Predicting Catalog Demand

Step 1: Business and Data Understanding

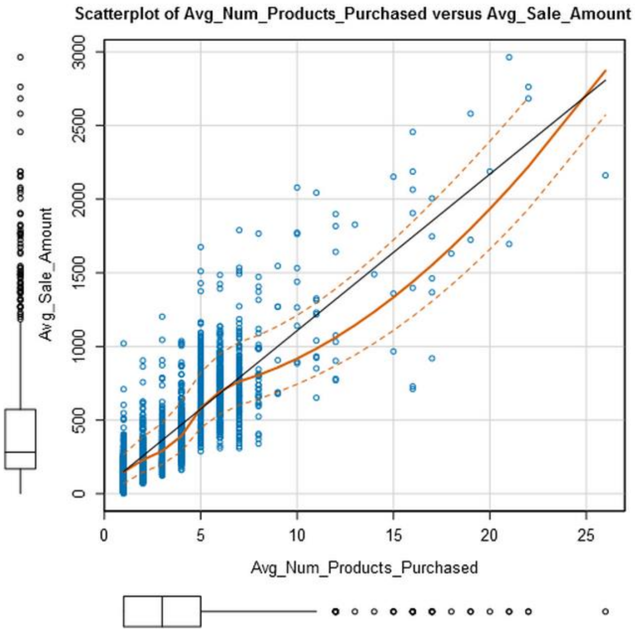
Key Decisions:

1. What decisions needs to be made?
   Should catalogs be sent to 250 new customers in the company's mailing list?
   This decision depends on whether the company can expect a profit from 250 new customers. If the expected profit is greater than $10,000, catalogs will be mailed to the new customers.

2. What data is needed to inform those decisions?
   To make the decision on whether to send catalogs to new customers we need data on the following:
   The expected profit if catalog is sent to new customer
   The Average number of products bought by each customer
   The Average Sale Amount spent by each customer for products bought through catalog
   The likelihood that the customer will buy through catalog
   The gross margin on products sold through catalog to each customer
   The cost of printing and distributing catalog to each customer

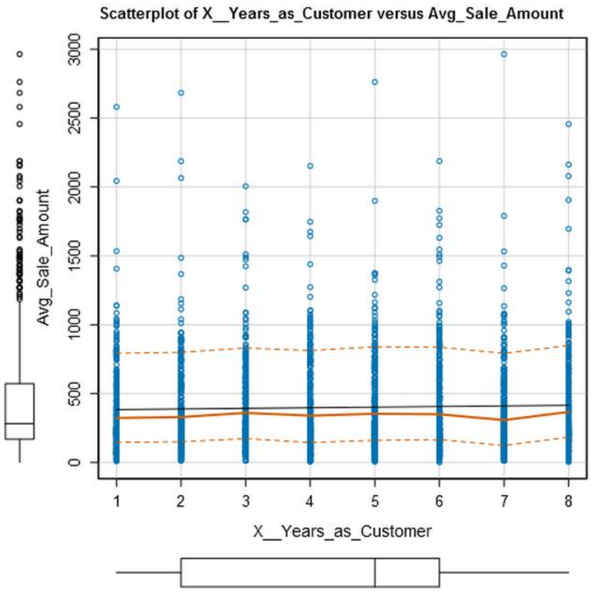
Step 2: Analysis, Modeling, and Validation

