Project 1: Predicting Catalog Demand

The Business Problem

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Details

The costs of printing and distributing is $6.50 per catalog.
The average gross margin (price - cost) on all products sold through the catalog is 50%.
Make sure to multiply your revenue by the gross margin first before you subtract out the $6.50 cost when calculating your profit.
Write a short report with your recommendations outlining your reasons why the company should go with your recommendations to your manager.

Step 1: Business and Data Understanding

1. What decisions needs to be made?

If it is profitable to send out catalogues to the 250 new customers.

2. What data is needed to inform those decisions?

Information of existing customers, their segment and the average product number of they purchased. Same data points of the new customers from the mailing list.

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.

First, I did a Linear regression with all parameters and I observed the statistical significance by looking at the Pr value in the output (red box below), I noted that Customer_ID, ZIP, Store_Number and Years_as_Customer are above 0.05, thus they are insignificant. The parameters which are significant and below 0.05, are the Customer_Segment and Avg_Num_Products_Purchased, thus I took a further look on their scatterplots.

Report for Linear Model Linear_Regression_5

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Customer_ID + ZIP + Store_Number + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-667.40	-67.94	-2.06	71.85	969.04

Coefficients:

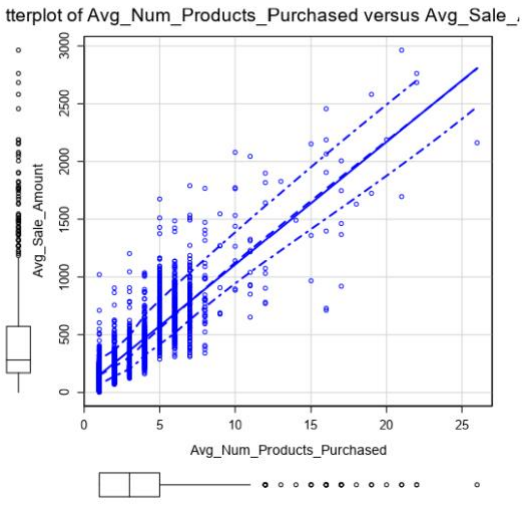
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.379e+03	2.149e+03	-0.64	0.52118	
Customer_SegmentLoyalty Club Only	-1.497e+02	8.980e+00	-16.6	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.824e+02	1.193e+01	23.6	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.459e+02	9.774e+00	-25.1	< 2.2e-16	***
Customer_ID	-1.373e-03	2.941e-03	-0.4	0.64063	
ZIP	2.248e-02	2.660e-02	0.8	0.39814	
Store_Number	-1.011e+00	1.007e+00	-1.0	0.31539	
Avg_Num_Products_Purchased	6.700e+01	1.517e+00	44.1	< 2.2e-16	***
X._Years_as_Customer	-2.345e+00	1.223e+00	-1.9	0.0554	.

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

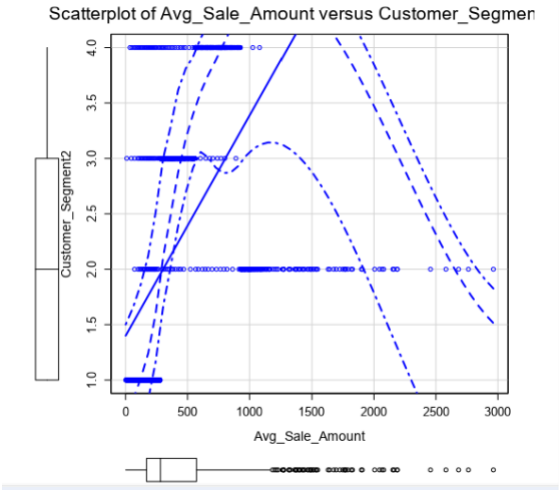
Residual standard error: 137.43 on 2366 degrees of freedom

Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8367

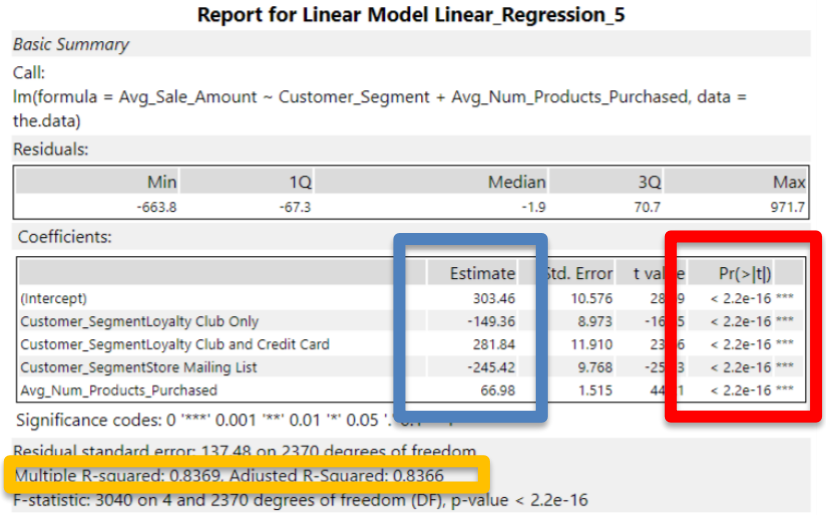
Clear correlation between Avg_Sale_Amount and Avg_Num_Products_Purchased



Clear correlation between Avg_Sale_Amount and Customer Segments, where segments are: 1 Store Mailing List, 2 Loyalty Club and Credit Card, 3 Loyalty Club Only, 4 Credit Card Only



After that I have rerun the regression only with the parameters which are statistically significant:



Report for Linear Model Linear_Regression_5

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 '.'

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 Pr values (red box above) in the above table are below 0.05, thus they are statistically significant for correlation. Since R-squared value is 0.84 (yellow box), it is high enough for ensuring the quality of the model.

3. What is the best linear regression equation based on the available data?

Based on the coefficients estimates from the table above (blue box), the linear regression equation is the following:

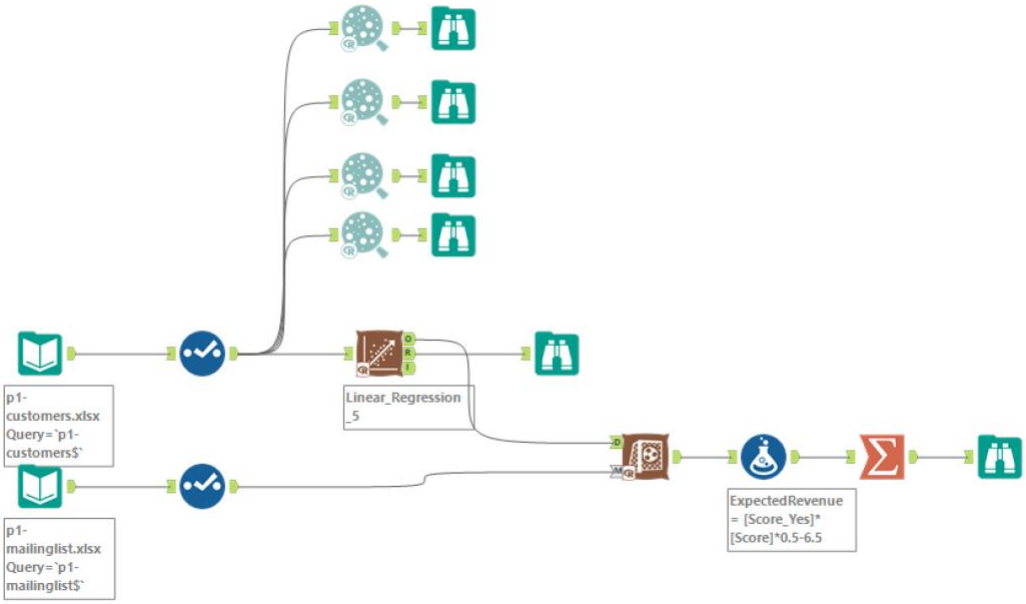
Avg_Sale_Amount = 303.46 – 149.36 (If Customer_Segment: Loyalty Club Only)
+ 281.84 * (If Customer_Segment: Loyalty Club and Credit Card)
– 245.42 (If Customer_Segment: Store Mailing List)
+ 66.98 * Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

1. What is your recommendation? Should the company send the catalog to these 250 customers?
   It is profitable for the company to send the catalogues for the new customers.

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

I used the previously described linear regression results from above, based on the Alteryx calculations, which then I feed to a Score calculation, together with the mailinglist.xlsx data, to predict the expected revenue per customer. Which then is scaled with 0.5 (average gross margin is 50%) and the price of catalogue 6.5 is deducted. This final expected revenue per customer is then summed up for all customers from the mailing list file to get the total expected revenue. See the Alteryx flow below.



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

The expected profit is 21987.44, which is way above the expected minimum of 10000.