Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

We want to predict the expected revenue from the 250 customers in order to get expected profit.
The management has to decide whether to send the catalog out to these new customers .
The management will send the catalog to new customers if the expected profit from them exceeds $10,000 .

2. What data is needed to inform those decisions?

We need to know the cost , profit and the data of all customers . The data must include whether the customers bought from us and if they bought from us , how many products they bought and what is the probability that the new customers will also buy from us . We need to know the average gross margin (price - cost) on all products sold through the catalog . We need to know the costs related to the printing and distribution of each catalog . And we need data on the average profit we make for each catalogue we send to the customer . By using Linear Regression technique , we can predict the expected profit and take better decision .

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.

At the minimum, answer these questions:

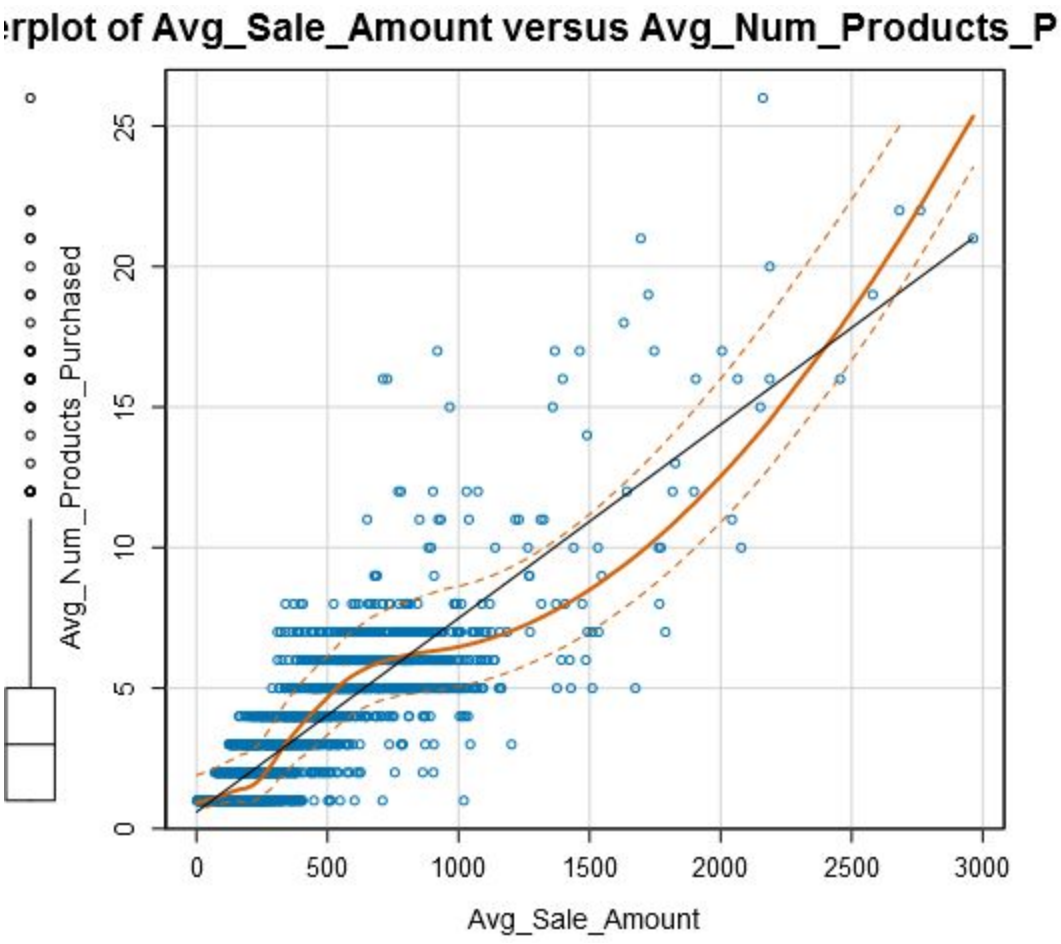
1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

I selected Avg_Num_Products_Purchased as a predictor variable due to linear relationship between Avg_Num_Products_Purchased and Avg_Sale_Amount as you can see from the scatterplot .

I selected Customer_Segment which is a categorical value as a predictor variable because its Pr value is below 0.05 and is considered as statistically satisfiable (see the image below where Pr value is highlighted) .

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			



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.

As the image below shows that the R-squared value and Adjusted R-squared value of my regression model is 0.839 and 0.8366 respectively (above 0.07) .Also , the predictor variables - Customer_Segment and Avg_Num_Products_Purchased which I have used to predict the value of Avg_Sale_Amount have Pr values less than 0.05(Pr value is <2.2e-16), so my predictor variables are statistically satisfiable(***) . So , I believe that my linear model is a good model .

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 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

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 + 66.98 * Avg_Num_Products_Purchased -149.36 (If Type: Loyalty Club Only) + 281.84 (If Type : Loyalty Club and Credit Card) - 245.42 (If Type: Store Mailing List) + 0 (If Type : Credit Card Only)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

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

The expected profit is $21,987.44 which is greater than $10,000 , so the company should send the catalog to the new 250 customers .

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

I applied Linear Regression on the p1-customers data by setting Avg_Sale_Amount as the target variable and Customer_Segment and Avg_Num_Products_Purchased both as predictor variables. Then I passed Linear Regression output and p1-mailinglist data as the inputs to the SCORE . Then I use FORMULA in which I applied this expression : [Score]*[Score_Yes] .
After that , I use another FORMULA in which I applied this expression : ([Score]*0.5)-6.5 .
In the end , I use SUMMARIZE that sum the all the score variables and the result gives us the expected profit .

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

The expected profit is $21987.44