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 key decision to be made is whether or not the company should send the catalog to the new customers. And to make that decision, the company needs to predict the expected profit of the 250 new customers then see if the profit is greater than $10,000.

2. What data is needed to inform those decisions?

First, the company should have data of both the new customers and historical data of existing customers. To get the expected profit, one would first need to know the sale amount of the existing customers. Then the company's gross margin (%) as well as the cost of printing the catalog. In addition to those, knowing the probability of the new customers making a purchase would help make the estimate more accurate.

Step 2: Analysis, Modeling, and Validation

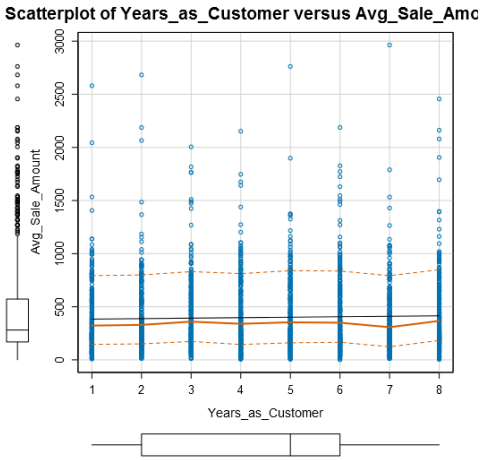
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

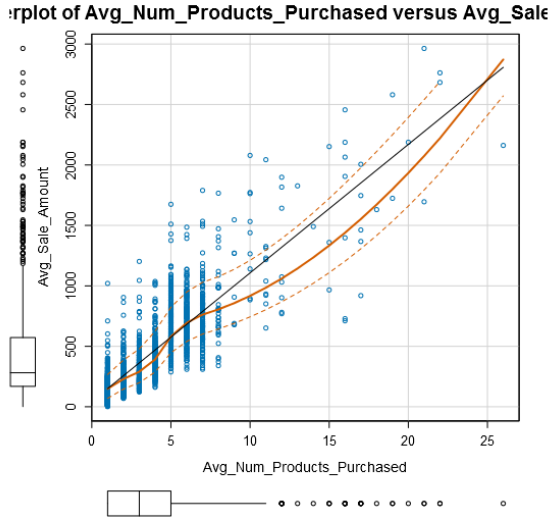
At the minimum, answer these questions:

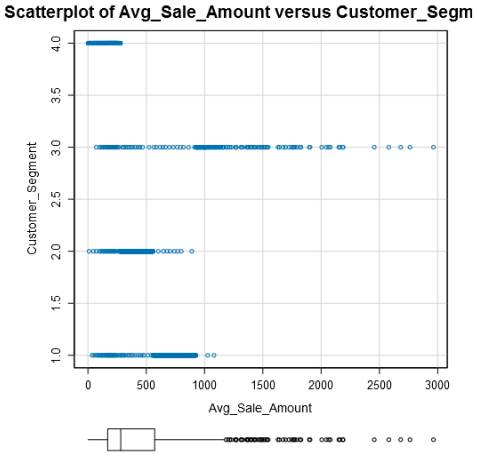
1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

I used 2 ways in selecting the predictor variables. The first way was by trying various combinations in the model to see what significance(p-value) it would give. The better the result the more likely I would use it for the next test. The second way was by looking at the relationship before inputting it in the model. I tried plotting several scatterplots to see which predictor variables had a relationship with the target variable Average Sale Amount. When I tried with variables such as store number, years as customer and customer ID, the scatter plot would show no correlation as seen below.



This is not the case when average number of products purchased is plot with the target variable as it gives a positive correlation. Customer Segment on another hand does not have a linear relationship but it does have a good p value when it is divided into the 3 categories(loyalty club only, loyalty and credit card , store mailing list) <2.2e-16





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.

   There were only 3 variables from the data that were statistically significant and those were customer segment, Average products purchased and Responded to Last Catalog. However, responded to last catalog variable could not be used as the mailing list data did not have the field. The P values of customer segment (loyalty club, loyalty club and credit card, Store Mailing List) and average products purchased are each <2.2e-16. They are noticeable lower than 0.05 and therefore make them reliable. The R-squared value is 0.8369 and the adjusted R-squared value is 0.8366. As one can see the R-squared value does not have a very high explanatory power but it is the best one could get with the data obtained. If possible I would ask for more data just so I could make the model more significant.

	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

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)

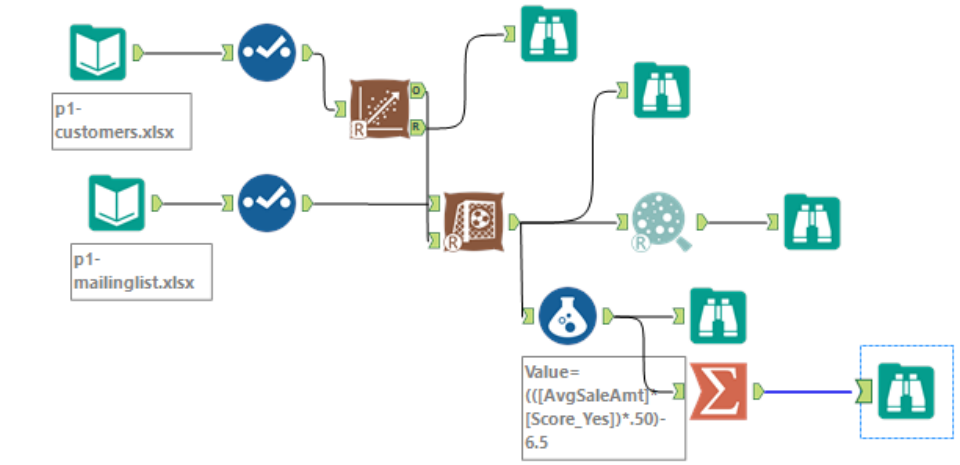
Avg Sale Amount= *303.46+66.98*AvgProductsPurchased-149.36**(If Type: *LoyaltyOnly)+281.84**(If Type: *LoyaltyAndCreditCard)-245.42**(If Type: *StoreMailingList)*

Step 3: Presentation/Visualization

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

Based on the result of the model, I believe the company should send the catalog to the 250 customers as the expected profit exceeds $10,000 with the estimated profit being $21987.44. The first thing I did was try to understand the problem. Then I tried to understand the data, what I had and what I didn't. Then I tried several tests on the regression model as to see what data I could make use to achieve a good result. Once I was satisfied with the result, I applied the model to get the Average Sale Amount of each of the 250 new customers. I multiplied the Average Sale amount to the probability that the customer would purchase after which I multiplied the result to the gross margin and subtracted the cost of printing and distributing the

catalog. Then to get the total profit I simply summed up the profit of all the 250 customers. Profit = (([AvgSaleAmt]*[Score_Yes])*.50)-6.5. For simplicity, I added the Alteryx workspace with the final result screenshot below.



ılts - Browse (20) - Input

1 of 1 Fields | Cell Viewer | 1 record displayed, 825 bytes

Record #	Sum_Value
1	21987.435687