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?

   The company sent out its first print catalog last year and this year, it is preparing to send out a new catalog. In the current year, the company has acquired 250 new customers. Management has decided that catalogs will be sent out to these new customers if the profit associated with them exceeds $10,000. We need to determine the expected profit from these new customers so that management can decide whether the current year's catalog should be mailed to these new customers.

2. What data is needed to inform those decisions?

   We are required to calculate the expected profit for the new customers. In order to do that we need to predict the sales associated with these customers. To forecast sales, we create a regression model based on last year's customer data. See step 2 below for an explanation of this model. The projected sales are then multiplied by the probability that a customer would make a purchase Additionally, we account for the gross margin of 50% which applies to all products sold through the catalog. Finally, we subtract ($6.50) the cost of printing and distributing the catalog. To get the total expected profit we add the predicted profit across all customers.

Step 2: Analysis, Modeling, and Validation

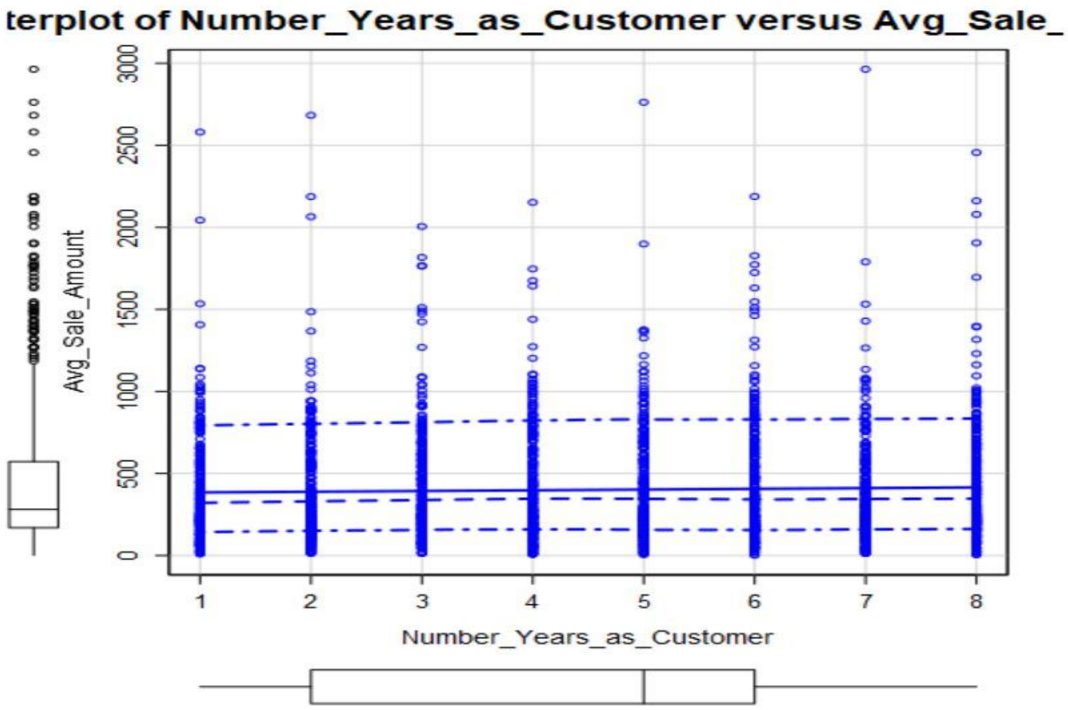
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

