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 management of the company that manufactures and sells high-end home goods want to determine if they should send print catalog to the 250 new customers in the coming months. If the expected profit contribution from these new customers exceed $10,000, the company will send the catalog. Otherwise, the company will not send the catalog to these new customers.

2. What data is needed to inform those decisions?

The expected profit is required to be calculated from these 250 new customers. For this, we need the historical data of the customers to build the linear regression model. After building the model, we need to make predictions on the mailing list dataset to come up with the predicted sales amount from the 250 new customers. Then, we need to use the Score_Yes column to calculate the worth of products actually bought by customers.
To calculate the final expected profit, we need the costs of printing and distributing is catalog which is $6.50. The average gross margin on all products sold through the catalog is 50%. Hence, the expected profit will be calculated as SUM(Predicted Sale Amount * Score_Yes) * 50% – $6.50 * 250.

These information are critical to inform the decision of whether the catalog will reap any profits for the company from the 250 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)

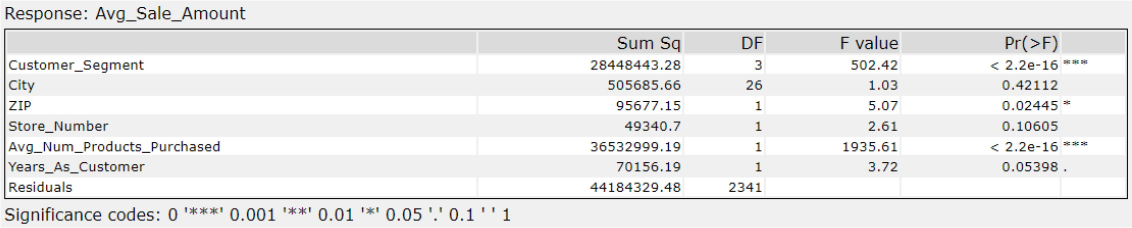
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with

Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

First, I analyzed all the possible predictor variables such as Customer Segment, City, Zip, Store Number, Average Number of Products Purchased and Years_As_Customer. I built a linear model taking all the above predictor variables and Average Sale Amount as target variable to find out the statistical significance of each of these variables.



Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28448443.28	3	502.42	< 2.2e-16 ***
City	505685.66	26	1.03	0.42112
ZIP	95677.15	1	5.07	0.02445 *
Store_Number	49340.7	1	2.61	0.10605
Avg_Num_Products_Purchased	36532999.19	1	1935.61	< 2.2e-16 ***
Years_As_Customer	70156.19	1	3.72	0.05398 .
Residuals	44184329.48	2341		

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

Figure 1

Based on the regression results above, only Customer_Segment, Zip and Avg_Num_Products_Purchased are statistically significant to establish a linear regression relationship with Average Sale Amount. From the below figure 2, we observe that Avg_Num_Products_Purchased is highly correlated to Average Sale Amount. It has a high coefficient of determination (R^2) = 0.73.

