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)

There is mailing list of 250 new customers and we need to ascertain how much profit company will make by sending catalogues to these 250 customers based on the data provided. Management wants to send catalog to these new customers only when the profit is greater than $10,000

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   a) Is the predicted profit more than $10,000 if catalog is sent to 250 customers
   b) Should the catalog be sent to new 250 customers (if predicted profit is more than $10,000 answer is yes, else No)

2. What data is needed to inform those decisions?

   First, we need to make sure that data is clean.

   We need to know the predicted average sales if we send catalog to 250 customers. We can predict average sales based on the last year's data from the existing customers. We know the actual sales from the previous year. We need to select the predictor variables which are relevant to target variable (predicted average sales).

   We have the probability that the new 250 customers will respond to the catalog and make a purchase. Based on the probable average sales, we can calculate gross margin which is 50% of this value. Profit will be gross margin minus the total cost in providing these catalogs.

   Finally we should know the profit value above which management gives approval to provide catalogs 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 (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.

   I selected the following two predictor variables:
   a) Avg_Num_Products_Purchased
   b) Customer segment

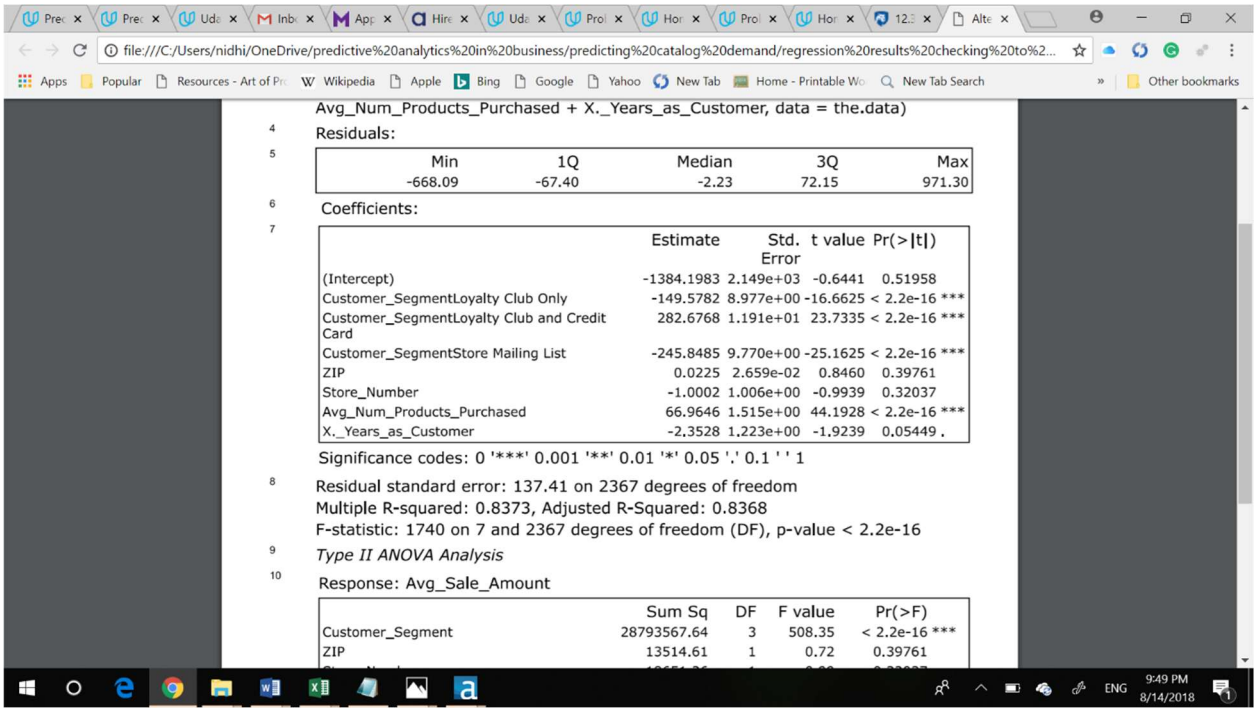
   The selection was based on the following three criterions:
   a) Predictor variables should be relevant to the target variable (Average sales):
      Avg_Num_products_purchased is a statistically significant predictor of the target variable. This can be seen from the results of a trial linear regression. The "p" vlues are $2.2X10^{-16}$ which are very close to zero for Customer_segments and Avg_Num_products_purchased. The lower the "p" value the higher the probability that a relationship exists between the predictor and target variable. Significance codes below the table signify the same.

