MAKE A COPY

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?

Whether catalogs should be sent to the 250 new customers, or not. A profit is expected ($10,000) from selling catalog products to these customers. If profit is below $10,000, catalogs would not be sent.

2. What data is needed to inform those decisions?

We require historical and new data regarding existing clients and their spending habits. These include: number of items bought, loyalty status, the amount spent, state, their inclination to place an order, profit gross margin, cost of catalog printing and distribution.

Based on some of the variables, we will start our analysis and notice if there is any connection between the predictor variables and the target variable.

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)

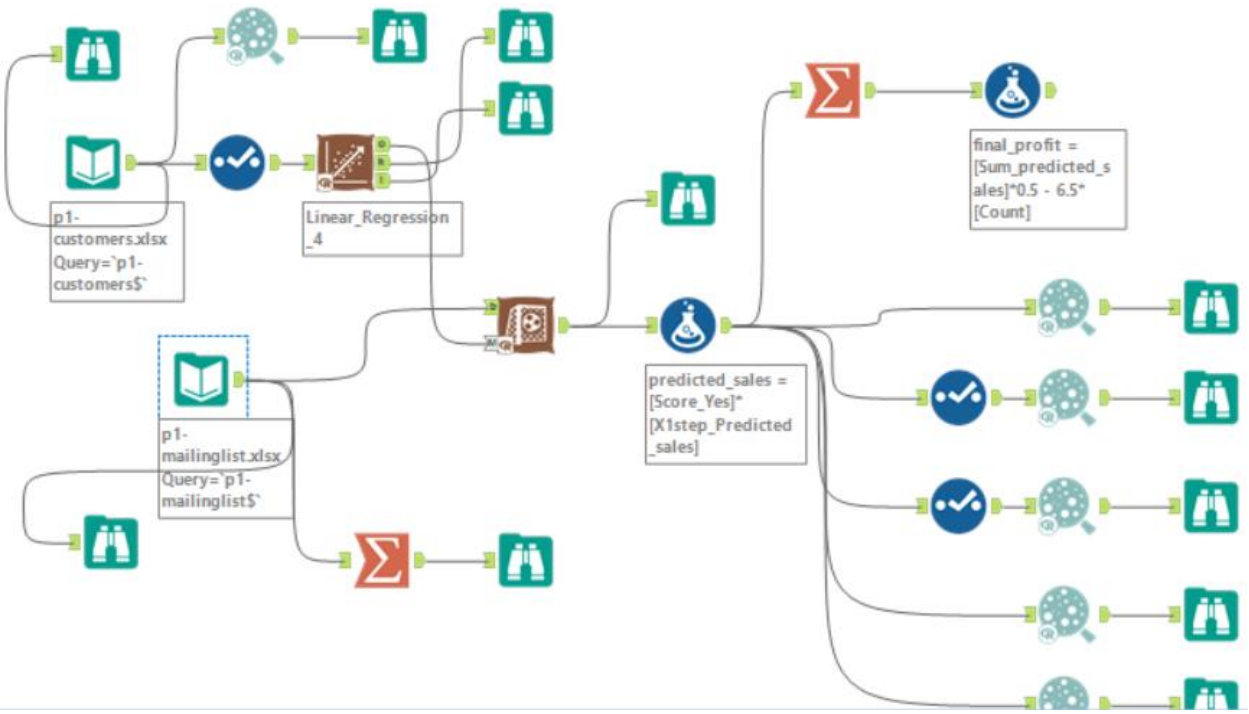
I used Alteryx to establish the analysis and modeling + validation.

Initially, my instinct was to deliberately eliminate some of the variables from the start. However, in this way, some data might slip our analysis and provide a model which is not exhaustive and foolproof.

As a result, I am introducing every variable to notice if there is any data that confer useful information to strengthen the model.

Without forgetting our scope, I am trying to find the variables that hold a connection between what are trying to predict and the variables taken into account.

Alteryx model



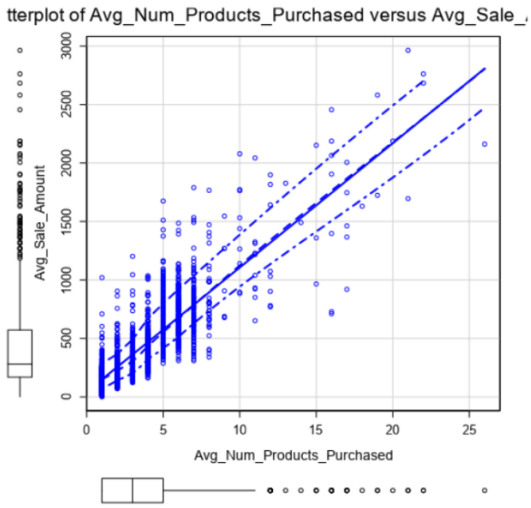
Important: Use the p1-customers.xlsx to train your linear model.

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.

From the start we can observe a linear relationship between the average number of products purchased and the average sale amount.

By applying what I have learned in this course, I ran the model multiple times to find the best variables for the prediction.



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.

