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?

A decision has to be made weather we should send product catalog to the 250 new customers. The catalog will be send to those new customers only if the expected profit contribution exceeds$10,000.

2. What data is needed to inform those decisions?

Some data needed to support the analysis includes Customer Segments, average of number of products purchased, response to catalog, average sales, costs of printing and distributing.

Step 2: Analysis, Modeling, and Validation

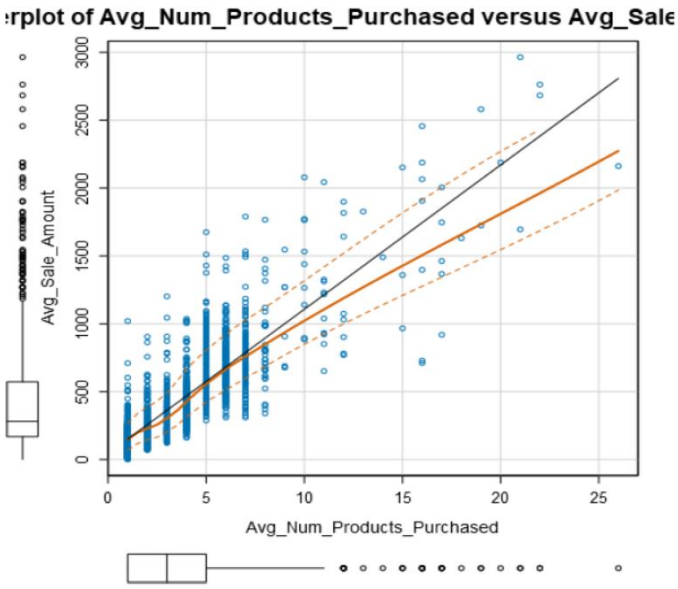
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

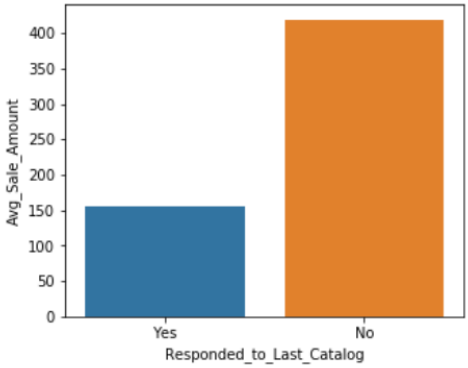
Important: Use the p1-customers.xlsx to train your linear model.

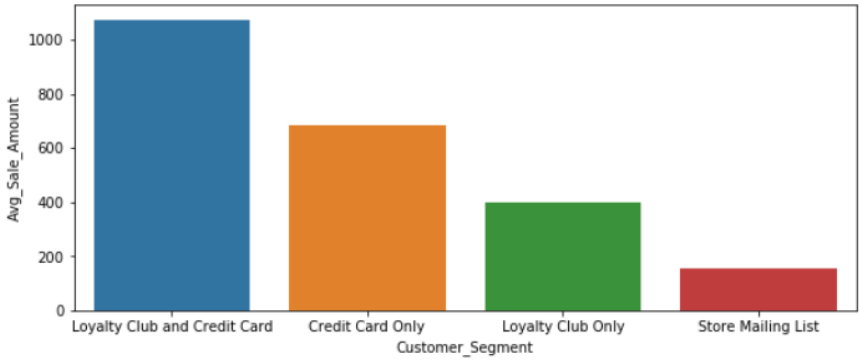
At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

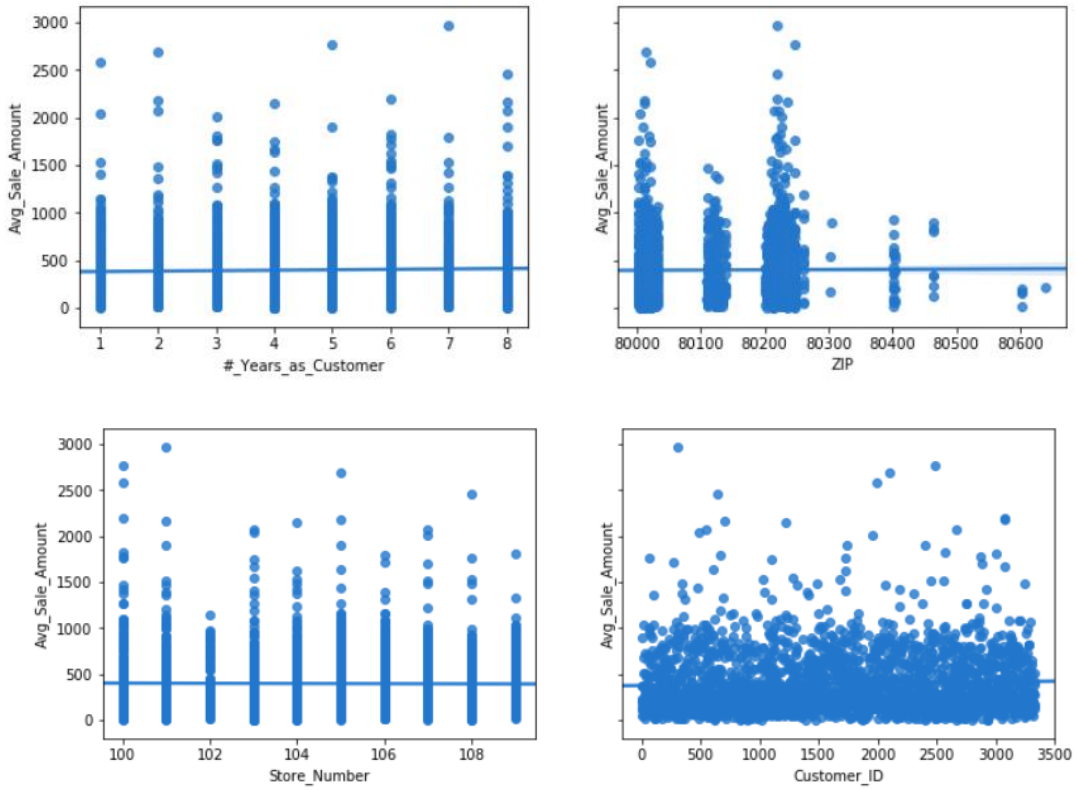
To select the predictor variables for a regression model, we first take look at the correlation between each variable against the target variable- average sale amount in pairs.







