Project 1: Predicting Catalog Demand

# Step 1: Business and Data Understanding

Our company is trying to make a decision on adding 250 new customers to the distribution list for this year's edition of our print catalog. In order for it profitable, the additions should generate at least $10,000 in profit. I will determine whether the total profit from the additional 250 customers exceeds the $10,000 mandate by leadership so the project can be approved. If the profit does not exceed the $10,000 minimum profit, the project will be rejected.

### Key Decisions:

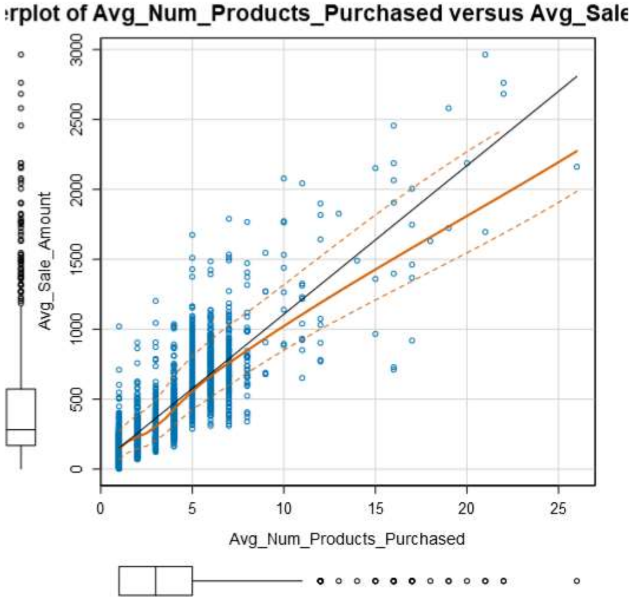
1. What decisions needs to be made? We need to decide if adding the additional 250 customers to the distribution list will result in an increase in profits of greater than $10,000 for the business. We know that the average gross margin is 50% and that there is also an additional cost of $6.50 per catalog for printing and distribution. These costs will be utilized to calculate an expected gross profit for the project. If we calculate the expected profit to be greater than $10,000 then we will move forward with sending the additional 250 catalogs. If the expected profit is less that $10,000 the we will not add the additional catalogs.

2. What data is needed to inform those decisions? The following data will be needed in order to make a good business decision:
    - We will need historical data on current customers' purchase history, type of customer, volume of purchases, and dollar volumes purchased.
    - We will need another dataset with matching variables on the 250 potential new customers that we want to add to the distribution list of the next catalog to include the probability that they would respond to the new catalog. This dataset would obviously be devoid of the target variable.
    - The other item we will need is the cost structure of the products produced as well as costs associated with distribution of the 250 additional catalogs.

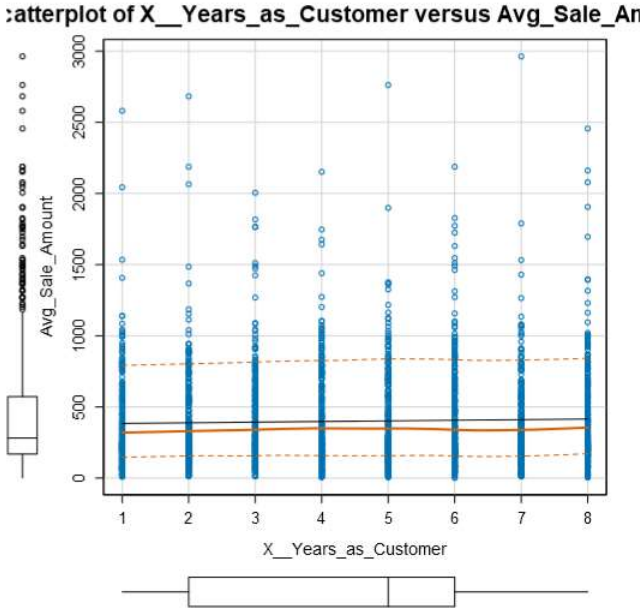
# Step 2: Analysis, Modeling, and Validation

Setting up my regression model I choose Avg_Sale_Amount as the target variable as we are trying to predict the additional revenue we will be able to generate with the new mailing list. As predictor variables I looked for continuous variables that were relevant to the target variable and that did not have a high number of null values. I then created scatterplots to determine if there was a linear relationship between the target variable and the predictor variables. This resulted in the following plots:

### Scatterplot of Average Number of Products Purchased vs. Average Sale Amount



### Scatterplot of Number of Years as a Customer vs. Average Sale Amount



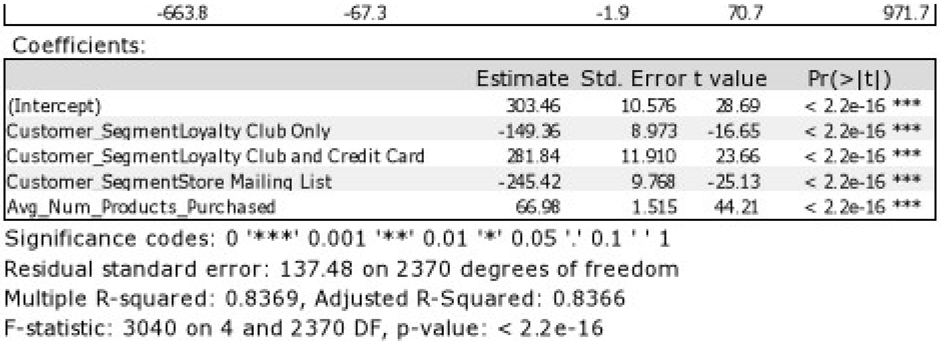
Plotting the potential predictor values on the x axis against the target variable on the y axis will give us an indication as to whether or not it has the potential to be a good predictor. From the two scatterplots we can see that while average number of products purchased appears to be a good predictor, the number of years as a customer does not. Given this I'll load the Avg_Num_Products_Purchased into the regression tool in Alteryx and run it against the target value, Avg_Sale_Amount.

When I run the tool, I see that products purchased is a good predictor of average sale amount as evident by its P value that is well below 0.05. The P value of <2.2e$^{16}$ tells us that there is a strong relationship between the number of products purchased and the average sale amount. We know that with P values this low between the target value and the predictor value that the relationship between the two is statistically significant and not occurring by chance.

Additionally I get an R Squared Value of 0.7323 and while this indicates a good fit, I may be able to find a better fit by testing some additional variables.

While looking through the categorical variables I randomly tested several categorical variables like customer segment, city, zip, and store number. I found that we can improve the model fit by loading the Customer_Segment variable. When I load customer segment variable into the tool and rerun it we see an improved fit with an Adjusted R Squared Value of 0.8366. R values range from 0 to 1 and give an indication of how much variance there is between variables being tested. In this instance we have added an additional variable to the model and increased the R squared value. This significantly improves the fit of the model.

With P vaules less than 0.05 and an Adjusted R squared value of 0.8366 the tool gives us the follow coefficients that can be used to create a model:



```
-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
```

Using the coefficients give by the regression tool we now have the following model to make our predictions:

**Y = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149 (if Customer_Segment = Loyalty Club Only) + 281.84 (if Customer_Segment = Loyalty Club and Credit Card) – 245.42 (if Customer_Segment = Store Mailing List) + 0 (if Customer_Segment = 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? The recommendation I would make in this case is to proceed with adding the 250 customers to the distribution list and send the catalogs.

2. How did you come up with your recommendation? After creating the model using regression analysis in Alteryx I utilized the score tool to project the predicted average sale for each of the 250 customers on the mailing list. I then used the formula tool to create three additional columns:

- Possible Sales - [Score_Yes]*[Pred_Avg_Sale_Amount]
- Gross Margin - [PossSales]*.5
- Profit - [Gross_Margin]-6.50

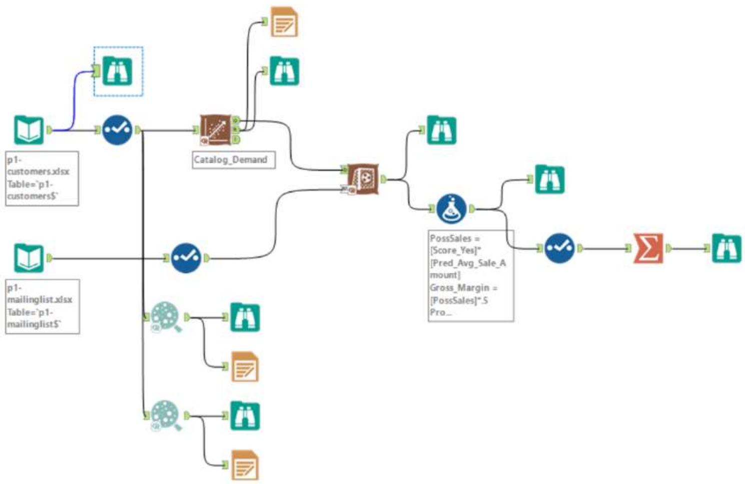
Creating these columns provided me with the profit per customer for the 250 customers on the mailing list. Utilizing the summarize tool I calculated the profit for the mailing list.

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

**Expected profit** = $21,987.44

(Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)

## Alteryx Workflow:



p1-customers.xlsx Table=`p1-customers$`
Catalog_Demand
p1-mailinglist.xlsx Table=`p1-mailinglist$`
PossSales = [Score_Yes]*[Pred_Avg_Sale_Amount] Gross_Margin = [PossSales]*.5 Pro...