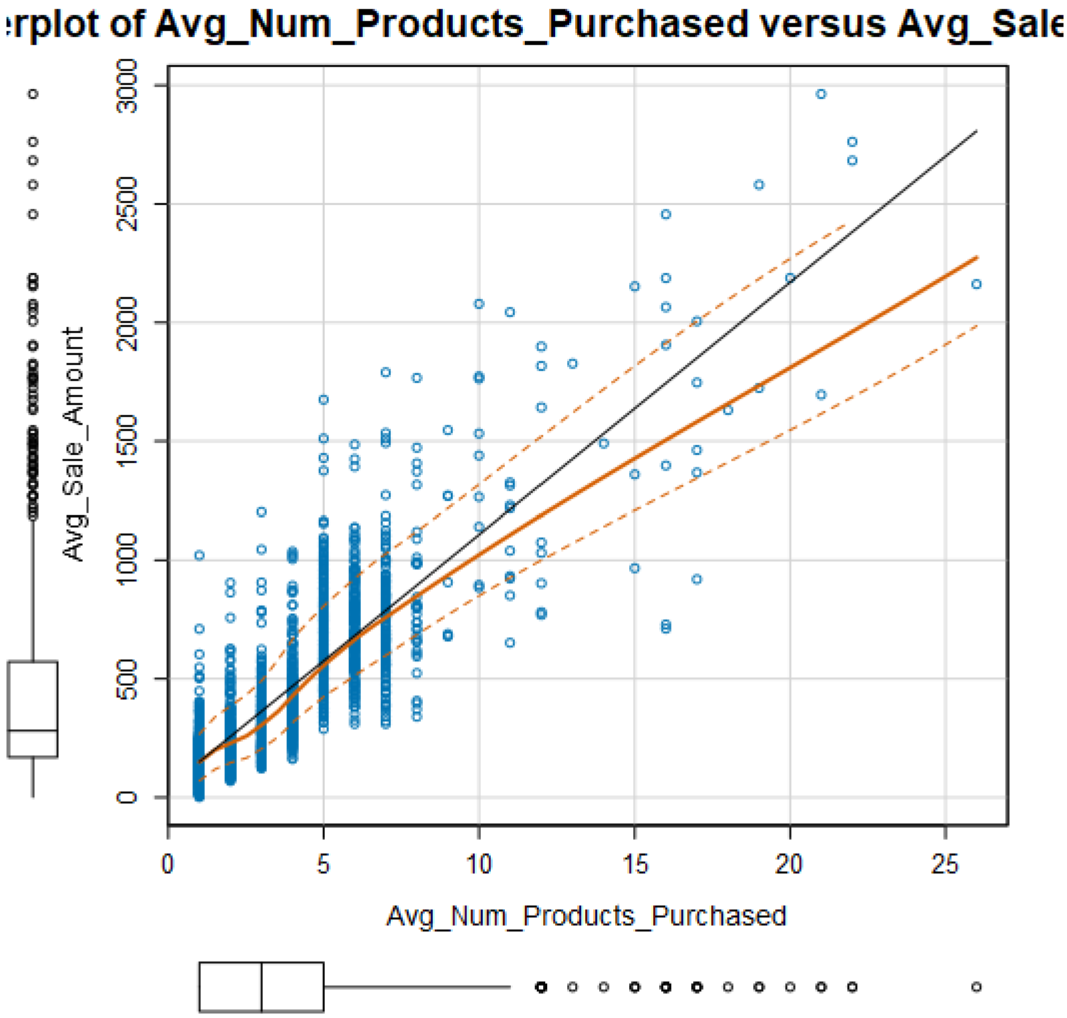
Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The company has 250 new customers from the mailing list. We are interested to send them the catalog and get extra business. We used existing customer data to predict the expected profit from these 250 new customers if we sending them the catalog. We are only interested to send them the catalog if the expected profit exceeds $10,000.

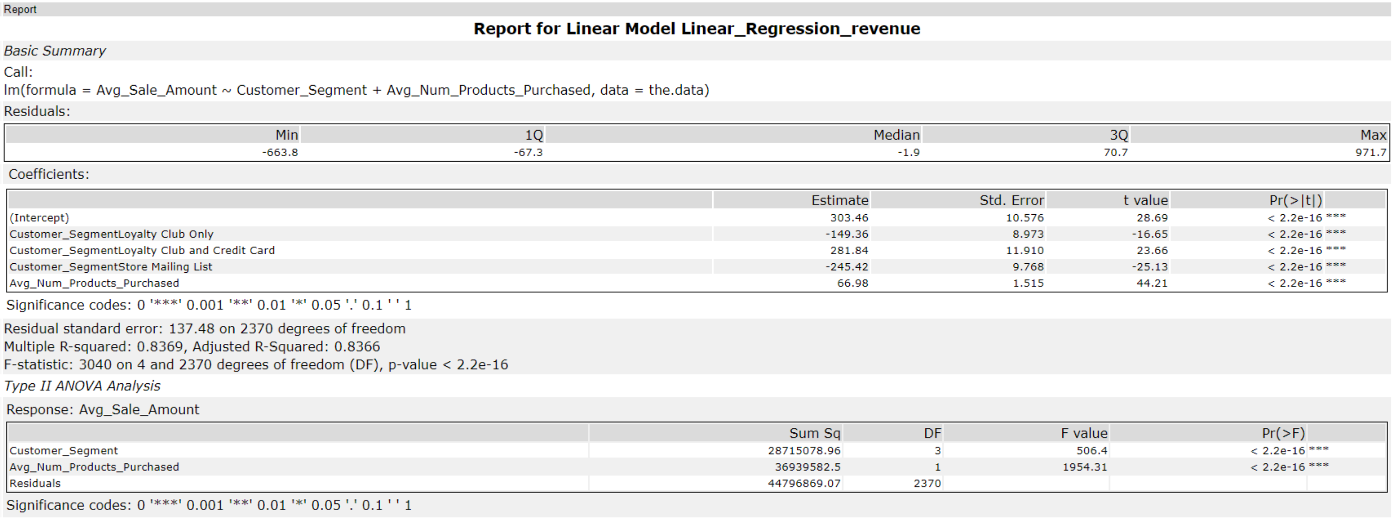
Step 2: Analysis, Modeling, and Validation

Existing customer dataset have both numeric and categorical data.
For numeric data, we used the scatterplots to see if the variable is a good candidate for a predictor variable. I tried ZIP code, Store Number, Year as customer and Avg number products



purchased. And it turns out that only avg number of products purchased have a positive relationship with sales amount.

For categorical data, we used trial and error to see if they are statistically significant. I checked P value on each variable and to see if they have a strong relationship with Sales amount. It turns out that customer segment is an important categorical variable to predict sales amount.



Report

Report for Linear Model Linear_Regression_revenue

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

After we build our model, we find our adjusted R-squared is 0.8366 and all P value are below 0.05. A high R-squared and low P-values indicate it's a good model.

The linear regression equation is below:

Ave sales amount = 303.46 + 66.98 (Avg_Num_Products_Purchased) – 149.36 (If Customer Segment : Loyalty Club only) + 281.84 (If Customer Segment : Loyalty Club and Credit Card) - 245.42 (If Customer Segment : Store Mailing List)

Step 3: Presentation/Visualization

I will recommend to send out the catalog to these 250 new customers. Based on our analysis, we will have an estimated profit for $ 21,987, which exceeds $10,000.

The way we calculate the estimated profit is below:

Expected profit for each new customer = Predicted avg sales amount * Score_Yes * 0.5 - 6.5

Total Expected Profit = Sum (expected profit for each customer) = $ 21, 987