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?
   Whether to send out a mail-order catalog or not, based on expected profit (greater than 10,000).

2. What data is needed to inform those decisions?
   Historical data of previous customers. One data set for predicting and the other one for validating results of prediction.

Step 2: Analysis, Modeling, and Validation

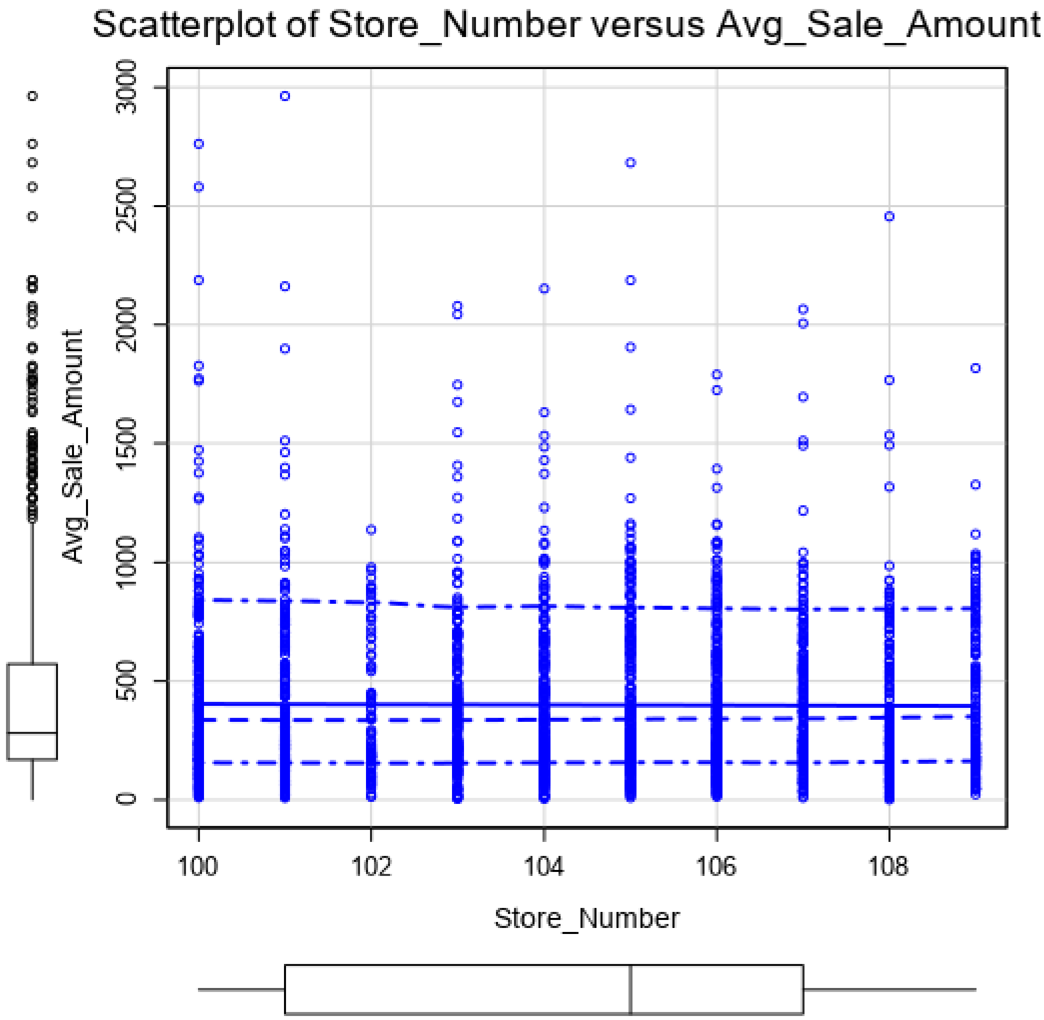
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

