

## 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 decision that we need to make is, “Do we have enough revenue that the company expects if they send out the catalog to the new customers?” and “If not, how much revenue can we expect from these customers?” and “whether we will send the catalogs to these customers or not”.

2. What data is needed to inform those decisions?

We need to know the cost incurred and profit earned for the customers. The data should include how much they have bought from the company and the what is the sale amount and also the probability that the new customers will buy from us. Once we have this information, we need to calculate the expected profit earned by the company by multiplying it with the average gross margin (price-cost) and subtracting the cost related to the printing and distribution of the catalogs and check whether the total profit earned is more than what the company expects that is, \$10,000.

### Step 2: Analysis, Modeling, and Validation

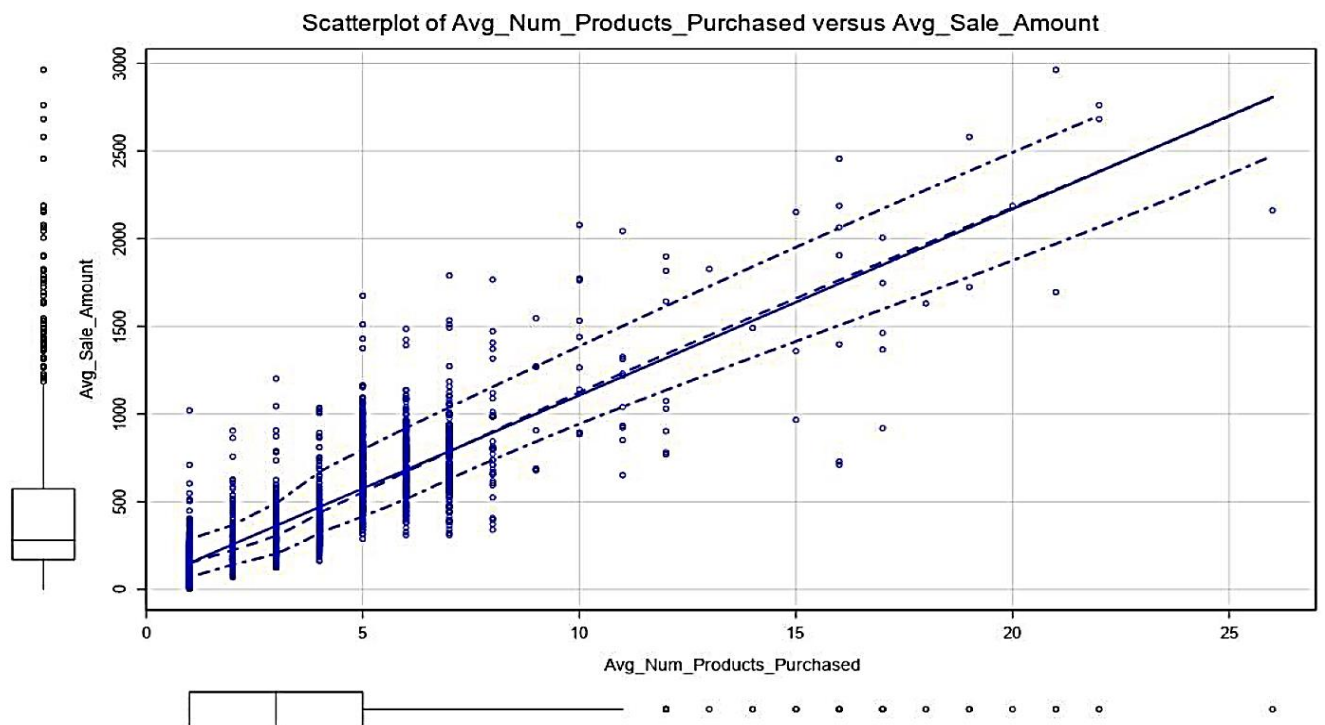
*Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)*

