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?
2. What data is needed to inform those decisions?

Answer:

Question 1.

The aim of this analysis is to make decision about whether or not we sent a catalog to 250 new customers in the company. This decision will be judged based on the expected profit will be generated from these new customers. Company will be sending the catalog if the expected profits exceed $10,000. So, we need to predict the potential profits for each new customer to help with that decision. This can be done by performing linear regression with Average Sale Amount as a target variable.

Question 2.

To make the decision, some data are needed to do this analysis. These includes:

- Cost related for preparing the catalog
- Average gross margin on all products sold through the catalog
- Historical data about previous sales
- Information related to new customers

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.

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.

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)

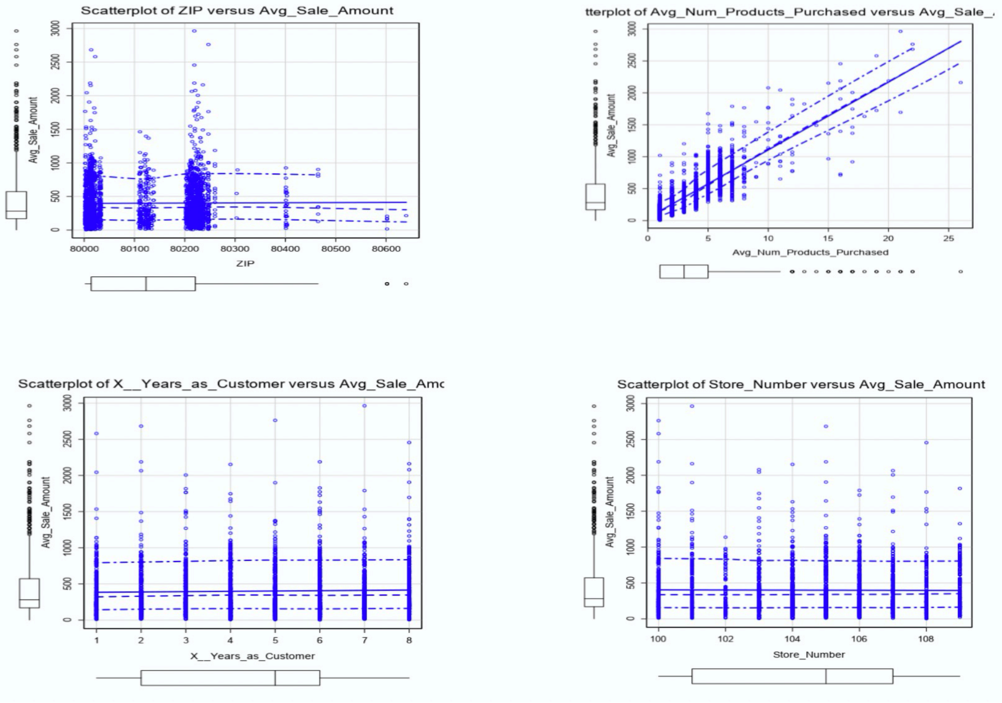
Note that we **must** include the 0 coefficient for the type Cash.

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.

Answer:

Question 1.

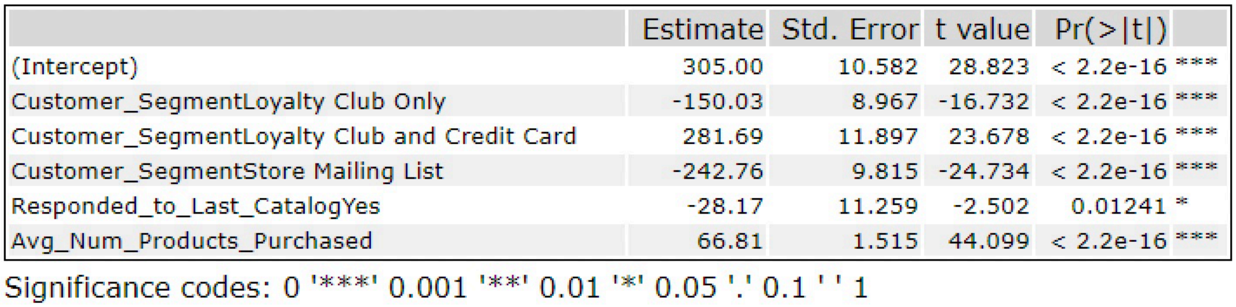
There are two steps that I took to find good candidate of predictors. First, I make a scatterplot of all numeric variables compared to our target variable. By doing this we can check which variable can be a good candidate for a predictor variable



Looking at the scatterplot, the variables that can be a good candidate as a predictor is only Avg_Num_Products_Purchased. This variable has positive correlation with Avg_Sum_Amount.