It seems that the Avg_Num_Products_Purchased has a strong positive linear relation with Avg_Sales. And the Customer_Segment seems to correlate to Ave_Sales as well. However, other predictor variables did not show a correlation from the following graphs.



Then a linear coefficient test is made to check if there is a significant correlation between a variable and the predicted outcome. From the coefficient table, some predictor variables are statistically significant (p value <0.05). Thus the valid predictor variables are Customer_Segment(three types:Loyalty Club Only, Loyalty Club and Credit Car, Store Mailing List), Ave_Num_Products_Purchased

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	327.0135	13.976	23.39859	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.0280	8.982	-16.70353	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	283.6100	11.916	23.80030	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.9831	9.827	-24.72667	< 2.2e-16	***
Store_Number101	-5.5294	11.235	-0.49214	0.62267	
Store_Number102	-8.6893	16.743	-0.51897	0.60383	
Store_Number103	-4.4862	11.903	-0.37688	0.70629	
Store_Number104	-21.2748	11.303	-1.88229	0.05992	.
Store_Number105	-20.9124	10.951	-1.90956	0.05631	.
Store_Number106	-18.1956	11.175	-1.62823	0.10361	
Store_Number107	-14.7112	11.899	-1.23631	0.21647	
Store_Number108	-12.0088	12.158	-0.98773	0.32339	
Store_Number109	-0.1426	13.024	-0.01095	0.99127	
Responded_to_Last_CatalogYes	-29.1449	11.277	-2.58455	0.00981	**
Avg_Num_Products_Purchased	66.7485	1.517	43.99951	< 2.2e-16	***
X._Years_as_Customer	-2.3737	1.224	-1.93886	0.05264	.

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

Residual standard error: 137.28 on 2359 degrees of freedom
Multiple R-squared: 0.8381, Adjusted R-Squared: 0.8371
F-statistic: 814.2 on 15 and 2359 DF, p-value: < 2.2e-16

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 the significant predictor variables, we did the linear regression and produced the following results. The adjusted R-squared is 0.8371, which relatively high and can explain some parts of the variation in the average sales. And each p value is far less than 0.05, which shows a significance of the coefficient. Therefore the linear model is a good model.

Residuals:

Min	1Q	Median	3Q	Max
-662.60	-67.17	-2.96	69.88	973.90

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	305.00	10.582	28.823	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.03	8.967	-16.732	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.69	11.897	23.678	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.76	9.815	-24.734	< 2.2e-16	***
Responded_to_Last_CatalogYes	-28.17	11.259	-2.502	0.01241	*
Avg_Num_Products_Purchased	66.81	1.515	44.099	< 2.2e-16	***

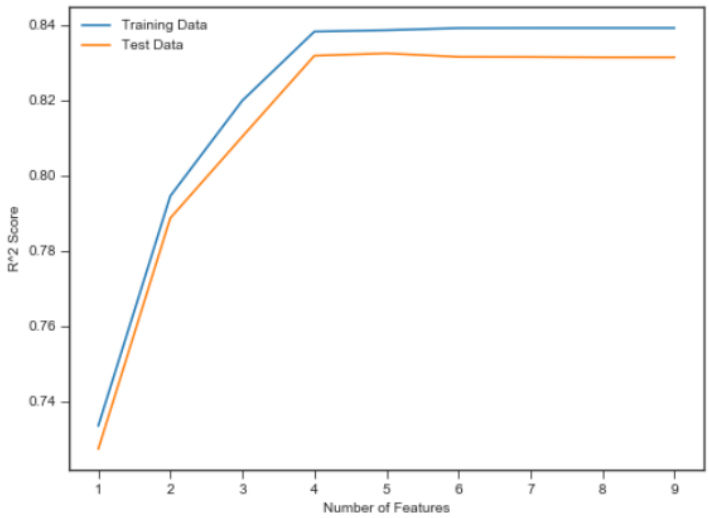
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

