Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions need to be made?

   The management is supposed to decide whether or not to send this year's catalog out to the 250 new customers from their mailing list.
   The management wants to send the catalog out to these new customers only if it could result in generation of profit in excess of $10,000.

2. What data is needed to inform those decisions?

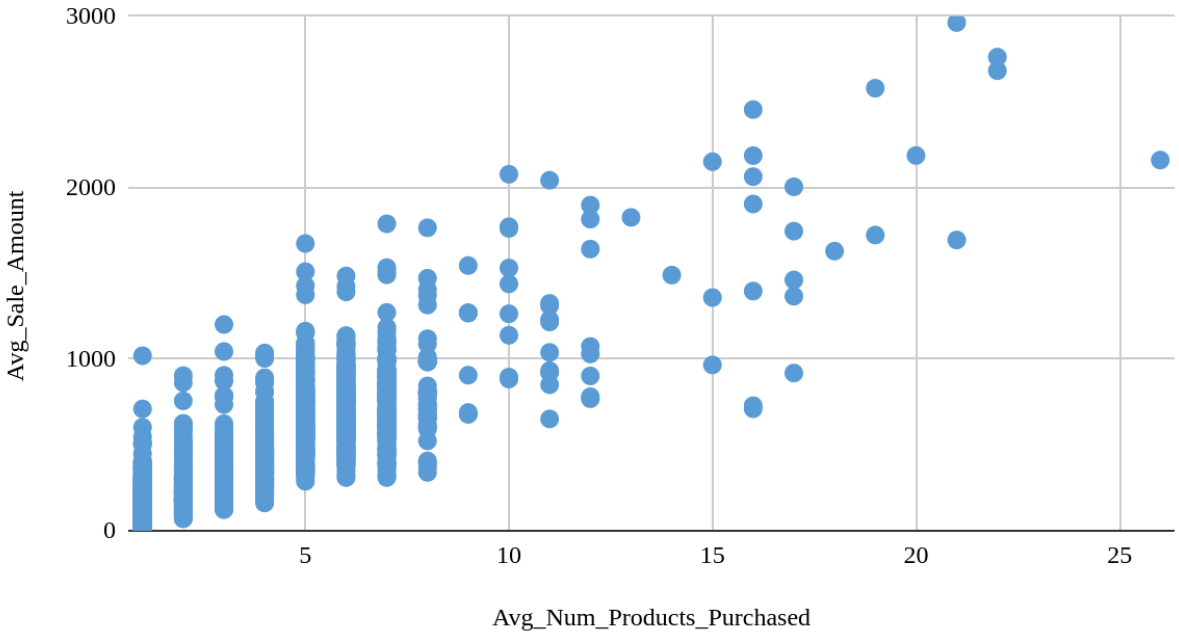
   We need to calculate the **expected revenue** from these 250 customers in order to get the **expected profit**. To do so we've to predict the average sales amount from each of the 250 customers and then multiply that amount with the probability that a customer is likely to make a purchase inorder to get the expected revenue. With the average gross margin being 50% we multiply the expected revenue by 0.5 then subtract $6.50 cost inorder to get the expected profit from a respective customer. We will then aggregate/sum up the profit amounts to find out if the total exceeds the $10,000 mark.

Step 2: Analysis, Modeling, and Validation

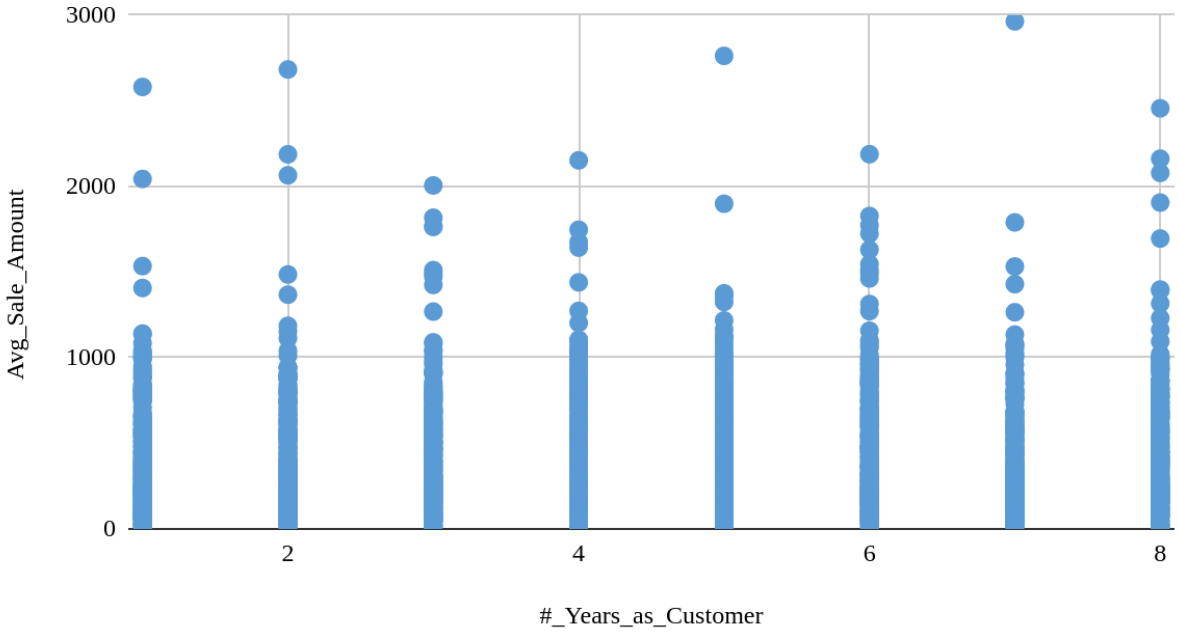
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatter plots to search for linear relationships. You must include scatterplots in your answer.

