Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

We need to predict the sales for 250 customers to whom the company has decided to send a sales catalog. We use the provided datasets and predict the estimated revenue and profit. Based on that projected profit a decision must be made to justify the sending of these catalogs. The cutoff profit margin is $10,000.

2. What data is needed to inform those decisions?

To make a business decision we will have to utilize the two data sets provided: -

p1-customers.xlsx - This dataset includes the information on about 2,300 prior customers. This will be used as the training data set. We will have to determine the predictor variable from this data set to predict the Target variable.

p1-mailinglist.xlsx - This dataset contains 250 customers that we need to predict sales.

Step 2: Analysis, Modeling, and Validation

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.

The predictor variable was selected based on P- value which is significant (P-value <= 0.05).

We also use the R-squared as a statistical measure of how close the data is to the fitted regression line. Here we see that the R-squared is 0.8369. In this exercise we see that the predictor variable Avg_Num_Products_Purchased has a linear relation with the Avg_Sale_Amount. We also see that the Avg_Sale_Amount has some correlation with the customer segment

Report for Linear Model Linear_Regression_10

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 '.' 0.1 ' ' 1

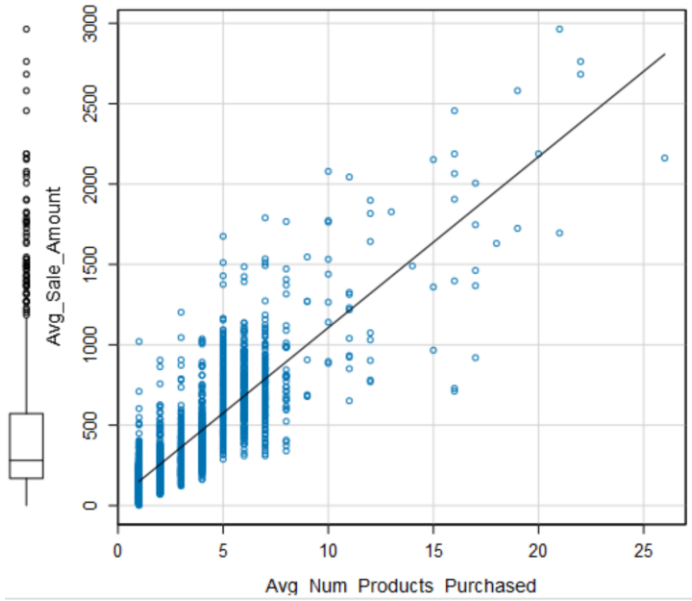
Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

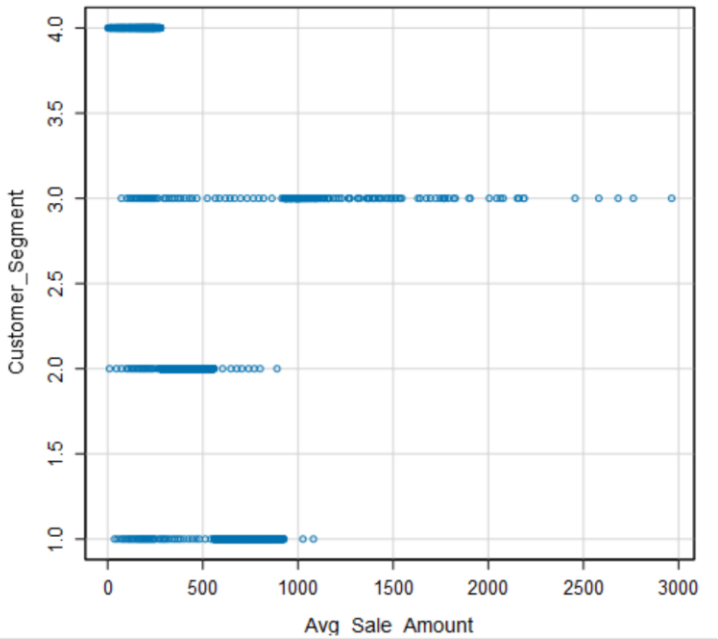
Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

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





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.

