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? :
    - Take a predictive analysis because prediction is going to be made from business problem.
    - Data is given in project so data analytical technique will be used.
    - The result will be in dollars (numbers) so we can use numeric techniques for analysis.
    - Use liner regression model so we can get the best prediction model.
    - Liner regression model will help with the equation and we can use the data from the email list for prediction.
    - If we can profit more than 10000 then company will send the catalog to the new customers.

2) What data is needed to inform those decisions?
    - From the mailing list, customer Id, number of years as customer.
    - From customer list, customer id, location, number of years as customer, purchase and sales data.
    - Data from previous year will be essential to make decision model.
    - In the end, we will calculate the expected profit > 10000 to make decision whether to send catalog or not.

Step 2: Analysis, Modeling, and Validation

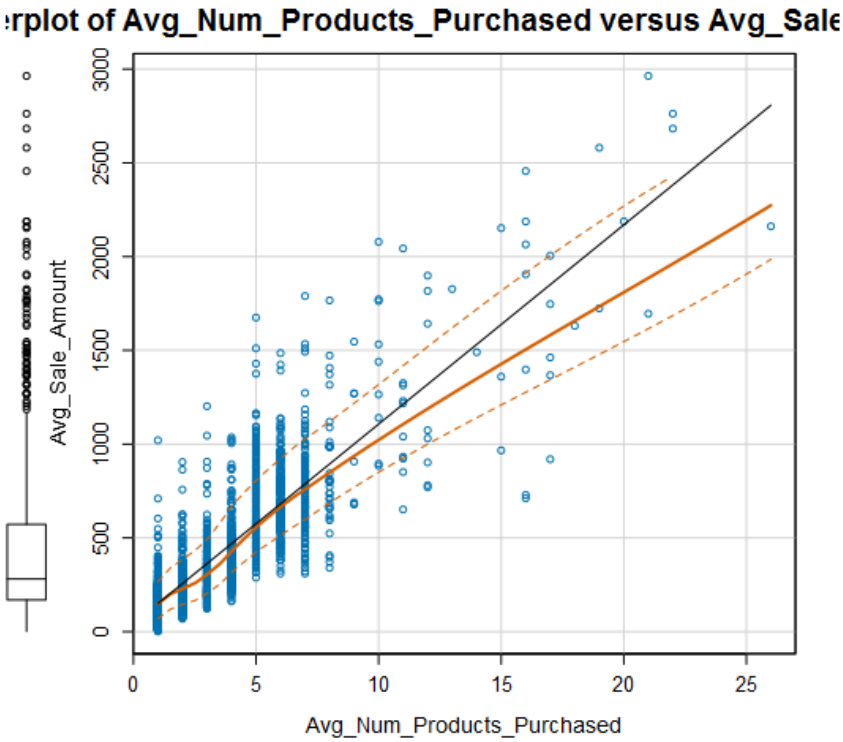
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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 in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.
    - First for linear model, use customers file as an input.
    - Create different scatter plot to get the best liner model.

- Avg_Sale_amount is main variable here, so generated different scatter plot with the input file customers.
- The most reliable liner model was found with the Avg_sales_amount on Y axis to Avg_Num_Products_Purchased.
- The output from the regression model is going to use with score_yes in score tool.



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.

Model has R-squared value > 0.5 which makes it strong model.

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 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)

Important: The regression equation should be in the form:

Y = 303.46 + (281.84 X Customer_SegmentLoyalty Club and Credit Card) + (-149.36 X Customer_SegmentLoyalty Club Card Only) + (-245.42 X Customer_SegmentStore Mailing List) + (66.98 * Avg_Num_Products_Purchased) + Credit Card * 0

Here, Model is strong as we have R-Squared value : 0.83 and Adjusted R-squared value: 0.83 which are highly significant per p-values.

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? – Yes, Profit margin is greater than 10,000

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

- Output from score tool gives the predicted sale amount for the mailing list customers.
- Mailing list customers are 250.
- 250 customers multiplied with the score_yes from the mailing list as well.
- After getting the result of it, the total values will be formatted for individual customers.

- Now, multiply it for the 50% of gross margin.
- Total $ 1625 is going to deducted from it.
- $1625 is result of 250 customers and $6.50 per catalog multiplication.
- Profit is more than $20000 which is obviously greater than expected profit ($10000).

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

- Expected profit is : $21987.43
- Expected revenue sum – 47224.9
- Margin- 23612.43
- Cost- 1625