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 Company would like to decide on whether or not to send out catalogues to 250 customers. The management decides to only send out catalogues if the expected profit contribution exceeds $10000.

2. What data is needed to inform those decisions?

Analyse and predict the potential profit from sending 250 catalogues. If the profit predicted exceeds $10000only then will catalogue will be sent.

Profit = ProjectedCustomerRevenue * Probability that customer will buy – CostOfPrinting and sending

Projected customer revenue depends on two main factors customer segment and the number of products. This is based on multiple linear regression.

Step 2: Analysis, Modeling, and Validation

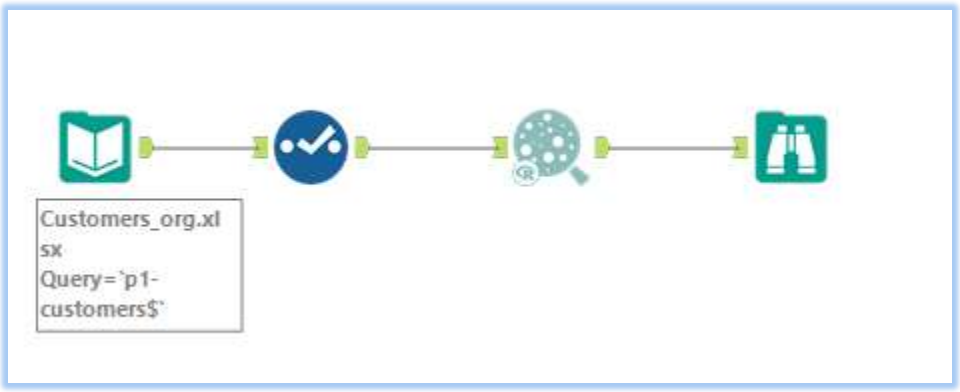
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

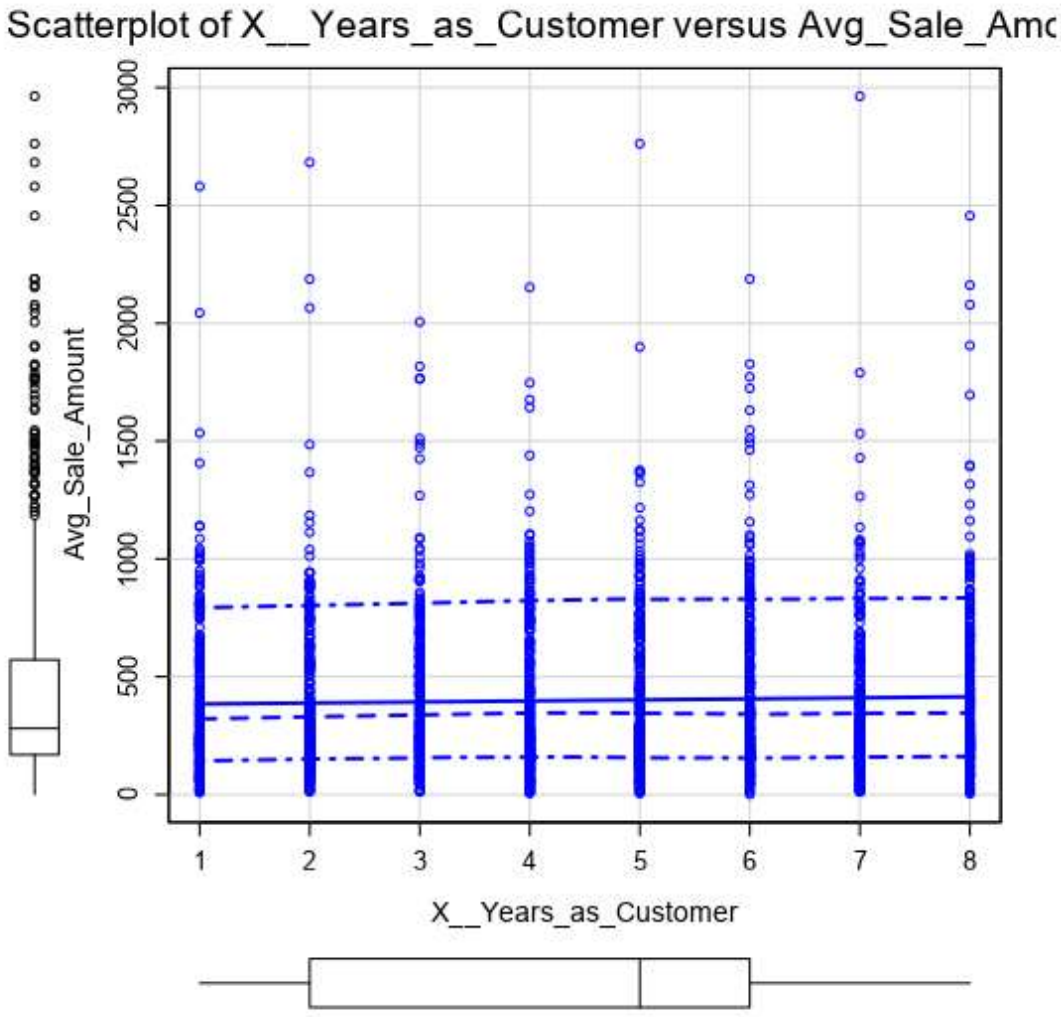
At the minimum, answer these questions:

