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?
    a. Shall we send catalog to these new 250 customers?
    b. How much profit can we make from these new 250 customers?
    c. Can expected profit from sending catalog to 250 new customers exceed $10,000?
2. What data is needed to inform those decisions?
    a. The gross margin for each catalog sending to customers
    b. The cost of printing and sending catalog to customers
    c. Data of customers that bought something from the catalog in the past including but not limited to:
        i. Bought an item from a past catalog
        ii. Average amount of items the customer buys from the company
        iii. The total dollar amount that the customer spent ordering from our catalog

Step 2: Analysis, Modeling, and Validation

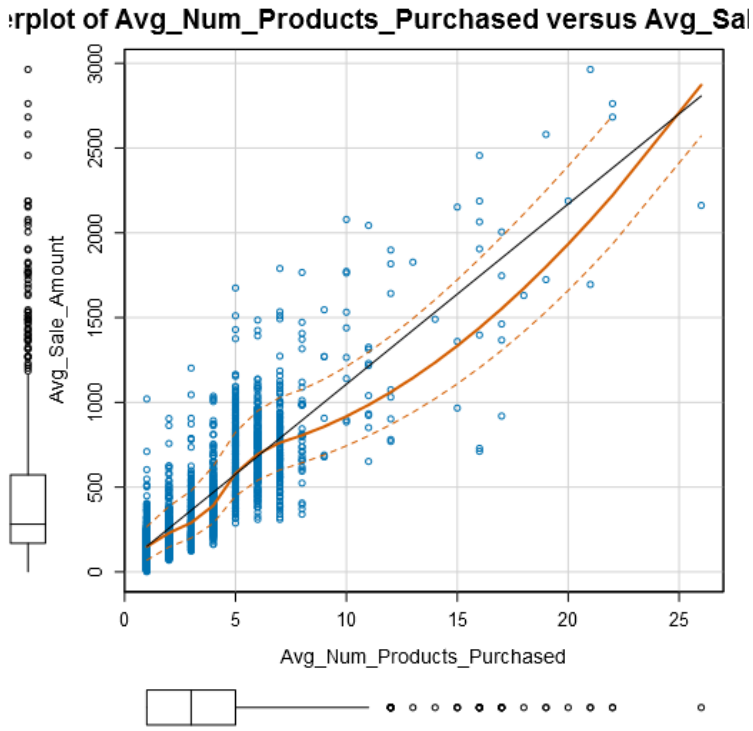
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

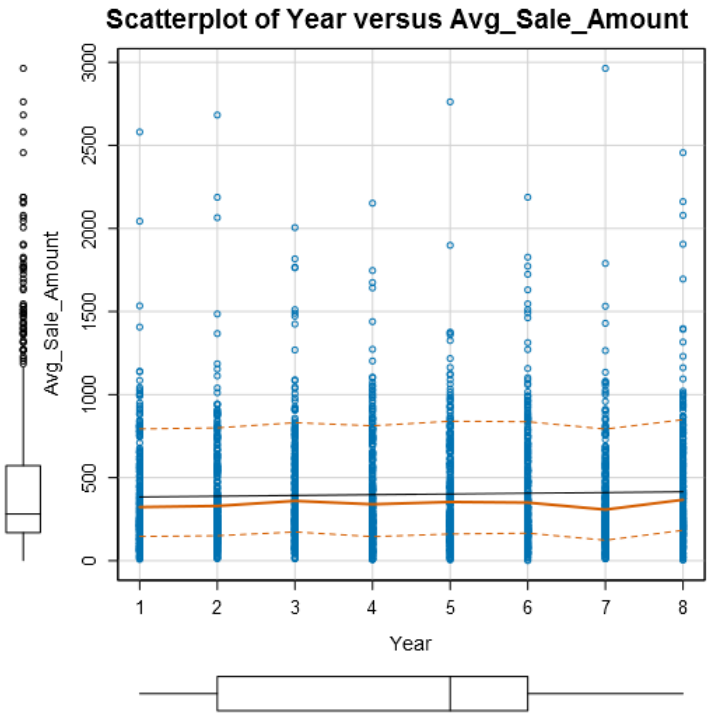
Important: Use the p1-customers.xlsx to train your linear model.

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.

    - Dataset contains
        - 3 continuous features and
        - 9 categorical features.
    - My target variable is Avg_Sale_Amount.
        - First I have to check the correlation between 2 continuous variables and the target variables.





- There is some positive relationship between
  Avg_Num_Products_Purchased and Avg_Sale_Amount variables.

- There is no visible relationship between
  #_Years_as_Customer variable and Avg_Sale_Amount variable.

- I have to exclude Name, CustomerID and Address variables from predictor variables, as this two features are not likely to have any relationship with the target variable.

- And I'll also exclude the State variable in this dataset since there is only one value in its field "CO".

- Since the customers to predict are new, they don't have any value in Responded_to_Last_Catalog variable. I have to exclude this variable too.

- Now, I have selected
    - Avg_Sale_Amount as my target variables and
    - Avg_Num_Products_Purchased and the remaining 4 categorical Variables as predictor variables.

- To find the relationship between these 4 categorical variables and the target variable, I have to run linear regression test model and check the correlation between these variables.

- I ran my linear regression model with 5 predictor variables and found that only 2 predictor variables are statistically significant for my training model.
    - Customer_Segment and Avg_Num_Products_Purchased
- So I decide to run my training model with
    - Avg_Sale_Amount as target variable and
    - Customer_Segment and Avg_Num_Products_Purchases as predictor variables.

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	27413020.63	3	480.37	< 2.2e-16 ***
City	478475.88	26	0.97	0.51066
ZIP	1290040.03	77	0.88	0.76054
Store_Number	194978.46	9	1.14	0.3312
Avg_Num_Products_Purchased	35308379.63	1	1856.17	< 2.2e-16 ***
Residuals	42952075.43	2258		

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.
    - I have removed all other variables which are not statistically significant predictors of target variables to increase accuracy of my training mode.
    - The p value of all remaining coefficients of my model are less than 2.2×10^{-16} with multiple R-squared value of 0,8369 and adjusted R-Squared value of 0.8366.
    - Any predictor variables with p value less than 0.05 is considered statistically Significant. Any model with R-squared value above 0.70 is considered a "strong" model.

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_Amont = 303.46 – 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) + 0 (If Customer_Segment: Cash) + 66.98 * Avg_Num_Products_Purchased

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. My recommendation is the company should send the catalog to these 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 first selected the predictor variables and built the linear model. Then scored the data on the mailing list, then multiplied “score” with “Score_Yes” (where Score_Yes is theprobability of purchase.) Then added up all values to get a combined value for 250 customers.

   - To get the expected profit from each customer, we have to subtract the cost of catalog ($6.50) from expected gross margin for each customer (expected average sale amount * 0.50).
   - The expected gross margin from these 250 customers is $23,612.
   - The total cost for printing and distributing 250 catalog is $1,625.
   - The expected profit contribution will be $21,987

   - The management want to send the catalog only if the expected profit contribution exceeds $10,000. According to the training model, the expected profit contribution will be 2 time greater than the minimum expected profits. So, I’ll recommend that the company should send the catalog.

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

   Assuming the catalog is sent to these 250 customers, expected profit is $21,987.44