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

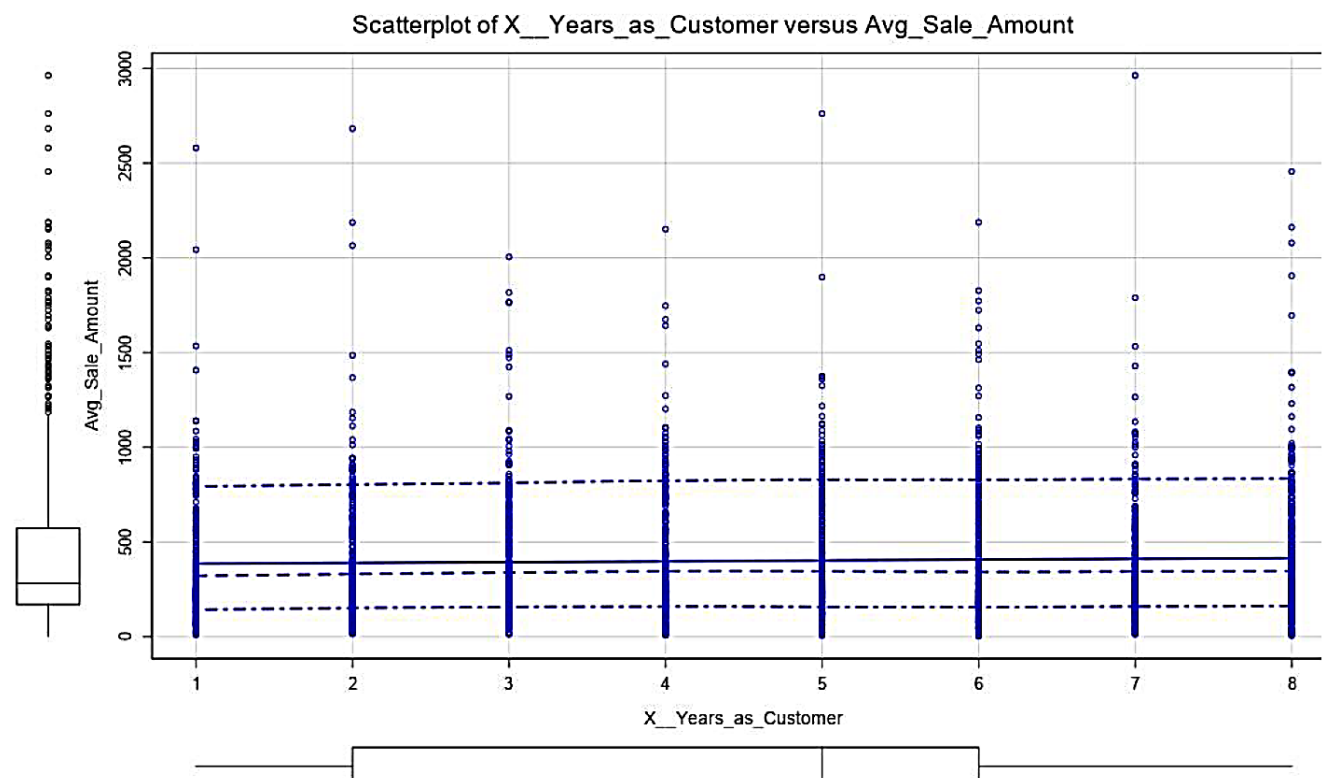
*At the minimum, answer these questions:*

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.

The predictor variable chosen is Avg\_Num\_Products\_Purchased due to the linear relationship between the predictor variable and the target variable Avg\_Sale\_Amount which we can see from the scatterplot below.



Also, we are not taking the #\_Years\_as\_Customer as predictor variable because there is no linear relationship between the predictor variable and Avg\_Sale\_Amount which is depicted in the scatterplot below.



From the correlation matrix, we can see that the correlation between Avg\_Sale\_Amount and Avg\_Num\_Products\_Purchased is quite higher that is, 0.85 which  $>0.7$  and therefore considered to have a strong relationship between them, whereas #\_Years\_as\_Customer

with the target variable has very poor relationship that is  $<0.3$  and therefore cannot be taken as useful.









