Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   Determine how much profit the company can expect from sending a catalog to each of the 250 new customers. The expected profit should exceed $10,000 otherwise the catalogs should not be sent.

2. What data is needed to inform those decisions?

   A dataset containing information on 2300 customers, used to build the model.
   A dataset on 250 new customers used to predict sales.
   Score_Yes which is the probability the customer will respond to the catalog and make a purchase. The probability to buy (Score_Yes) will be multiplied by the output (predicted sales amount) of the linear regression model produced in Alteryx to provide the expected revenue.
   The costs of printing and distributing is $6.50 per catalog.
   The average gross margin (price - cost) on all products sold through the catalog is 50%.

Step 2: Analysis, Modeling, and Validation

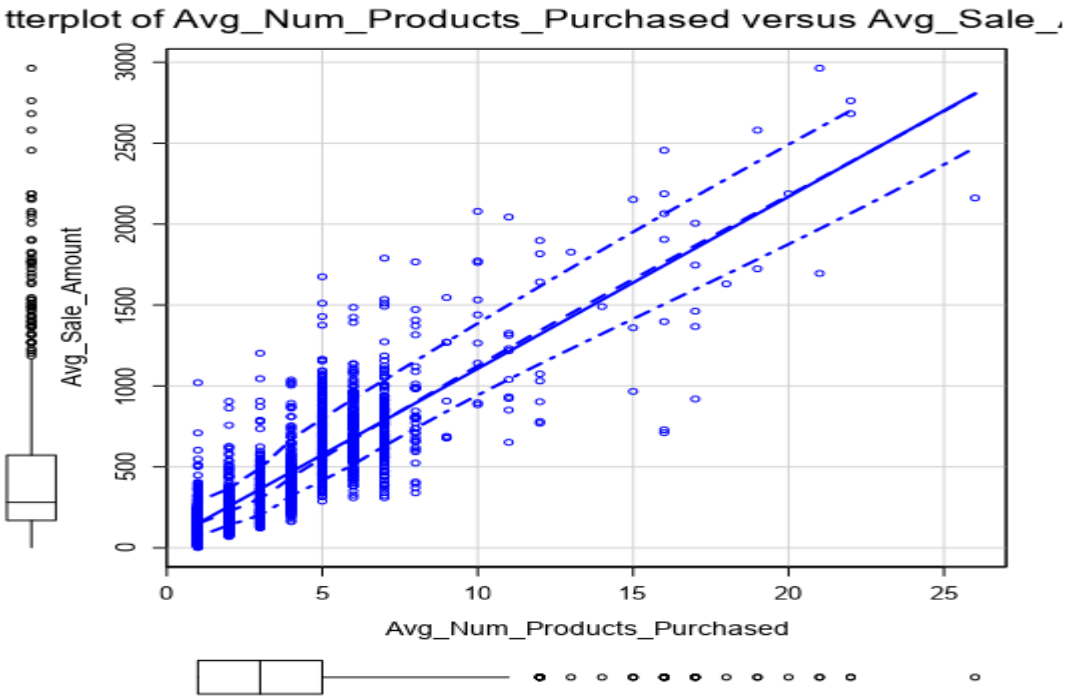
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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.

The continuous predictor variable (Average number of products purchased) was found to be the only numerical predictor variable to have a linear relationship with the target variable (Average sales amount).



The predictor variable X Years as Customer does not show a linear relationship with the target variable which suggests that this is a bad candidate as a predictor variable.

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 Alteryx linear regression model was used to find strong relationships between the predictor variables. The summary below shows that the most statistically significant P values are displayed by the variables Customer Segment and Average Number of Products Purchased. The P values for these predictor variables are below 0.05 and display three asterisks this means that it is highly unlikely that the two variables are not related.
The R squared value is 0.8369 and the adjusted R squared value is 0.8366 suggesting that the model is highly predictive (>0.7).