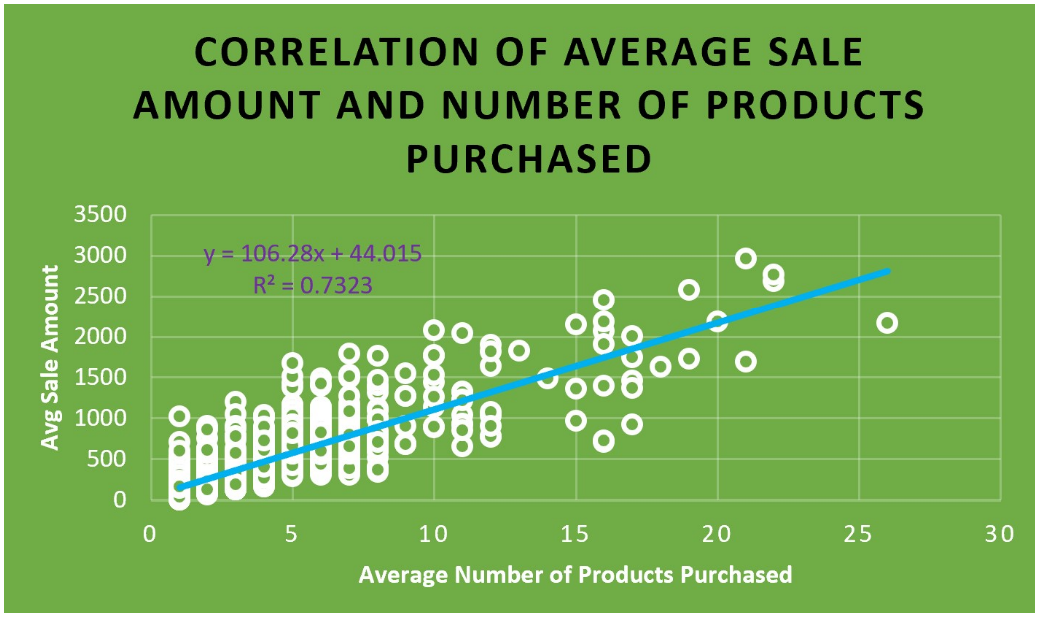


Figure 2

The remaining variables such as Years as Customer, Zip, City, Address and Store Number did not have a linear relationship with Average Sale Amount. This can be observed from Figure 3-7. The scatter plots and line charts demonstrate the absence of linear relationship between the target variable and predictor variables.

