Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

The decision of if we should send out the catalog to the new 250 mailing list customers. This decision will be based on the predicted profit we might get from sending out the catalog. Thus, if the predicted profit exceeds $10,000 then we will decide on sending out the catalogs and if not, I will recommend not sending the catalogs.

2. What data is needed to inform those decisions?

To make an informed decision, we need to have records of Previous and Current sale data, Cost of printing, Revenue, Average Gross Margin, Customer Segments, and Average Purchased products. Which will help us in making the best prediction of the profit generated from the sales data we will gain from the predicted amount of the new customers' purchases.

Step 2: Analysis, Modeling, and Validation

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

By using the Alteryx tool, the datasheet has been imported, cleansed, and analyzed. A linear regression data analysis has been made on the given variables. The P-value of the imported data showed statistical significance on both the customer segments and the average number of products purchased. By using scatter plots, a major positive trendline can be noted between the sales amount and the average number of products purchased.

Report for Linear Model P1_Linear_Regression_3

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-665.62	-67.24	-2.43	71.05	970.91

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.352e+02	1.052e+02	4.1360	4e-05 ***
Customer_SegmentLoyalty Club Only	-1.496e+02	8.980e+00	-16.6629	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	2.823e+02	1.193e+01	23.6622	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-2.460e+02	9.773e+00	-25.1727	< 2.2e-16 ***
Customer_ID	-1.377e-03	2.941e-03	-0.4684	0.63956
Store_Number	-1.138e+00	9.955e-01	-1.1426	0.25331
Avg_Num_Products_Purchased	6.699e+01	1.517e+00	44.1575	< 2.2e-16 ***
X._Years_as_Customer	-2.345e+00	1.223e+00	-1.9167	0.0554 .

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

Residual standard error: 137.42 on 2367 degrees of freedom
Multiple R-squared: 0.8372, Adjusted R-Squared: 0.8368
F-statistic: 1739 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16

Figure 1: 1st Report for the Linear Model Regression from the Alteryx tool

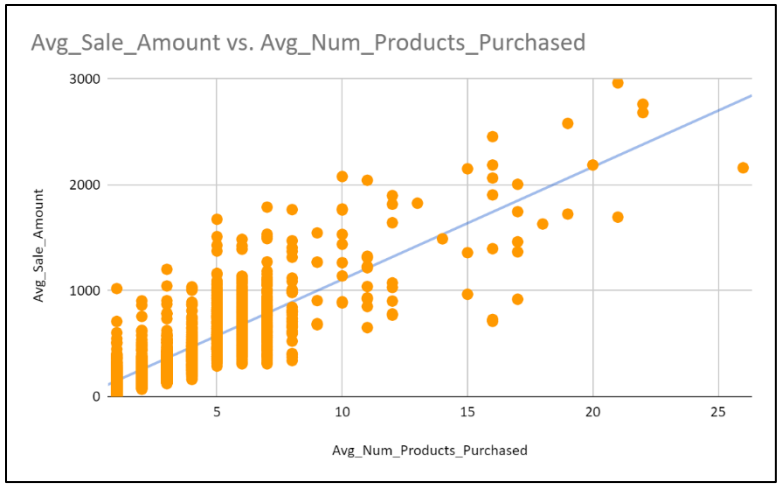

Figure 2: Scatter plot of average number of products purchased vs Average sale amount

Other data with low statistical significance has been tested and proven insignificant by using scatter plots and checking trendlines.