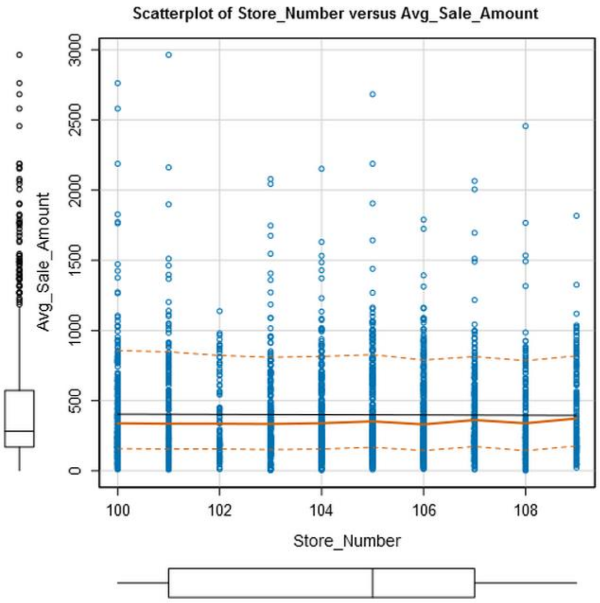
Since we need to predict the profit the Average Sale Amount is target variable. Scatterplots are used to check which numeric predictor variables that have a linear relationship with the target variable.

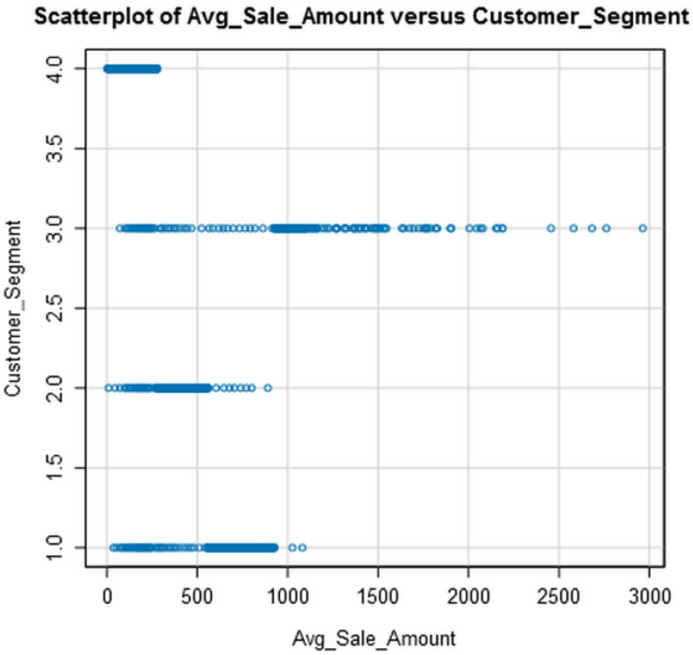


The trend lines in scatter plots below show that only the Average number of products purchased has a linear relationship with Average sale amount, and is thus a good variable to explore in the linear regression model.

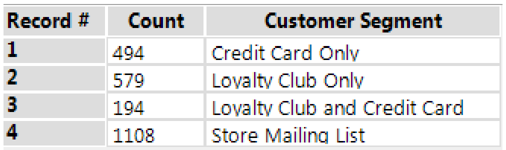
The target variable does not have a linear relationship with other continuous numeric variable as shown in the scatterplots below and need not be included in the regression model.