SCATTER PLOTS

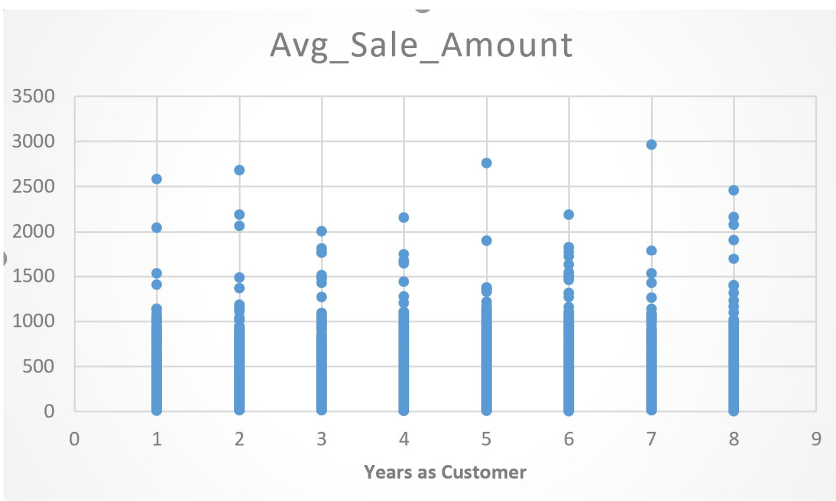


Figure 3- AVERAGE SALE AMOUNT AND YEARS AS CUSTOMER

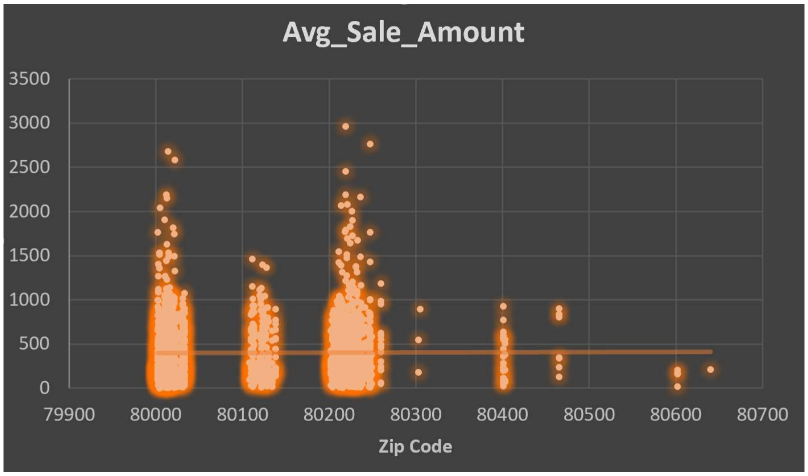


Figure 4- AVERAGE SALE AMOUNT AND ZIP

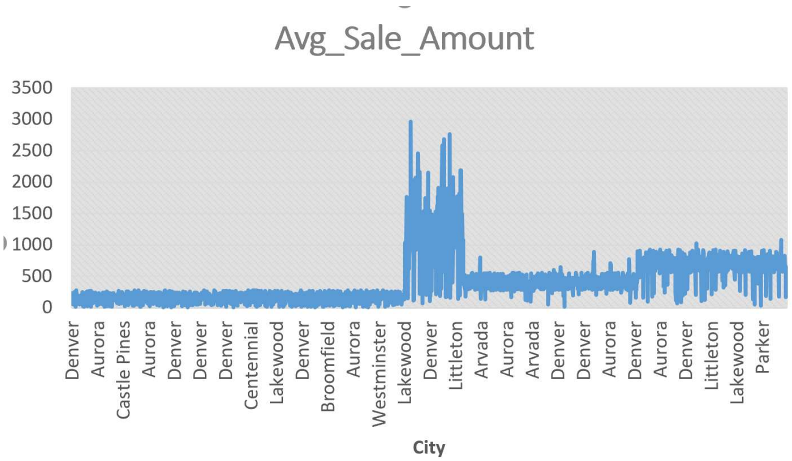


Figure 5- AVERAGE SALE AMOUNT AND CITY

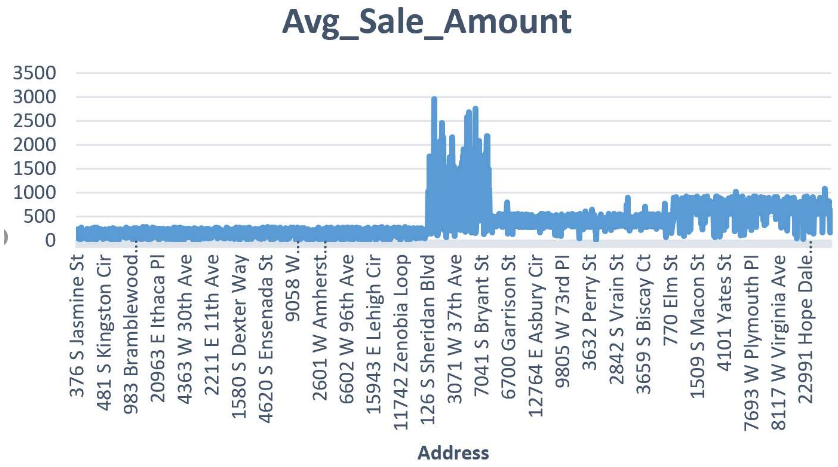


Figure 6- AVERAGE SALE AMOUNT AND ADDRESS

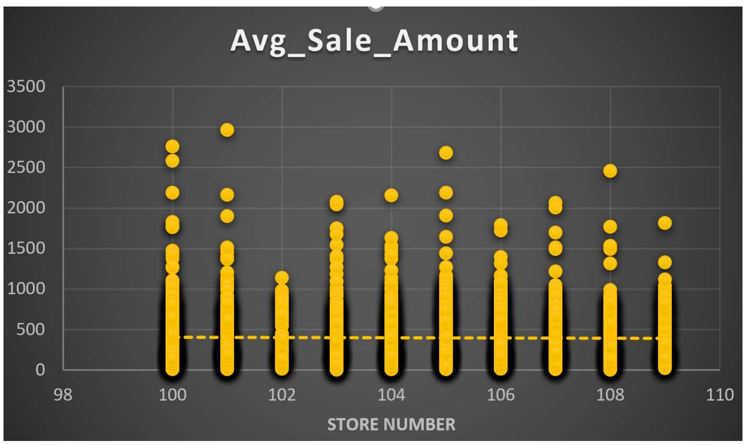
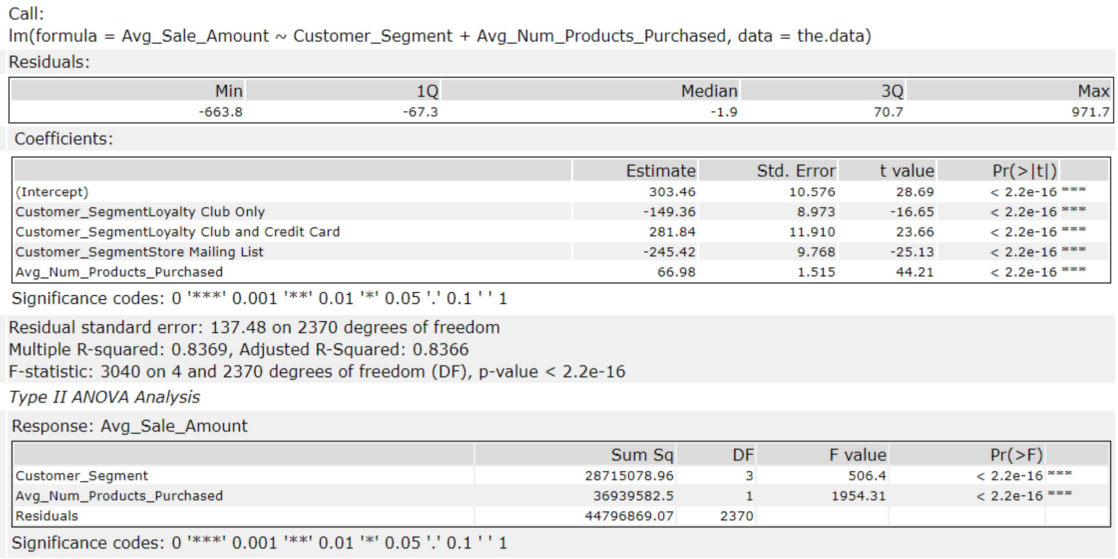


Figure 7- AVERAGE SALE AMOUNT AND STORE NUMBER

After performing the above analysis, I built the final linear regression model using the two predictor variables- **Customer Segment** and **Average Number of Products Purchased**. The target variable used is **Average Sale Amount**.

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.

The statistical results that the linear regression model created showed impressive results. As we can see, the p-value for both the variables is less than 2.2e-16. Since the predictor variables have a p-value below 0.05, the relationship between them and the target variable Average Sale Amount is considered to be statistically significant. Also, the R-square value is 0.8369, and the adjusted R-squared value is 0.8366, which is a high value. Considering these factors, the linear regression model is a good model. The below Figure 8 shows the final results of the linear regression model.



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

Figure 8

The below Normal Q-Q plot and Residual vs Fitted plots also demonstrates a good reliability of the linear model. We can clearly see that the data is quite evenly distributed across the fitted line.

