UDACITY
ND.Business Analytics

Francihelena Uzcategui

September 19, 2018.

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?

We should send the catalog to these 250 new customers? How much will be the profit from this predicted sales? We have a minimum value to profit, it will be $10,000 for those 250 new customers; if this condition meets, we will send the catalogues to new customers.

2. What data is needed to inform those decisions?

This business problem has data already collected because the last year the company sent catalogs to customers (data rich.)
We have to use this information given:

- The average sale amount to check the customer expenses.
- The customer's response to the last catalog, and the expectations for this new catalog.
- The customer's payment method.
- The number of new customers - 250.
- Expected profit must be high than $10,000.
- Gross Margin 50%.
- Cost of printing $ 6.5 per catalog.

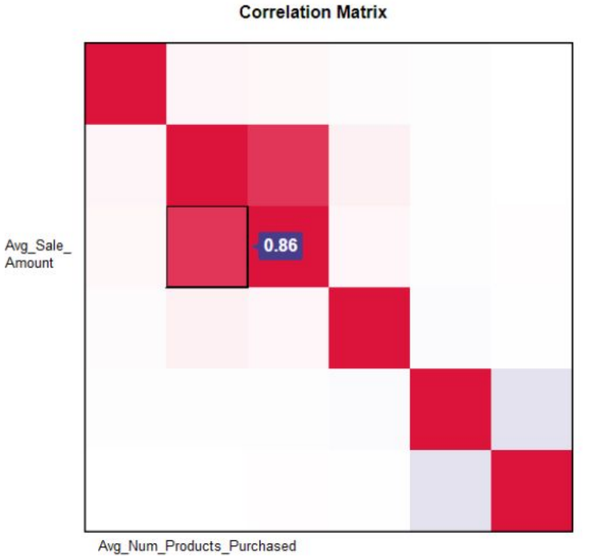
Step 2: Analysis, Modeling, and Validation

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)

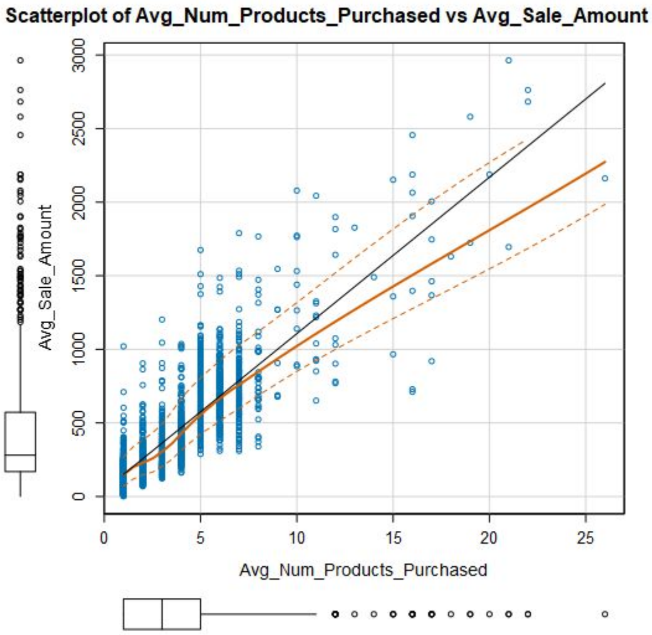
Important: Use the p1-customers.xlsx to train your linear model.

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 scatter plots to search for linear relationships. You must include scatterplots in your answer.

