Predicting Catalog Demand

Step 1: Business and Data Understanding

1. What decisions need to be made?
    a. We have a predictive business problem which requires linear logistical regression to solve
    b. The data which we'll use is rich
    c. The goal is to generate expected profit then decide whether the company will decide to send the catalogs or not to the 250 customers
2. What data is needed to inform those decisions?
    a. Data on the customers from this year and previous year such as number of years as customers and purchase of sales data and expected profit from each catalog

Step 2: Analysis, Modeling, and Validation

1. I've selected the Customer_Segments and Avg_Num_Products_Purchased as the predictors for the linear regression model
2. I chose them because they have high significant codes as well as P-Value codes below 0.05. As opposed to the last submission where I added Store Number and X_Years_As_Customer even though they had low significant code and P-Value higher than 0.05. Also, simple analysis using scatterplot shows that there is somewhat strong relationship between target variable and avg_sale_amount as well as customer_segments
3. To explain the p-values and r-squared values, the p-values first of all are all 2.2e-16 which is lower than 0.05 which signifies that the outcome was not by chance and thus there is a relationship between the variables. As for the r-squared, the value was 0.8366 which means that over 84% of variance is explained by model. So the fit is good.
4. Results of the linear model as well as scatterplot is listed below

5.

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

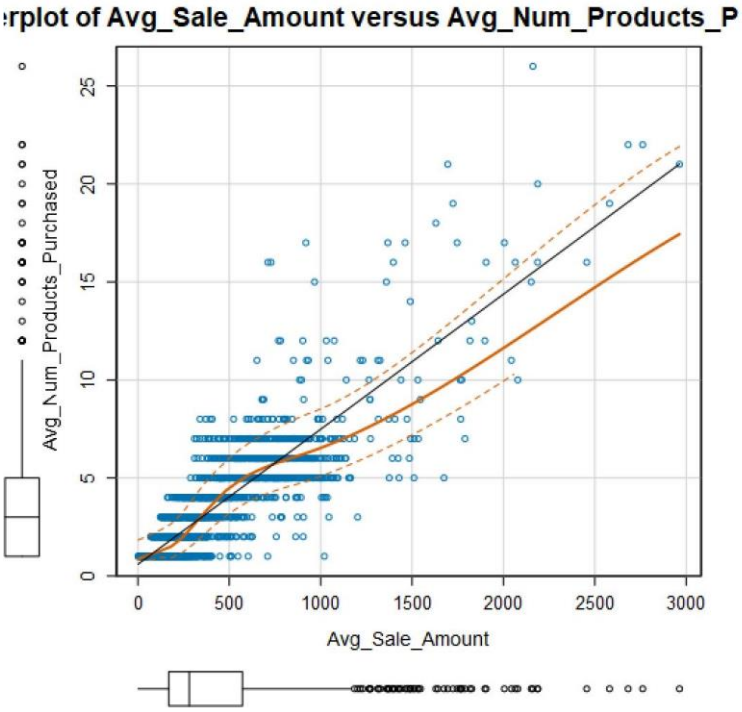
Type II ANOVA Analysis

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

Graph



6.

7. The linear regression equation ended up being Y = 303.46 + (-149.36 * Customer_SegmentLoyalty Club Card Only) + (281.84 * Customer_SegmentLoyalty Club and Credit Card) + (-245.42 * Customer_SegmentStore Mailing List) + (66.98 * Avg_Num_Products_Purchased) + (Credit Card * 0)

a. The model fits well because of R-squared is a high 0.8369 as well as the adjusted R-squared being high 0.8366. The variables chosen are also high significant as the p-values shows.

Step 3: Presentation/Visualization

1. What is your recommendation?
    a. My recommendation is that the company should send out the catalog to 250 customers as the profit is greater than $10,000
2. How did you come up with your recommendation?
    a. First, multiply the expected revenue with score_yes to get probable predicted revenue
    b. The values are then totaled by adding
    c. After total, multiply by gross margin of 50%
    d. Then subtract the cost of catalog per customer which is 250 * 6.50
    e. The profit is then $21,987.44 which is greater than the $10,000 limit
    f. Below are the results attached
    g.

Record #	Sum_avg_sales_probable	cost	margin	profit
1	47224.871373	1625	23612.4356865455	21987.435687