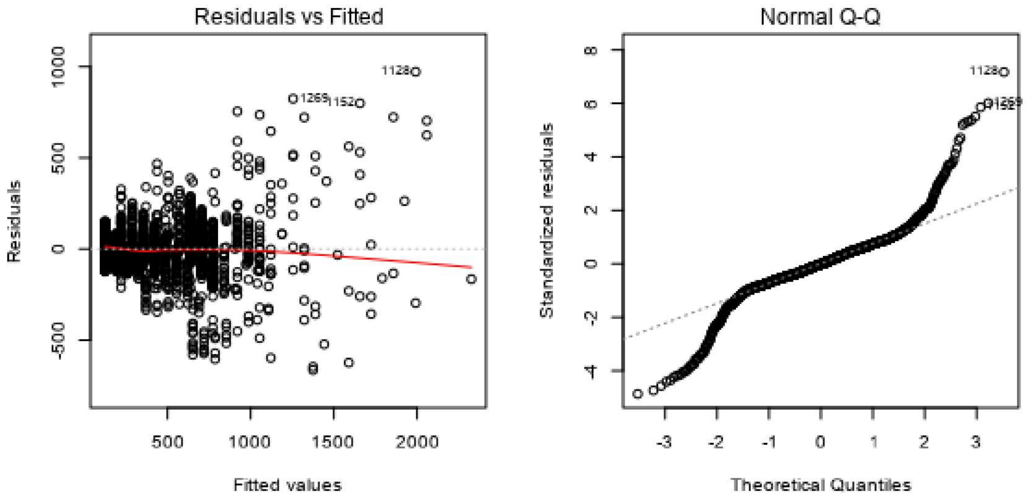


Figure 9

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)

Important: The regression equation should be in the form:

Regression equation is as follows:

Predicted Sales = 303.4635 + 281.84* 1 (If Customer_Segment = “Loyalty Club and Credit Card” - 149.36 * 1(If Customer_Segment =”Loyalty Club Only”) - 245.42* 1 (If Customer_Segment = “Store Mailing List”) + 66.98* Average Number of Products Purchased + 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?

Yes, I recommend the company to send the catalog to the new 250 customers. The expected profit from these customers is about $21987.44 which is quite higher than the threshold amount of $10000 which the company is targeting. The new customers would be profitable for the company.

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

   After building the model, I used the score tool to make predictions on the dataset of the 250 new customers. Then, I used a scatter plot tool in Alteryx to find the relationship between the predicted sales amount and average the number of products purchased. In the below figure 10, I validated the strong linear relationship that exists between the two variables which is similar to what was obtained during the analysis phase on the original dataset.