For the categorical variables, I checked these variables by try it on linear regression along with Avg_Num_Products_Purchased. So, I use Customer. I didn't use Address, City, and State because those already represented by ZIP code in our numerical variables.

The result of this linear regression are as below:



	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	305.00	10.582	28.823	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.03	8.967	-16.732	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.69	11.897	23.678	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.76	9.815	-24.734	< 2.2e-16	***
Responded_to_Last_CatalogYes	-28.17	11.259	-2.502	0.01241	*
Avg_Num_Products_Purchased	66.81	1.515	44.099	< 2.2e-16	***

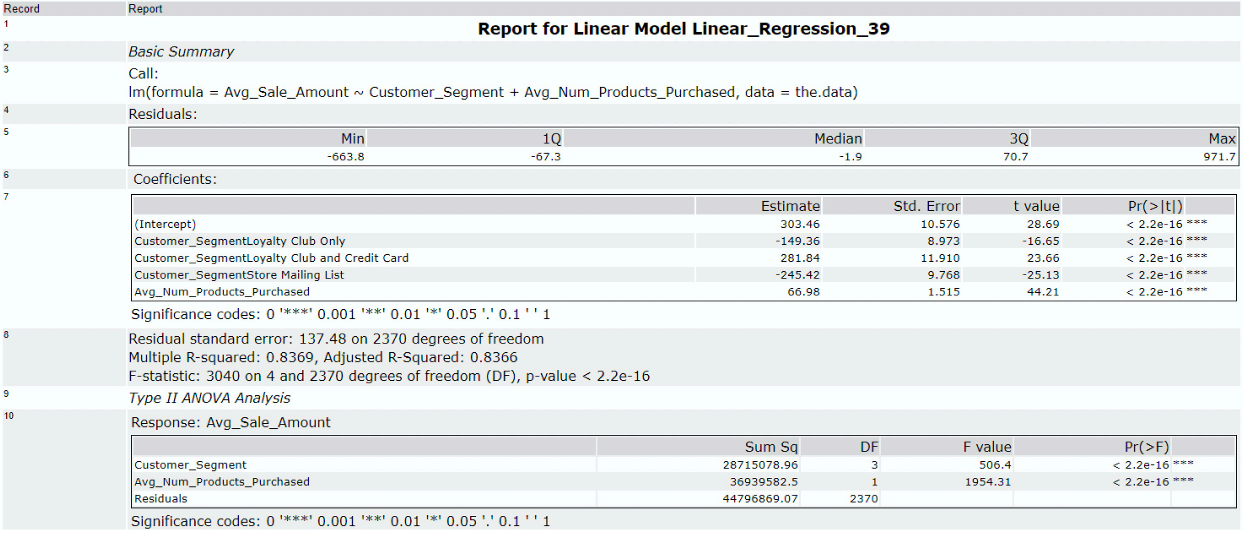
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Looking at the P-Values from above tables, I pick Customer_Segment as our predictor because it highly statistically significant with lower P-Values (<2.2e-16).

So that we have two variables that will be used to make the linear regression equation.

1. Avg_Number_Products_Purchased
2. Customer_Segments

Finally the results of this linear regression can be seen below:



Record	Report
1	**Report for Linear Model Linear_Regression_39**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 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

8 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

9 *Type II ANOVA Analysis*

10 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

Question 2.

This linear regression is a good model because we have chosen the predictor variables that have statistically significant with the target variables. This can be seen from max p value which less than $2.2x10^{-16}$. Moreover, this model also have both R-squared and Adjusted R-Square 0.837 which above 0.7 as a sign of good fit.

Question 3.
The result of this modelling is this equation:
Predicted_Avg_Sales = 303.46 -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) + 0 (if Customer_Segment: Credit Card Only) + 66.98*Avg.number_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?

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)?

