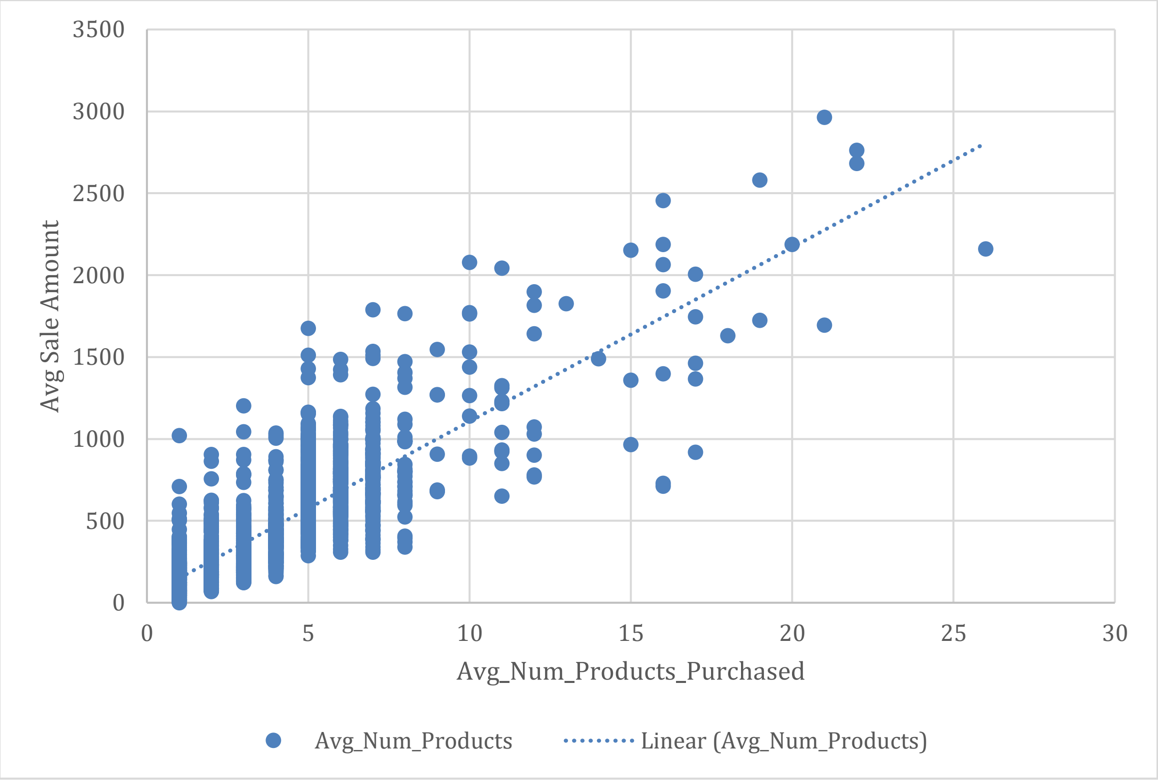
Step 1: Business and Data Understanding

The Business Problem tackled here is whether the company should carry out its catalog mailout campaign to its 250 customers. A prediction is required of the profit generated from this campaign. If the profit exceeds $10,000 the company should carry on with the campaign otherwise abort.

The data necessary for this analysis is customer's purchase history. The data present should contain the average amount of purchase the customer performed in the past, the items bought using the catalog and the amount spent buying products from the catalog. The mailing and fabricating cost per catalog and gross margin is also essential to compute the profit generated from the campaign.

