MAKE A COPY

Project 1: Predicting Catalog Demand

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

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 manager needs to know whether the company should send this year's catalog to its 250 new customers from their mailing list.
    - The business analyst needs to decide how much profit the company can expect from sending a catalog to the 250 new customers and makes sure it’s more than $10,000.

2. What data is needed to inform those decisions?
    - The records of customers that received the last year catalog. (p1-customers.xlsx)
    - The records of the 250 new customers that the new catalog will be sent to. (p1-mailinglist.xlsx)
    - How the company calculates the profit
    - The costs of printing and distributing
    - The average gross margin

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)

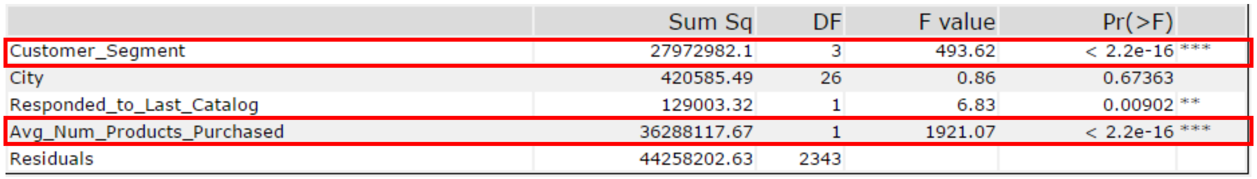
Important: Use the p1-customers.xlsx to train your linear model.

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

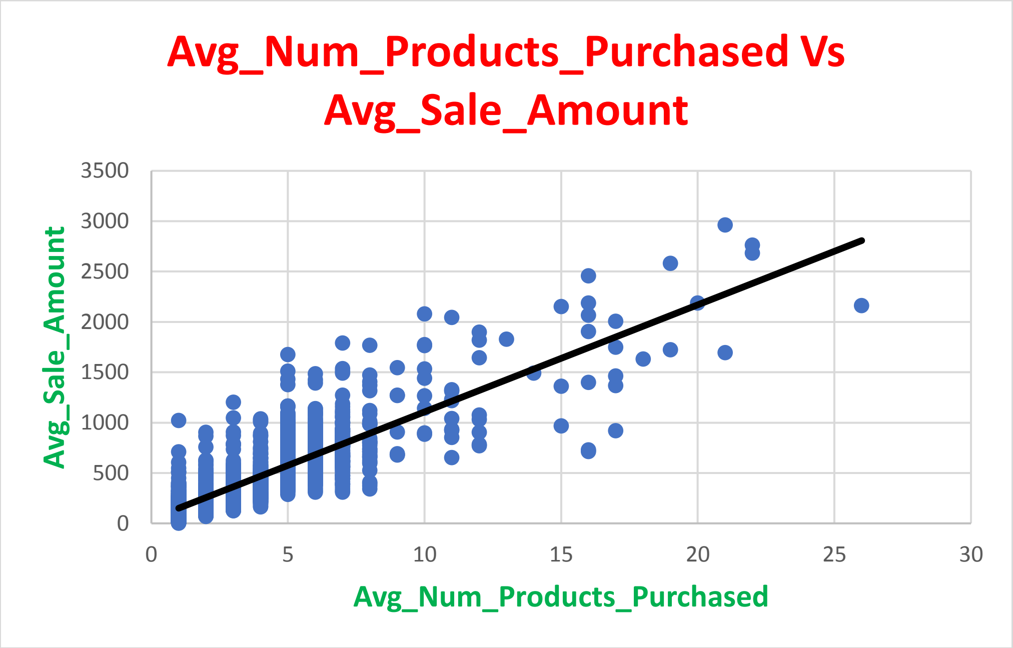
and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

- The selected continuous predictor variables are: Customer_Segment and Avg_Num_Products_Purchased.
    - They were selected because they have low p-values (< 2.2e-16) and they were considered to be the most significant than other variables.



	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	27972982.1	3	493.62	< 2.2e-16	***
City	420585.49	26	0.86	0.67363	
Responded_to_Last_Catalog	129003.32	1	6.83	0.00902	**
Avg_Num_Products_Purchased	36288117.67	1	1921.07	< 2.2e-16	***
Residuals	44258202.63	2343			

    - The Avg_Num_Products_Purchased variable has a linear relationship with the target variable, as shown in the below scatterplot.