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + ZIP + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-667.24	-67.79	-2.83	71.00	973.11

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.685e+03	2149.8261	-0.7836	0.43334	
Customer_SegmentLoyalty Club Only	-1.503e+02	8.9708	-16.7525	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.825e+02	11.8968	23.7483	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.431e+02	9.8171	-24.7664	< 2.2e-16	***
ZIP	2.627e-02	0.0266	0.9873	0.3236	
Store_Number	-9.991e-01	1.0052	-0.9939	0.32036	
Responded_to_Last_CatalogYes	-2.870e+01	11.2709	-2.5467	0.01094	*
Avg_Num_Products_Purchased	6.679e+01	1.5151	44.0838	< 2.2e-16	***
X._Years_as_Customer	-2.325e+00	1.2216	-1.9033	0.05712	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.25 on 2366 degrees of freedom
Multiple R-squared: 0.8377, Adjusted R-Squared: 0.8372
F-statistic: 1527 on 8 and 2366 degrees of freedom (DF), p-value < 2.2e-16

Basic Summary

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

Residuals:

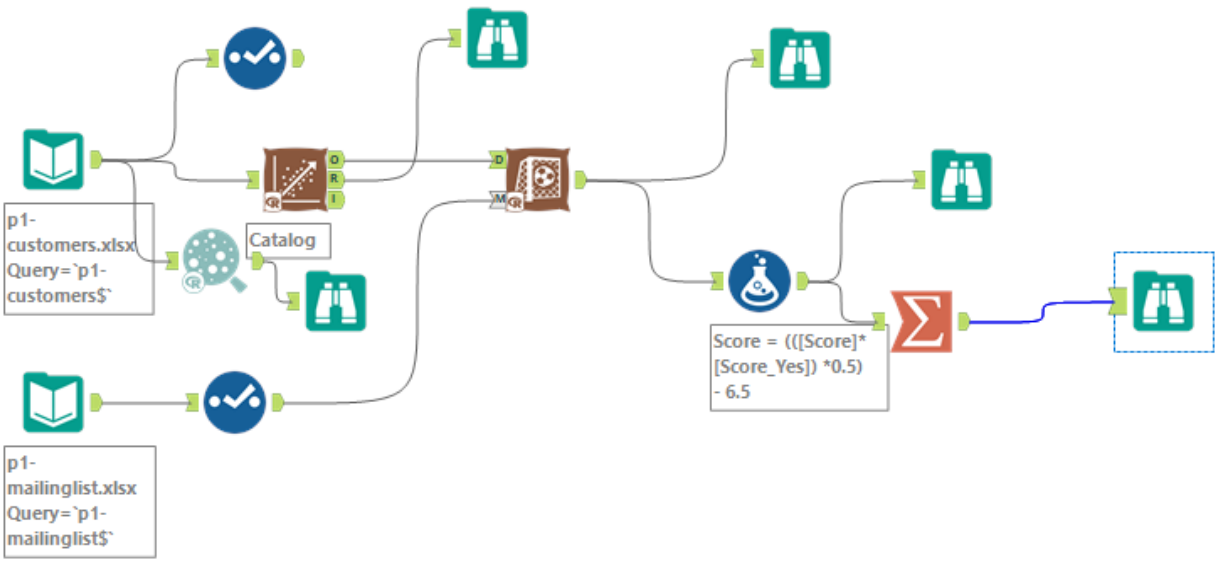
Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

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



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)

Intercept + b1*Avg_Num_Products_Purchased - b2(if Loyalty_club_only) + b3 (if Loyalty Club and Credit Card) – b4(if Store Mailing list) + b5 (If Credit card only)

Abby Pierson;
303.46 + (66.98 6) – (149.36*0) + (281.84*1) – (245.42*0) + (0*Credit card only) = 987.18*

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

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

The analysis shows that the catalog should be sent out to the 250 new customers as the expected profit is more than $10,000.

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

Using Alteryx to build a linear regression model I found two predictor variables, the numerical variable Average number of products purchased and the categorical variable Customer segment. These variables were used to predict the target variable Average sale amount for the

250 new customers. The output was multiplied by the probability of a customer making a purchase (Score_Yes) to give the expected revenue which was then multiplied by the gross margin (0.5). The cost of printing and distributing the catalog ($6.50 per catalog) was then subtracted for each customer and finally these values were added together to give a predicted average sale amount.

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

The expected profit is $21987.44.