Project 1: Predicting Catalog Demand

ansjayan@msn.com

Step 1: Business and Data Understanding

The project is analyzing a mail-order catalog business.

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   The company sells high-end home goods. The company is planning to send out catalog to 250 new customers. Before doing that, the company want to know, how much money she can earn from sending catalog to new customers. Only if the profit exceeds $10,000 the company sends catalog. The task is to predict the profit earned from sending out catalog. Is it exceeds $10,000 ?.

2. What data is needed to inform those decisions?

   Two excel files are provided for the project, p1-customers and p1-mailinglist. We need to build a linear regression model using the past purchase details of customers. Using the model parameters, we can estimate the profit from the new customers details. We have to use customer details like: - segments, average sale amount, average number of products purchased.

Step 2: Analysis, Modeling, and Validation

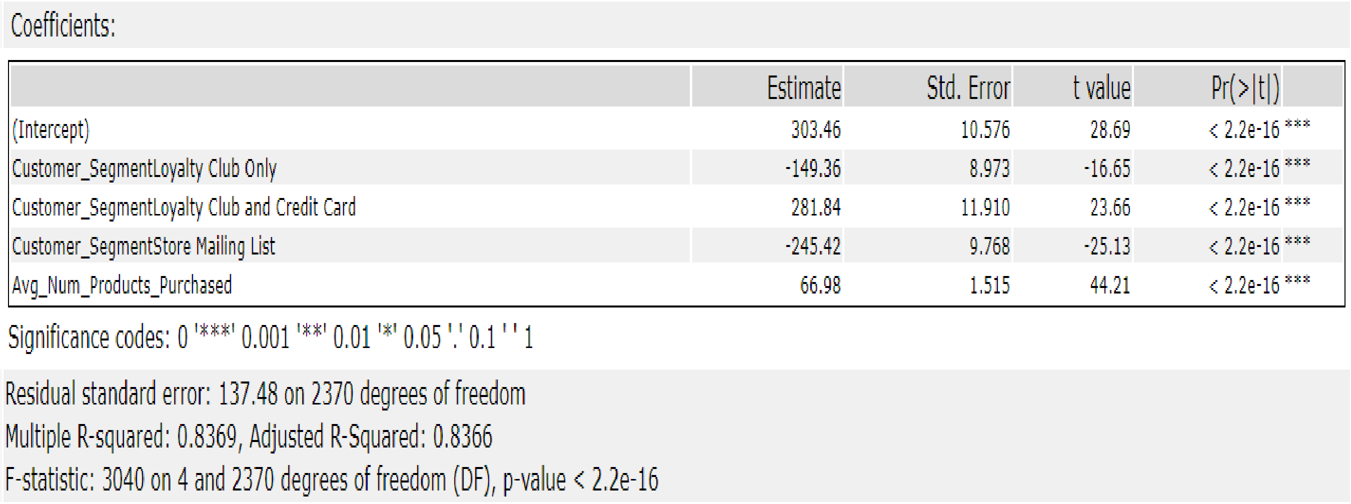
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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.

   First fit the model with all the variables and then omit variables with no significance. Use only variables with p values <= 0.05. Significance code * above can be used. Redo the fit by removing less significant variable and check change in the R and Adjusted R Squared. Final coefficients are given below.



Coefficients:

