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?
   To determine if the company will send the catalog to the 250 new customers based on the expected profits from them if it exceeded 10,000$, in other respects they won't send it.

2. What data is needed to inform those decisions?
   predicting the expected revenue by building the linear regression model by using the customer dataset which includes customer segment, average number of products purchased, score yes, and margin and cost of catalog and apply it on mailing list dataset.

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.

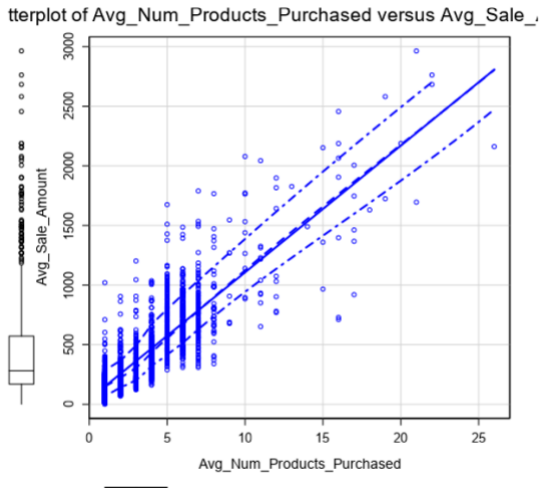
At the minimum, answer these questions:

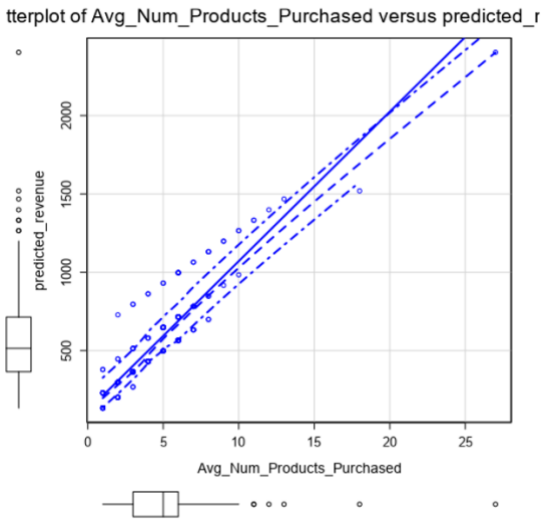
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

After understanding the data, I have eliminated Name, Customer ID, Address, State, Responded_to_Last_Catalog. Because personal information is not necessary in this model, state is fixed "Co", and Responded_to_Last_Catalog cannot be used because it doesn't exist in mailing list dataset

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28769501.17	3	507.92	< 2.2e-16	***
Avg_Num_Products_Purchased	36978219.27	1	1958.55	< 2.2e-16	***
X._Years_as_Customer	69132.67	1	3.66	0.0558	.
Residuals	44727736.4	2369			





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.

As shown below, all variables have a P value less than 0.05 and especially the predictor variables. R-squared is 0.8371 which indicates a strong relation. As a result, the linear regression model is a good model and statically significant.

Record | Report

Report for Linear Model Linear_Regression

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.04	-68.42	-1.69	71.58	976.10

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	313.76	11.861	26.454	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.11	8.969	-16.625	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.62	11.910	23.729	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.48	9.762	-25.146	< 2.2e-16 ***
Avg_Num_Products_Purchased	67.02	1.514	44.255	< 2.2e-16 ***
X._Years_as_Customer	-2.34	1.223	-1.914	0.0558 .

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

Residual standard error: 137.41 on 2369 degrees of freedom
Multiple R-squared: 0.8371, Adjusted R-Squared: 0.8368
F-statistic: 2435 on 5 and 2369 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	28769501.17	3	507.92	< 2.2e-16 ***
Avg_Num_Products_Purchased	36978219.27	1	1958.55	< 2.2e-16 ***
X._Years_as_Customer	69132.67	1	3.66	0.0558 .
Residuals	44727736.4	2369		

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 (ex: 1.28)

Avg_Sale_Amount = 303.76 – 149.11 x (Loyalty Club Only) + 281.62 x (Loyalty Club and Credit Card) – 245.48 x (If Type: Store Mailing List) + 0 x (Credit Card Only) + 67.02 x (Avg_Num_Products_Purchased) + -2.34 (X._Years_as_Customer)

Important: The regression equation should be in the form:

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

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (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.

Step 3: Presentation/Visualization

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

At the minimum, answer these questions:

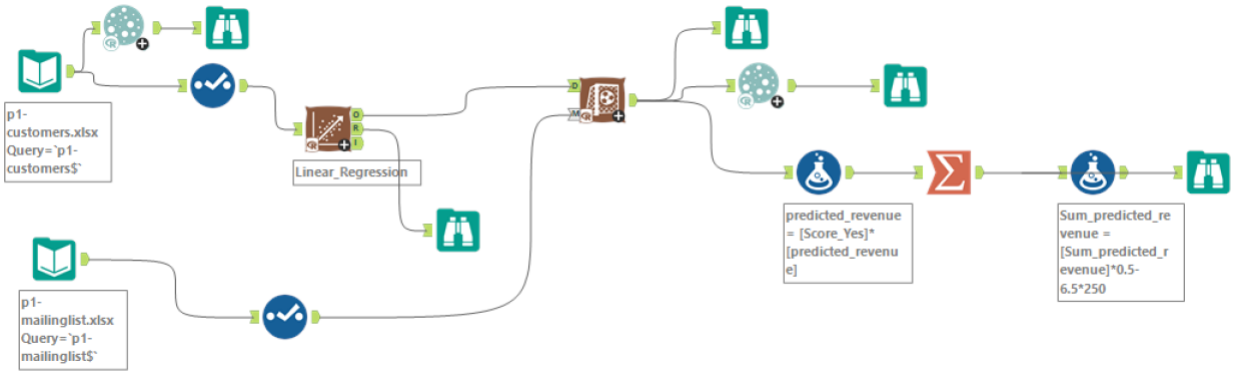
1. What is your recommendation? Should the company send the catalog to these 250 customers?

The company should send the catalog to these 250 new customers because the expected profit is $21,987.44, which is higher than $10,000

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

By Using linear regression model, the expected revenue from each customer is determined by multiplying expected sale amount with Score_Yes value, then having the sum of the amount and multiply it by gross margin of 50% - 6.50$ (catalogue price) and multiply it by 250 (new customers).

Process workflow:



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

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

References:

[1] https://github.com/kksieski/udacity_pand/blob/master/all_submissions/FINAL_kacper.ksieski.project1.pdf