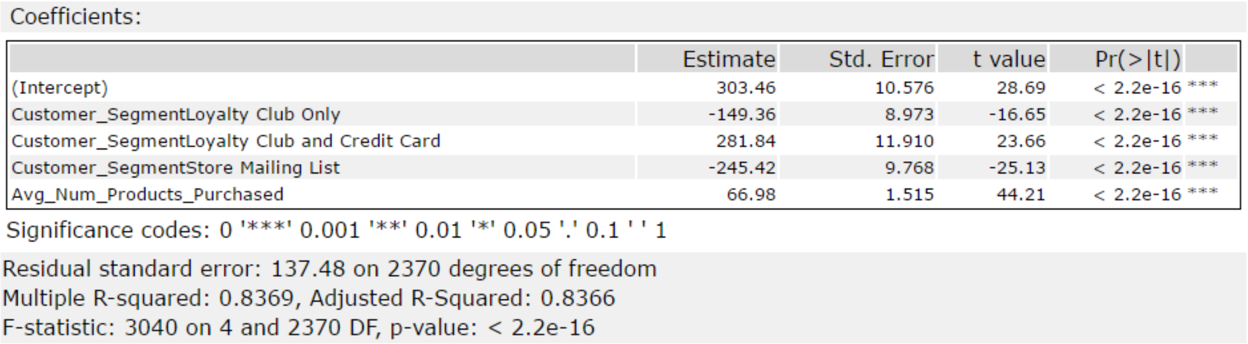


    - The Customer_Segment variable couldn’t be represented by a scatterplot as an independent variable, since it’s a categorical variable.

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.

- This linear model has a high adjusted R-squared value (0.8366), which is an indication of high explanatory power of the model.

- Both the Customer_Segment and Avg_Num_Products_Purchased variables are considered to be the most significant with low p-values (< 2.2e-16) .



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

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.

Y = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149.36 (If Customer_Segment:Loyalty Club Only) + 281.84 (If Customer_Segment:Loyalty Club and Credit Card) - 245.42 (If Customer_Segment:Store Mailing List)

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, I recommend the company to send this year catalog to the 250 new customers.

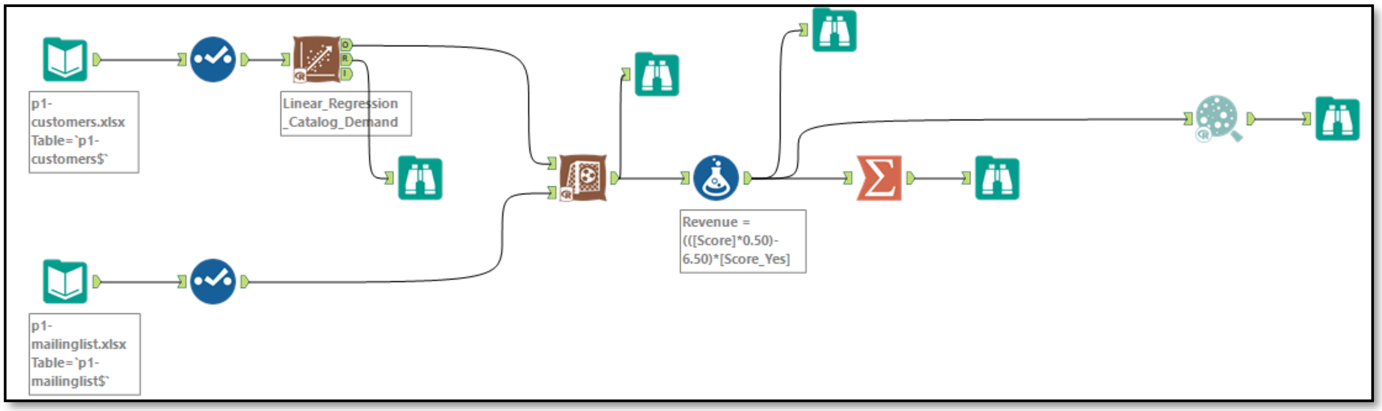
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