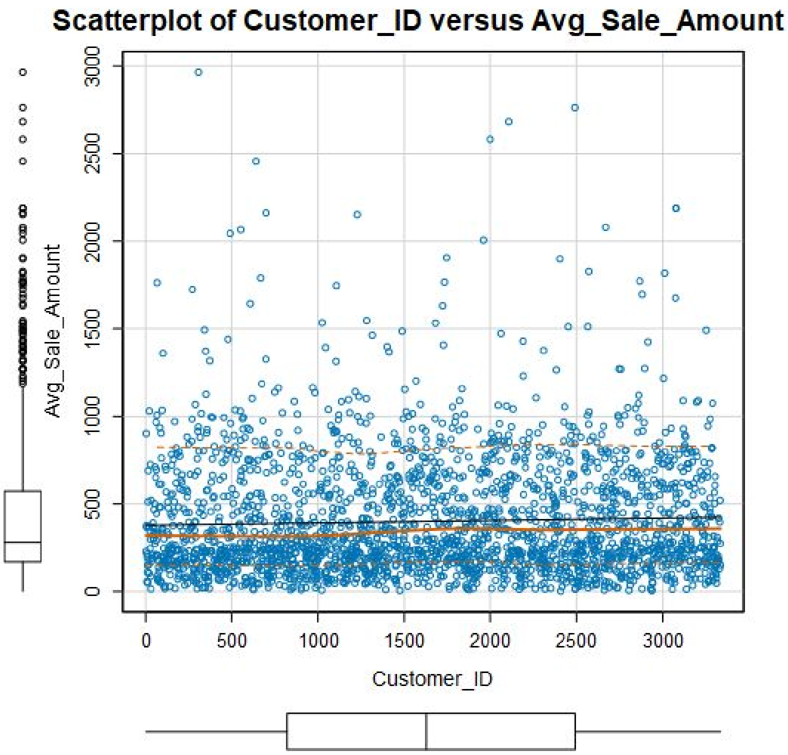
= By Association Analysis tool we obtained the "Correlation Matrix," below on the graph, a classification by color and values correlated. This red raspberry square shows 0.86 the strong relationship between Avg_Sales_amount and Avg_Num_products_purchased; as well the scatter plot shows the linear relationship between them.

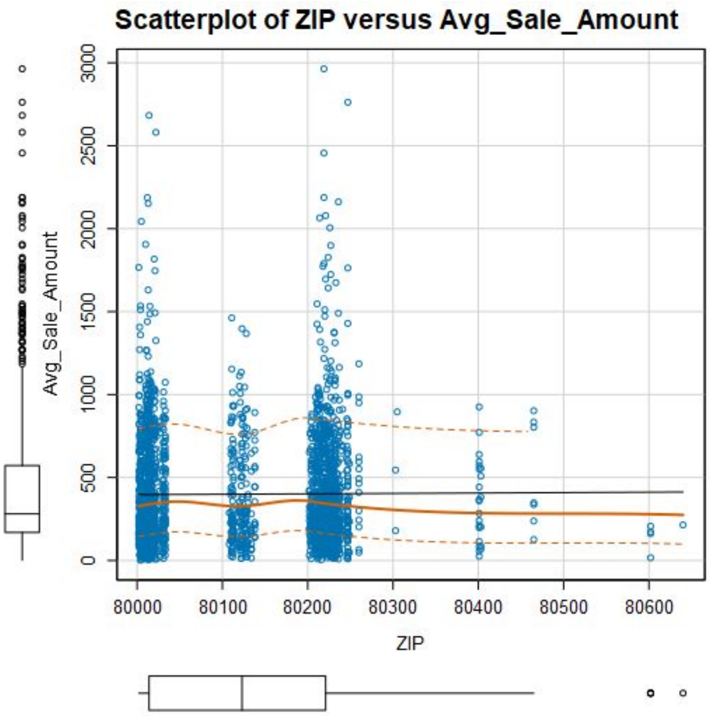


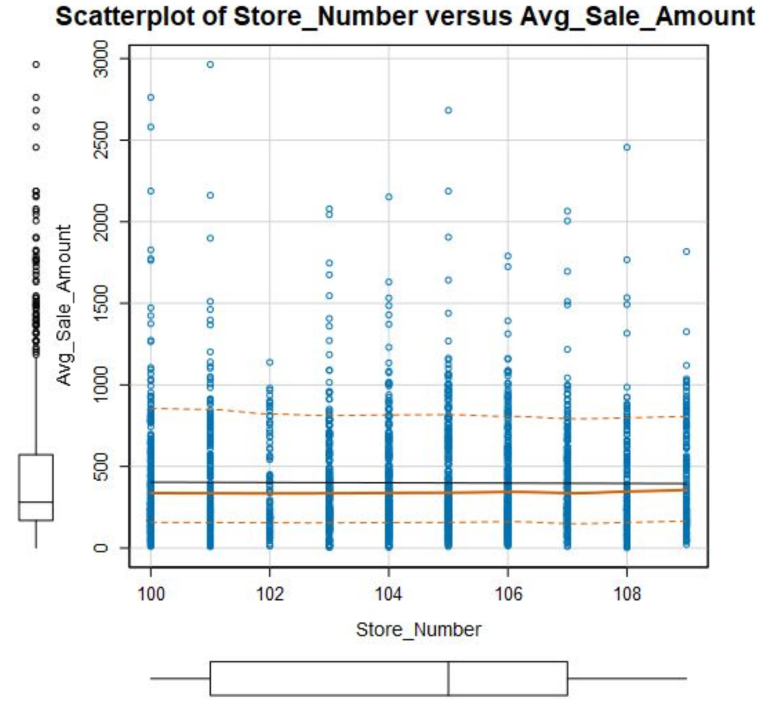
The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

