<u>Project 1: Predicting Catalog Demand</u>

Step 1: Business and Data Understanding

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers.

You’ve been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Key Decisions:

Answer these questions

1. What decisions needs to be made?

The key decision to be made is to predict how much profit the company can expect to gain from sending catalog to the 250 new customers from the mailing list, given Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

2. What data is needed to inform those decisions?

p1-customers: This dataset includes the following information on about 2,300 customers. We will be building our predictive model using this data.

p1-mailinglist: This dataset is the 250 customers that we need to predict sales.

Step 2: Analysis, Modeling, and Validation

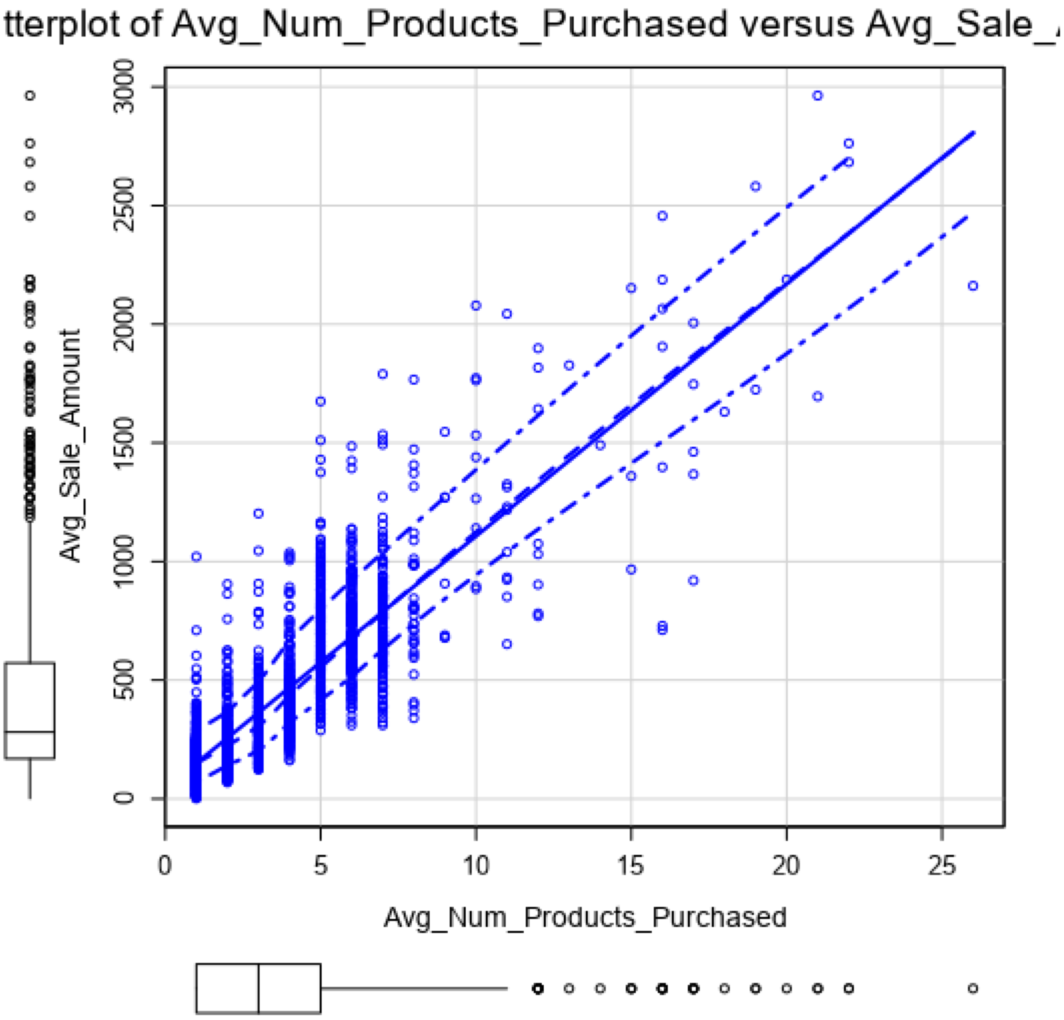
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

