Project 1: Predicting Catalog Demand

By: Fairoza Amira Binti Hamzah

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?

   Sales manager need to answer the question of whether to send the catalog can be sent to 250 customers or not, based on the expected profit

2. What data is needed to inform those decisions?

The previous sales data, which is from `p1-customers.xlsx`.The data is as below.

`Customer_Segment`

`Customer_ID`

`Address`

`City`

`State`

`ZIP`

`Avg_Sale_Amount`

`Store_Number`

`Responded_to_Last_Catalog`

`Avg_Num_Products_Purchased`

`#_Years_as_Customer`

So, we need to find which attributes (features) that actually contributed to the `Avg_Sale_Amount`. Then, we need to calculate the profit to ensure that we can get more than USD 10,000 when send the catalogs to 250 customers.

Step 2: Analysis, Modeling, and Validation

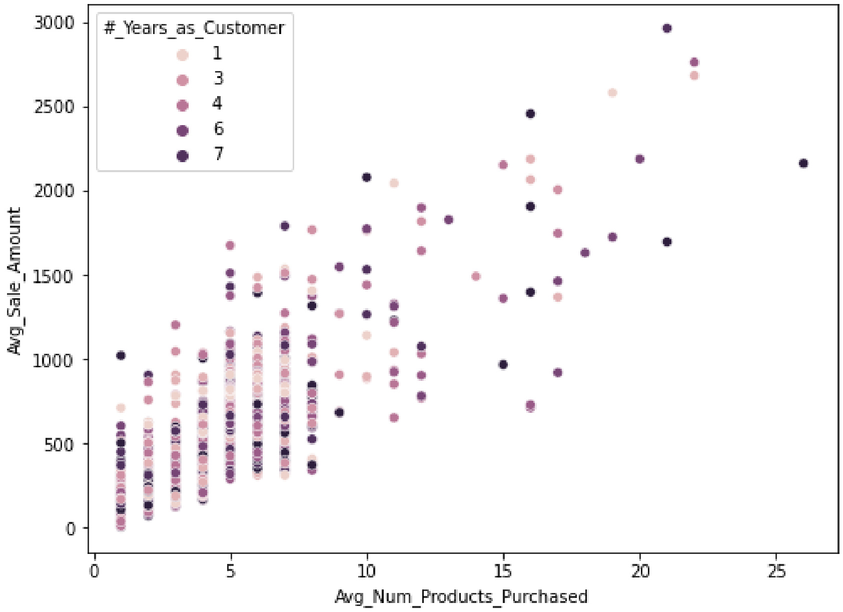
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

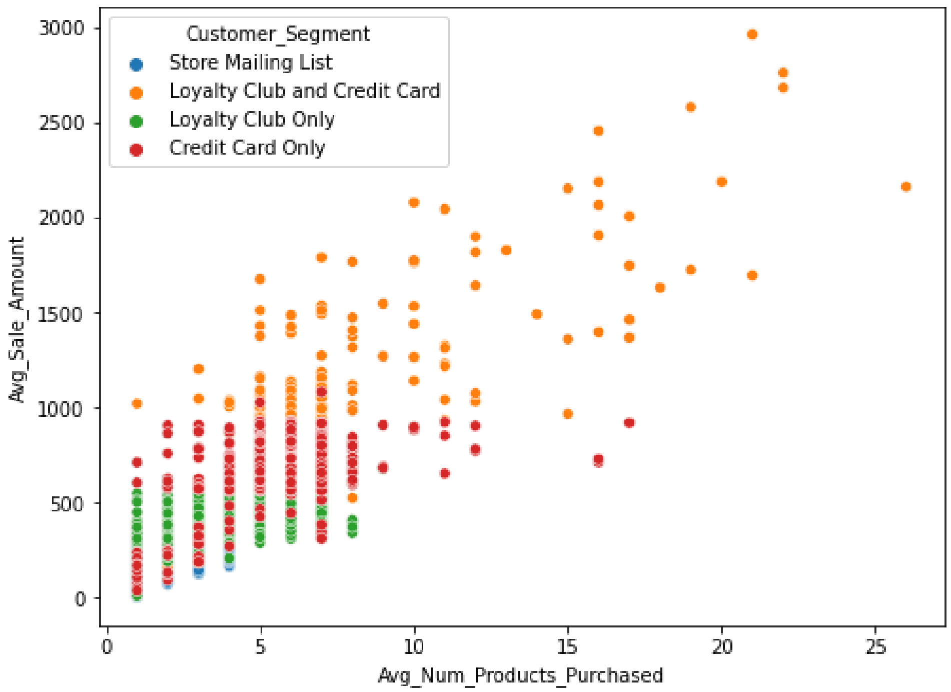
At the minimum, answer these questions:

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.

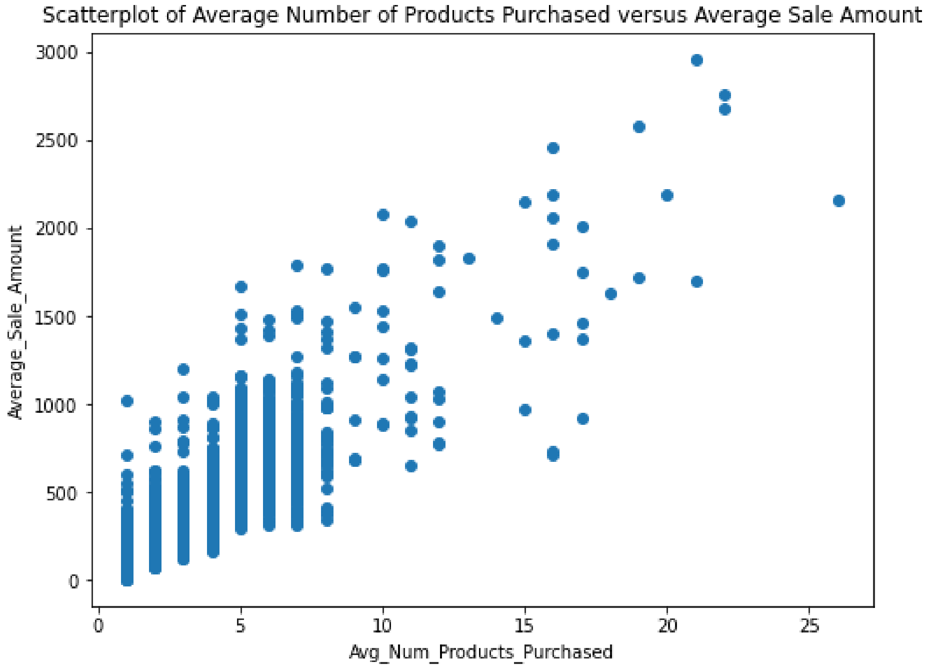
   Visualize the relationship by using scatterplots.