In Alteryx, the confidence given by a predictor variable is represented by an asterisk (*). All the variables kept have 3*.

As they have a p-value < 0,05, the chosen predictor values are:

- Customer segment – which is a categorical variable
- Average number of products purchased

Any variables that had a p-value > 0,05 were deducted from the model as they reduce the precision.

R-squared and the adjusted R-squared, both display a high value of 0,84 which gives me confidence in the model following the linear regression and the formula deduced. While the values might be high, there is still a percentage of 16% which cannot be explained by the model. Therefore, additional data is required to explain the variation in the target variable.

All the analyzed stats compose a model which can successfully predict the data we need: the average sales as a result of sending the catalogs to the new customers.

Report for Linear Model Linear_Regression_4

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

Variable	Estimate	Impact	Confidence
(Intercept)	303		★★★
Customer_SegmentLoyalty Club Only	-149		★★★
Customer_SegmentLoyalty Club and Credit Card	282		★★★
Customer_SegmentStore Mailing List	-245		★★★
Avg_Num_Products_Purchased	67		★★★

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 = Intercept + **b1** * Customer_SegmentLoyalty Club Only + **b2** * Customer_SegmentLoyalty Club and Credit Card + **b3** * Customer_SegmentStore Mailing List + **b4** * Avg_Num_Products_Purchased + **0** * Customer_SegmentCredit_Card_Only

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_SegmentCredit_Card_Only

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

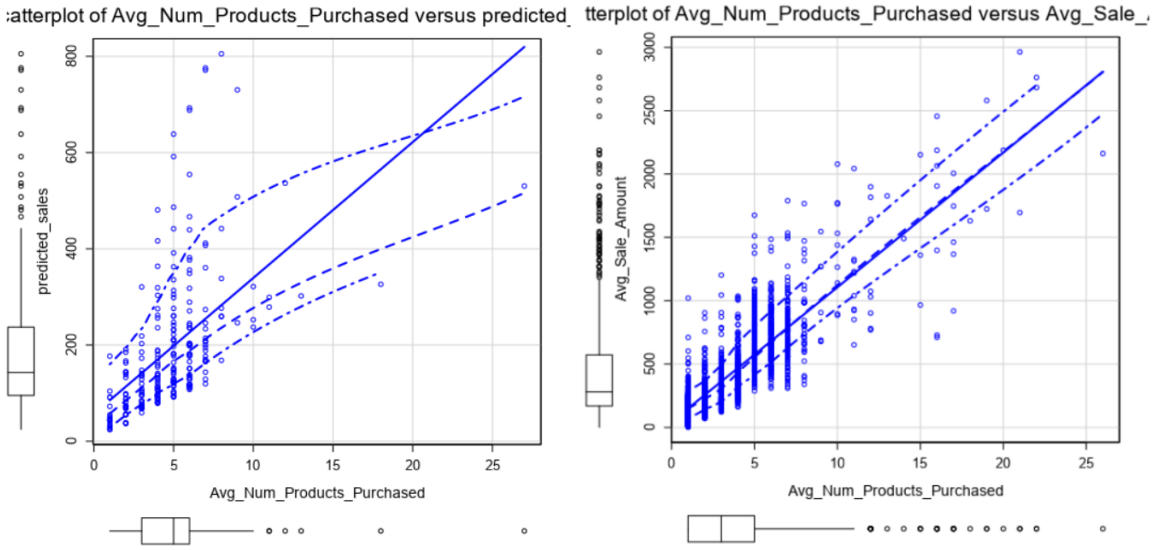
Use your model results to provide a recommendation. (500 word limit)

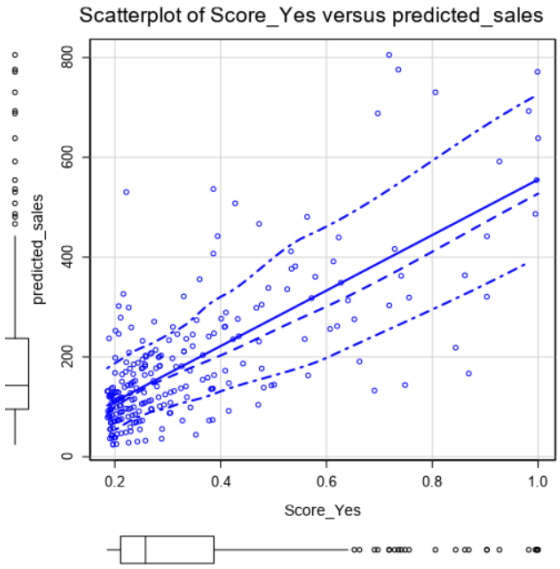
At the minimum, answer these questions:

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

Absolutely! Our model displays characteristics of a well-crafted statistics predictor.

By comparing the predicted graph (left) with the historical data (right), we can clearly see a connection as the trending lines follow the same behavior and the sales are direct proportional with the number of products sold.





