Project 1: Predicting Catalog Demand

Business Issue

You recently started working for a company that manufactures and sells high-end home goods. Last year, the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to. Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models. You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Details

- *The costs of printing and distributing is $6.50 per catalog.*
- *The average gross margin (price - cost) on all products sold through the catalog is 50%.*

Step 1: Business and Data Understanding.

Key Decisions:

1. *What decisions needs to be made?*
   The company's goal is to send their catalog to a group of 250 new customers provided that the expected profit pool generated through it by this mailing list exceeds $10,000. Based on this, we need to decide whether to send the catalog out or not.

2. *What data is needed to inform those decisions?*
   Generally, Profits are the result of Revenue minus COGS (cost of sold goods). We already know that COGS include $6.50 per catalog. We should therefore predict customers' revenue based on the given training set. We should also take the avg. gross margin into account when calculating the expected profit contribution. Additional factors may influence our calculations, so we need to analyze our datasets first.

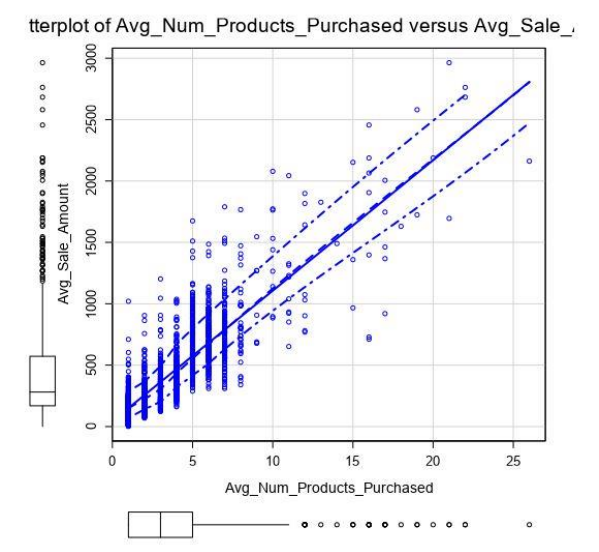
Step 2: Analysis, Modeling, and Validation

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.*

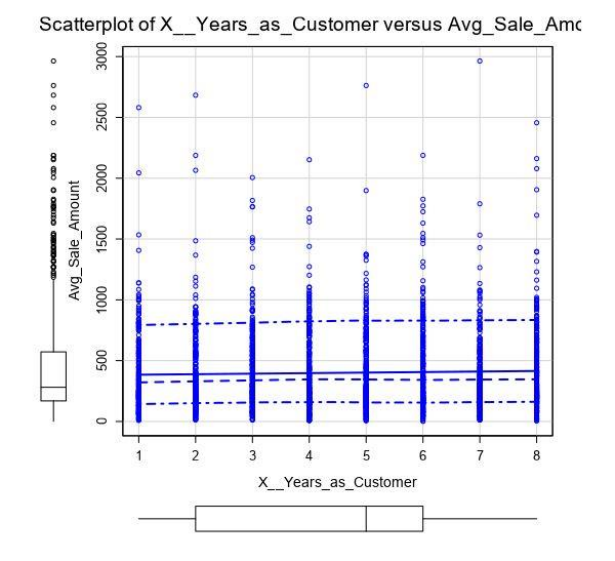
Since we are in a data-rich scenario and our target variable is numeric – continuous, we are going to apply a multiple linear regression model. The training dataset contains both numeric (double) and categorical variables (V_String):

	Field	Type	Size	Re
▶	Name	V_String	255	
	Customer_Segment	V_String	255	
	Customer_ID	Double	8	
	Address	V_String	255	
	City	V_String	255	
	State	V_String	255	
	ZIP	V_String	19	
	Avg_Sale_Amount	Double	8	
	Store_Number	Double	8	
	Responded_to_Last_Catalog	V_String	255	
	Avg_Num_Products_Purchased	Double	8	
	#_Years_as_Customer	Double	8	
	*Unknown	Unknown	0	