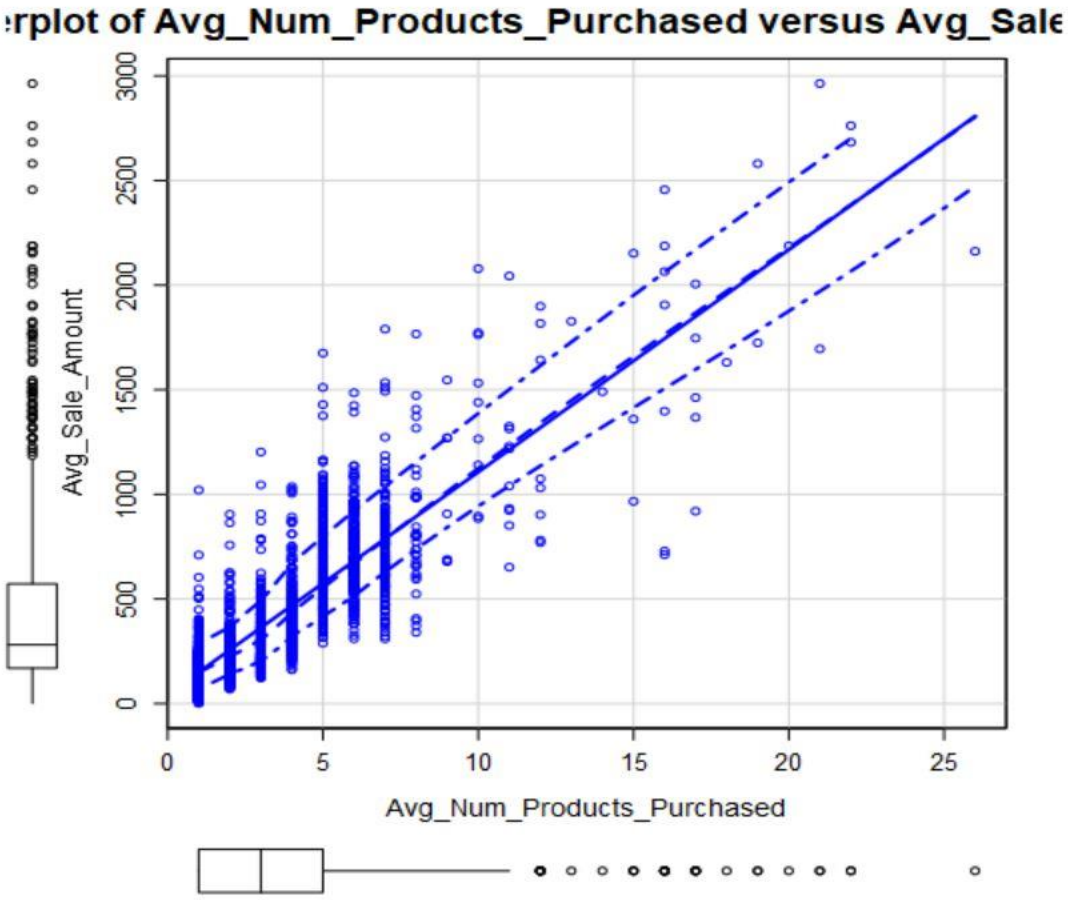
At the minimum, answer these questions:

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.

   After importing data into Alteryx from p1-customers.xlsx, I looked at the scatter plots to see the relationship between the target variable (Avg_Sale_Amount) and other numeric variables. First, I considered a scatter plot between Avg_Sale_Amount and Number_Years_as_Customers. I wanted to see whether customers who were longer with the company spent more. I didn't see a relationship between these two variables – see scatter plot below.



Next, I used Alteryx to create a scatter plot between Avg_Sale_Amt and Avg_Num_Products_Purchased. The scatter plot shows a positive relationship between the two variables – see below.



I didn't create scatter plots for the other numeric variables Customer_ID, ZIP and Store_Number as they have no relational impact on the target variable (Avg_Sale_Amount). For example, increasing or decreasing customer id will not impact sales.

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.

   Next I added the Linear Regression tool to the Alteryx workflow and selected the numeric predictor variable Avg_Num_Products_Purchased. The P value associated with it is very low – see chart below. In order to select the categorical variables, I selected one item at a time and ran the model. Then I checked the P values associated with the selected variable and the R squared values. The only categorical variable with a statistically significant P value was Customer_Segment. See below.