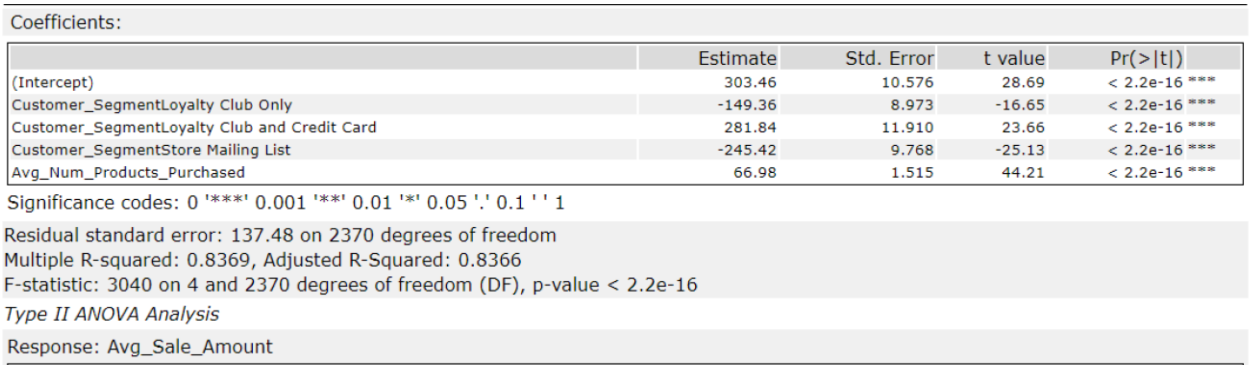
At the minimum, answer these questions:

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.

Target variable is Avg_Sale_Amount, checking for relationship between **Avg_Num_Products_Purchased** & the target variable using scatterplot tool. You find that there is a positive correlation between both variable as determined by the scatterplot below.



Next Variable I used for my Predictive model is the **Customer Segment**, seeing it’s a categorical variable, there’s no way to plot the relationship between it and the target variable using a scatterplot chart, method used is trial and error to see if they are statistically significant. by including the predictor variable in the model, running it and checking the P value.



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

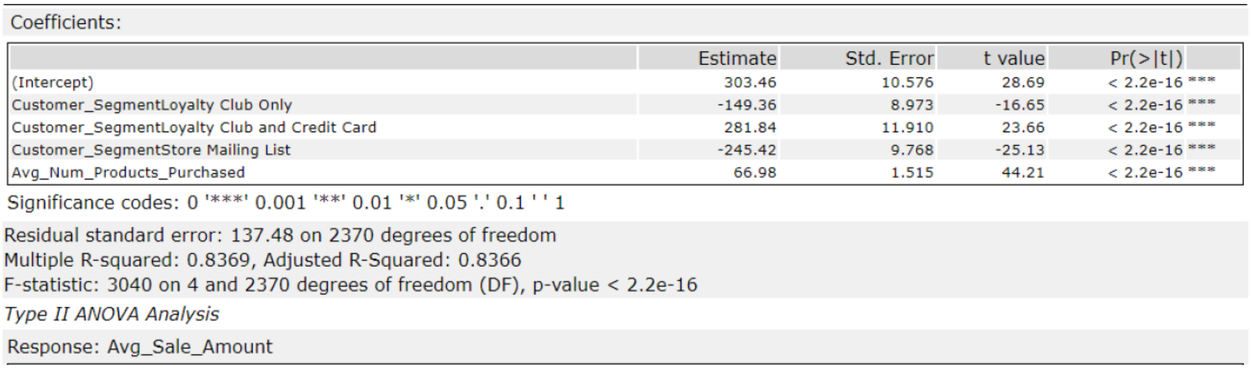
Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

Low P-values(< 0.05) of the customer segment variables suggest it’s good fit for my Linear regression model.

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.



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

From the P Value & R squared value from the model output above, there’s a good reason to believe the selected predictor variables are good for the model since Low P-values and a high R-squared suggest the model is highly predictive.

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)

Avg_Sale_Amount = 303.48 + Customer_SegmentLoyalty Club Only(-149.36) + Customer_SegmentLoyalty Club and Credit Card(281.84) + Customer_SegmentStore Mailing List(-245.42) + Avg_Num_Products_Purchased(66.98)

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, the company should send out the catalogs to the 250 customers since the expected profit predicted from the Model exceeds the $10,000 threshold set by Management.

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

Step 1: Using Customer_Segment and Avg_Num_products_purchased columns from the p1-customers data to train our linear regression model.

Step 2: Applying the linear regression equation on the p1 mailing list data to predict Avg_Sale_Amount for every 250 rows of data.

Step 3: Multiplying the Predicted Avg_Sale_Amount by Score_Yes (Probability that a customer will respond to catalog and make a purchase) and assigning the value to a variable called Revenue.

Step 4: Calculating profit from each sale by taking into factor 50% profit margin and $6.5 cost for producing each catalog.

i.e., Profit = Revenue * 0.5 - 6.5

Step 5: Summarizing the profit values for each row of data to see if it falls below or above the $10,000 threshold.

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

Expected profit from the new catalog is around $21,987.44