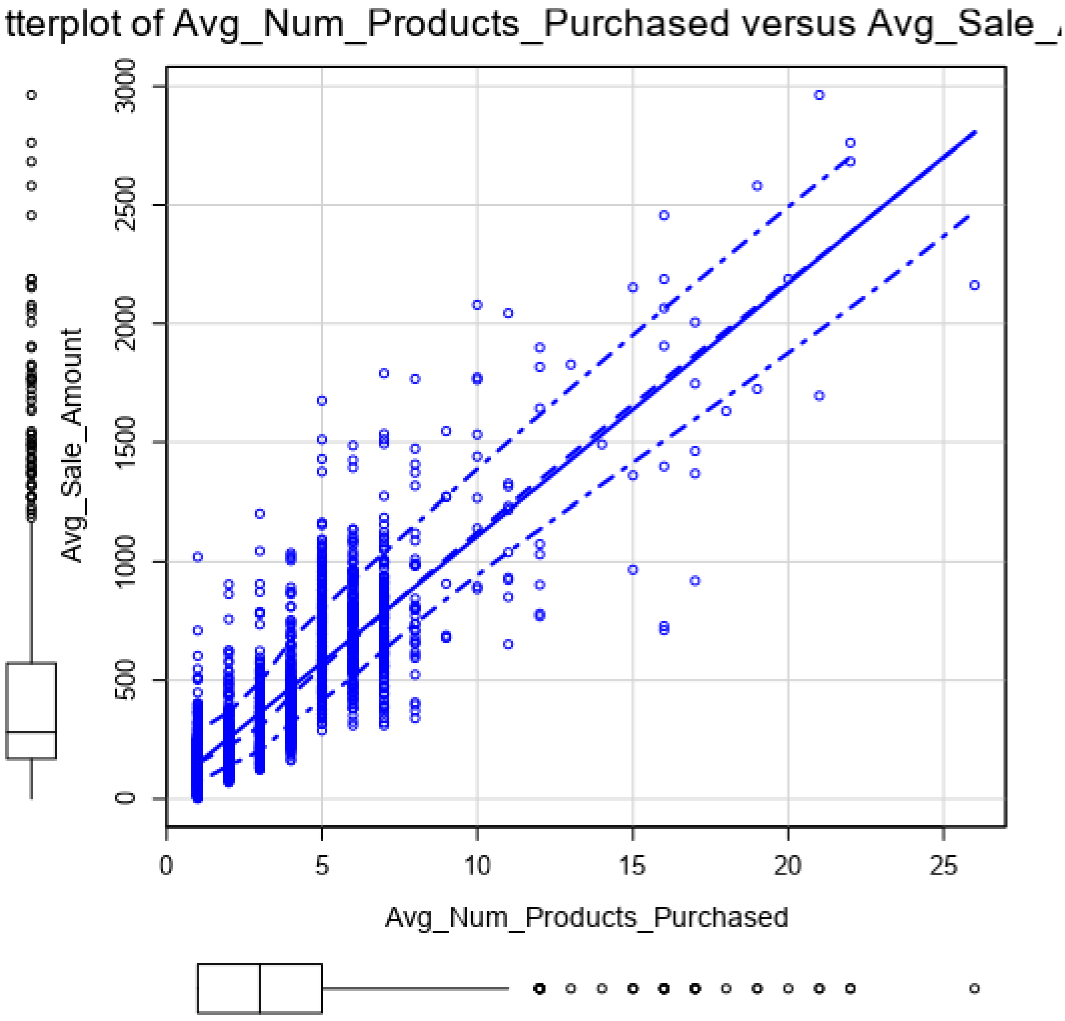
At the minimum, answer these questions:

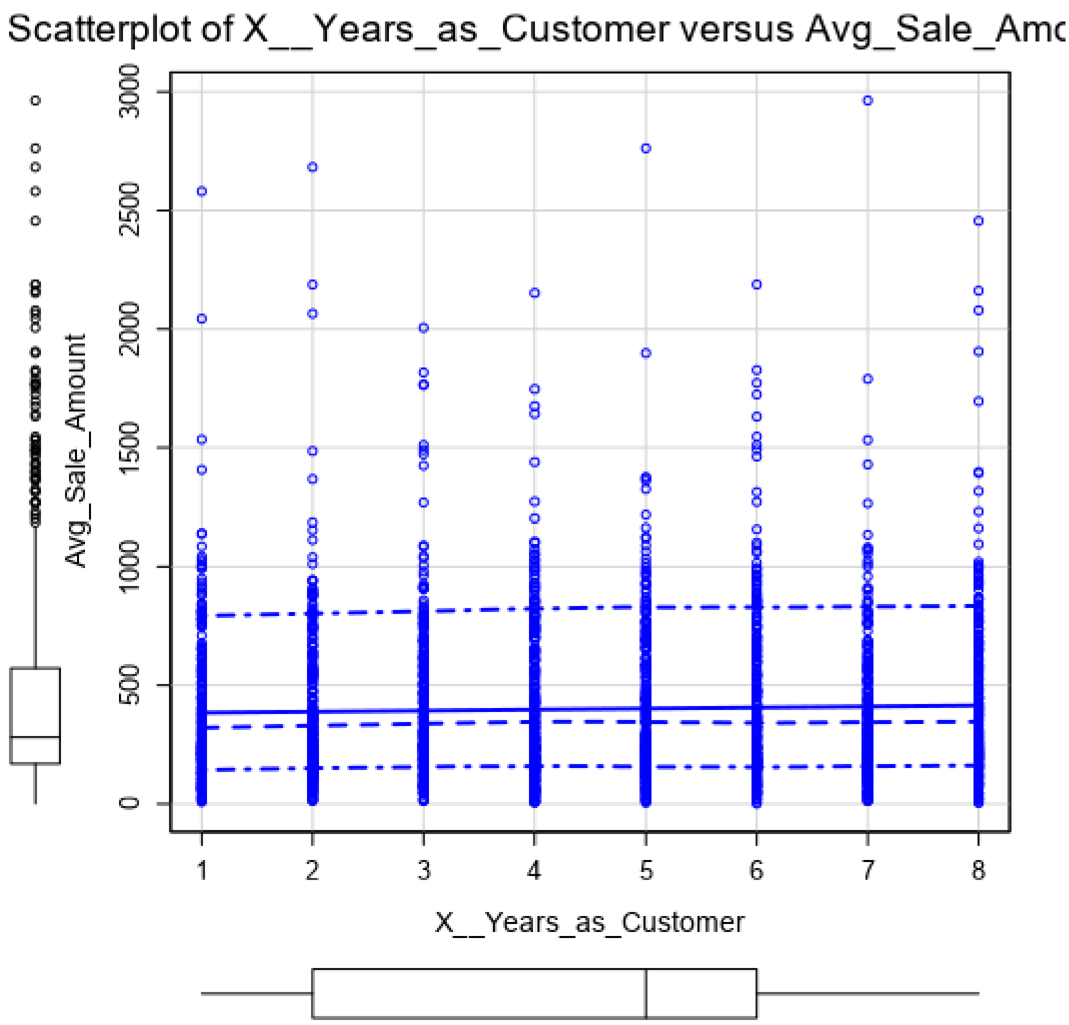
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

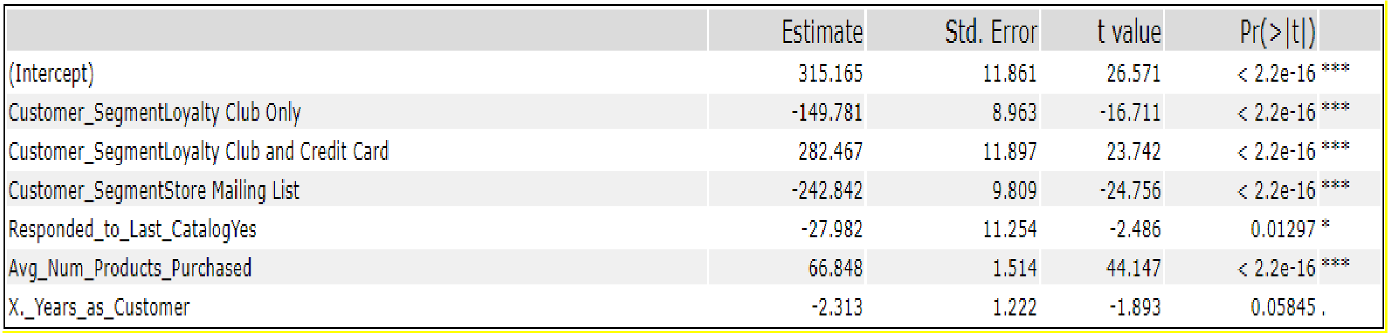
- My target variable was Avg_Sale_Amount because this what I have to predict in my project.
- I disregarded the customer information variables: Name, Customer_ID, Address, City, State, ZIP and Store_Number.
- I performed Linear Regression on Avg_Num_Products_Purchased, Responded_to_Last_Catalog, X_Years_as_Customer and Customer_Segment
- I chose Avg_Num_Products_Purchased and Customer_Segment based on p-value greater than 0.05.



Scatterplot of Store_Number versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
100
102
104
106
108
Store_Number