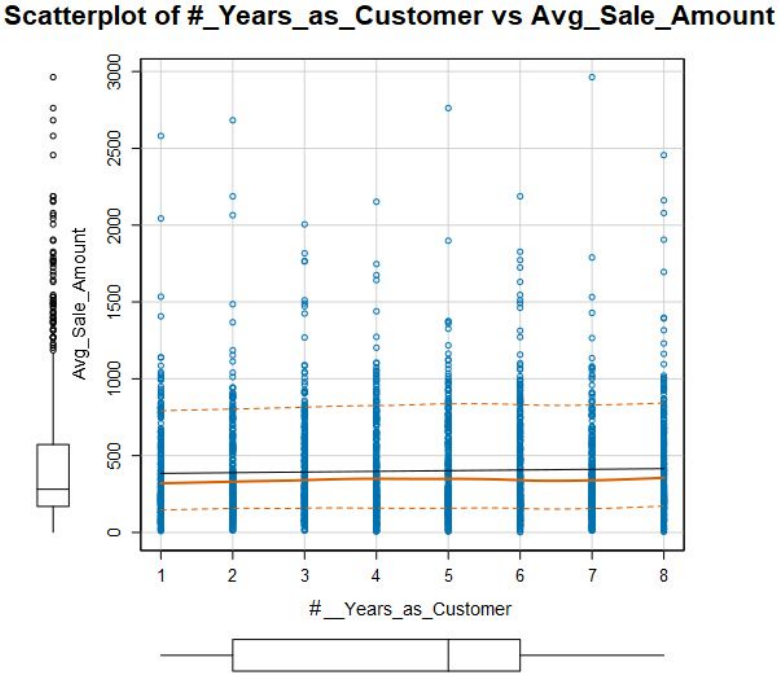


Since our target variable is Avg. Sale Amount, we verified the relationship between these variables by scatterplot.









Preceding scatterplots show the small relationship of these variables with the target variable - Avg_Sale_ Amount, thus there were not used as predictor variables[1].

Alteryx- scatterplots did not include all the categorical variables, although, by the Linear Regression report we observed the low p-values and strong relationship between the Avg_Sale_ Amount and Customer_Segment, so we chose this last variable as a predictor.

[1] See Annexes 1.

Finally, we omitted the variables Name, Address, City and State because include personal and irrelevant information for this report.

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.

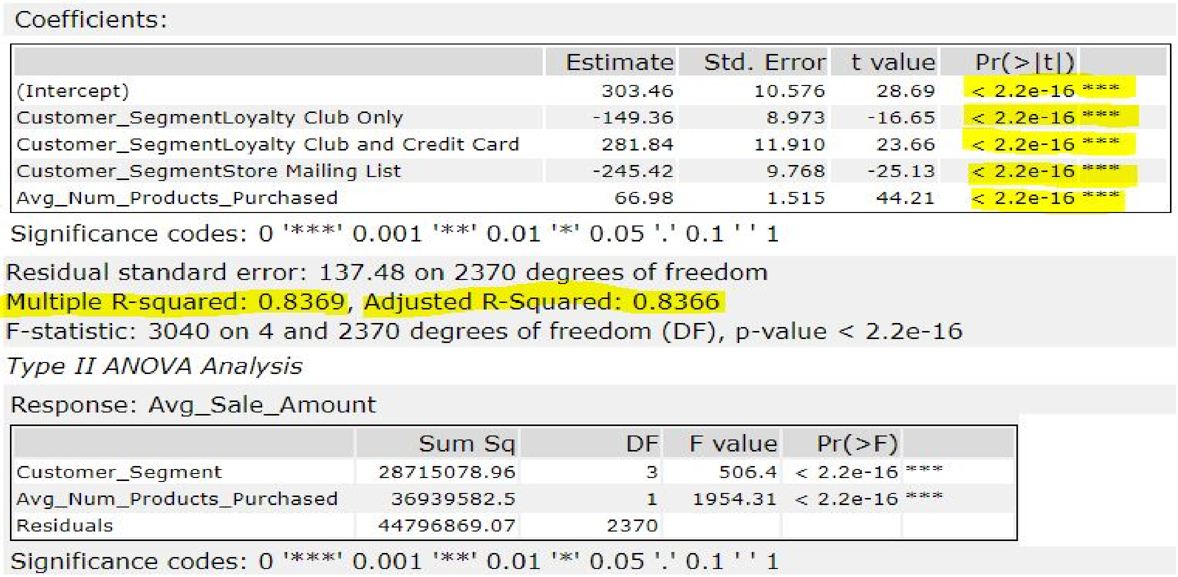
= After reviewed the Correlation Matrix with Scatterplot we selected as target variable the Avg_Sale_Amount (quantitative), and as predictable variables the Customer_Segment (categorical) and Avg_Number_Products_Purchased(quantitative).