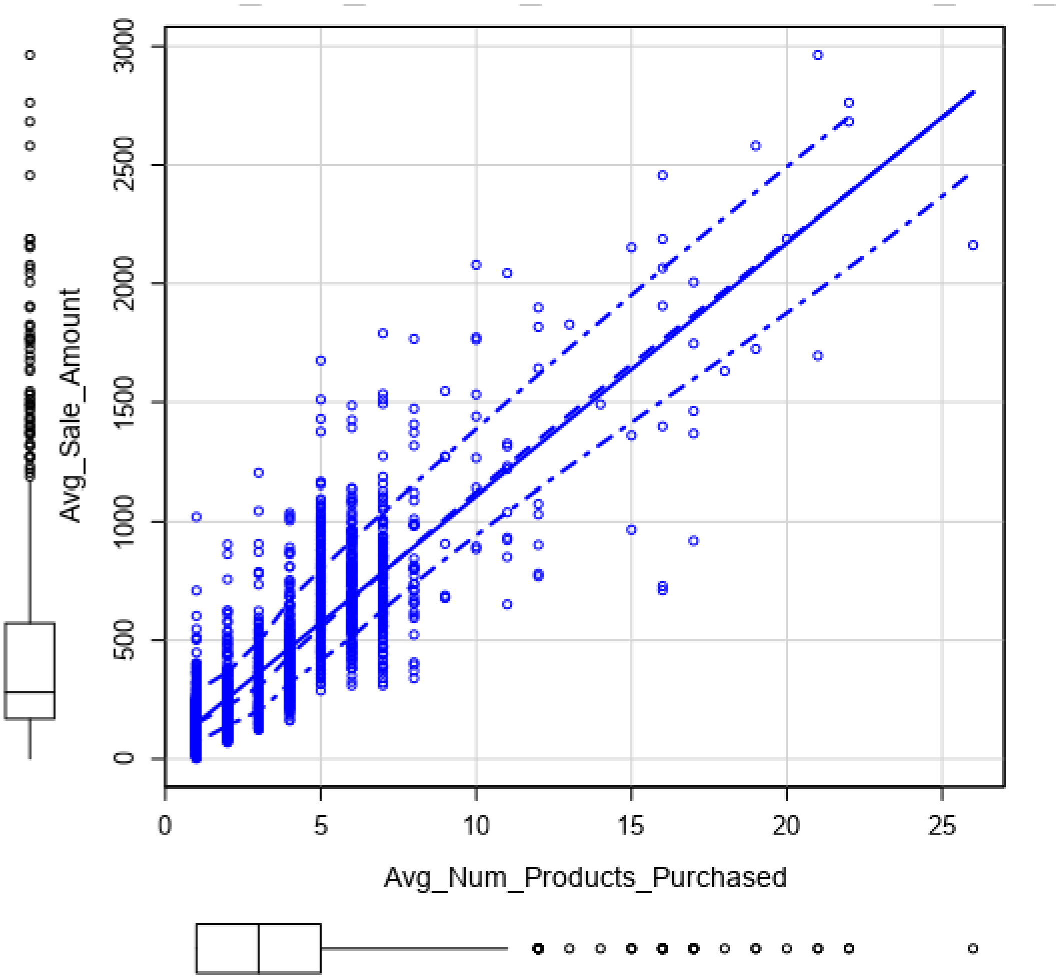
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
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

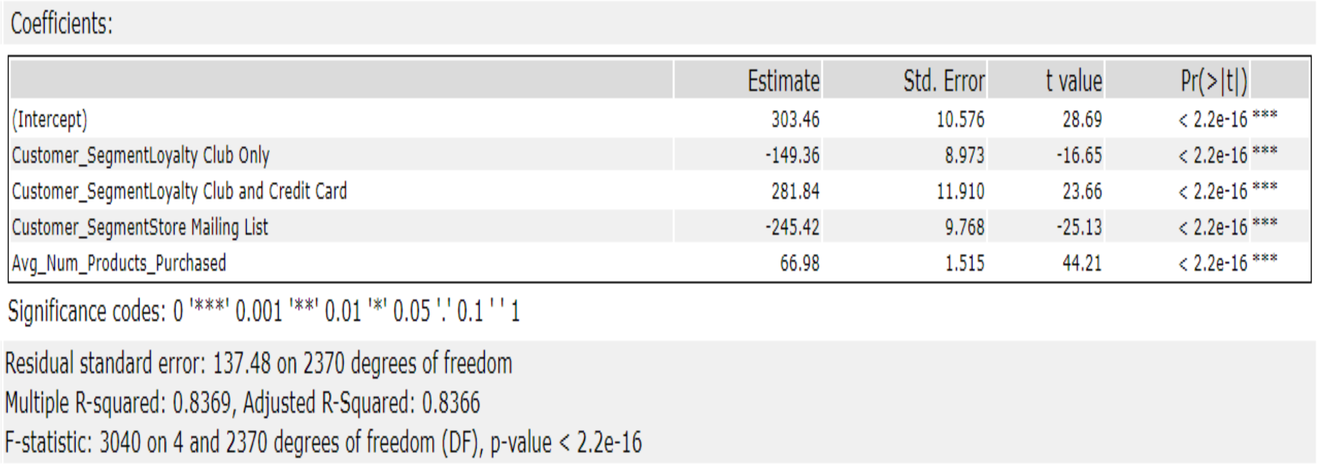
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366

