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 decision needs to be made is, "Is the company going to send this year's catalog to the new (250 customers) which will be done by predicting the expected profit it will generate by doing this, which must be more than $10,000"

2. What data is needed to inform those decisions?

We need to calculate the average sales amount by using both of the average number of products purchased and the customer segment. Then, we need to multiply it with the probability that a customer will buy, in order to get the predicted sales per customer. Once we have this information, we aggregate the predicted sales of each customer to get a total of predicted sales which eventually we can use to predict the expected profit of sending out these catalogs.

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.

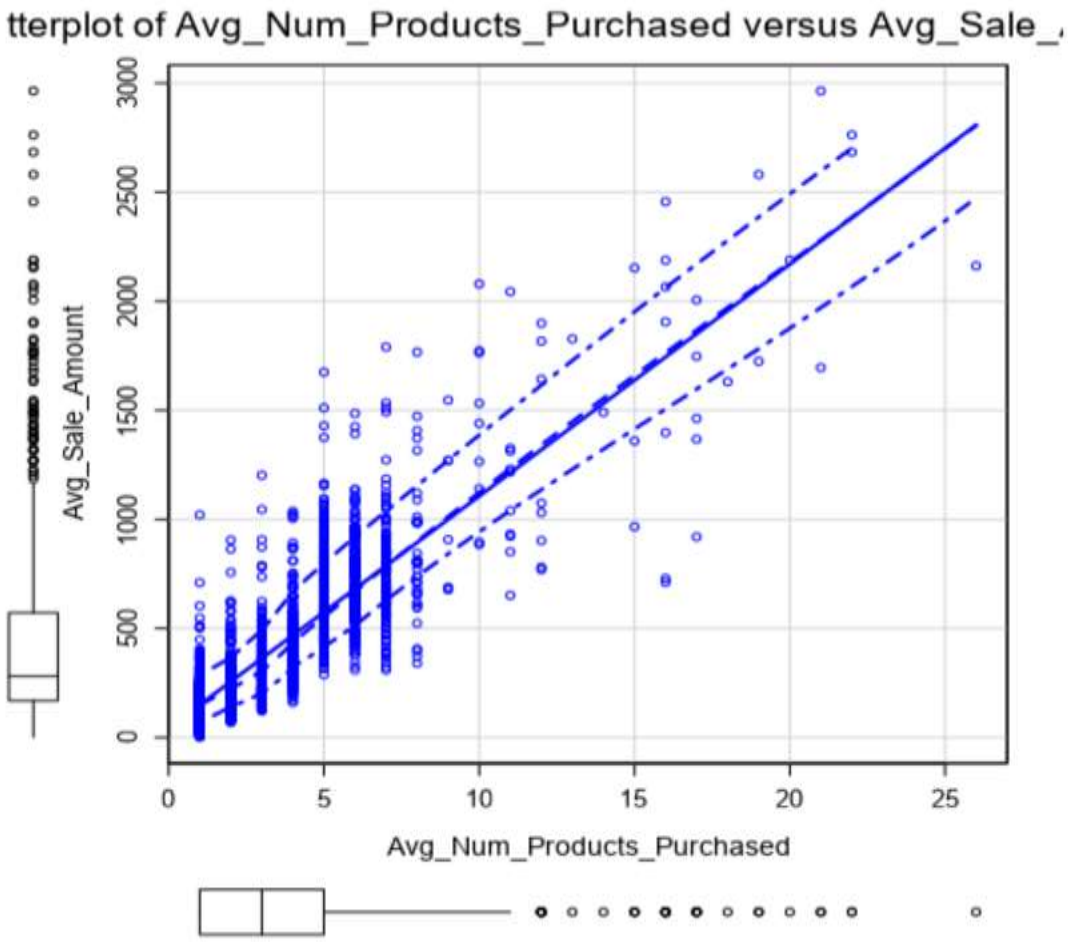
At the minimum, answer these questions:

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.

In order to create a linear regression model, first we need to understand the relationship between the predictor variables and the target variable. I began by using an input tool to bring the data from p1-customers Excel spreadsheet. Next, I needed to use a scatterplot in order to know which variables can be a good predictor variable for the target variable. Then, by using a linear regression tool, I built the model. After that, by using the Score tool I reached out to the target variable.

I chose the predictor variables by doing a scatterplot between the target variable and all the numeric individual variables, and by assessing the p-values for the categorical variables (which must be <= 0,05). By doing this, I found out that Avg_Num_Products_Purchased, Customer_Segment are the good predictor variables for the target variable which is the Avg_Sale_Amount.

First, the scatterplot of the Avg_Num_Products_Purchased and the Avg_Sale_Amount variables has a slope (As shown in the graph below). Therefore, there is a linear relationship between them which indicates that it is a good predictor variable for this target variable.



Second, the p- value of the customer segment categorical variable, as shown below, is less than 0.05 and since it has 3 asterisks, it is statistically significant variable. So, it is a good candidate for a predictor variable.