Below is the report Model Linear_Regression_sale, it shows the coefficients chose with low P-values: < 2.2e-16 and a high R-squared: 0.8366 ≈ 0.84; these values suggest the model is highly predictive. A reliable model must have a low P-value and a high R-squared value.

The column Pr(>|t|) shows the p-values for this specific t-test. For this model P-value is < 0.00000000000000022, it is significant because is close to zero.

R-Squared means is the coefficient of multiple determination for multiple regression when the number is close to one the model explains how well it fit with the datasets. In this case, we obtained ≈ 0.84 to close to one.

It's necessary to highlight the good or acceptable value of R-squared depends on the field of study, for example, social sciences - human behavior accept low values as 0.5.



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

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)

The regression equation formula:

Y = 303.46 -149.36*Customer_segment_Loyalty Club only + 281.84*Customer_segment_Loyalty Club and Credit Card - 245.42*Customer_segment_Store Mailing list + 0*Customer_segment_Credit card only + 66.98* Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

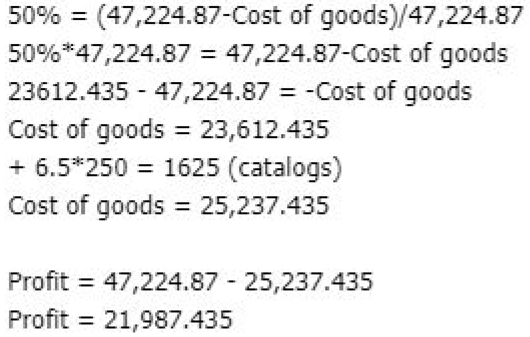
1. What is your recommendation? Should the company send the catalog to these 250 customers?
= According to the outputs from the model, we suggest sending the catalogs to the new clients because will be a profit high than $10,000.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process
= For the process, we used Alteryx's tools.
First, we chose the target variable Avg_Sales_Amount, because we must predict the profit to send the catalogs.
Further, we used Association analysis, Scatterplots and Linear Regression analysis tools to check and confirm which predictor variables have low P-values.
Then, by Score tool, we connected both datasets (customer and mailing list) to the model already done. Next, by the Formula tool, we multiplied the Score_Yes by Sale_amount.
At this point, we had enough results to work with the formula Gross margin percentage and their components. By hand, we isolated variables and computed simple math. Then, by Formula tool, we incorporated these calculations into the workflow. Further explanation is below the formula Gross margin percentage.

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

After carefully review the project's guidelines and hints, as well as preliminary calculations, we highlight some useful values to use:
Avg gross margin = 50%
Sum_Sale_Amount=47,224.87 or Total revenue
Cost of goods = ?

Avg gross margin= (Total revenue - Cost of goods) / Total revenue



50% = (47,224.87-Cost of goods)/47,224.87
50%*47,224.87 = 47,224.87-Cost of goods
23612.435 - 47,224.87 = -Cost of goods
Cost of goods = 23,612.435
+ 6.5*250 = 1625 (catalogs)
Cost of goods = 25,237.435

Profit = 47,224.87 - 25,237.435
Profit = 21,987.435

The profit will be $ **21,987.44** the double amount expected as a minimum condition to send catalogs to new customers.