F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16



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.

All variables p values are less than 0.05, and significance code is ***. R squared value is 0.8369 and Adjusted R-Squared is 0.8366 are high values. Those are the required conditions for a good model. So, regression model is good.



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

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:

*Avg_Sale_Amount = 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: Store Mailing List)*
*+ 0 * (If Customer_Segment: Credit Card Only)*

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.

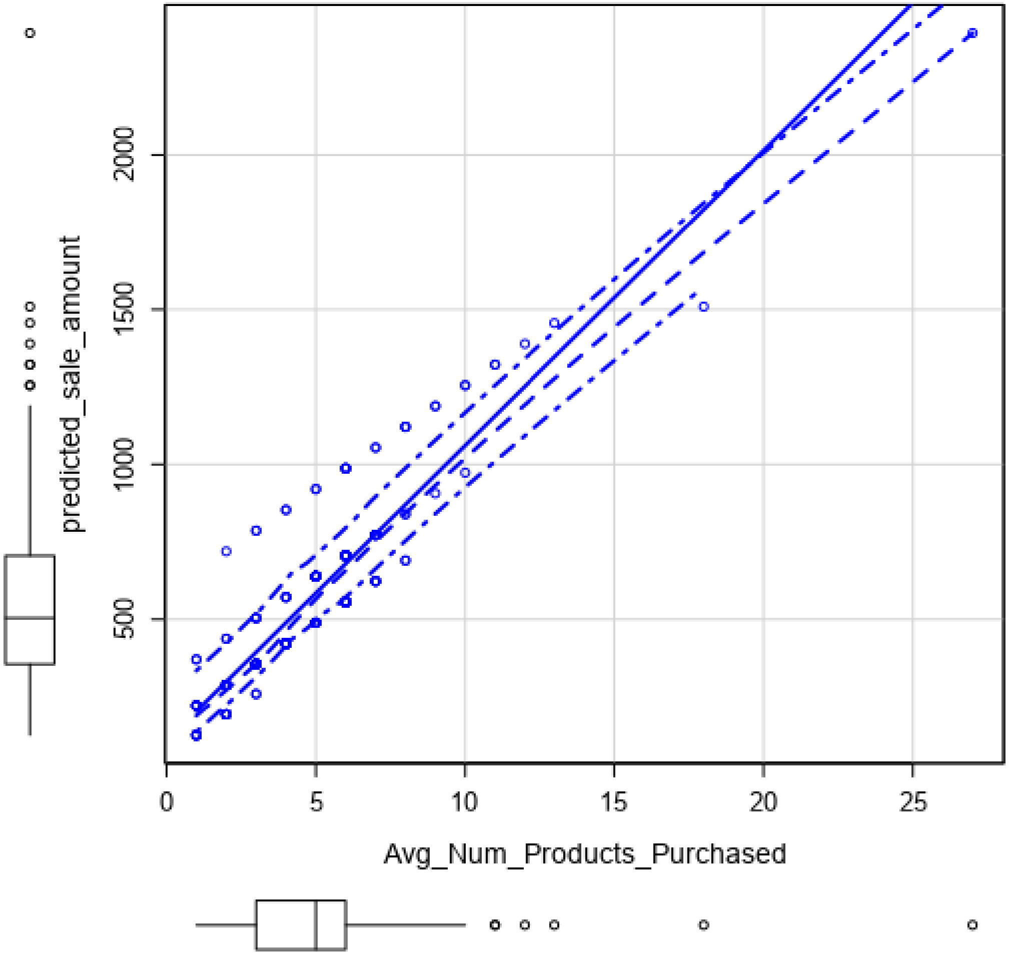
Step 3: Presentation/Visualization

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?

Yes, the company should send the catalog to these 250 customers.

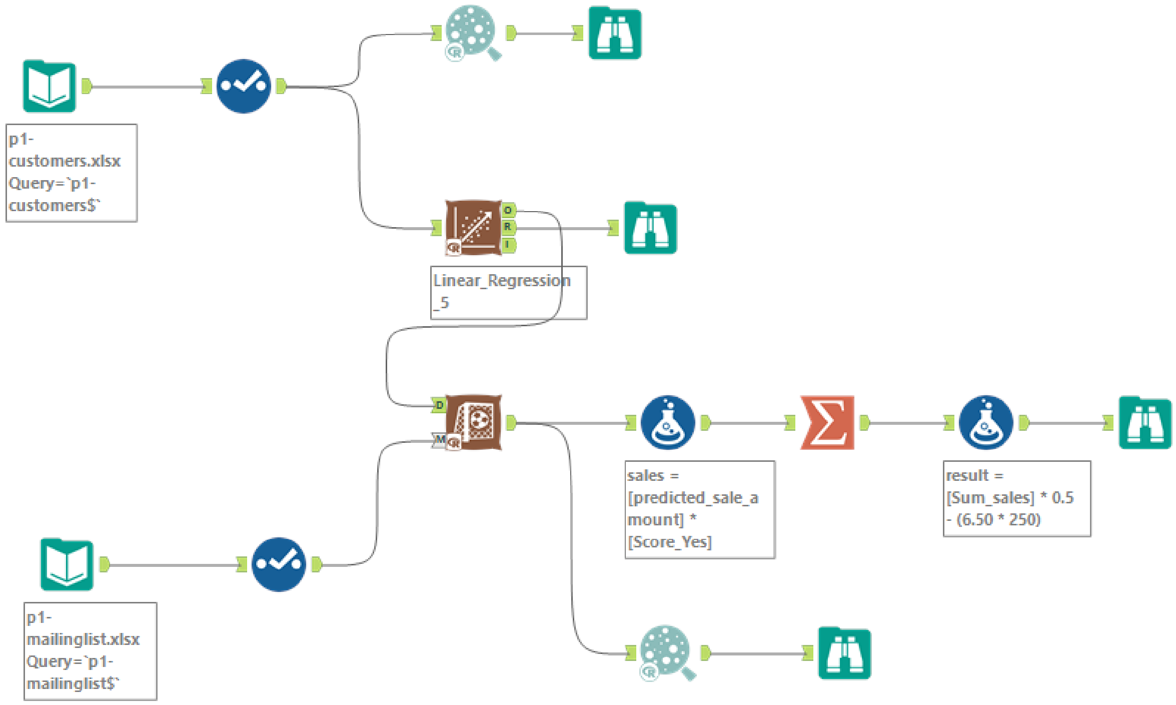


2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

Using the regression model and p1-mailinglist.xlsx, i.e. the details of the 250 customers, the predicted sale amount is calculated. Which is then multiplied with the Score_Yes variable and summed. The summed value is then multiplied with 0.5 and subtracted from (6.50 * 250).

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 is $21987.4357.



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.