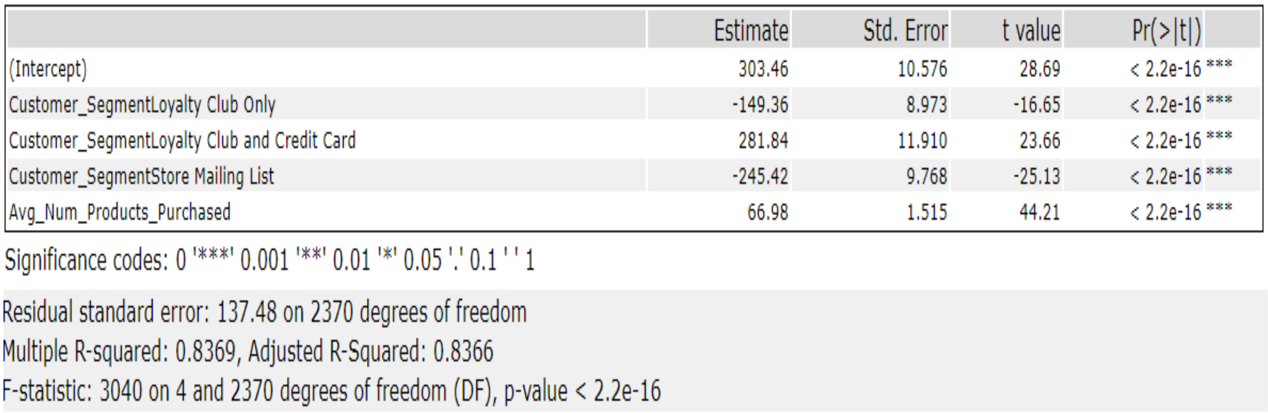

tterplot of Avg_Num_Products_Purchased versus Avg_Sale_
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
Avg_Num_Products_Purchased





	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	315.165	11.861	26.571	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.781	8.963	-16.711	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.467	11.897	23.742	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-242.842	9.809	-24.756	< 2.2e-16 ***
Responded_to_Last_CatalogYes	-27.982	11.254	-2.486	0.01297 *
Avg_Num_Products_Purchased	66.848	1.514	44.147	< 2.2e-16 ***
X._Years_as_Customer	-2.313	1.222	-1.893	0.05845 .

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.



	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

With Avg_Num_Products_Purchased and Customer_Segment as predictor variables for predicting Avg_Sale_Amount, R^2 comes out 84%(predictive power of predictor variables). Also, the p-values are much smaller than 0.05(high significance of predictor variables). Both R^2 and p-values are suggesting a good model to be used for decision making.

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:
*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.

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

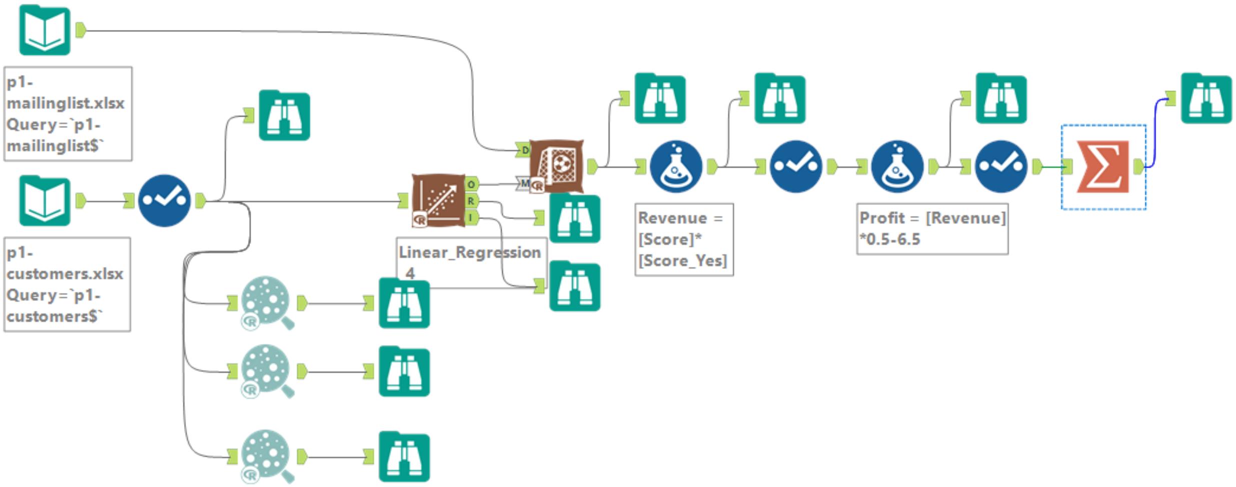
1. What is your recommendation? Should the company send the catalog to these 250 customers?

Yes, Company should send the catalog to these 250 customers

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

Expected profit(21,987) from the Alteryx model came out to be much higher than the targeted profit(10,000). So, it will be beneficial for manager to send out a mail-order catalog to customers.

Also, p-values of variables are much greater than 0.05, which shows model can be well generalized.


p1-mailinglist.xlsx Query=`p1-mailinglist$`
p1-customers.xlsx Query=`p1-customers$`
Linear_Regression 4
D
M
O
R
I
Revenue = [Score]* [Score_Yes]
Profit = [Revenue] *0.5-6.5

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

My predicted profits came out to be: 21,987.43

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.