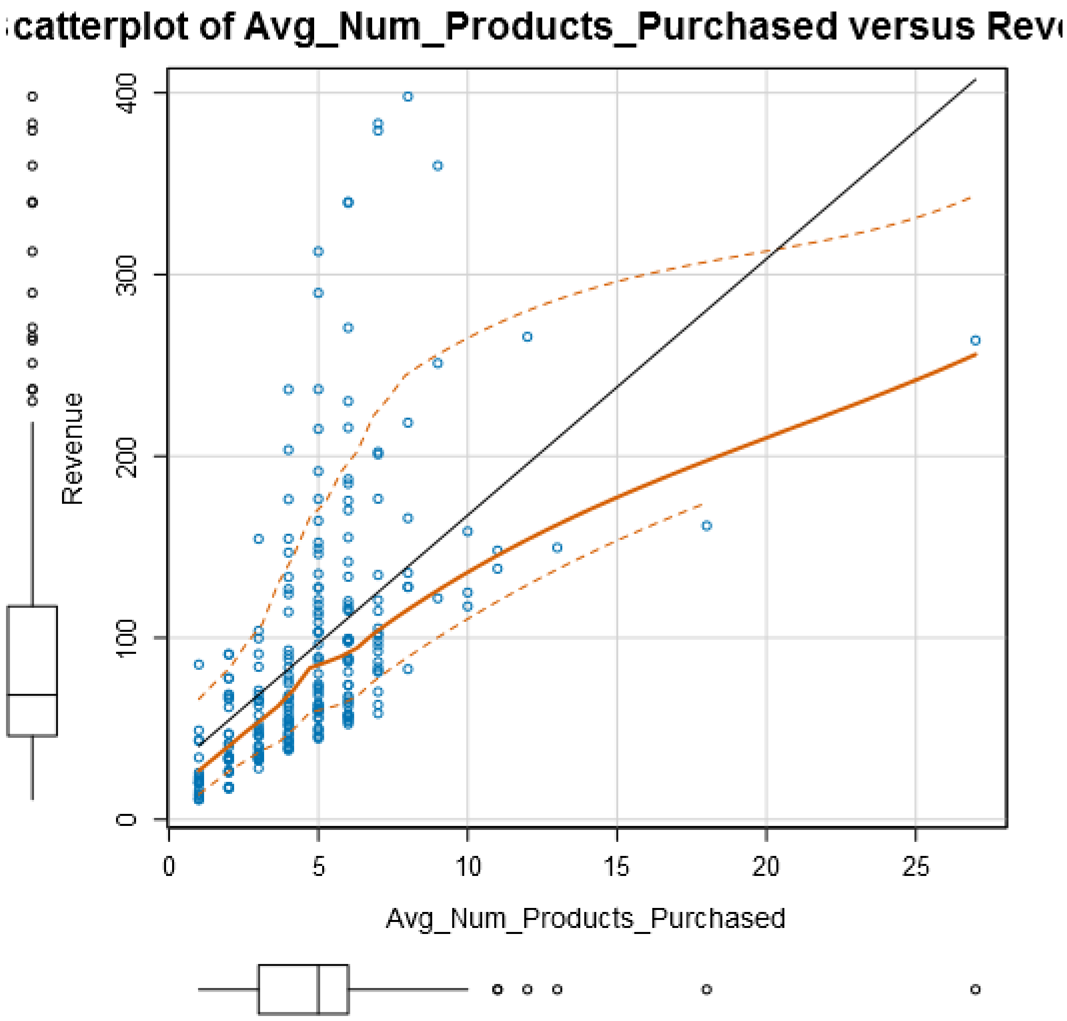
Alteryx software was used to complete this task. The followed steps were:

1. Specify the predicator variables (Customer_Segment and Avg_Num_Products_Purchased) using scatterplot and linear regression tools
2. Create the linear regression model and verify it's a good model by the reported statistical results (p-values and R-squared)
3. Deploy the linear regression model on the test data (p1-mailinglist.xlsx) using Score tool
4. Calculate the customer’s revenue using Formula tool, the used formula is

   Revenue= ([score] * [score_yes] * 0.5 - 6.5)

5. Calculate the expected profit by adding all the revenues of customers using Summarize tool
6. Represent the result using scatterplot tool


p1-customers.xlsx Table=`p1-customers$`
Linear_Regression_Catalog_Demand
O
R
I
Revenue = (([Score]*0.50)-6.50)*[Score_Yes]
p1-mailinglist.xlsx Table=`p1-mailinglist$`



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

The expected profit is **$21987.435687**

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.