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?

Is it worth it to send the catalog out to these new 250 customers? The threshold for a good profit is $10,000.

2. What data is needed to inform those decisions?

We need several numeric and categorical variables from our previous customers, to build the model, focusing on the previous sales and the relationships with our store (loyalty cards). We need to calculate the expected revenue from these 250 people in order to get the expected profit. After building a linear multiple regression model from our customers and getting the predicted money spent for each of the 250 people, we should multiply it by the probability of that person to actually buy from our store. The next step is to add up all those amounts and calculate the benefit we obtain from that purchase: 1) multiply by that gross margin of 50%, price minus cost; and 2) subtract the expenses for printing the catalogs ($6.50 * 250 people).

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.

First, it is worth mentioning that I had to rename the variable #_Years_as_Customer as Years_as_Customers without the "#_".

I explored with scatterplots the relationship between the average sale amount and the continuous numeric variable avg_Num_Products_Purchased (Figure 1) and Years_as_Customer (Figure 2), respectively. The first variable is the only one that shows a non-zero linear relationship.

Afterwards, in the linear regression model it is seen that the p-value is much smaller than 0.05 (2.2 e-16).

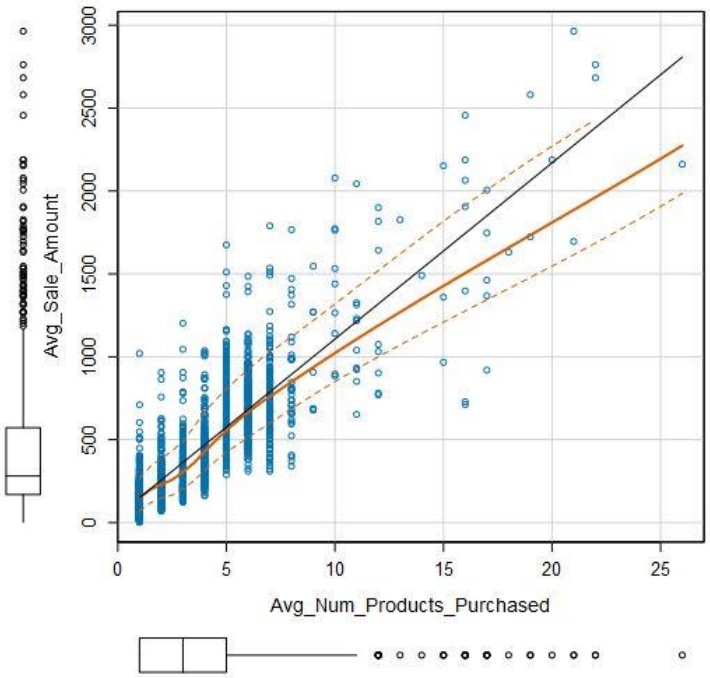

Figure 1. Scatterplot of average sale amount versus average number of products purchased. There is a positive linear relationship (positive slope).

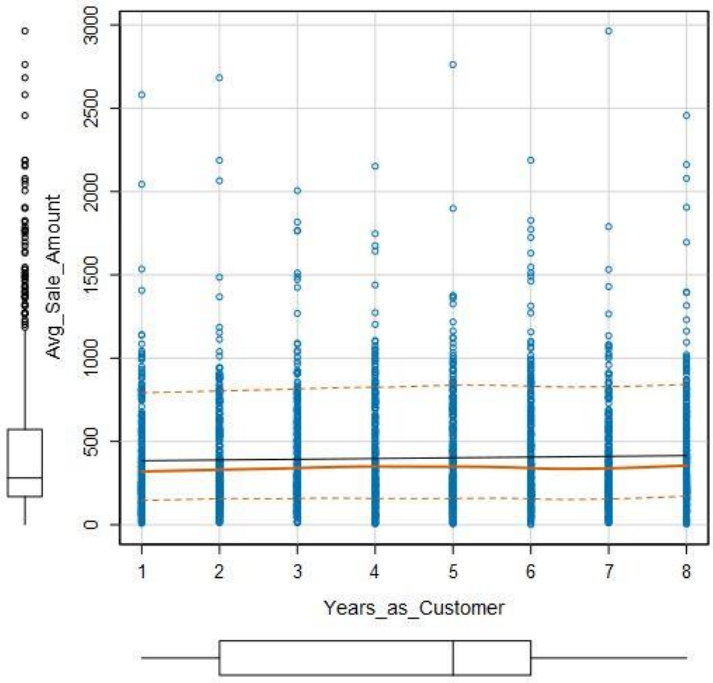

Figure 2. Scatterplot of average sale amount versus years as customer. There is almost no relationship between the variables, as it is indicated by a slope close to zero.

Then I explored the categorical variables Customer_Segment, with four different categories, in the linear model, adding this variable made the model to increase in R-squared value, so it is a good variable to include.

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.

According to the results in Figure 3 the model has a good R-squared value (>0.7), concretely the adjusted R-squared value (used for linear multiple regression) is 0.8366 (in a blue rectangle). So the 83.66% of the variance is considered by the model. All the coefficients have a significant p-value (< 0.05), very close to zero, as dashed in green. There is a very high probability that there is a true relationship between the predictor variable and each target variable, and that did not occur by chance. Considering only the numeric variable I got a R-squared value of 0.7323, so in combination with the categorical variable custom segment more variance in the data is captured.