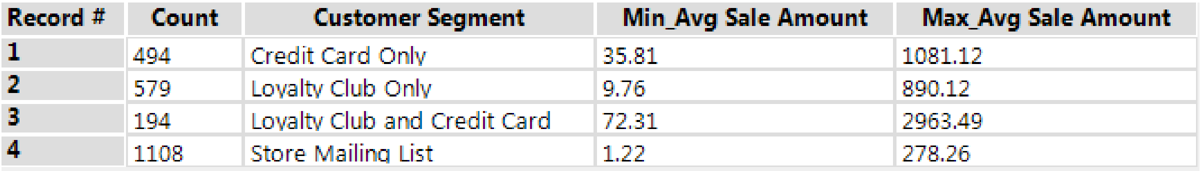




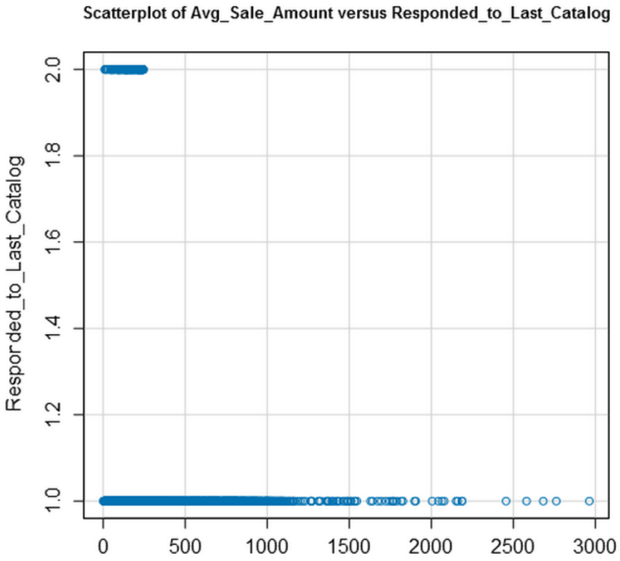
Avg Sale Amount vs Customer Segment shows 4 groups of customers.



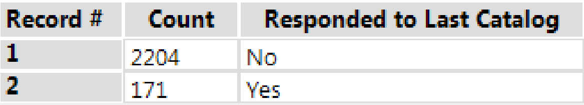
Record #	Count	Customer Segment
1	494	Credit Card Only
2	579	Loyalty Club Only
3	194	Loyalty Club and Credit Card
4	1108	Store Mailing List



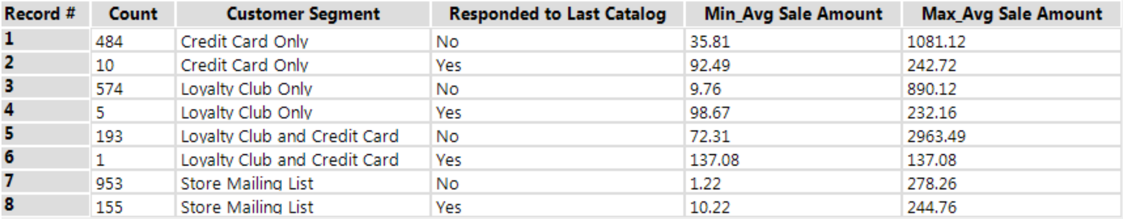
Record #	Count	Customer Segment	Min_Avg Sale Amount	Max_Avg Sale Amount
1	494	Credit Card Only	35.81	1081.12
2	579	Loyalty Club Only	9.76	890.12
3	194	Loyalty Club and Credit Card	72.31	2963.49
4	1108	Store Mailing List	1.22	278.26



Avg Sale Amount vs Responded to Last Catalog shows 2 groups of customers (Yes/No)



Record #	Count	Responded to Last Catalog
1	2204	No
2	171	Yes



Record #	Count	Customer Segment	Responded to Last Catalog	Min_Avg Sale Amount	Max_Avg Sale Amount
1	484	Credit Card Only	No	35.81	1081.12
2	10	Credit Card Only	Yes	92.49	242.72
3	574	Loyalty Club Only	No	9.76	890.12
4	5	Loyalty Club Only	Yes	98.67	232.16
5	193	Loyalty Club and Credit Card	No	72.31	2963.49
6	1	Loyalty Club and Credit Card	Yes	137.08	137.08
7	953	Store Mailing List	No	1.22	278.26
8	155	Store Mailing List	Yes	10.22	244.76

From the detailed study at the relationship between variables in the data, the numeric variable Avg_Num_Products_Purchased and the categorical variables Customer_Segment, Responded_to_Last Catalog were used as predictor variables to explore in the linear regression model for the target variable Avg_Sale_Amount.

