Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions need to be made?

The company has 250 new customers from their mailing list that they want to send the catalog to. The decision to be made is whether to send the catalog to these 250 new customers based on expected profit. The condition being, only if the expected profit is greater than $10000, the company will send out the catalogs

2. What data is needed to inform those decisions?

A few of the columns from the dataset are needed to predict sales and calculate the expected profit. They are *Customer Segment*, *Average Number of Product Purchased*.

The values for *Gross Margin* and *Cost of Catalog* have been provided for calculation

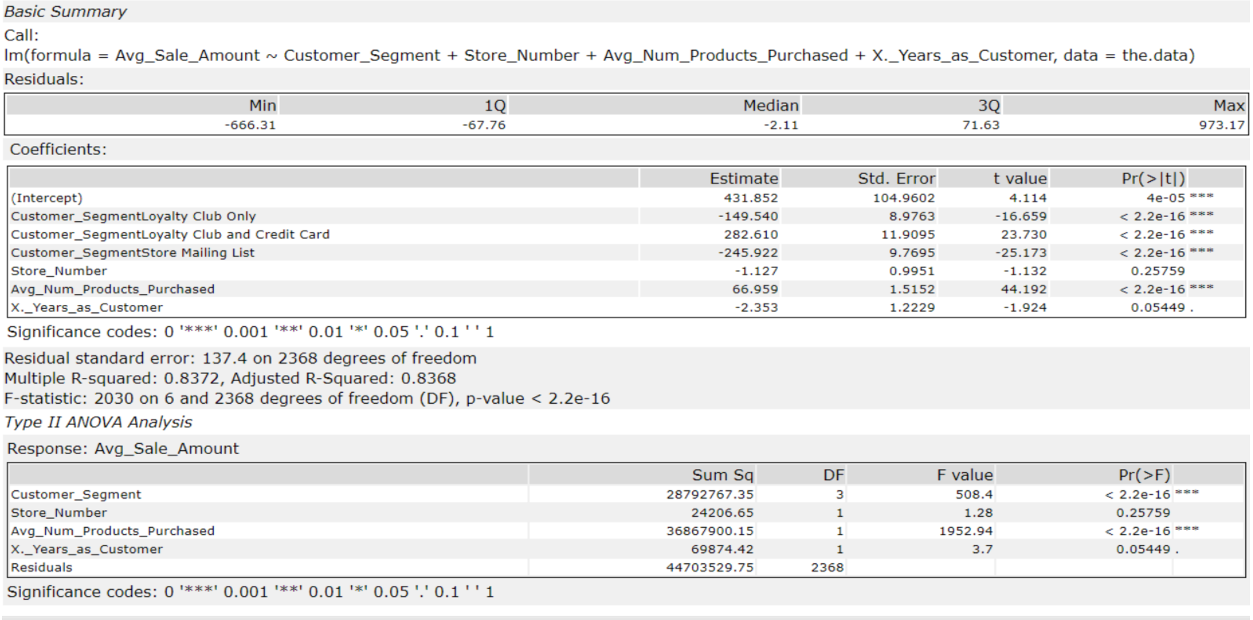
Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

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

The linear regression result table below performed on all variables against Average Sale Amount. Only Average Number of Product and Customer Segment have a p-value of less 0.05 which implies statistical significance.



Basic Summary

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-666.31	-67.76	-2.11	71.63	973.17

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	431.852	104.9602	4.114	4e-05 ***
Customer_SegmentLoyalty Club Only	-149.540	8.9763	-16.659	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.610	11.9095	23.730	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.922	9.7695	-25.173	< 2.2e-16 ***
Store_Number	-1.127	0.9951	-1.132	0.25759
Avg_Num_Products_Purchased	66.959	1.5152	44.192	< 2.2e-16 ***
X._Years_as_Customer	-2.353	1.2229	-1.924	0.05449 .