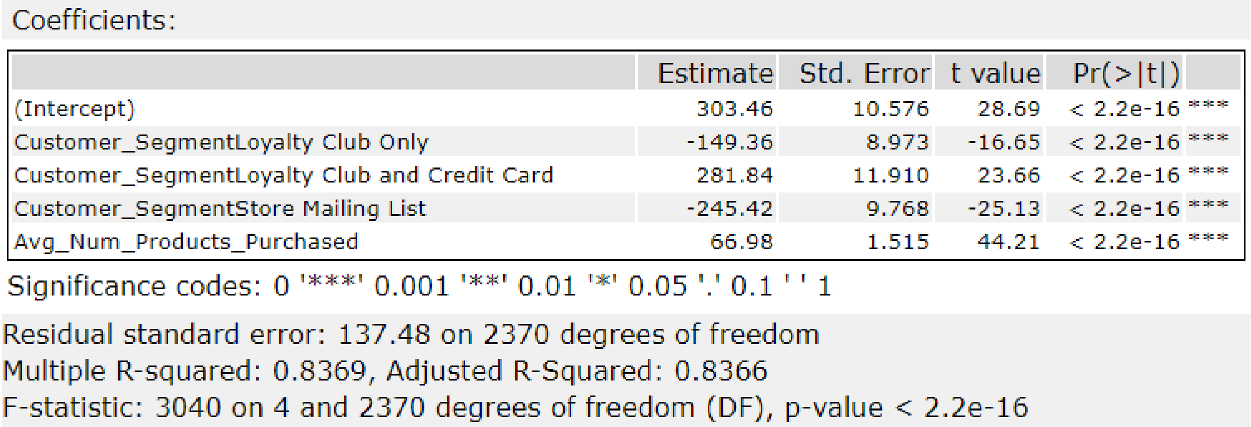
Step 2: Analysis, Modeling, and Validation

The variables were finalized after series of analysis of different linear regression. The most stable and Statistically significant model was discovered by using two predictive variables, Average number of products purchased .and Customer Segment. The targeted variable used was Average Sale Amount as discussed in the Data Understanding section. The predicted variable "Average number of products purchased", is a continuous variable so a scatterplot using Excel was used to determine the strength of the relationship with the target variable as shown below.



Although a continuous variable, “Years as Customer”, also show a promising scatterplot but after analysis its p-value was determined to be near 0.50. The validation for categorical variables was run using Alteryx which determined that variable such as, “Zip” and “Store Number” also had a p-value near 0.5.

The p-value and r squared value for Average number of products purchased alone were determined to be, 1.75314785927468E-14 and 0.732315 respectively. The validation which was earlier run on Excel data analysis taught in “Multiple Linear Regression using Excel” was later carried out on Alteryx. The Results of the Alteryx, shown below, also justify the use of the categorical variable “Customer Segment”.



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

As shown above, the results obtained from the linear regression highlights adjusted r squared value and p value that are favorable to the stability of the model. The r squared value, which is percentage of the variance in the targeted and predictive variables, should denote a good model if its value is above 0.7. Since this is a multiple linear regression, we use adjusted R-squared for validation. The adjusted r-squared value is 0.836, which is greater than 0.7.

Similarly, the p-value, which tests the null hypothesis that the coefficients are equal to zero, shows a significant relationship described by the model if its value is below 0.05-the lower the better. In the figure above all variables portray a small p-value, much smaller than 0.05, hence showing a Statistic Significance in the model.

Equation:
Revenue = 303.46 + 66.98 * Avg_Num_Products_Purchased - 149.36(Loyalty Club Only) + 281.84(Loyalty Club and Credit Card) - 245.42(Store Mailing List) + 0 (Credit Card Only)

Step 3: Presentation/Visualization

Yes, according to the Problem-Solving Framework, CRISP-DM, followed in this project, the company should send the catalog to the 250 customers. The Liner regression predicted the Average Sales which was multiplied by the probability of each customer “Score_Yes”. The

Average gross margin is 50% of the revenue for the products from the catalog. The revenue is subtracted from 250*6.50, where 6.50 is the price per catalog mailed, to find the profit.

The profit predicted to be generated is $21,987. An amount of $47224.87 was predicted from the model. With the 50% gross margin considered and total cost of mailing catalogs, which is 250*6.50 = $1,625, the price of $23,612.43 is subtracted from the catalog cost to obtain the profit produced. The profits exceed $10,000 so the company should conduct the mailout campaign.