Rohit

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?
    - How much profit can be earned by sending out a catalog to new customers, and management needs to make a decision on whether or not to send out the catalog to these new customers if the total expected profit exceeds $10,000.

2. What data is needed to inform those decisions?
    - Who the new customers are - the expected profit based on the probability of each customer responding to the catalog and their expected sales.
    - The amount of sales existing customers have purchased in order to train the model.
    - Other data such as customer segment and average number of products purchased to use as predictor variables in the linear regression model.

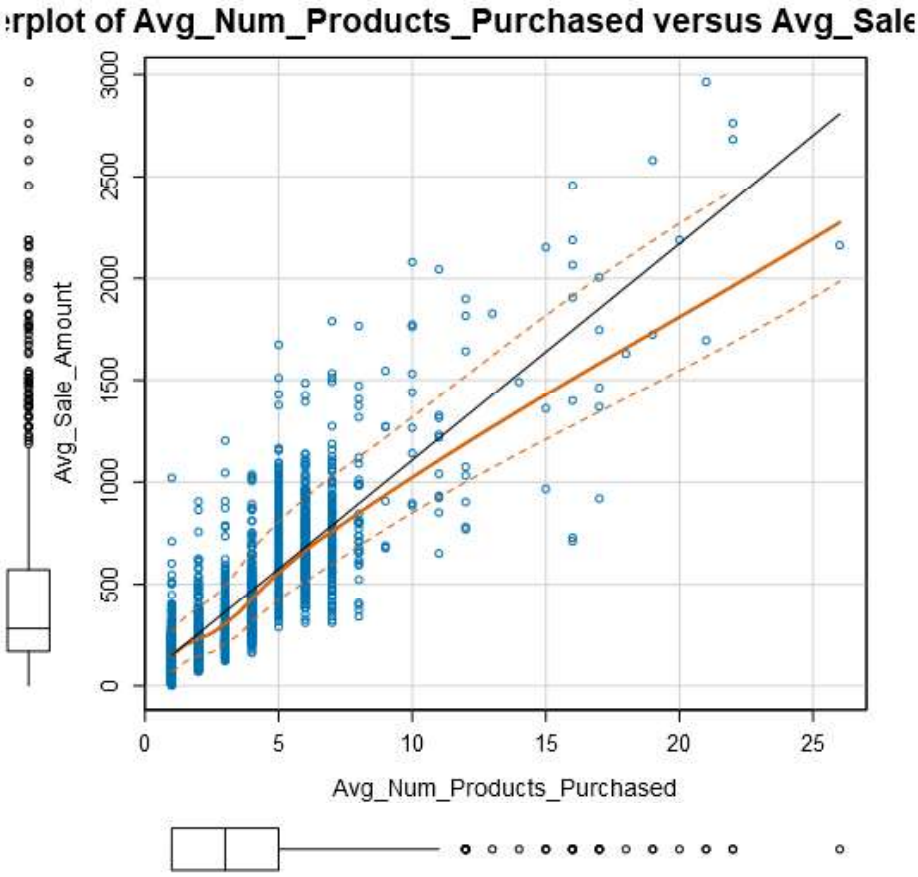
Step 2: Analysis, Modeling, and Validation

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)

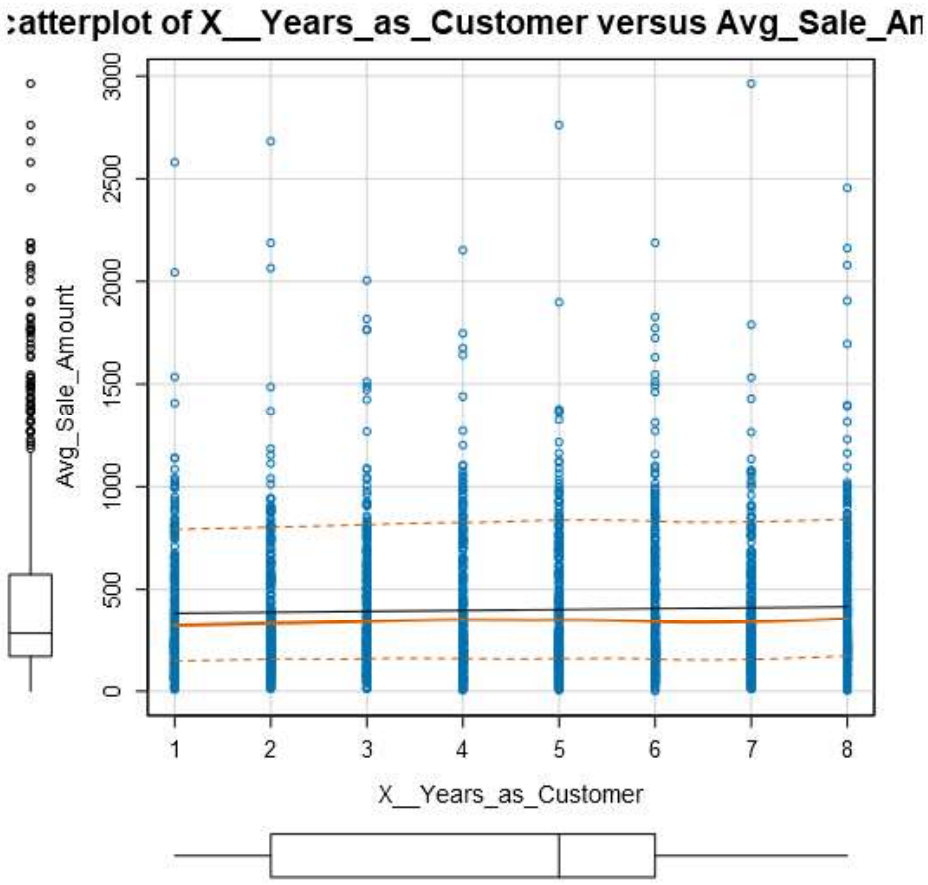
Important: Use the p1-customers.xlsx to train your linear model.