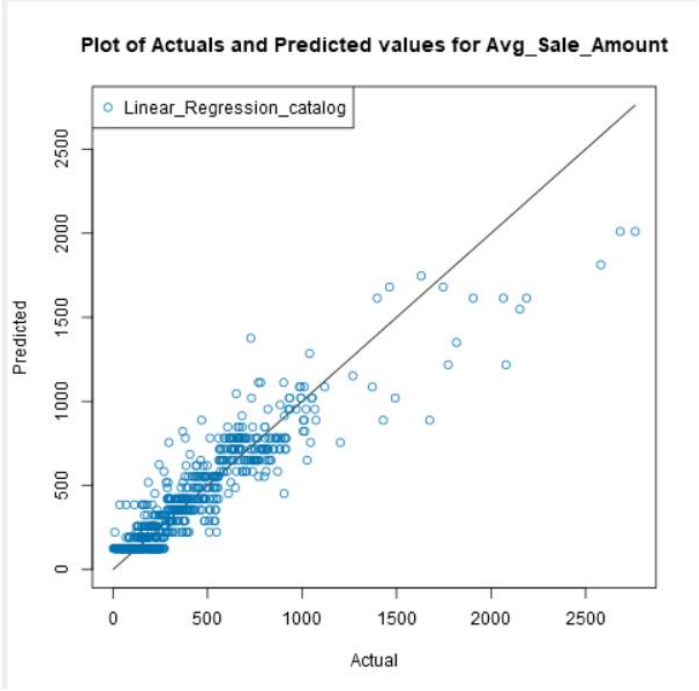
Residual standard error: 137.33 on 2369 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.837
F-statistic: 2438 on 5 and 2369 DF, p-value: < 2.2e-16

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)
   From the coefficent table ,we can find that the three customer segment factors and average number of products purchased are significant correlated to the target value.

   To test the model validation, we select 70% of sample from the dataset as the training set and 30% as the validation set. When we check the number of features and the R-square values, we find that the differences between R-square increase as the number of features increases. The gap between two lines strongly increases after the number of value turns to 5. So we decide to use 4 predictor variables:

   Customer_Segment Loyalty Club Only
   Customer_Segment Loyalty Club and Credit Card
   Customer_Segment Stoe Mailing List
   Avg_Num_Products_Purchased





The linear model is as following:

Avg_Sales = 305+

-150.03* Customer_Segment (IF type: Loyalty Club Only)

+281.69* Customer_Segment (IF type: Loyalty Club and Credit Card)

-242.76* Customer_Segment (IF type: Store Mailing List)

+66.81*Avg_Product

(Score_Yes = Responded_to_Last_CatalogYes\

Avg_Product = Ave_Num_Products_Purchased)

Step 3: Presentation/Visualization

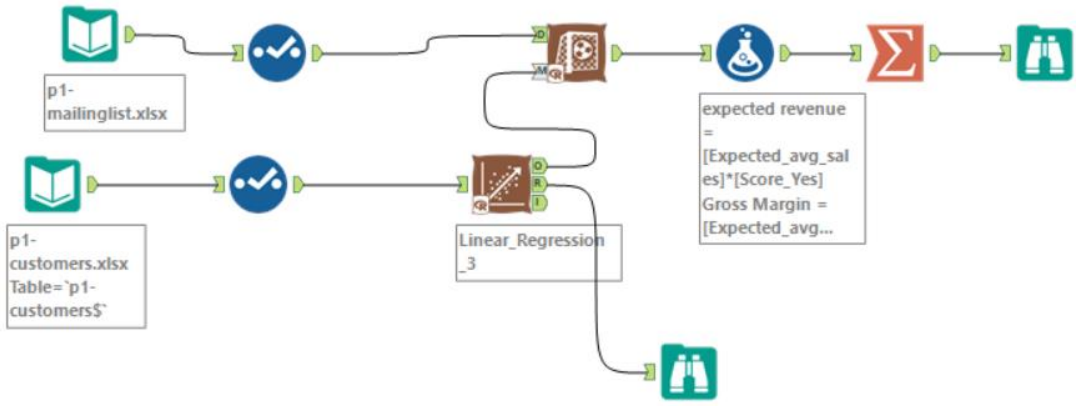
Use your model results to provide a recommendation. (500 word limit)

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 customers.

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



The linear model is used to predict the average sales amount based on a batch of 250 new customers. After setting the datatypes, we used the score tool to produce the expected average sales for each customer. Then we built a formula to calculate the gross margin and expected revenue

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

Sum_Expected Profit = Sum_Expected_avg_sales* Gross Margin – 250*6.5=$21987.44

The expected profit from the new catalog is $21987.44.