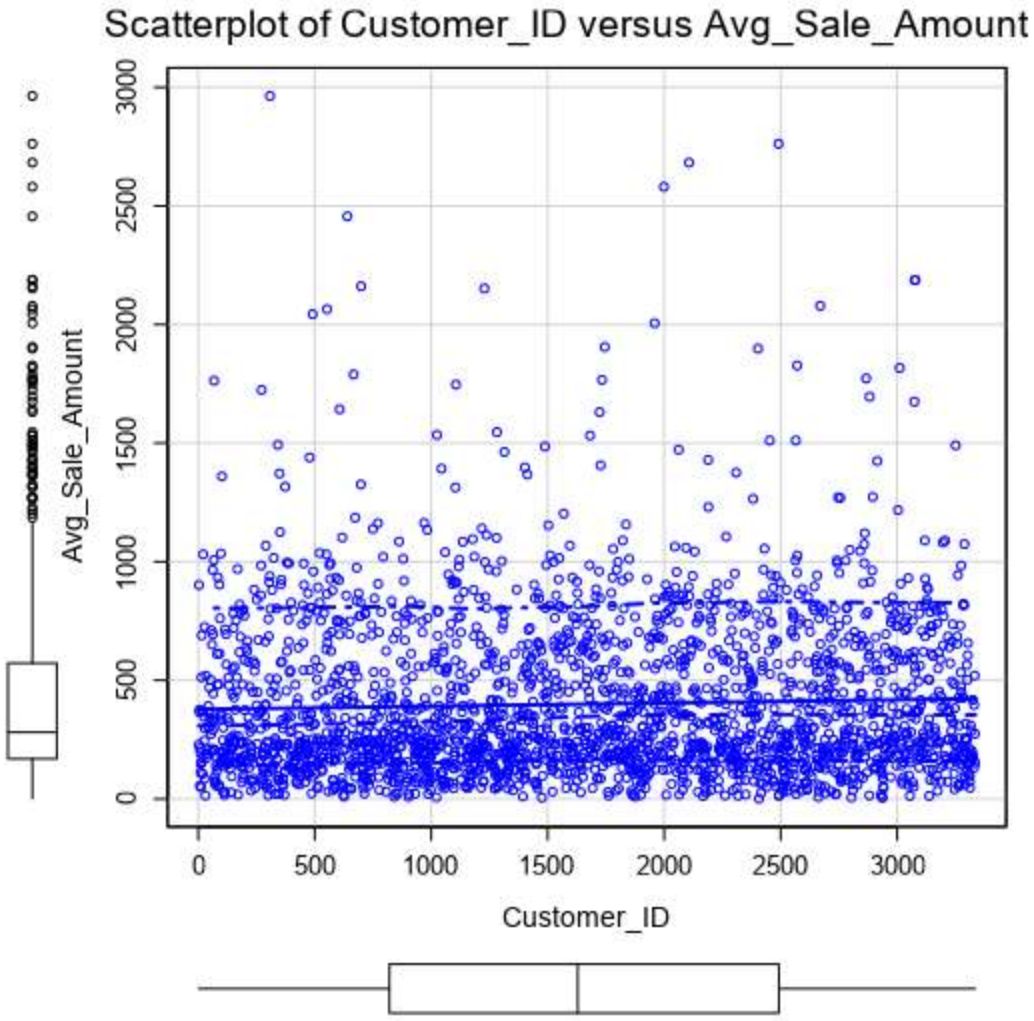
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