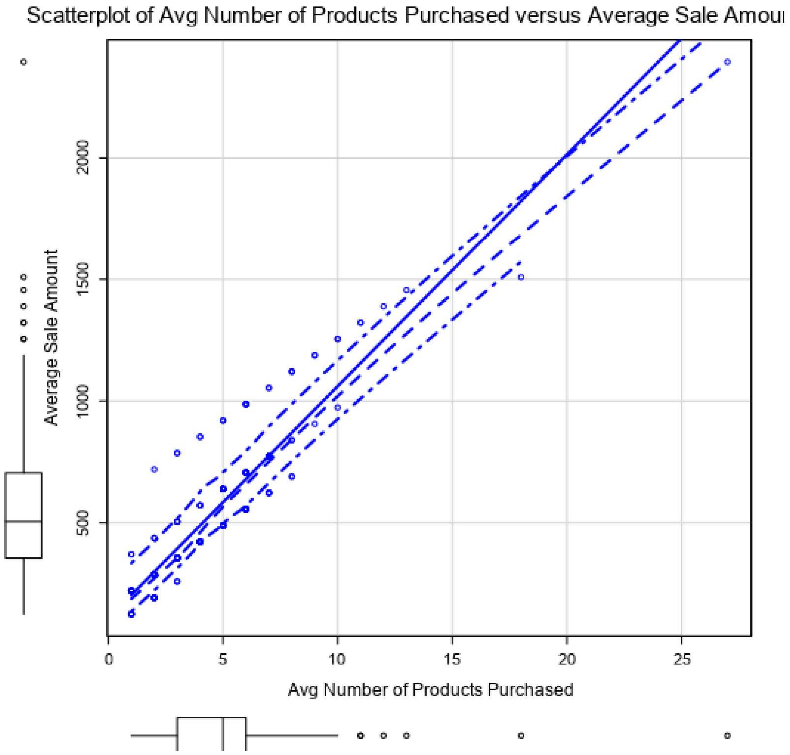


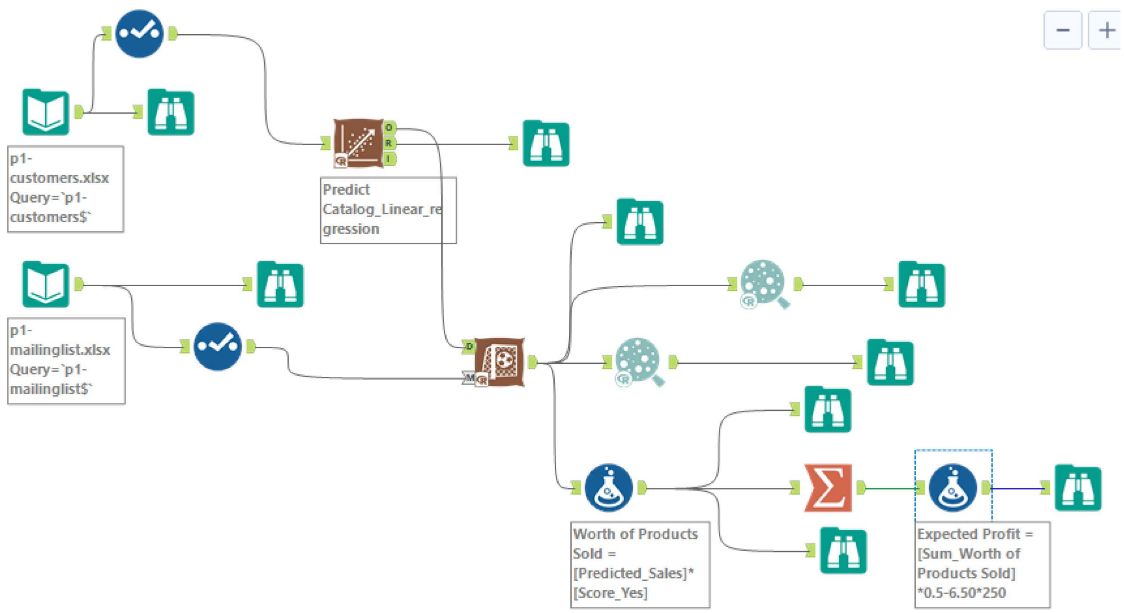
Figure 10

The below are the detailed steps that were followed:

1. I used the following linear regression equation obtained.
   Predicted Sales = 303.4635 + 281.84* 1 (If Customer_Segment = ”Loyalty Club and Credit Card” - 149.36 * 1(If Customer_Segment =”Loyalty Club Only”) - 245.42* 1 (If Customer_Segment = ”Store Mailing List”) + 66.98* Average Number of Products Purchased + 0 (If Customer_Segment= “Credit Card Only”)

2. The used the score tool on the output from the model to the mailing list data set to get the predicted sale amount.
3. Then multiplied **Predicted Sale Amount** by Score_Yes (which is the probability to buy) for each customer and named as **Worth of Products Sold.**
4. Finally, summed **Worth of Products Sold** and multiplied by 50% (gross margin) and subtracted the catalog cost ($6.5 *250) for all 250 customers. This will give us the **Expected Profit** from the 250 customers.

Below is the Alteryx Workflow and formula for Expected Profit:



p1-customers.xlsx Query=`p1-customers$`
Predict Catalog_Linear_regression
O
R
I
p1-mailinglist.xlsx Query=`p1-mailinglist$`
D
M
Worth of Products Sold = [Predicted_Sales]* [Score_Yes]
Expected Profit = [Sum_Worth of Products Sold] *0.5-6.50*250

Expected profit = **SUM(Predicted Sales * Score_Yes) * 0.5 – $6.50 * 250**

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

The expected profit = SUM(Predicted Sales * Score_Yes) * 50% – $6.50 * 250
The predicted Revenue = SUM(Predicted Avg_Sale_Amount * Score_Yes) = $47,224.87
The expected profit = $47,224.87 * 50% – $6.5 *250 = **$21,987.44**

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.