The predictor variable was selected based on P- value which is significant (P-value <= 0.05).

We also use the R-squared as a statistical measure of how close the data is to the fitted regression line. Here we see that the R-squared is 0.8369.

In this exercise we see that the predictor variable Avg_Num_Products_Purchased has a linear relation with the Avg_Sale_Amount.
We also see that the Avg_Sale_Amount has some correlation with the customer segment

Report for Linear Model Linear_Regression_10

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 '.' 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

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

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)

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 -149.36 * (Customer_SegmentLoyalty Club Only) + 281.84 * (Customer_segmentLoyalty Club and Credit Card) -245.42 * (Customer_segmentStore Mailing List) – 0 * (Credit card Only).

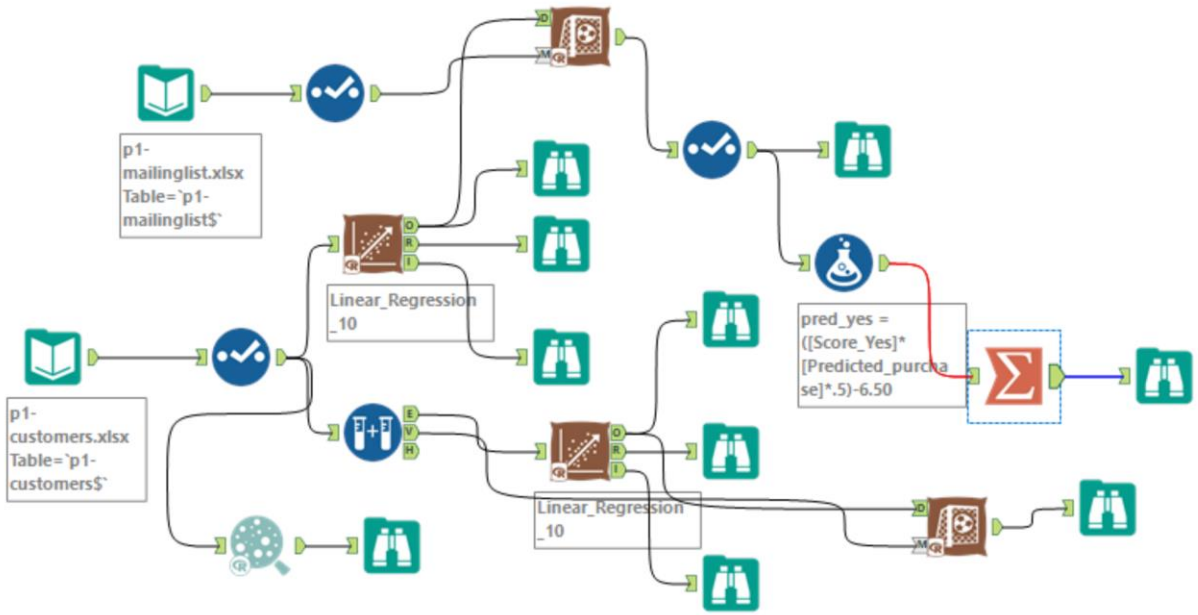
Step 3: Presentation/Visualization

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

Yes, the company should go-ahead and send out the catalogs since the expected profits exceed the $10,000 cutoff.

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

The following regression model was built in Alteryx.



The Two data sets used are the Mailinglist.xls and the Customers.xls.

We first read in the xls files using the “Inputdata” tool and then the “select” tool was used to assign the appropriate type to each filed. Once that was completed the Liner regression tool was used to input the two data streams to determine the best possible predictor variable to predict the target variable. The P and R-squared values were used to determine the best possible outcomes. The Score tool was used to make predictions based on our model. Once we determined the “Predicted_purchase” for each customer the “Formula tool” was used to calculate the profit per customer.

([Score_Yes]*[Predicted_purchase]*.5)-6.50

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

The projected profit from the new catalog is ~ $ 21987.43.

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.