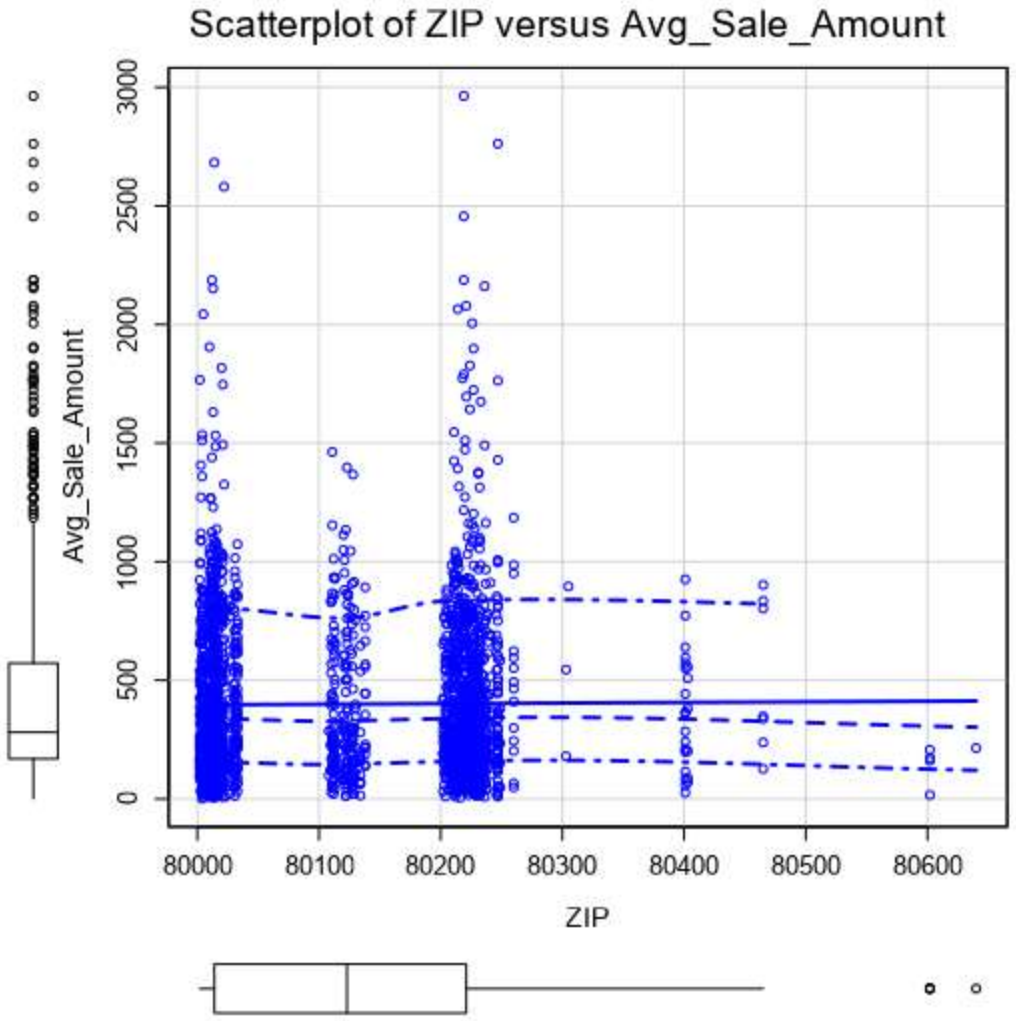
For Numeric predictor variables scatter plots were used.
Numeric variables scatter plots are used to predict the target variable.
From the below scatter plot we observe that only the average number of products purchased VS Avg sales have a direct positive relation.