      Moreover, from the scatter plot of Avg_Num_Products_Purchased vs Average sales we can see that as the number of products purchased increases the average sales increases too. This shows change in target variable can be explained by the change in this predictor variable.

      All other numerical predictor variable (Cust ID, Zip, Store_number, #years as customer) do not show any linear relationship with target variable (average sales)

   b) Predictor variable should not be correlated to another predictor variable
      The Predictor variable Avg_Num_Products_Purchased is not correlated tp Customer_segment as with the increase or decrease of one predictor variable the other variable does not increase or decrease or vice versa.

   c) Predictor variable should not have high number of missing values
      Our data is clean with no missing values, so we don't have to worry about this



Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-668.09	-67.40	-2.23	72.15	971.30

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958	
Customer_SegmentLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.8485	9.770e+00	-25.1625	< 2.2e-16	***
ZIP	0.0225	2.659e-02	0.8460	0.39761	
Store_Number	-1.0002	1.006e+00	-0.9939	0.32037	
Avg_Num_Products_Purchased	66.9646	1.515e+00	44.1928	< 2.2e-16	***
X._Years_as_Customer	-2.3528	1.223e+00	-1.9239	0.05449	.

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

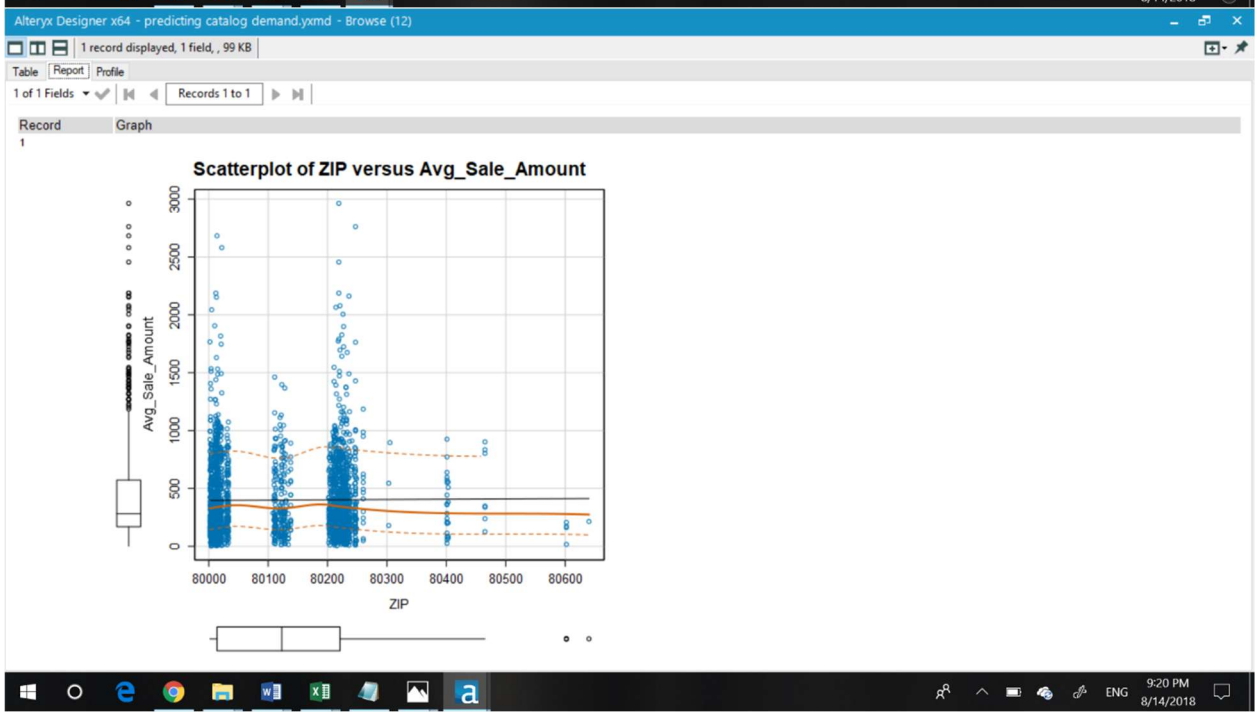
Residual standard error: 137.41 on 2367 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8368
F-statistic: 1740 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16

