Predicting Catalog Demand

Business and Data Understanding

Key Decisions:

What decisions needs to be made?

The company needs to decide whether or not to send catalogs to 250 new customers on the mailing list. If doing so would render a profit that exceeds $10,000, then the company should send out the catalogs.

What data is needed to inform those decisions?

Data for existing customers contained in the *p1-customers.xlsx* file to use in our linear regression model in order to predict the **Average Sale Amount per customer**.

The **Average Sale Amount per new customer** in the *p1-mailinglist.xslx* file can be extrapolated from linear regression model.

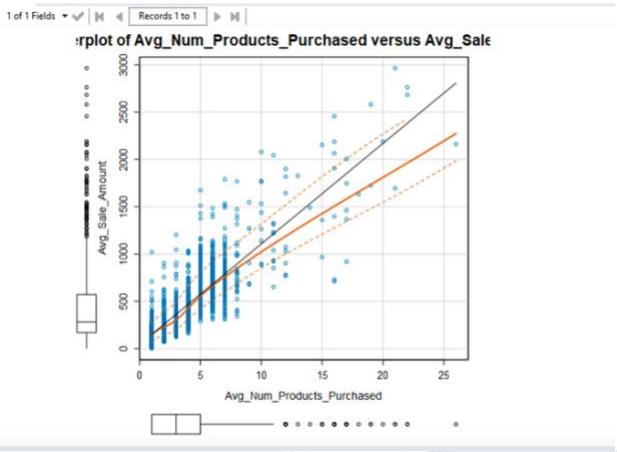
Our average gross margin on all product sold through the catalog is 50%.

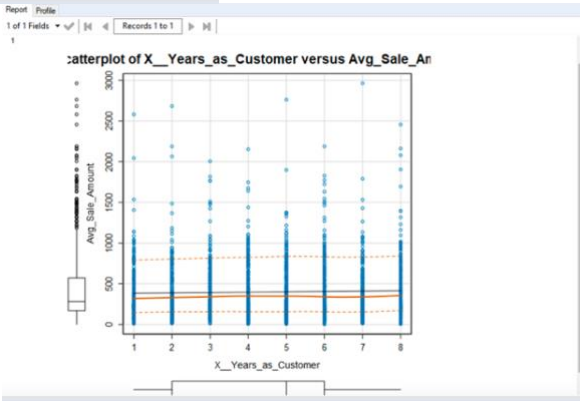
Our fixed cost is $6.50 per catalog * 250 new customers = $1,625.

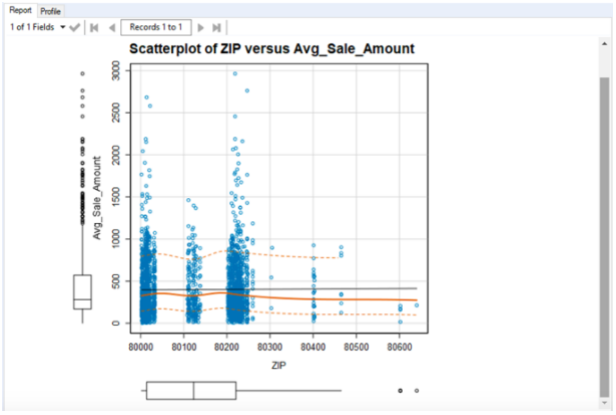
Analysis, Modeling, and Validation

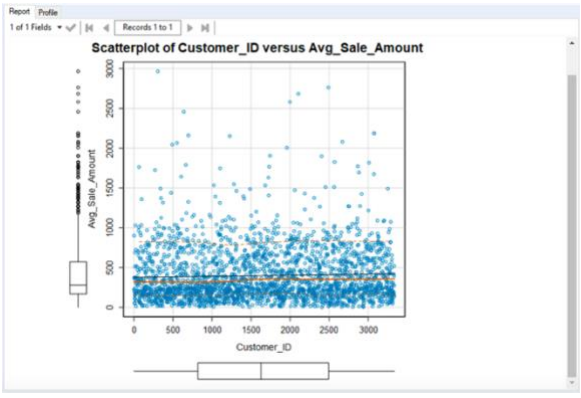
How and why did you select the predictor variables in your model?

Using Alteryx and the data from *p1-customers.xlxs*, I ran a scatter plot for each of the numeric predictor variables compared to our target variable, *average number of sales*, to verify a linear relationship between any of the variables. Among the numeric predictor variables, I chose to select *average number of products purchased* as a predictor variable because it rendered the strongest linear relationship while the others showed nearly no discernable relationship to our target variable.