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

Residual standard error: 137.4 on 2368 degrees of freedom
Multiple R-squared: 0.8372, Adjusted R-Squared: 0.8368
F-statistic: 2030 on 6 and 2368 degrees of freedom (DF), p-value < 2.2e-16

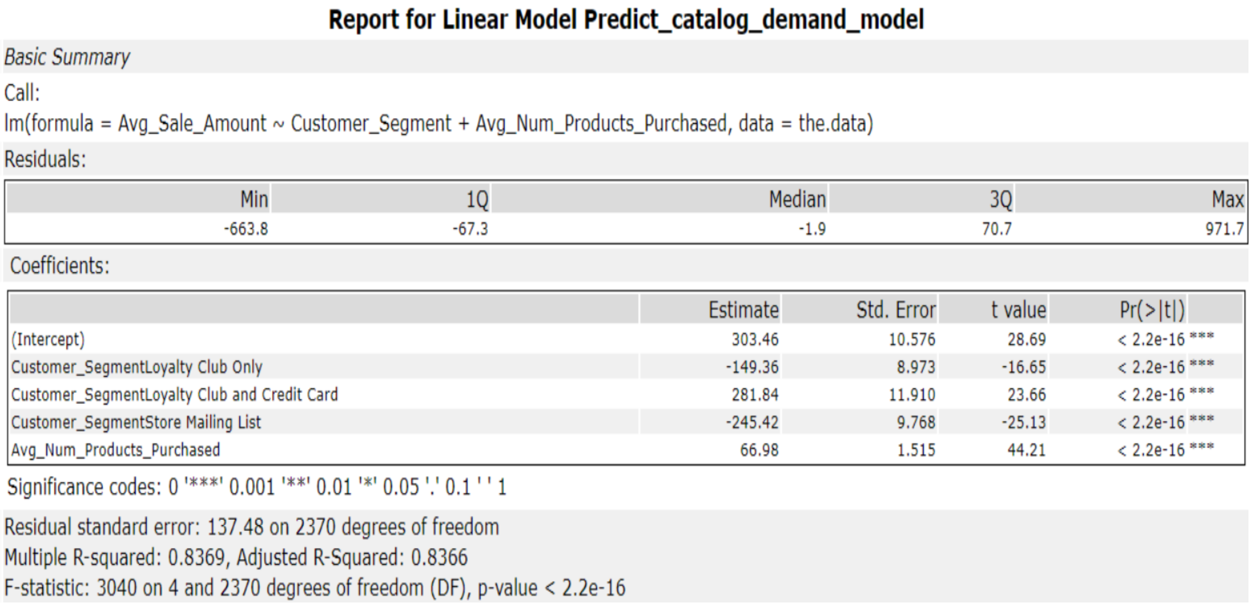
Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28792767.35	3	508.4	< 2.2e-16 ***
Store_Number	24206.65	1	1.28	0.25759
Avg_Num_Products_Purchased	36867900.15	1	1952.94	< 2.2e-16 ***
X._Years_as_Customer	69874.42	1	3.7	0.05449 .
Residuals	44703529.75	2368		

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.



Report for Linear Model Predict_catalog_demand_model

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

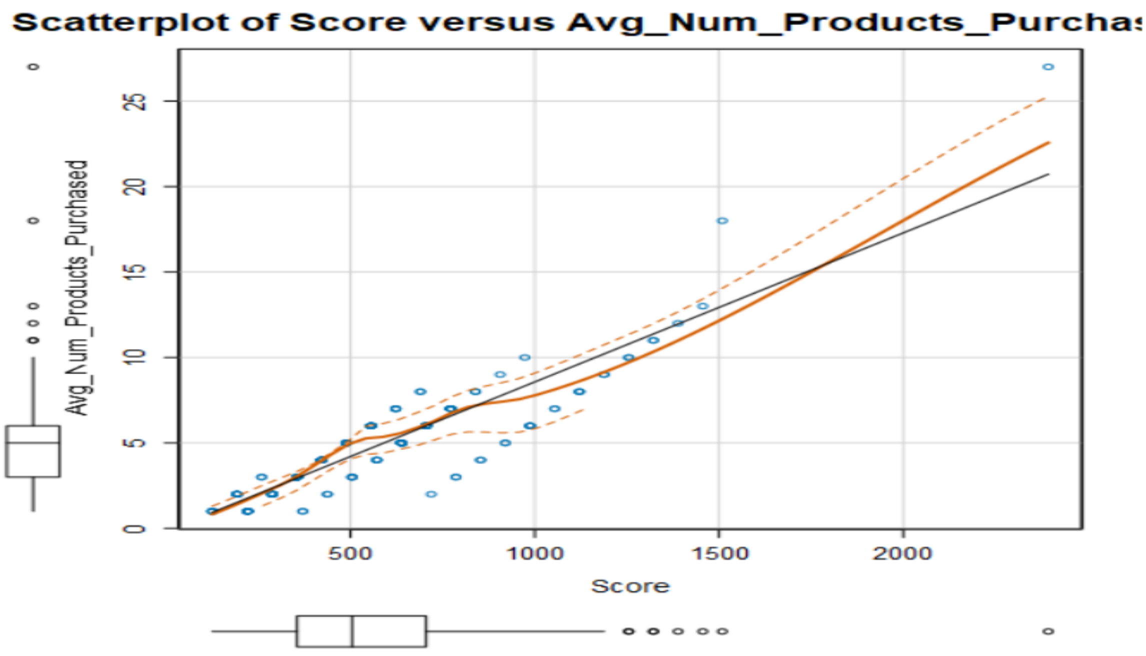
Coefficients:

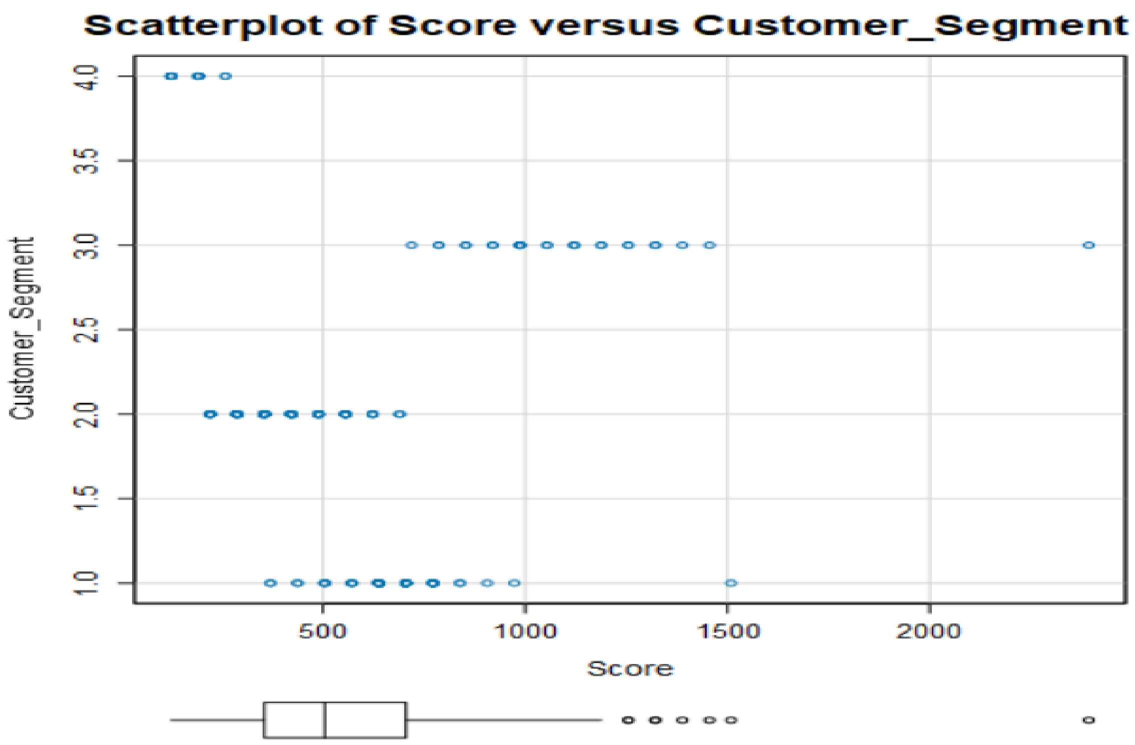
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
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

Scatterplots of Average Number of Product and Customer Segment versus Average Sale Amount are also plotted below. There is positive linearity associated with both the predictor variables.