Report for Linear Model Old_Mailing_List

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

Residuals:

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 degrees of freedom (DF), p-value < 2.2e-16

   The low P values (< 2.2e – 16) indicate that there is a relationship between the predictor and target variables. Moreover, the high R squared (0.8366) depicts that variation in the target variable is explained by variation in the predictor variables. The 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)

   Y = 303.46 + 66.98 * Avg_Num_Products_Purchased -149.36 * (if Loyalty Club Only) + 281.84* (if Loyalty Club and Credit Card) – 245.42 * (if Store Mailing List) + 0 (if Credit Card Only)

Step 3: Presentation/Visualization

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

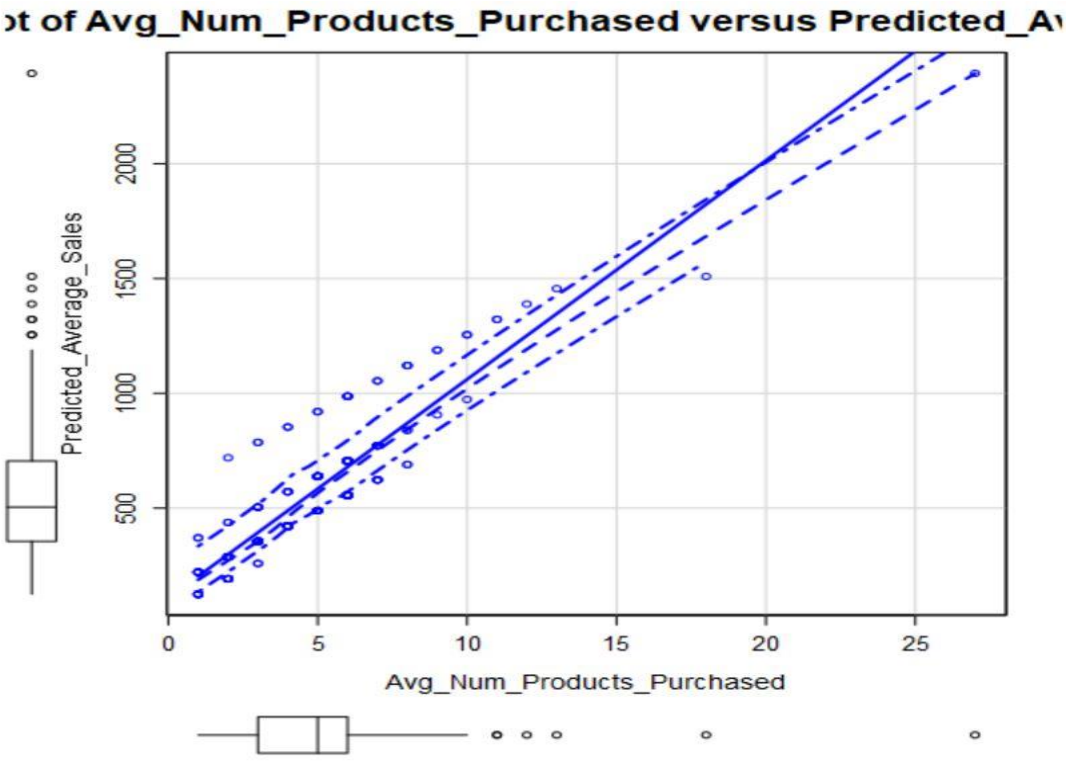
At the minimum, answer these questions:

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

   Yes, the company should send 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)

   After viewing results from the linear regression tool, I imported the data from p1 – mailing list and used the score tool to predict the target variable (Predicted_Average_Sales). I used the scatter plot tool to see the relationship between the Predicted_Average_Sales and Avg_Num_Products_Purchased. See below.



Then I added the formula tool to the workflow and incorporated three more fields:

Pred_Avg_Sales_Score = Predicted_Average_Sales * Score_Yes (probability that a customer would respond to a catalog and make a purchase)

Rev_gross_margin = Pred_Avg_Sales_Score * 0.5 (average gross margin on all products sold through the catalog is 50%)

Profit = Rev_gross_margin – 6.50 (cost of distributing and printing is $6.50 per catalog)

Finally, I added the Summary tool to the workflow to calculate the total predicted profit for these 250 customers = $21,987.44. The management required that the catalog should only be sent if the expected profit from these new customers is greater than $10,000. Since our prediction is more than double of $10,000, catalogs should be sent to these new customers.

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

   The expected profit from the new catalog is $21,987.44.