1. Linear Models

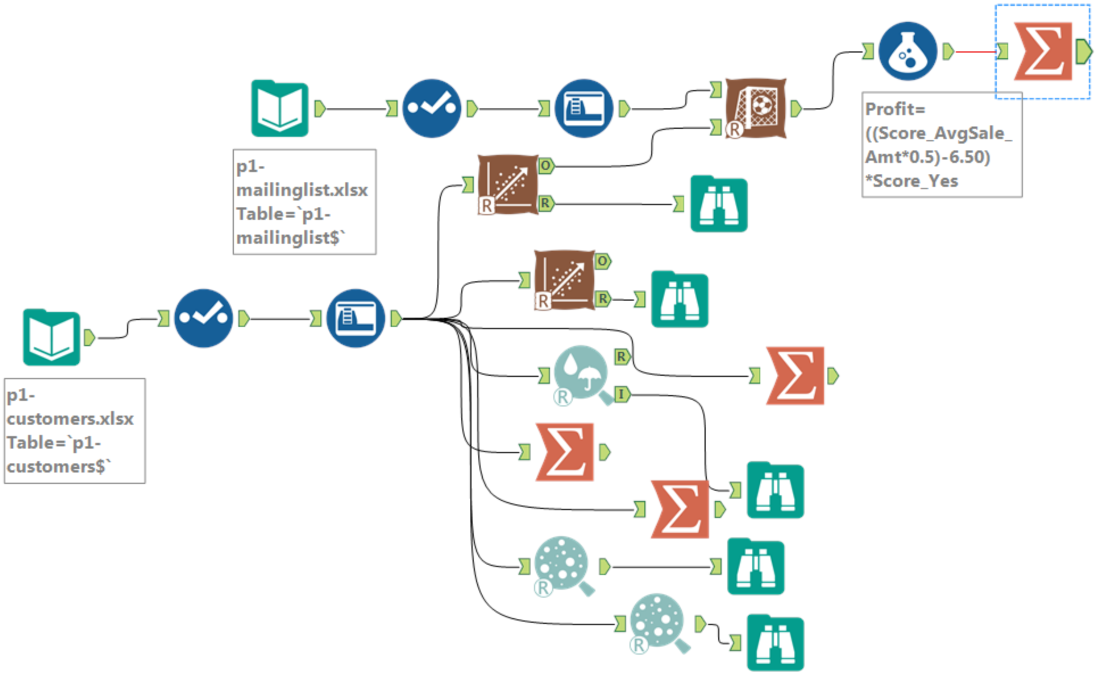
Two linear regression models were built to explore the importance of “Responded _to_last_Catalog. This variable does not change the change the Adjusted R-squared value, it is dropped from the model.