Answer:

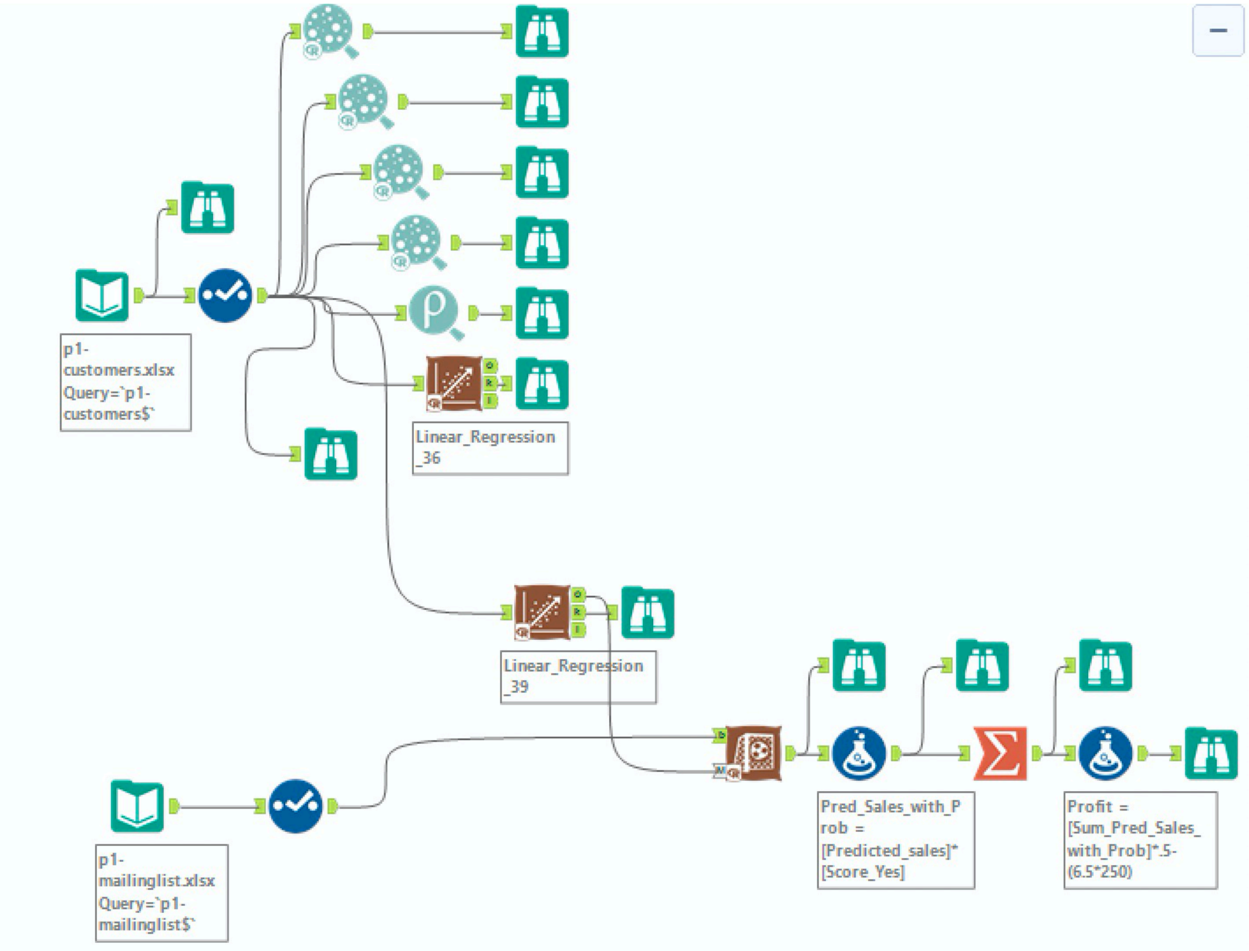
Question 1.
My recommendation is company should send the catalog to these 250 customers.

Question 2.
This recommendation based on the calculation of predicted profit from these 250 customers. First, I predict the expected Average Sales Amount of each these customers using the linear regression formula that we create before. Second, I multiply predicted sales for each customer with the probability of this customer will make purchase (Score_yes). Then totaling all those values to get the predicted revenue which is 47224.87. Last step was calculated expected profit by multiply the revenue with margin 50% and subtract that value with the cost of catalog and get 21,987.43 for expected profits. This profit are higher than the minimal profit that management want which is 10,000. So, it is recommended to send the catalog.

Question 3.
The expected profit from the new catalog is 21,987.43.

APPENDIXES:

Alteryx Design:


p1-customers.xlsx Query=`p1-customers$`
Linear_Regression_36
Linear_Regression_39
Pred_Sales_with_Prob = [Predicted_sales]*[Score_Yes]
Profit = [Sum_Pred_Sales_with_Prob]*.5-(6.5*250)
p1-mailinglist.xlsx Query=`p1-mailinglist$`

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.