The Alteryx linear regression function used to determine the statistical model shows an adjusted R-squared value of 0.8366 which is higher than the 0.7 value. This means that the model is strong. Additionally, the customer segment and Average Number of Products also have a p-value lower than 0.05. This implies that their values are statistically significant. We can strongly say that the model is a good one.

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……*

Avg_Sale_Amount = **303.46** – **149.36** x (If Type: Loyalty Club Only) + **281.84** x (If Type: Loyalty Club and Credit Card) – **245.42** x (If Type: Store Mailing List) + **66.98** x (Avg_Num_Products_Purchased) + **0** x (If Type: Credit Card Only)

Step 3: Presentation/Visualization

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

(Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000)

Since the expected profit is **$21,987.44**. The recommendation is to go ahead and send the catalogs to the 250 new customers

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

Once I inspected the data in Alteryx, I then used a linear regression model. The two predictor variables chosen were an average number of products and the customer segment based on the p-value and R-squared value.

The expected profit from each customer was calculated by as follows.

The expected revenue from each customer is determined by multiplying 'score' with 'Score_Yes' value.

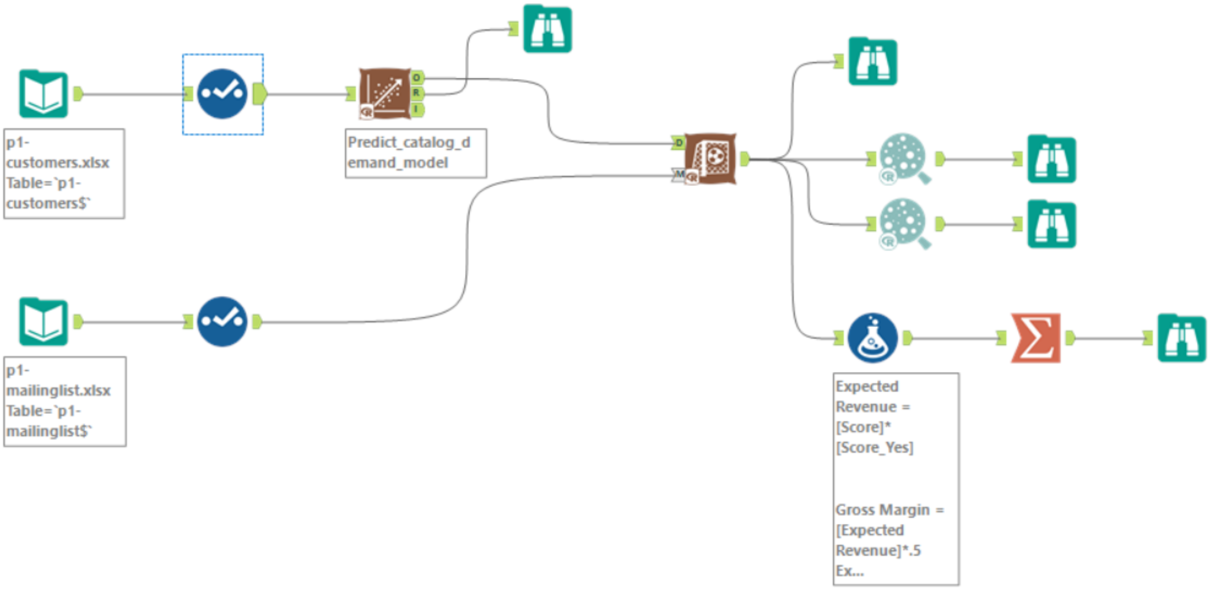
The gross margin is given as 50%. So, 50% is deducted from the sum of expected revenue.

The expected profit is calculated by subtracting the cost of the catalog ($6.50) from the gross margin.

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

Expected Profit = *(Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)*
= *(47,225.87 x 0.5) – (6.50 x 250)*
= *23,612.44 – 1,625*
= *$21,987.44*

My Alteryx Workflow


p1-customers.xlsx Table=`p1-customers$`
Predict_catalog_demand_model
O
R
I
D
M
p1-mailinglist.xlsx Table=`p1-mailinglist$`
Expected Revenue = [Score]*[Score_Yes]
Gross Margin = [Expected Revenue]*.5 Ex...