Works cited

Stackexchange.*Meaning of Pr(>|t|)*
https://stats.stackexchange.com/questions/49939/interpreting-summary-function-for-lm-model-in-r

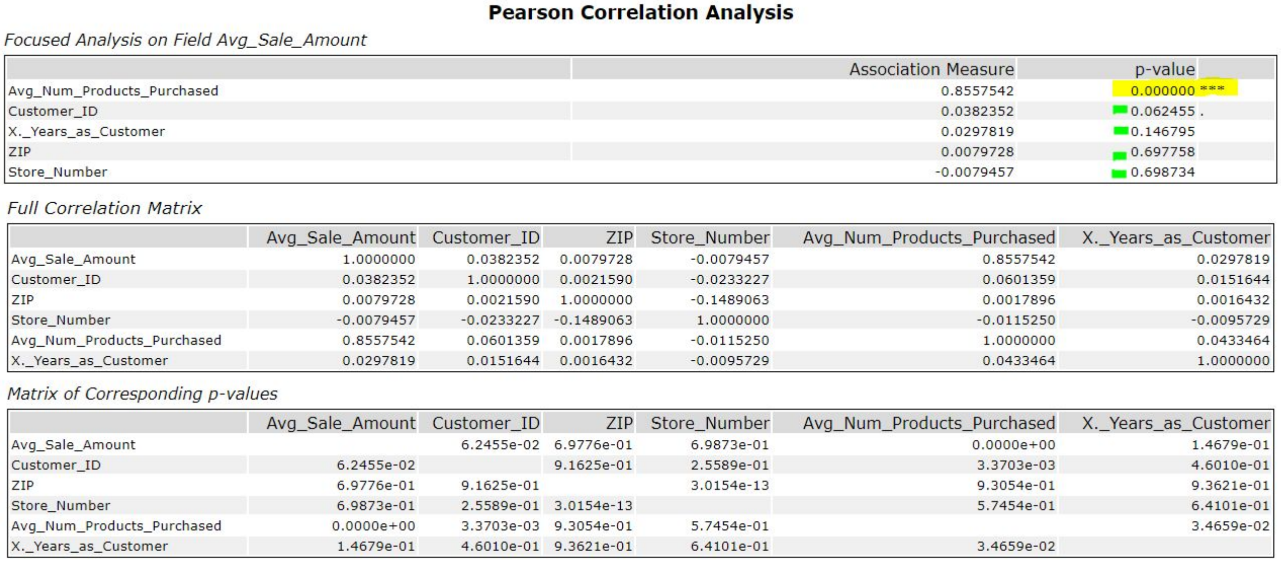
Minitab.*Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*

http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Panalysis*. How to Calculate Gross Margin Percentage*
https://www.panalysis.com/resources/articles/how-to-calculate-gross-margin-percentage

Annexes

1)