Graph: the predicted sales per customer in accord with their inclination of buying (1 = yes; 0 = no)

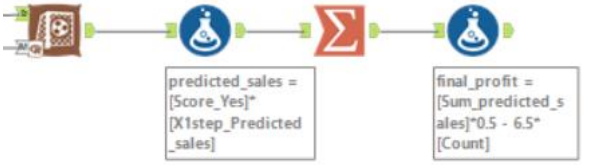
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

The model is trustworthy, so we can continue the final calculation.

I summed up all the predicted average sale per customer and took into consideration:

- the gross margin of 50%
- a $6,5 cost per catalog
- the probability of a person to buy from the catalog (score_yes)

The result is above manager's starting limit of $10.000.



Sum_predicted_sales	Count	final_profit
47224.871373	250	21987.4356865455

Output Column	Data Preview
final_profit	21987.4356865455

```
[Sum_predicted_sales]*0.5 - 6.5*[Count]
```

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

The expected calculated profit is $21,987.43.

Thank you for the thorough feedback! It has helped me a lot.

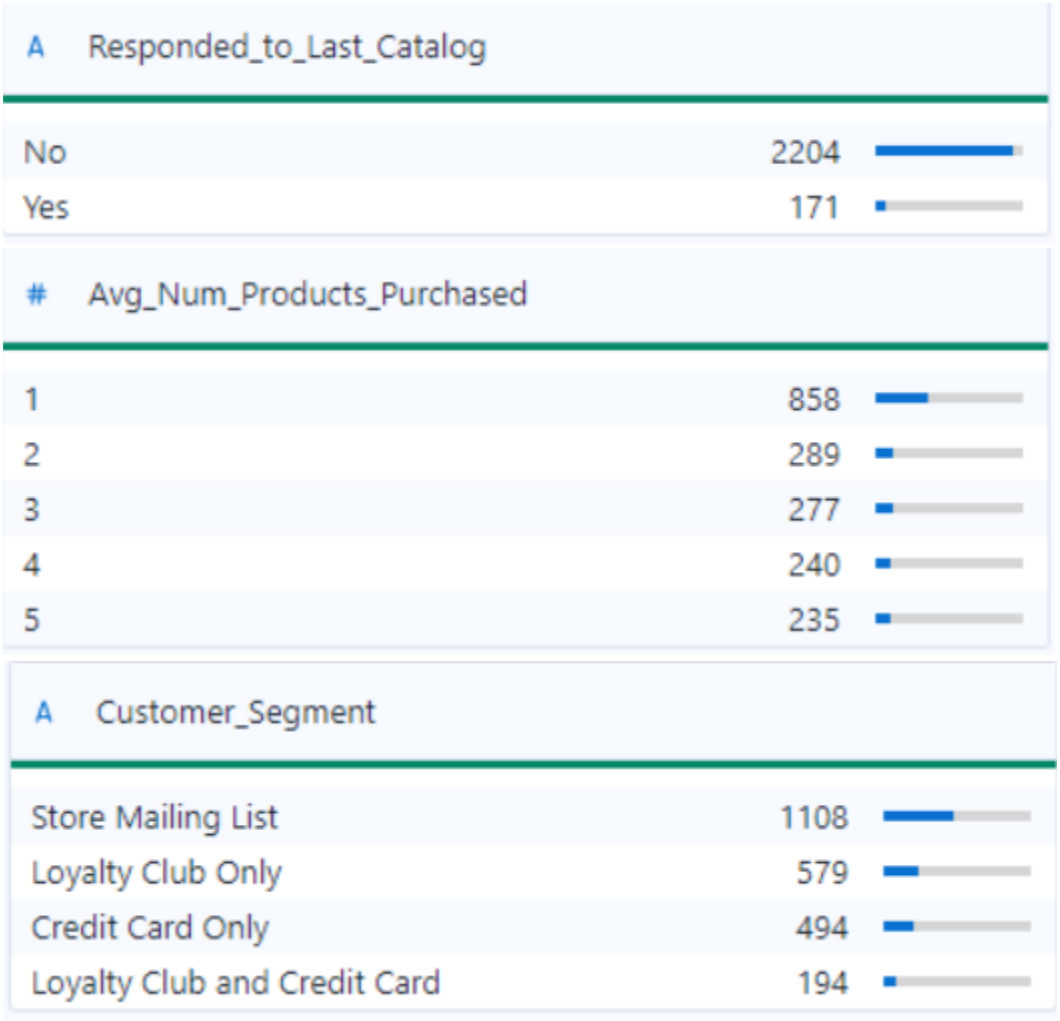
I had noticed something fishy with the last calculation but could not put my finger on the issue.

Additional thoughts:

While we added the 30% chance of buying (from) the catalog, ‘customers’ dataset shows an inclination towards NOT responding to the last catalog.

Although we may expect a profit, more data would help make a better model.

Also, most of the clients buy only 1 product and they might not know of all the products the company sells – innovative marketing might work to promote the products (besides catalogs).



A Responded_to_Last_Catalog	
No	2204
Yes	171

# Avg_Num_Products_Purchased	
1	858
2	289
3	277
4	240
5	235

A Customer_Segment	
Store Mailing List	1108
Loyalty Club Only	579
Credit Card Only	494
Loyalty Club and Credit Card	194

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.