Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The company needs to know how much profit it should expect from sending a catalog to 250 new customers and check if the expected profit contribution exceeds $10,000 for the management to support such decision.

There are 2 data sets provided:

- p1-customers.xlsx contains the data used to build the regression model
- p1-mailinglist.xlsx contains the data related to 250 customers to be used for the prediction.

The data needed to inform the decision is

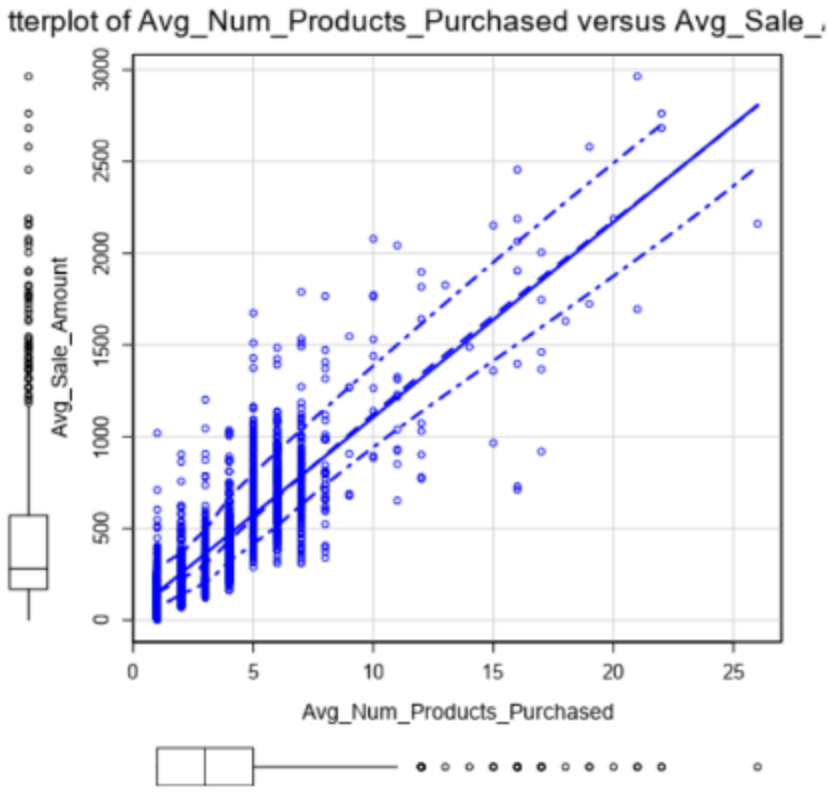
- Previous sales data which will be used to build regression model
- Avg_Sale_Amount from the 250 new customers to be predicted by applying the model
- Predicted revenue calculated by multiplying Avg_Sale_Amount by Score_Yes
- Cost to print and distribute the catalog
- Average gross margin
- Expected profit which will be calculated

Step 2: Analysis, Modeling, and Validation

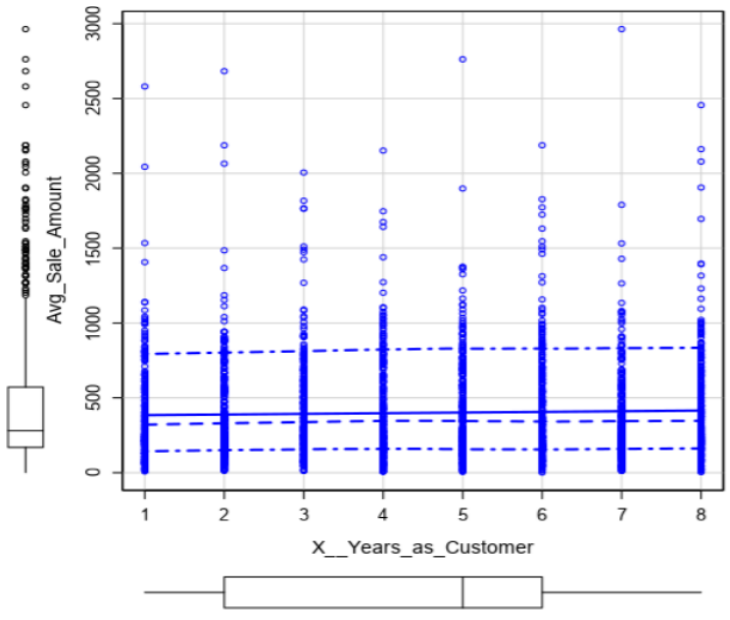
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

To validate management decision, I calculated the predicted profit will be calculated which was derived from predicted Average sales amount. Therefore, the target variable is the Average Sale Amount. I used the Scatter plot tool in Alteryx to show the relationship between variables and their correlation.

From the scatter plot blow: this shows Average Sale Amount has a linear dependence on Average number of Products purchased. For this reason, Average number of Products purchased was selected as continuous predictor variable for building the linear regression model.



Also from the scatterplot below, Average Sale Amount did not show a clear linear dependence on Years as Customer. Therefore, Years as Customer was not selected as continuous predictor variable for building the linear regression model.



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.

From the statistical results shown below, the predictor variables that would be fit for the regression model are Customer_Segment and Avg_Num_Products_Purchased. Here, the p-values are far below the accepted threshold p-value (0.05). Also the R-squared value (0.8369) and Adjusted R-squared value (0.8366) shows a strong correlation.

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

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

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

Y = 303.46 + 66.98B1 – 149.36B2 + 281.84B3 – 245.42B4
Where: Y= Avg_Sale_Amount
B1= Ave_Num_Products_Purchased
B2=Customer_Segment-Loyalty Club Only
B3= Customer_Segment-Loyalty Club and Credit Card
B4=Customer_Segment-Store Mailing List

Step 3: Presentation/Visualization

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

Yes, the company should send the catalog to these customers since the expected profit exceeds $10,000 which the Management insists on.

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

First, I did a scatter plot for the numerical variables to check the relationships between them. Only two had a linear relationship; Avg_Sales_Amount and Avg_Num_Products_Purchased. To check if they are statistically significant, I built a linear regression model. From the results, Customer_Segment and Avr_Num_Products_Purchased has their p-values far below the accepted threshold p-value (0.05) which is good. Then in Alteryx, I used a score tool to predict sales for each new customer. Then, I multiplied the predicted sales times the probability (score_yes data) since this shows the probability for a customer to buy if they received the store catalog. This gives me the expected revenue, which I multiplied by 50% (per management's average gross margin (price - cost) to get the expected profit. Finally, I subtracted the cost of mailing catalogs to the new 250 customers, which gave me the final predicted profit for these new customers.

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

Expected Profit = (Sum of expected revenue * Gross Margin) – (Cost of Catalog * 250)
= $21,987.44