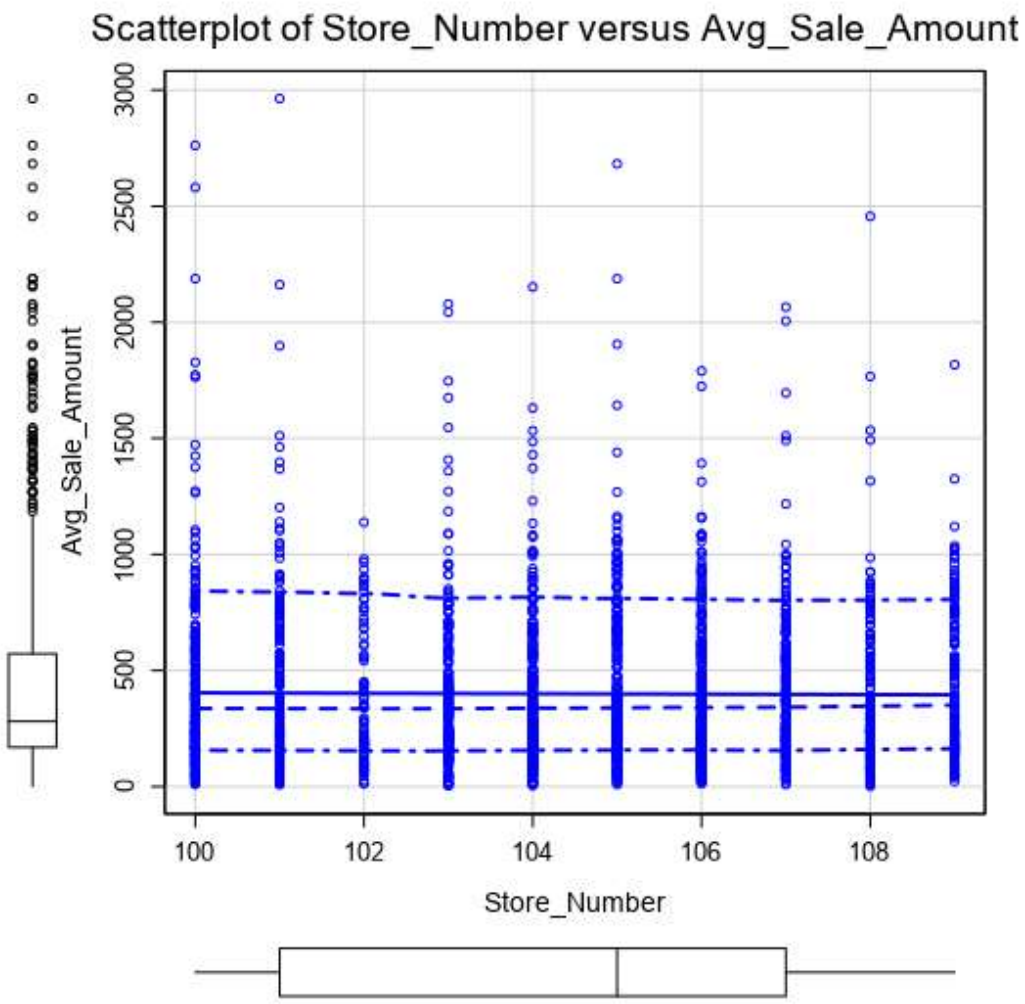


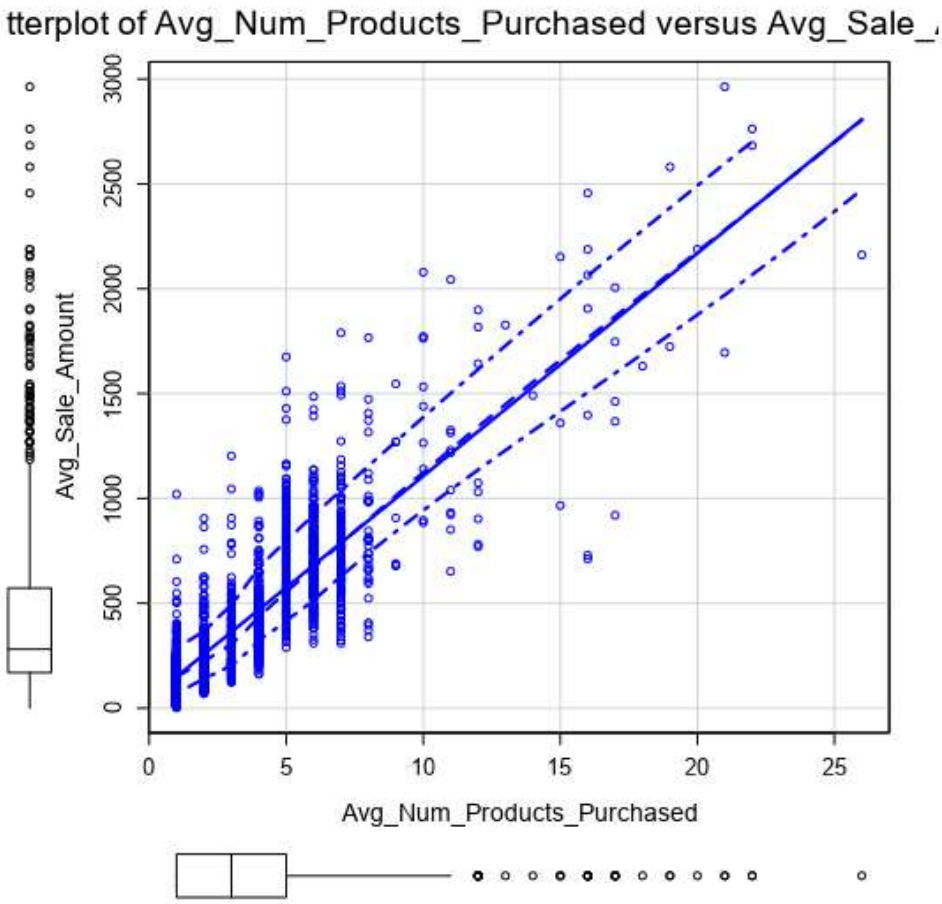
Various Scatter plots for numeric variables included below:


Scatterplot of X__Years_as_Customer versus Avg_Sale_Amc
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
X__Years_as_Customer


Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
Customer_ID


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


Scatterplot of Store_Number versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
100
102
104
106
108
Store_Number



For categorical (non-numeric variables) linear regression was used to and selected those variables where p<=0.05. Two variables a) various customer segments and the average number of products purchased were selected. P-value is extremely small as indicated below.

	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 ***

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 model is good since it has a high r squared value of 0.8369; the selected variables both numeric and non-numeric can be used in the model.

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)

Record | Report

1

Report for Linear Model Linear_Regression_6

2 *Basic Summary*

3 Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

4 Residuals:

5

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

6 Coefficients:

7

	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

Based on the above we can establish the best linear regression equation as:

Avg_Sale_Amount = 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

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?

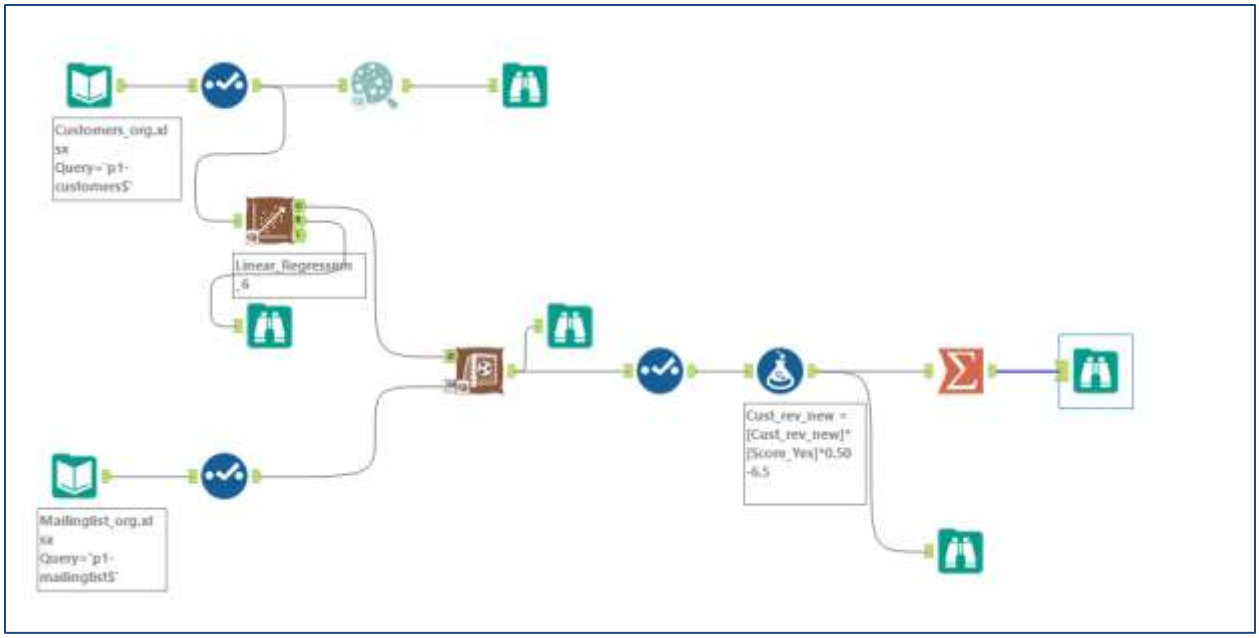
Yes, the company should send the catalogues. This is based on the fact that the estimated profit is: $21987, and is greater than $10000.

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

After the selection of the numeric and non-numeric variables; the predicted revenue for the for each customer in the mailing list is calculated. This predicted revenue is then used to calculate the predicted profit using the formula :

predicted Profit = ProjectedCustomerRevenue * Probability that customer will buy – CostOfPrinting and sending

The sum of all the predicted profits is $21987 which is greater than $10000.



Record	Sum_Cust_rev_new
1	21,987.435687

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

As per the model, the expected profit is: $21987

Record	Sum_Cust_rev_new
1	21,987.435687