Record	FieldName	Avg_Sale_Amount	Avg_Num_Products_Purchased	#_Years_as_Customer
1	Avg_Sale_Amount	1	0.855754	0.029782
2	Avg_Num_Products_Purchased	0.855754	1	0.043346
3	#_Years_as_Customer	0.029782	0.043346	1





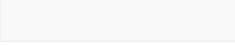
Also, Customer\_Segment is taken as a Predictor variable which is a categorical variable. The P-value of the variable is  $<2.2\text{e-}16$  which is  $<0.05$  and therefore considered to be statistically significant. The below report shows the P-values of the predictor variables.

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	$< 2.2\text{e-}16$ ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	$< 2.2\text{e-}16$ ***
Residuals	44796869.07	2370		

- 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 linear model is a good model because it fits the data accurately and explainable over large variances. The adjusted R-squared value is 0.837 which is  $>0.7$  and therefore considered to be a good model. Also, the p-values of Customer\_Segment and Avg\_Num\_Products\_Purchased are  $<2.2\text{e-}16$  both which is  $<0.05$  and therefore the linear model is considered to be statistically significant.

	R SQUARED <b>0.837</b>		ADJUSTED R SQUARED <b>0.837</b>
	MEAN ABSOLUTE ERROR <b>93.068</b>		MEAN ABSOLUTE PERCENT ERROR <b>0.58</b>
	MEAN SQUARED ERROR <b>18861.84</b>		ROOT MEAN SQUARED ERROR <b>137.338</b>
	F-STATISTIC <b>3039.74 on 4 and 2370 degrees of freedom</b>		RESIDUAL STANDARD ERROR <b>137.483 on 2373 degrees of freedom</b>

Variable	Estimate	Impact	Confidence	Pr(> t )
(Intercept)	303		***	1.12e-155
Customer_SegmentLoyalty Club Only	-149		***	6.35e-59
Customer_SegmentLoyalty Club and Credit Card	282		***	2.58e-111
Customer_SegmentStore Mailing List	-245		***	1.05e-123
Avg_Num_Products_Purchased	67		***	7.99e-312

- 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:**

$$Y = \text{Intercept} + b1 * \text{Variable}_1 + b2 * \text{Variable}_2 + b3 * \text{Variable}_3 \dots$$

**For example:**  $Y = 482.24 + 28.83 * \text{Loan\_Status} - 159 * \text{Income} + 49 (\text{If Type: Credit Card}) - 90 (\text{If Type: Mortgage}) + 0 (\text{If Type: Cash})$

Note that we **must** include the 0 coefficient for the type Cash.

**Note:** For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

The best linear regression equation based on the data is,

$$Y = 303.46 - (149.36 * \text{If Type: Loyalty Club Only}) + (281.84 * \text{If Type: Loyalty Club and Credit Card}) - (245.42 * \text{If Type: Store Mailing List}) + (66.98 * \text{Avg\_Num\_Products\_Purchased})$$

**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

## Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

- What is your recommendation? Should the company send the catalog to these 250 customers?

The recommendation is to send the catalog to the 250 new customers, since the predicted profit is \$21,987.44 which is more than the expected profit \$10,000.

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

The Linear Regression model was applied on the p1-customers dataset with Avg\_Sale\_Amount as the target variable and Customer\_Segment and Avg\_Num\_Products\_Purchased as the predictor variables. The model was then used to predict the scores for the p1-mailinglist dataset by passing it through SCORE tool and the p1-mailinglist dataset as the inputs. The scores were then passed through a FORMULA tool in which the following expression was used:  $([Score] \times [Score\_Yes]) \times 0.50 - 6.50$ , where the [Score] is the predicted scores output by the SCORE Tool, [Score\_Yes] is the probability that the customer will respond to the catalog and make a purchase, and multiplied the product with 0.50 which is the average gross margin and subtracted 6.50 which is the cost incurred in printing and distribution of the catalogs. Next, we summarized the total profit for the 250 customers by passing it through a SUMMARIZE tool. The result is the predicted profit for the 250 new customers.

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

The expected profit from the new catalog by sending it to the 250 new customers is \$21,987.44

<b>1<sup>2</sup></b> Sum_Cost_of_catalog	
21,987.4356	1

### 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.