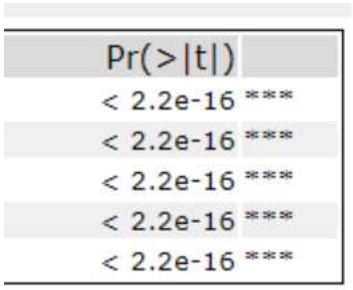
Coefficients:

	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

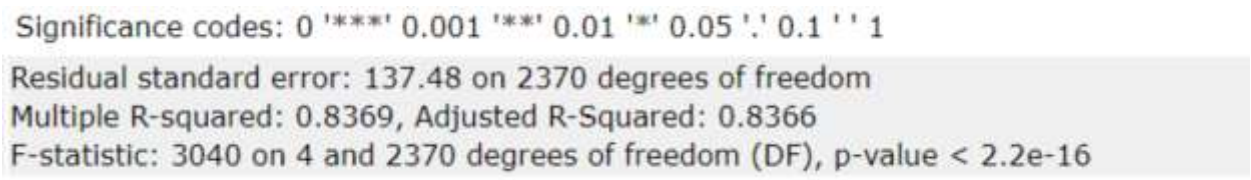
2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

The p-values and R-squared values indicate how well the linear model is. First, as you can see in the following figure, the p-values of the selected predictor variables (which are the Avg_Num_Products_Purchased, Customer_Segment) in the report is less than 0.05. Therefore, it represents that there is a relationship between the predictor variables and the target variable and it is statically significant since they have more than one asterisks.



Pr(>|t|)	
< 2.2e-16	***
< 2.2e-16	***
< 2.2e-16	***
< 2.2e-16	***
< 2.2e-16	***

Second, as you can see in the following figure, the R-squared values range from 0.8369 to 0.8366 which shows the amount of variation that represents a high explanatory of the model.



```
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
```

Therefore, the low P-values and high R-squared values indicates that the model is highly predictive.

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.

As you can see in the following figure, the coefficients (the b's) are the estimates.

Coefficients:

	Estimate
(Intercept)	303.46
Customer_SegmentLoyalty Club Only	-149.36
Customer_SegmentLoyalty Club and Credit Card	281.84
Customer_SegmentStore Mailing List	-245.42
Avg_Num_Products_Purchased	66.98

Therefore, the regression equation is:

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

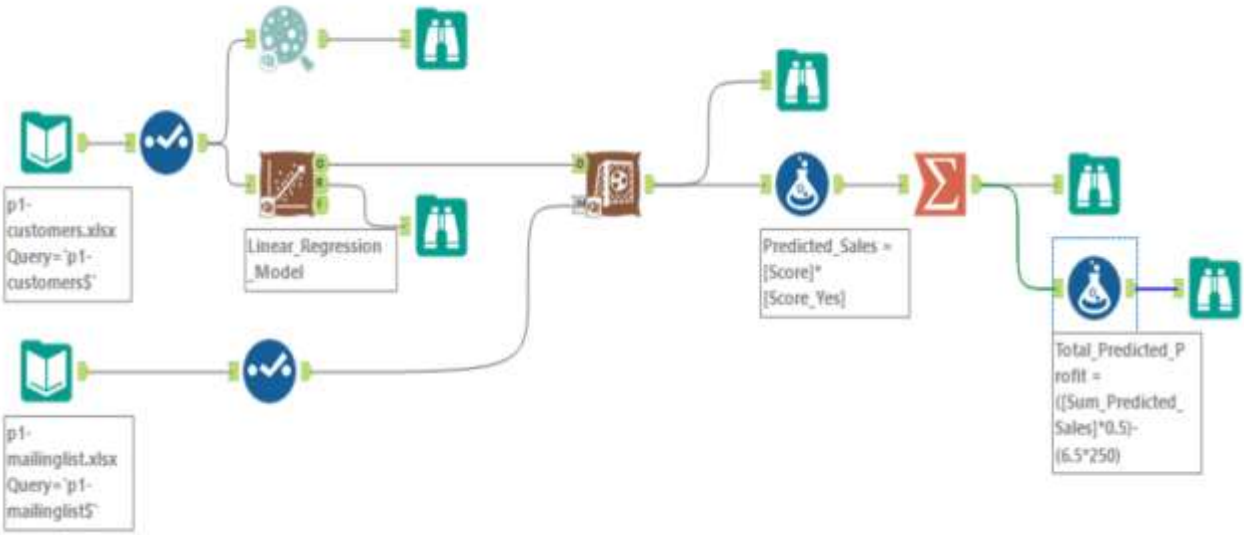
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

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

First, I need to calculate the Predicted_Sales for every customer by multiplying [Score] by [Score_yes], where [Score_yes] is the probability that every customer will buy. Then, in order to get the total predicted sales for all of the 250 customers, I have to make a summation of the

Predicted_Sales. Next, in order to get the gross margin, I multiply the [Sum_ Predicted_Sales] by 0.5. After that, I subtract the total expenses of the catalogs for 250 customers ([Sum_ Predicted_Sales *0.5] -(6.5* 250)). By doing this, I got the Total predicted profit which is: 21,987.4356865455. Therefore, I recommended the management to send the catalogs to the 250 new customers, since the predicted profit will be more than $10,000.

Alteryx's workflow:



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.