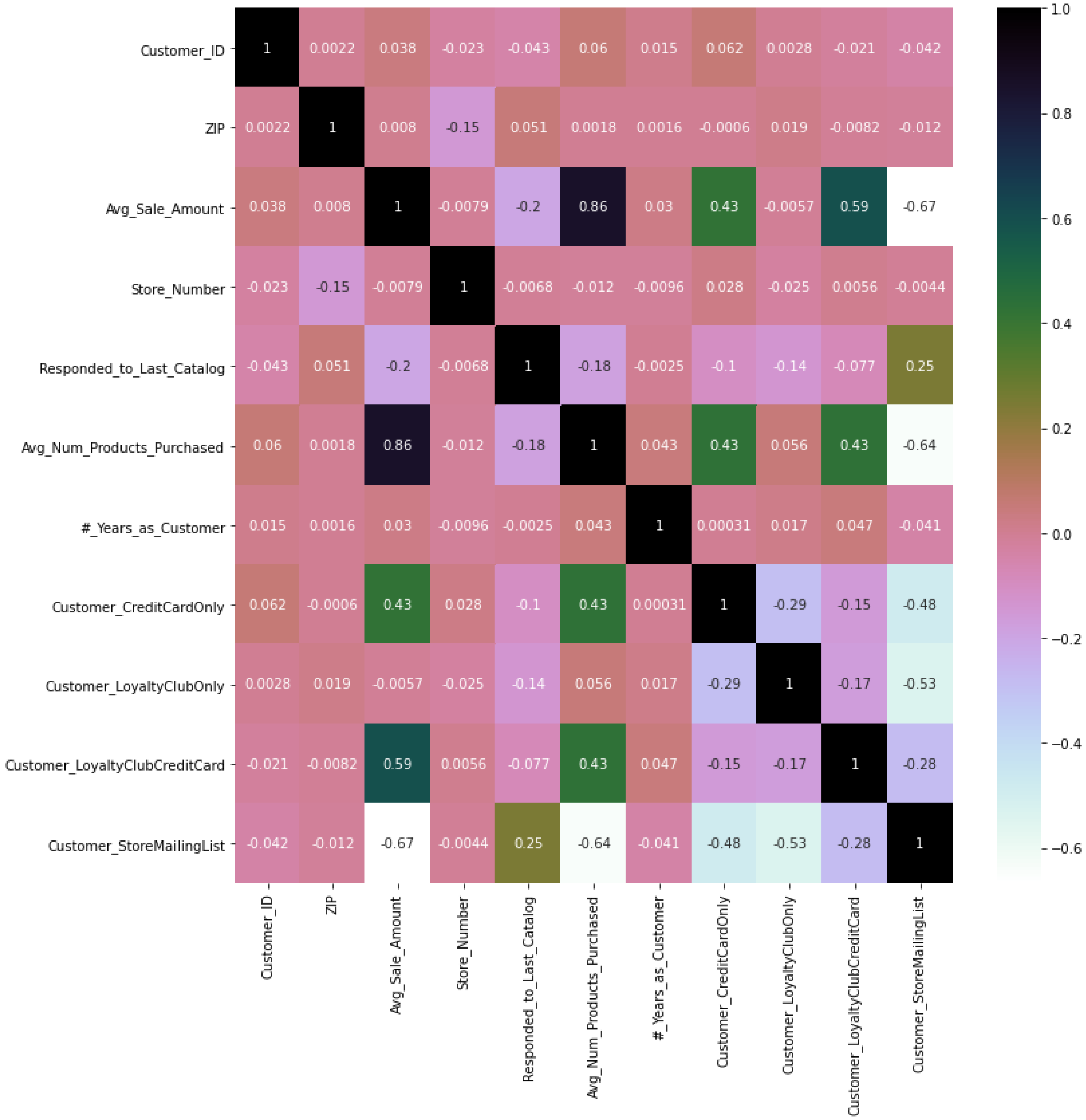
   We can see that only customers with more than 3 years, purchased more than 20 products. The Avg_Num_Products_Purchased is also linearly related to Avg_Sale_Amount.



The linear relationship can be found for Loyalty Club and Credit Card (customer segments).



Next, to find the most correlated features.