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

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28715078.96	3	506.4	< 2.2e-16	***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16	***
Residuals	44796869.07	2370			

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

Figure 3. Coefficients and statistics for the linear multiple regression model, inputting the average number of products purchased and the customer segment.

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)

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.

The regression equation, according to the Coefficients table in Figure 3 is

Profit = 303.46 - 149.36 * Customer_SegmentLoyalty Club Only + 281.84 * Customer_SegmentLoyalty Club and Credit Card - 245.42 * Customer_SegmentStore Mailing List + 0 * Customer_SegmenCredit Card Only + 66.98 * Avg_Num_Products_Purchased

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?

My recommendation is that the company should send the catalog to these 250 people, because the expected profit is more than the double than the $10,000 wanted, concretely $21,987.44.

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

First I thought of the variables from the csv files that made sense to include in the model. I selected the numeric variable (Avg_Num_Products_Purchased) for the model because it had a positive linear relationship with the customers average sale and excluded the years as customers because it did not have a relationship. I built a linear regression model and got an R-squared value of 0.7323. Then, I entered this variable and the categorical value customer_segment to the linear multiple regression which made the R-squared value to increase in 0.1, so I kept both variables in the

model (although the categorical variable was divided into 3 dummy variables). Afterward, I got the predicted money spent for each of the 250 people and multiplied it by the probability of that client to buy from our story. Then I added the quantities for each of the 250 people, multiply it by the gross margin (50%, price minus cost) and subtracted the expenses for printing the catalogs ($6.50 * 250 people). The Alteryx schematic is shown in Figure 4.

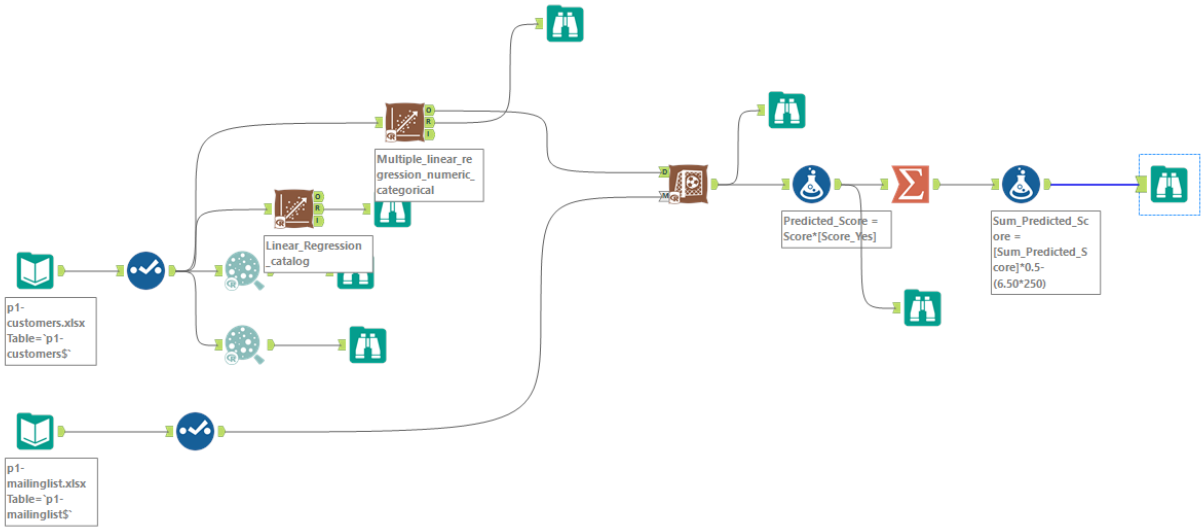

Figure 4. Schematic of the steps to get the expected profit from sending the new catalog.

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.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

Extra

Figure 5 shows that most people purchased between 0-4 products, and some of them 5-10 products. In Figure 6 we can see that most people had the loyalty club card, meaning that loyalty and confidence in this store is a main characteristic for a good profit.

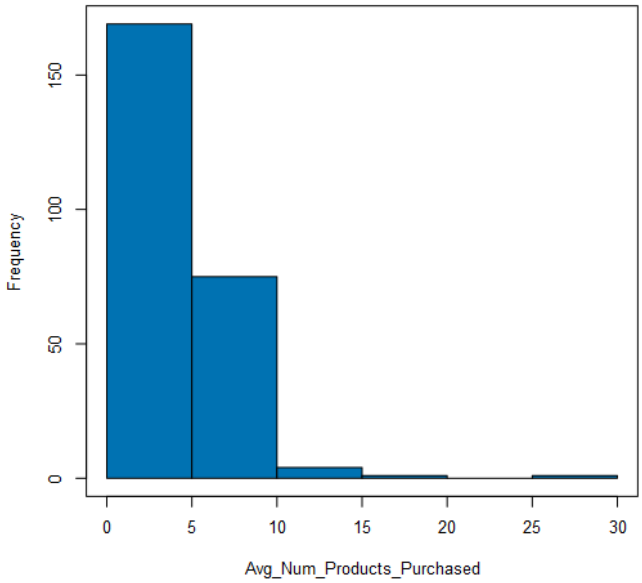

Figure 5. Histogram of the average number of products purchased.

Value	Frequency	Percent
Loyalty Club Only	122	48.80
Credit Card Only	82	32.80
Loyalty Club and Credit Card	26	10.40
Store Mailing List	20	8.00

Figure 6. Frequency table of the customer segment.