Our target variable must be “Avg_Sale_Amount” since we want to predict customers’ revenue. When plotting “Avg_Sales_Amount” against “Avg_Num_Products_Purchased” we can identify a positive linear relationship:



The same does not apply to “Avg_Sales_Amount” vs. “#_Years_As_Customer”. No sloped line is visible:



Meanwhile, we can exclude the "responded_to_last_catalog" and "customer_id" because new customers can't be influenced by these variables: they are new clients and have never had the chance to respond to a catalog. Eventually, the "state" field is the same throughout the file, so it won't play any role as predictor. When plugging in with the remaining variables, the fields "city", "ZIP" and "Store_Number" are shown as not statistically significant. As foreseen, the "#_Year_As_Customer" is not a good predictor either.

Here below is the final report:

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

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.*
   With this linear model we can confidently predict the expected sales amount. All variables have 3 stars which means that we can reject the null (there is no relationship between the variables) at the significance level of .001. In fact, the p value can be written as 2.2 * 10^16 which corresponds to a value smaller than .00000000000000022. The adjusted R (.8366) is close to 1 which generally suggests that nearly all variance in the target variable is explained by the model.

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)*
   Y = 303.46 -149.36 * (If Type: Loyalty Club only) +281.84 (If Type: Loyalty Club and Credit Card) -245.42 (If Type: Mailing List) + 0 (If Type:Credit Card Only) + 66.98

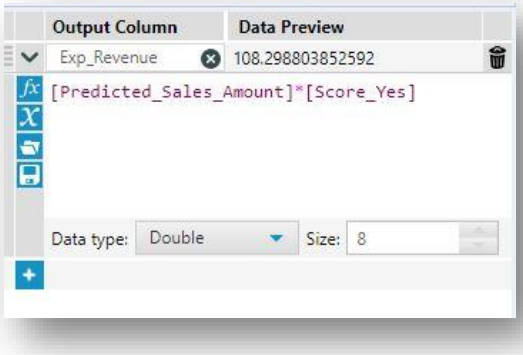
Step 3: Presentation/Visualization

1. *What is your recommendation? Should the company send the catalog to these 250 customers?*
   Yes, the company should send its catalog to the selected customer group.

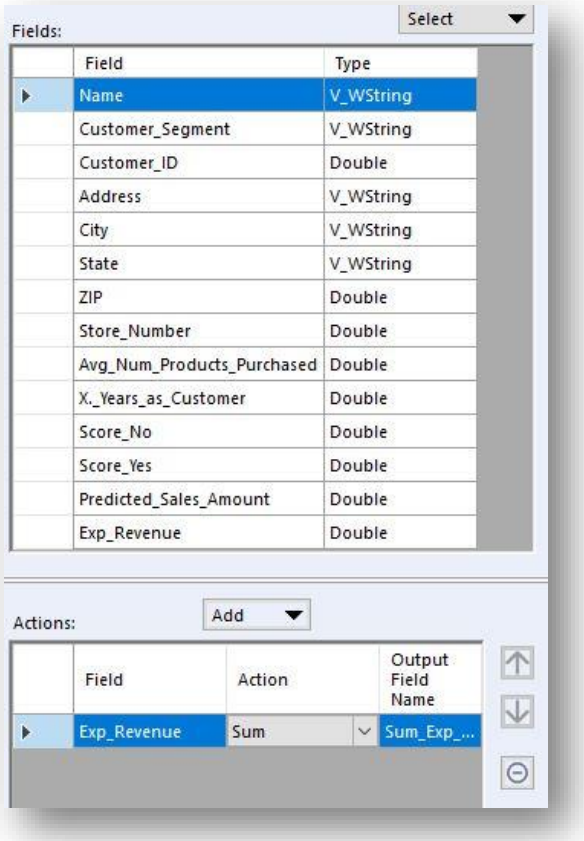
2. *How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)*

Step 1: Calculating the Total expected Revenue

When scoring the model, we can predict the expected sales amount. Taking into account the probability that a person will buy the firm's catalog, I have multiplied the expected revenue for the probability level:

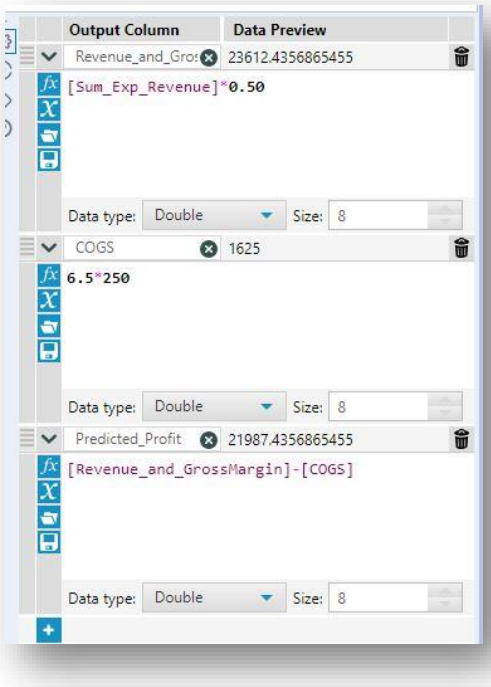


We can find the total revenue by using the "Summarize tool":



Step 2: Calculating the profit pool

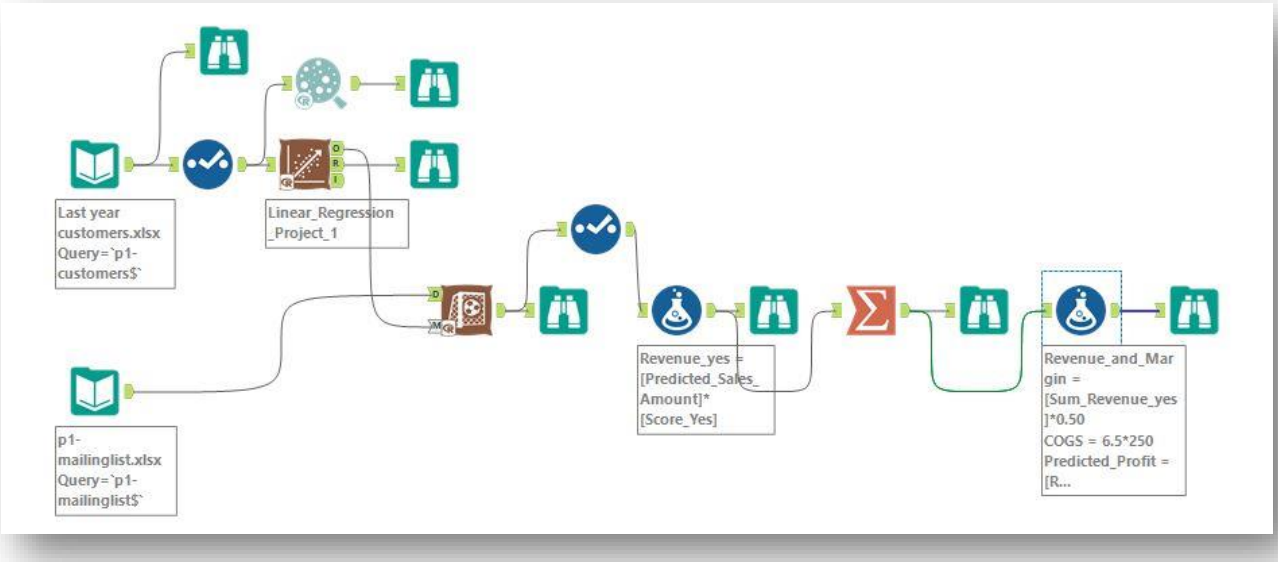
The final formula should be: (Expected Revenue * Gross Margin) – (Cost per Unit * Unit No.). By using the formula tool, we can make the required calculations:



Step 3 – Conclusion:

The result indicates that the company's condition has been satisfied as profits exceed $10,000.

Here below you may find the complete workflow:



3. *What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?*
   The expected profit from the mailing list is approx. $21,987.44.

Record	Sum_Exp_Revenue	Revenue_and_GrossMargin	COGS	Predicted_Profit
1	47,224.871373	23,612.435687	1,625	21,987.435687