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.
    - In Alteryx, I used the "Scatterplot" tool on the p1 customers data set that is being used to train the model to look at which variables have a strong relationship with the average sales. I noticed that the "avg number of products purchased" generated a line of best fit with the average sales:



I also generated a scatterplot selecting the “years as a customer” variable to see how it is related to the average sales:



Based on the scatterplot, there doesn't seem to be as strong of a relationship between the years as a customer with the average sales as there was with the number of products purchased, so I can throw this variable out of my model. The only continuous variable I used in the model was the avg number of products purchased.

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.

- Here is the output from my linear regression model:

Record | Report

Report for Linear Model Linear_Regression_8

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$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 DF, p-value: < 2.2e-16

The customer segment and avg_num_products_purchased both have a very small p-value (less than .05), so I can infer that both of these variables are statistically significant in predicting the average sales per customer. My adjusted r-squared is .8366 which is close 1, so the model is showing that these variables can reasonably predict the average sales per customer.

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 - 149.36*customer_segment_loyalty_club_only + 281.84*customer_segment_loyalty_club_and_credit card - 245.42*customer_segment_mailing_list + 66.98*avg_num_products_purchased

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

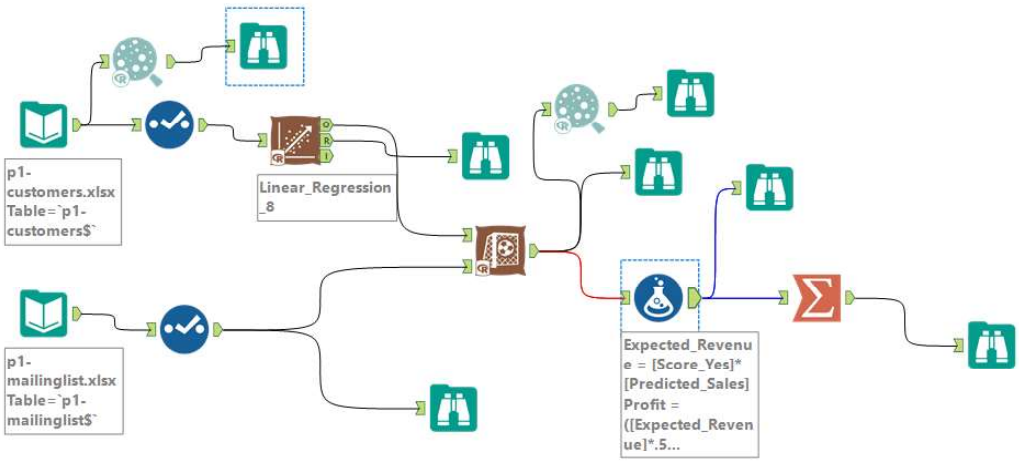
Step 3: Presentation/Visualization

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 these 250 customers because based on the predicted sales that the linear regression model outputted, the company can expect to generate $21,987.44 in profit which is more than the $10,000 minimum expected profit set by management.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process) I ran a linear regression model in Alteryx and selected avg_num_products_purchased and customer_segment as predictor variables and avg_sale amount as the target variable because we are trying to predict how much in sales sending a catalog to new customers would generate. I added a score tool to show what the predicted sales would be for each customer and used the formula tool to calculate the "Expected Revenue" which is Predicted Sales * Score_Yes which shows how much predicted sales the customer would generate based on the probability the customer would respond to the catalog. I added another formula for "Profit" which would be the ("Expected Revenue" * .5) - 6.5 cost per catalog. I added a summarize tool to sum up all the Profit rows to get the final expecting profit number management was looking for. Here is a screenshot of my Alteryx workflow:



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