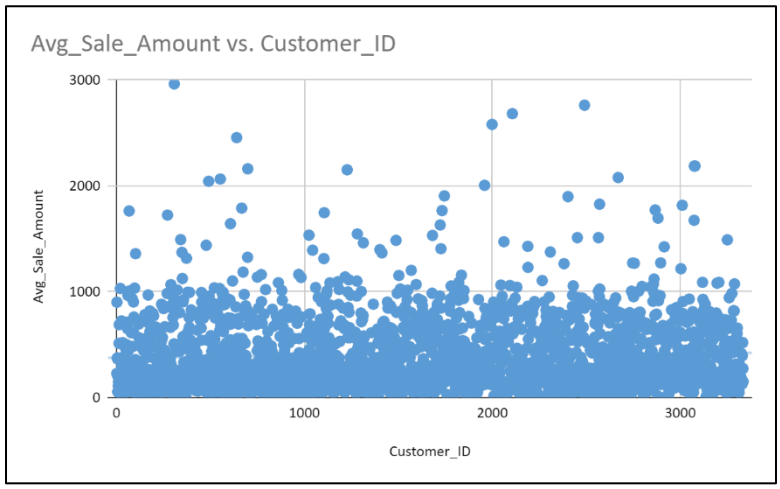

Figure 3: Scatter plot of Customer ID vs Average sale amount

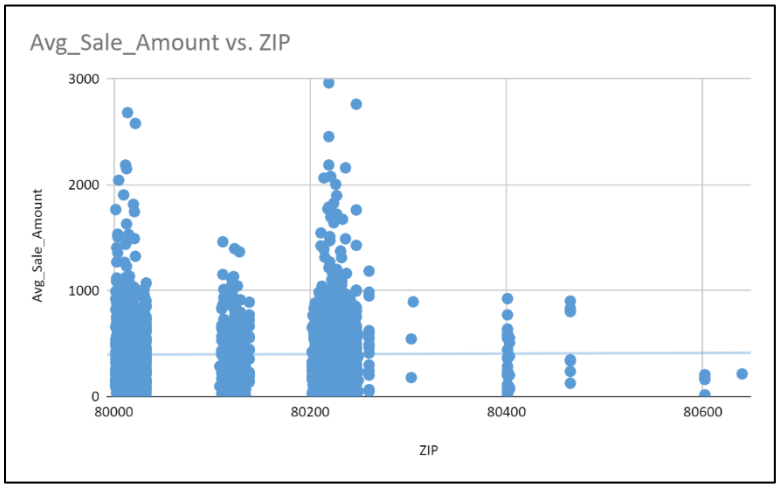

Figure 4: Scatter plot of Zip vs Average sale amount

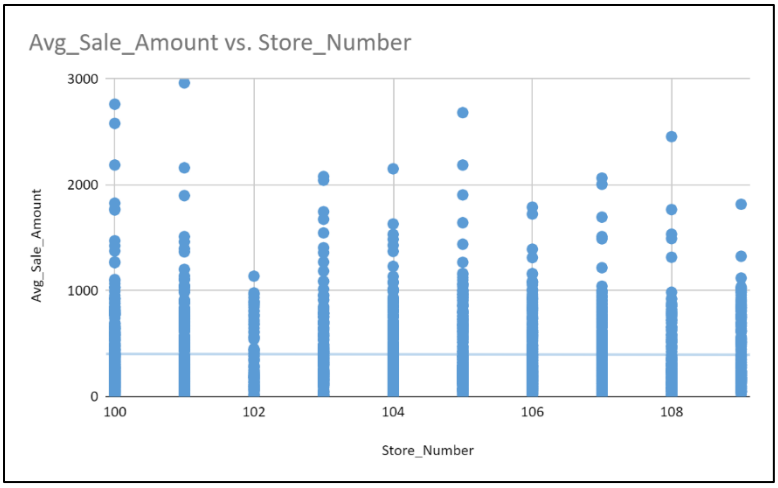

Figure 5: Scatter plot of Store Number vs Average sale amount

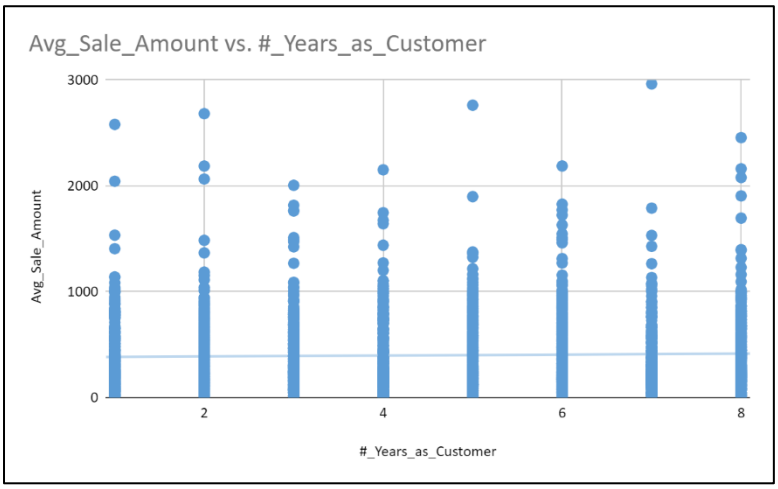

Figure 6: Scatter plot of Number of years as a customer vs Average sale amount

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.

By using and adjusting the data in the Alteryx model to only obtain the significant data, the Multiple R-squared equaled 0.8369, Adjusted R-Squared equaled 0.8366 which is a high value considering that the higher the r-squared, the higher the explanatory power of the model with r-squared ranging from 0 to 1. All p-values are below 0.05 which proves statistical significance.

Record | Report

Report for Linear Model P1_Linear_Regression_3

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

Figure 7: 2nd Report for the Linear Model Regression from the Alteryx tool

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)

Y=303.46-149.36*(Loyalty Club Only) +281.84*(Loyalty Club & Credit Cards)-245.42*(Mailing List) +0*(Credit Cards only) +66.98*(average number of products purchased)

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 catalog because the profit exceeds $10,000 which is the goal.

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

After applying the linear regression model, cleansing, modeling, and analyzing the data, I have used the following formula to calculate the expected profit of sending the catalog:

Expected profit = (sum expected revenue * gross margin)- (cost of printing * number of customers) once the total exceeded the expected, I made the decision to recommend sending the catalogs.

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

Cost of printing catalog = $ 6.50

Number of new subscribed to mailing list customers= 250

Gross Margin = 50% = 0.5

Expected Revenue = score (expected sale amount) * score_yes

Expected profit = (sum expected revenue * gross margin)- (cost of printing * number of customers)

Sum Expected Profit = (sum expected revenue*0.5) – 6.50*250 =

Profit= (47,224.87*0.5) -(6.50*250)

= (23,612.43) -(1625)

= $ 21987.43

For the 250 new customers the expected profit = $ 21987.43 which is more than $10,000