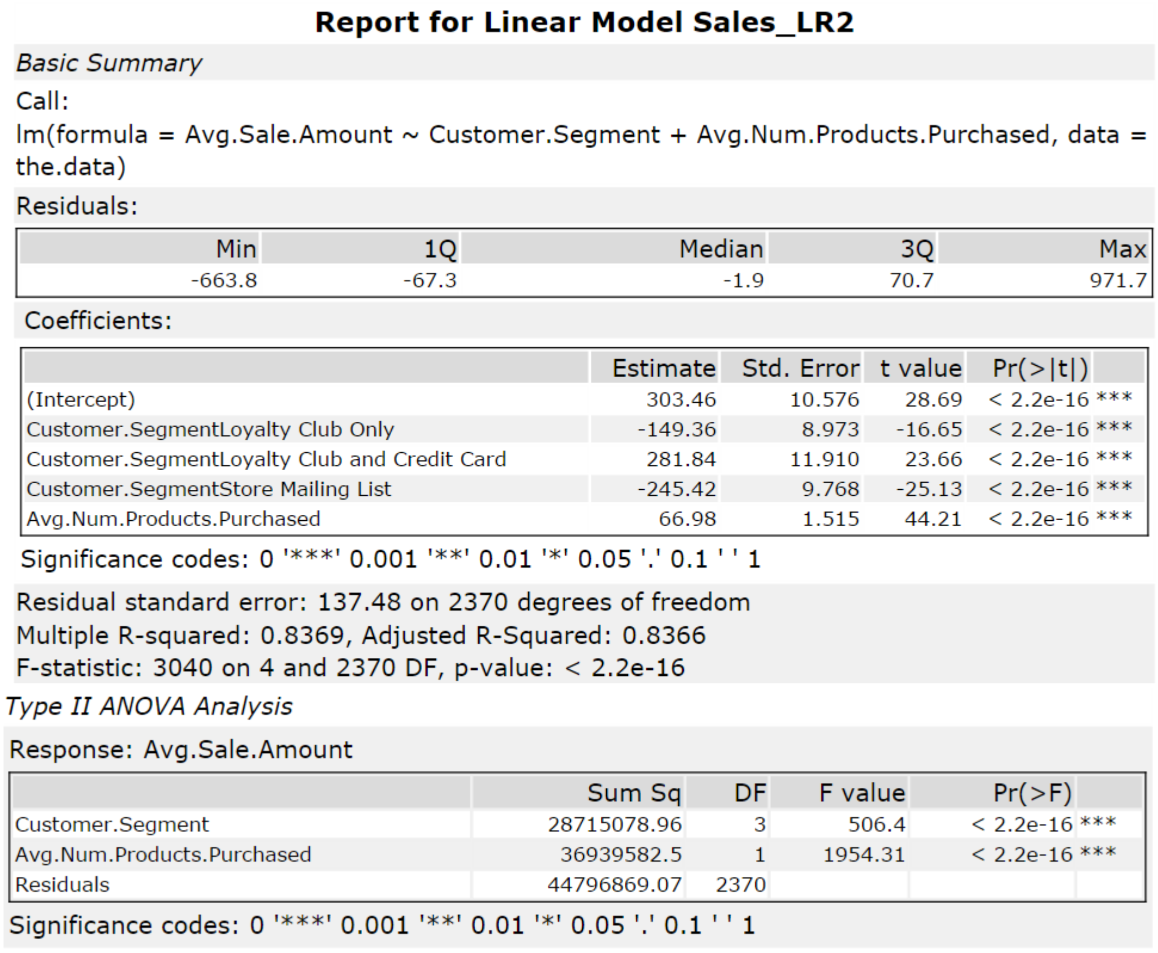
The numeric variable Avg_Num_Products_Purchased and the categorical variable Customer_Segment were used as predictor variables in the linear regression model for the target variable Avg_Sale_Amount.

From R-squared and Adjusted R-squared values of the regression model are both about 0.837. This indicated that all predictor variables are important. About 83.7% variation in the Avg_Sale_Amount is explained by the regression model.

The extremely small p values shows that both predictors are statistically significant.

The report for the linear regression model is shown below.


p1-mailinglist.xlsx Table=`p1-mailinglist$`
p1-customers.xlsx Table=`p1-customers$`
Profit=((Score_AvgSale_Amt*0.5)-6.50)*Score_Yes
O
R
I



Report for Linear Model Sales_LR2

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 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. **Linear regression equation based on the available data:**

   Avg_Sale_Amount = 305 + 66.81 * Avg_Num_Products_Purchased – 150.03 (If Custome_Segment: Loyalty Club Only) +281.69 (If Custome_Segment: Loyalty Club an Credit Card) – 242.76 (If Custome_Segment: Store Mailing List) + 0 (If Custome_Segment: 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 send the 250 new customers.
   Since the predicted net profit taking into account the probability the customer will buy the catalog, the $6.50 cost of printing and distributing catalog, and the 50% average gross margin on products sold through catalog is $21,987.44
   Since the predicted net profit > $10,000 requirement, company can go ahead and send the catalogs

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   Linear regression built with data from past 2375 customers is used to predict the average sale amount of the 250 new customers.
   Net expected profit calculated taking into account the probability the customer will buy the catalog (Score_Yes), the 50% average gross margin on products sold through catalog and the $6.50 cost of printing and distributing catalog by the formula

   Profit = ((Score_AvgSale_Amt * Score_Yes *0.5)-6.50)

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

   Profit = $21,987.44