Pearson Correlation Analysis

Focused Analysis on Field Avg_Sale_Amount

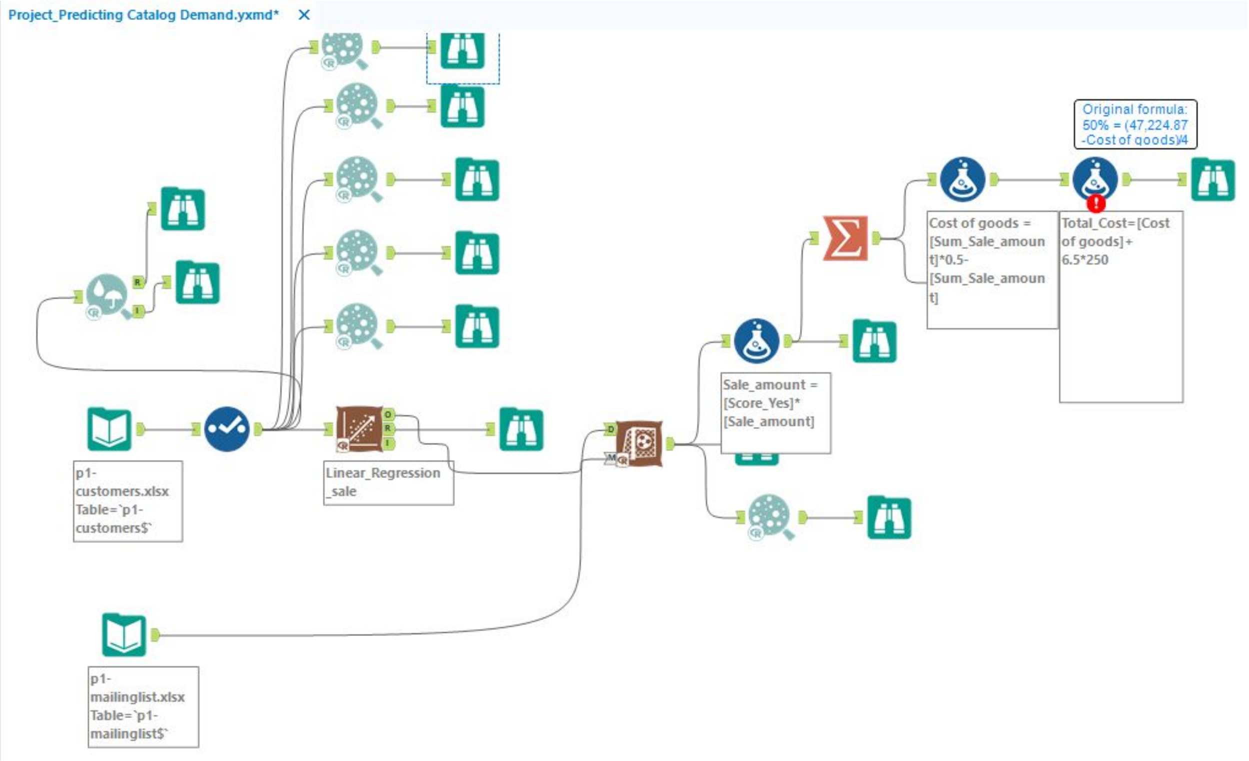
	Association Measure	p-value	
Avg_Num_Products_Purchased	0.8557542	0.000000	***
Customer_ID	0.0382352	🟩0.062455	.
X._Years_as_Customer	0.0297819	🟩0.146795	
ZIP	0.0079728	🟩0.697758	
Store_Number	-0.0079457	🟩0.698734	

Full Correlation Matrix

	Avg_Sale_Amount	Customer_ID	ZIP	Store_Number	Avg_Num_Products_Purchased	X._Years_as_Customer
Avg_Sale_Amount	1.0000000	0.0382352	0.0079728	-0.0079457	0.8557542	0.0297819
Customer_ID	0.0382352	1.0000000	0.0021590	-0.0233227	0.0601359	0.0151644
ZIP	0.0079728	0.0021590	1.0000000	-0.1489063	0.0017896	0.0016432
Store_Number	-0.0079457	-0.0233227	-0.1489063	1.0000000	-0.0115250	-0.0095729
Avg_Num_Products_Purchased	0.8557542	0.0601359	0.0017896	-0.0115250	1.0000000	0.0433464
X._Years_as_Customer	0.0297819	0.0151644	0.0016432	-0.0095729	0.0433464	1.0000000

Matrix of Corresponding p-values

	Avg_Sale_Amount	Customer_ID	ZIP	Store_Number	Avg_Num_Products_Purchased	X._Years_as_Customer
Avg_Sale_Amount		6.2455e-02	6.9776e-01	6.9873e-01	0.0000e+00	1.4679e-01
Customer_ID	6.2455e-02		9.1625e-01	2.5589e-01	3.3703e-03	4.6010e-01
ZIP	6.9776e-01	9.1625e-01		3.0154e-13	9.3054e-01	9.3621e-01
Store_Number	6.9873e-01	2.5589e-01	3.0154e-13		5.7454e-01	6.4101e-01
Avg_Num_Products_Purchased	0.0000e+00	3.3703e-03	9.3054e-01	5.7454e-01		3.4659e-02
X._Years_as_Customer	1.4679e-01	4.6010e-01	9.3621e-01	6.4101e-01	3.4659e-02	


Project_Predicting Catalog Demand.yxmd*
R
I
O
R
I
D
M
p1-customers.xlsx Table=`p1-customers$`
Linear_Regression_sale
Sale_amount = [Score_Yes]*[Sale_amount]
Cost of goods = [Sum_Sale_amount]*0.5-[Sum_Sale_amount]
Original formula: 50% = (47,224.87 -Cost of goods)/4
Total_Cost=[Cost of goods]+ 6.5*250
p1-mailinglist.xlsx Table=`p1-mailinglist$`

Project_Predicting Catalog Demand.yxmd"