Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

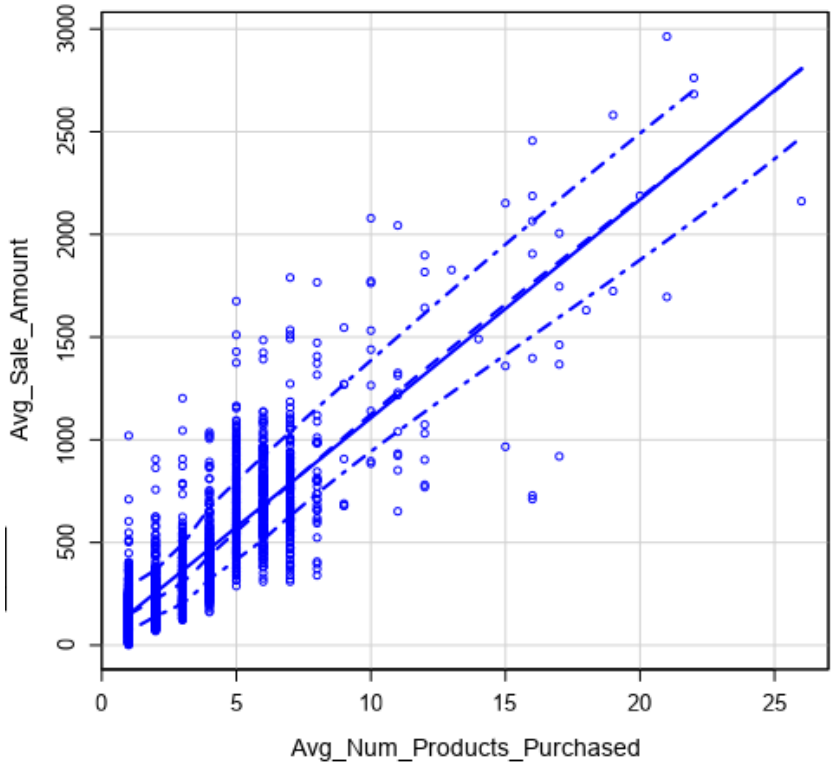
Key Decisions:

1. What decisions needs to be made?
   The decision that needs to be made is whether to send a catalog to the new 250 customers or not. This decision will be made is the expected profit from sending the catalog to the new 250 customers exceed $10,000.

2. What data is needed to inform those decisions?
   The data that is needed to inform the decisions is data from the previous catalog and data of existing customers including their average sale amount. The data from existing customers ia to be used to build a linear regression model to predict the expected sales, gross margin and profit of the new 250 customers.

Step 2: Analysis, Modeling, and Validation

1. To select the predictor variables for the model I have plotted scatterplots (shown below) for the numerical variables and noted that only the variable *Avg_Num_Products_Purchased* (average number of products purchased) has a linear relationship with *Avg_Sale_Amount* (average sale amount). After training the model, I have observed that the variable *Avg_Num_Products_Purchased* has a p-value < 2.2e-16 which suggests it is a good predictor.



For non-numerical variables I did a trial and error adding them one by one and observed that the variable *Customer Segment* with a p-value < 2.2e-16 is also a good predictor. This variable showed multiple R-squared of 0.702, which is more than the required 0.701

2. I believe that my linear regression model is good because it has achieved a Multiple R-squared of 0.8369 and an Adjusted R-Squared of 0.8366 which are relatively high, meaning about 84% of the average amount of sales values can be explained by the input values using the linear regression model. Also, the p-values of the predictor variables I have selected are <2.2e-16, which is less than the required 0.05, meaning that predictor variables are good and significant predictors. The coefficients of the linear regression model are as follows:

	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	***

3. The best linear regression equation based on the available data is:

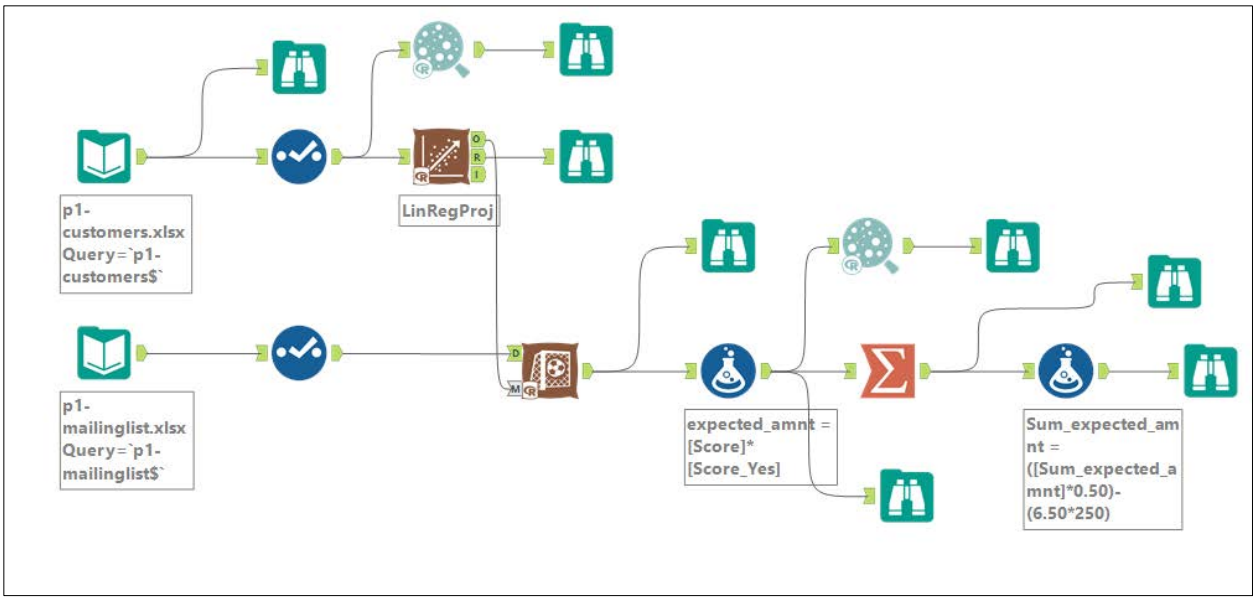
ExpectedSales = 303.46 + (Customer_SegmentLoyalty Club Only * -149.36) + (Customer_SegmentLoyalty Club and Credit Card * 281.84) + (Customer_SegmentStore Mailing List * -245.42) + (Avg_Num_Products_Purchased * 66.98) + 0

Step 3: Presentation/Visualization

1. I highly recommend the company to send the catalog to the 250 new customers as the expected profit contribution exceeds $10,000.

2. How did you come up with your recommendation?
I came up to this recommendation by using the linear regression model to predict the amount of sales expected for each new customer. I then applied the probability that the customer will respond to the catalogue and make a purchase to the amount of expected sales to produce the expected revenue by customer. I multiplied this expected revenue by customer with 50% gross margin to produce the predicted gross margin and summed the predicted gross margin for all the new customers. From the sum of the predicted gross margin I subtracted the total cost of printing and distributing the catalogs (($6.50 * 250) for the new customers.

My solution on Alteryx is as follows:



3. The expected profit from the new catalog is $21,987.44