   In the p1-customers data there were two more continuous numerical variables(Avg_Num_Products_Purchased and #_Years_as_Customer) aside from the target variable(Avg_Sale_Amount). The following are the scatter plots of the target variable and the continuous numerical variables.

Avg_Num_Products_Purchased vs. Avg_Sale_Amount



#_Years_as_Customer VS Avg_Sale_Amount



The scatter plots depict that there's significant correlation between Avg_Sale_Amount and Avg_Num_Products_Purchased while there there's very little/insignificant correlation between Avg_Sale_Amount and #_Years_as_Customer variables.
Furthermore on using the Alteryx linear regression tool to create a model on a trial and error basis the p-value of Avg_Num_Products_Purchased as a predictor variable was much less than 0.05 indicating significant correlation with target variable whereas the p-value of #_Years_as_Customer variable was higher than 0.05.
Therefore #_Years_as_Customer was ruled out as a viable predictor variable.

For the remaining categorical variables I used the Alteryx linear regression tool to create models on a trial and error basis and emerged with the CustomerSegment dummy variables as the only categorical variables with p-values less than 0.05. The p-values are illustrated in the next section(Q2)
Therefore from analyzing the p1-customer data I ended up with Avg_Num_Products_Purchased and CustomerSegment dummy variables as my predictor variables.

I also utilized the Alteryx stepwise tool and fed all the variables into the linear regression and stepwise tools.
The stepwise tool suggested the use of Avg_Num_Products_Purchased, CustomerSegment dummy variables and also included #_Years_as_Customer as a predictor variable despite it having a p-value of higher than 0.05.

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.
The following are the p-values of the predictor variables that I included in the model

Predictor Variable	**p-value**
Customer_SegmentLoyalty Club Only	2.2e-16
Customer_SegmentLoyalty Club and Credit Card	2.2e-16
Customer_SegmentStore Mailing List	2.2e-16
Avg_Num_Products_Purchased	2.2e-16

The following are the R-Squared values of the models:
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366

Based on the p-values of the predictor variables and R-squared values we can conclude that this is a viable model that can be used to predict average sales amounts of the new customers.

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)

Equation:
Avg_Sales_Amount = 303.46 + Avg_Num_Product_Purchased x 66.98 + CustomerSegmentLoyalty Club Only x -149.36 + CustomerSegmentLoyalty Club and Credit Card x 281.84 + CustomerSegmentStore Mailing List x -245.42

Step 3: Presentation/Visualization

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

> I recommend that the company should send the catalog to the 250 customers because the predicted expected profit exceeds the $10,000 mark.

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

> I used past data of last year's customers to generate a linear regression model that can predict the average sales amount of the 250 new customers.
> Using the Alteryx linear regression tool I found out that only the Avg_Num_Product_Purchased and CustomerSegment dummy variables had high correlation with the target variable(Avg_Sales_Amount). Using these variables to generate a model resulted in a model with an Adjusted R_Squared value of 0.8366 that indicates it to be a viable model.
>
> Using the Alteryx score tool I applied the model to the data of the 250 new customers in order to get the expected Avg_Sales_Amount of each customer.
> Using the Alteryx function I then multiplied the average sales amount from each of the 250 customers with the probability that a customer is likely to make a purchase inorder to get the expected revenue.
> With the average gross margin being 50% I multiplied the expected revenue by 0.5 then subtracted $6.50 cost inorder to get the expected profit from a respective customer.
> Using the Alteryx Summarize tool I then summed up the profit amounts to get the total which resulted in $21,987.44.

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