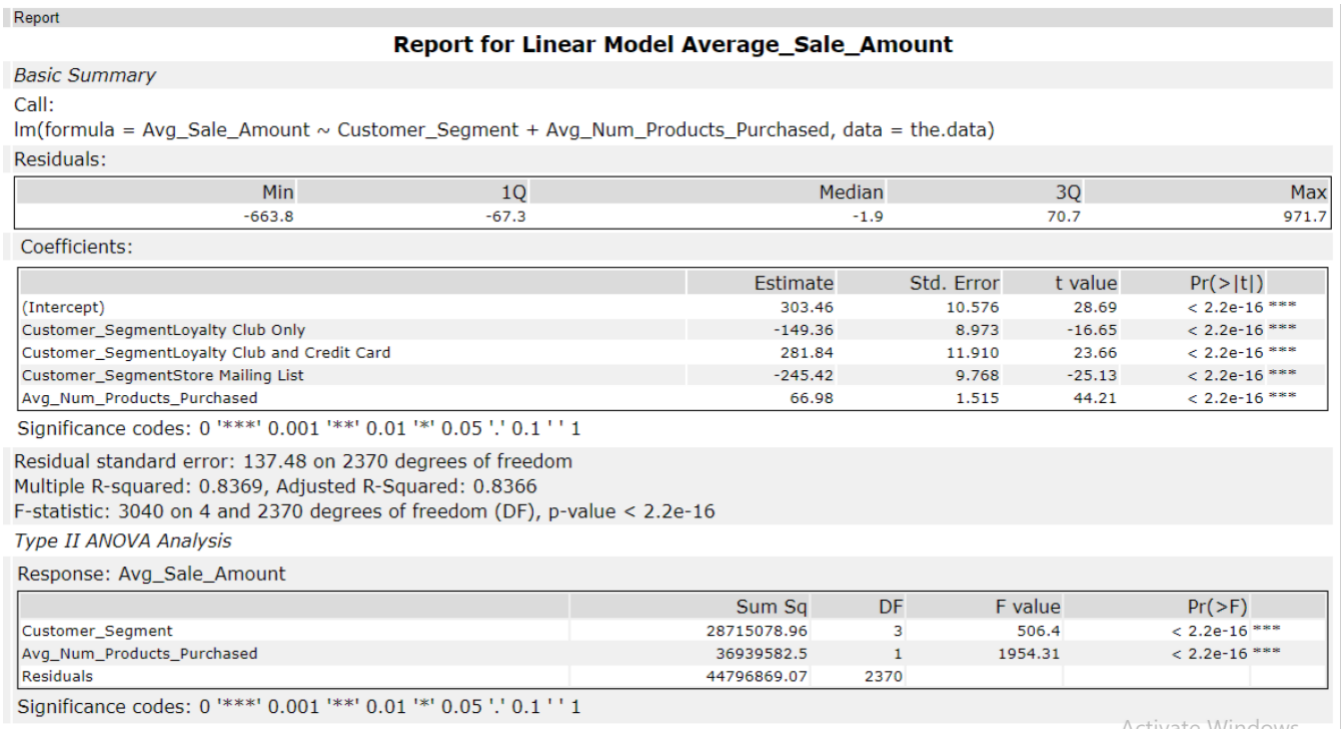






Then, I ran several linear regressions with various combinations of our categorical variables with our chosen numeric predictor variable, **Average Number of Products Purchased**.

Among our categorical variables, I decided to include **Customer Segment** and exclude **Response to Catalog. Response to Catalog** had relatively higher p-value (<.05), but including it with the other predictor variables had a statistically insignificant effect on the R-squared value for the model.



Report

Report for Linear Model Average_Sale_Amount

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

Explain why you believe your linear model is a good model.

This linear model is a good fit for predicting our target variable because the low p-values (<.01) for our predictor variables indicate that they are statistically significant and closely related to our target variable. Furthermore, the R-squared value of .8366 indicates that 83.66% of the variance in *Average Sale Amount* can be explained by the predictor variables used in our model.

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

Average Sale Amount = 303.46 - 149.36 * [**Loyalty Club Only**] + 281.84 * [**Loyalty Club and Credit Card**] – 245.42 * [**Mailing List**] + 66.98 * [**Avg_Num_Products_Purchased**]

Presentation/Visualization

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.

How did you come up with your recommendation?

After determining that **Avg_Num_Products_Sold** and **Customer Segment** from our existing customers in the *p1-customer.xlsx* file were appropriate to use in our linear regression model, I extrapolated these findings to our new customers in the *p1-mailinglist.xlxs* file to calculate the predicted **Average_Sale_Amount** for each of the new customers using the Score tool in Alteryx.

Multiplying the predicted **Average_Sale_Amount** for our new customers by their **Score_Yes** gave us the **Average_New_Sale_Amount** for each new customer.

To calculate the expected revenue generated from sending the catalog to our 250 new customers, I used the Summarize tool in Alteryx to sum the **Average_New_Sale_Amount** from each new customer.

Expected Profit = Expected Gross Revenue * .5 - $6.50 * 250

where .5 is gross profit
$6.50 per catalog
250 new customers

What is the expected profit from the new catalog?

The expected profit from the new catalog is **$21,987.44**.