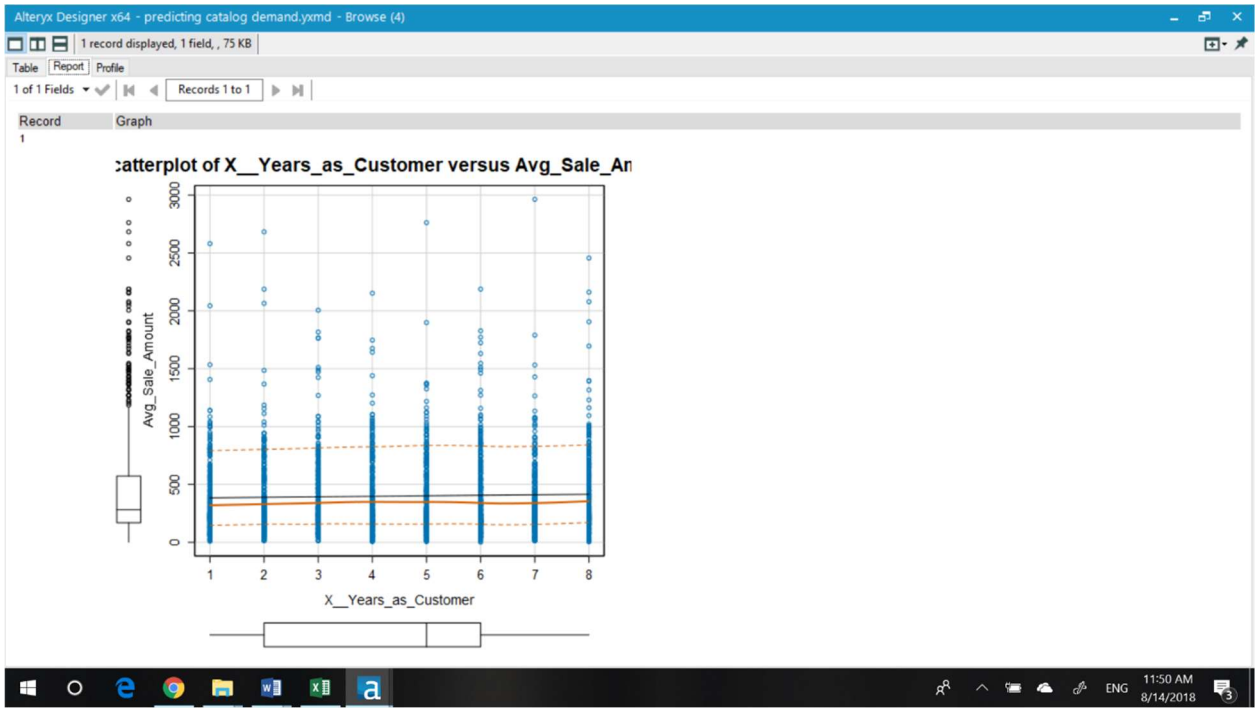
Type II ANOVA Analysis

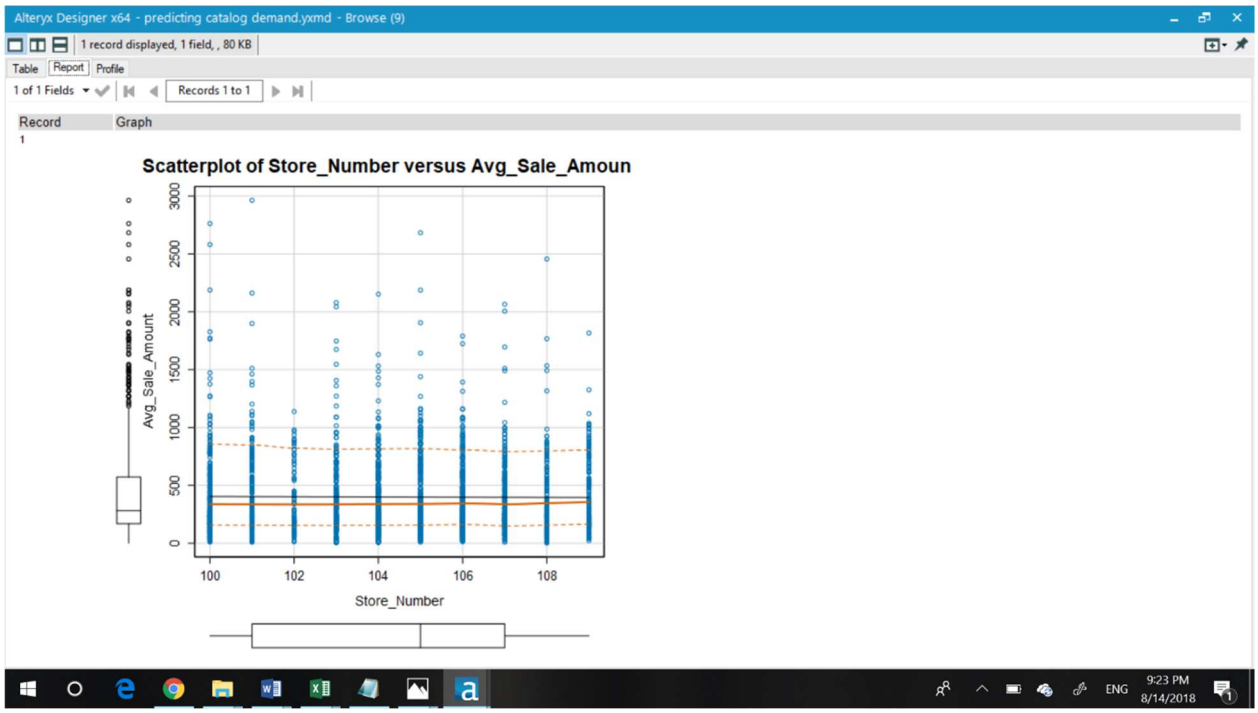
Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28793567.64	3	508.35	< 2.2e-16	***
ZIP	13514.61	1	0.72	0.39761	


Alteryx Designer x64 - predicting catalog demand.yxmd - Browse (4)
1 record displayed, 1 field, , 236 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 1
Record
Graph
1
Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
Customer_ID
ENG
11:46 AM
8/14/2018


Alteryx Designer x64 - predicting catalog demand.yxmd - Browse (12)
1 record displayed, 1 field, , 99 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 1
Record
Graph
1
Scatterplot of ZIP versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
80000
80100
80200
80300
80400
80500
80600
ZIP
ENG
9:20 PM
8/14/2018





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.