	Customer_ID	ZIP	Avg_Sale_Amount	Store_Number	Responded_to_Last_Catalog	Avg_Num_Products_Purchased	#_Years_as_Customer	Customer_CreditCardOnly	Customer_LoyaltyClubOnly	Customer_LoyaltyClubCreditCard	Customer_StoreMailingList
Customer_ID	1	0.0022	0.038	-0.023	-0.043	0.06	0.015	0.062	0.0028	-0.021	-0.042
ZIP	0.0022	1	0.008	-0.15	0.051	0.0018	0.0016	-0.0006	0.019	-0.0082	-0.012
Avg_Sale_Amount	0.038	0.008	1	-0.0079	-0.2	0.86	0.03	0.43	-0.0057	0.59	-0.67
Store_Number	-0.023	-0.15	-0.0079	1	-0.0068	-0.012	-0.0096	0.028	-0.025	0.0056	-0.0044
Responded_to_Last_Catalog	-0.043	0.051	-0.2	-0.0068	1	-0.18	-0.0025	-0.1	-0.14	-0.077	0.25
Avg_Num_Products_Purchased	0.06	0.0018	0.86	-0.012	-0.18	1	0.043	0.43	0.056	0.43	-0.64
#_Years_as_Customer	0.015	0.0016	0.03	-0.0096	-0.0025	0.043	1	0.00031	0.017	0.047	-0.041
Customer_CreditCardOnly	0.062	-0.0006	0.43	0.028	-0.1	0.43	0.00031	1	-0.29	-0.15	-0.48
Customer_LoyaltyClubOnly	0.0028	0.019	-0.0057	-0.025	-0.14	0.056	0.017	-0.29	1	-0.17	-0.53
Customer_LoyaltyClubCreditCard	-0.021	-0.0082	0.59	0.0056	-0.077	0.43	0.047	-0.15	-0.17	1	-0.28
Customer_StoreMailingList	-0.042	-0.012	-0.67	-0.0044	0.25	-0.64	-0.041	-0.48	-0.53	-0.28	1

The most correlated features are:

	index	Corr
0	Avg_Sale_Amount	1.000000
2	Avg_Num_Products_Purchased	0.855754
5	Customer_StoreMailingList	0.666655
4	Customer_LoyaltyClubCreditCard	0.591488
3	Customer_CreditCardOnly	0.426358

1	Responded_to_Last_Catalog	0.199358

So, only those features above are considered to build the linear regression model. Since the Responded_to_Last_Catalog feature is not in the Mailing List dataframe, this feature is excluded in the list of features for the modeling.

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.

```
OLS Regression Results
==============================================================================
Dep. Variable:        Avg_Sale_Amount   R-squared:                       0.837
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                     3040.
Date:                Thu, 03 Jun 2021   Prob (F-statistic):               0.00
Time:                        14:20:44   Log-Likelihood:                -15061.
No. Observations:                2375   AIC:                         3.013e+04
Df Residuals:                    2370   BIC:                         3.016e+04
Df Model:                           4
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          154.1077      7.916     19.467      0.000     138.584     169.631
Avg_Num_Products_Purchased      66.9762      1.515     44.208      0.000      64.005      69.947
Customer_CreditCardOnly        149.3557      8.973     16.645      0.000     131.760     166.951
Customer_LoyaltyClubCreditCard 431.1945     12.697     33.962      0.000     406.297     456.092
Customer_StoreMailingList      -96.0620      7.756    -12.386      0.000    -111.271     -80.853
==============================================================================
Omnibus:                      359.638   Durbin-Watson:                   2.045
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4770.580
Skew:                           0.232   Prob(JB):                         0.00
Kurtosis:                       9.928   Cond. No.                         21.7
==============================================================================
```

The p-value for each feature is below than 0.05 and the R-squared value is 0.837. The lower the p-value, the more significant it is to the model. R-squared if defined as explained variation divided by the total variation. The closer the R-squared 1.0 indicates that the model explains all the variability of the response data around its mean, thus the better the model fits to the data. Therefore, as our current p-value are all less than 0.05 and our R-squared value is 0.837, indicates that it fits well to our 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)

$$AvgSaleAmount = 154.11 + 66.98\ \times AvgNumProductsPurchased + 149.36\ \times CustomerSegmentCredictCardOnly + 431.19\ \times CustomerSegmentLoyaltyCreditCard - 96.06\ \times CustomerSegmentStoreMailingList$$

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, I will recommend to 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)

   The profit is calculated as
   $$totalProfit = totalRevenue \times 0.5 - 6.5 \times 250$$
   0.5 refers to the gross margin, 6.5 is the cost of catalog and 250 is the number of customers we want to send the catalogue.

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

   The expected profit will be USD 21851.15 and we are able to send to the 250 customers.