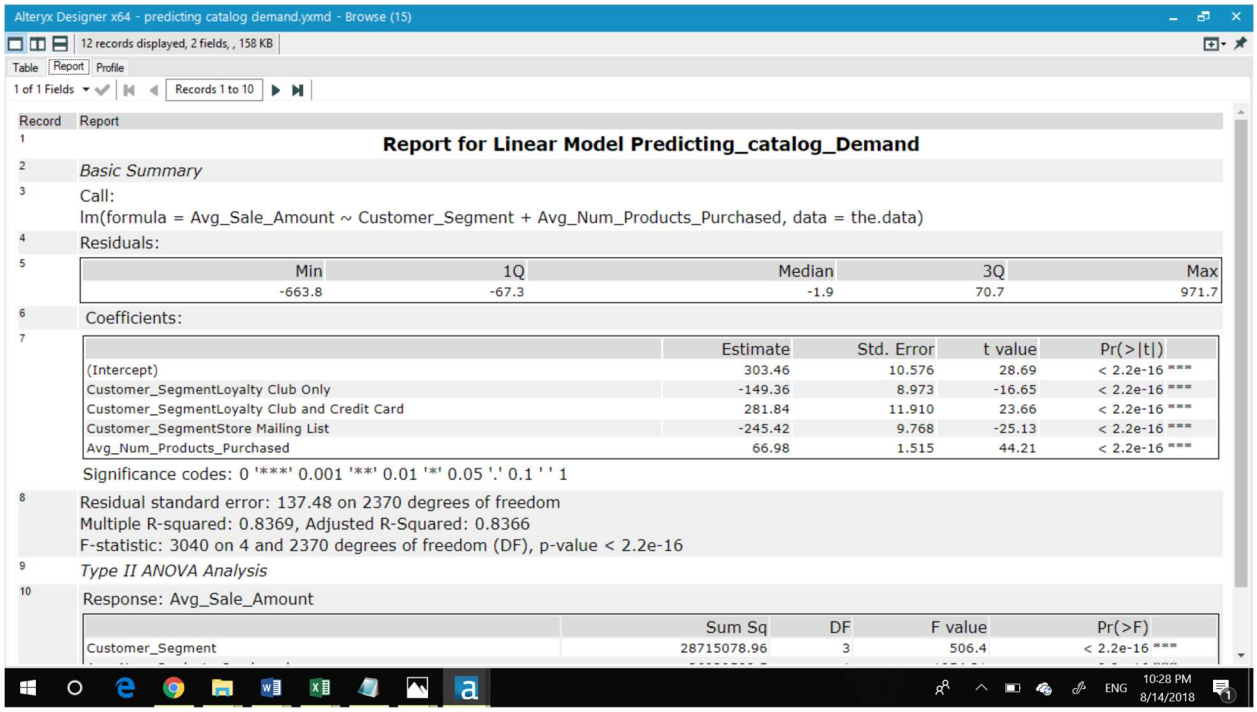
   This linear model (pls. refer image below) is a good model because the predictor variables chosen are linearly related to target variable and both the selected predictor

variables do not have any correlation between them. The same can be seen from max p value which is less than or equal to $2.2X10^{-16.}$ Low p values indicate that the variables are statistically significant.

R-squared: 0.8369, Adjusted R-Squared: 0.8366

Both the regular R-squared and Adjusted R-square are approximately around 0.84(which is more close to 1 than zero). Closer this value is to 1 the better the model is because it indicates the percentage of variation explained by only the independent variables that actually affect the dependent variable. The adjusted R squared gives a pretty good indication of the validity of model, because if the quality of predictors was poor then it would have reduced also(but, it didn't reduce in this case). The best model can be only as good as the variables measures by the study.

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:

Predicted_Avg_Sales = 303.46 -149.36(if Type: Customer_segment_loyalty_club) +281.84 (if Type: Customer_segment_loyalty_club_credit_card) – 245.42 (if Type: Customer_segment_mailing list) + 66.98 * Avg.number_products_purchased

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?
   Yes, the company should 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)

   Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000. Hence, we predicted the expected profit by first creating the linear regression model from the past customers and thereafter computing the profit by fitting the data of new 250 customers on to this model. Since the predicted data is much more than the minimum required by the management, we can confidently say that sending catalogs will increase profitability.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   Assuming the catalog is sent to these 250 customers, the expected profit is $21,987.

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.