Dhuha Baqarish
PAND

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. **What decisions needs to be made?**

The decision that needs to be made is to send the catalog to 250 clients or not, based on the profit that will be calculated .

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

We are given two files of dataset (customers.xlxs and mailing.xlsx.), From this two files we need :
Avg_Num_Products_Purchased, Customer Segment, Score_Yes.
In addition to :
- cost of catalogue and that equals($6.50)
- gross_margin (50%) to find the profit.

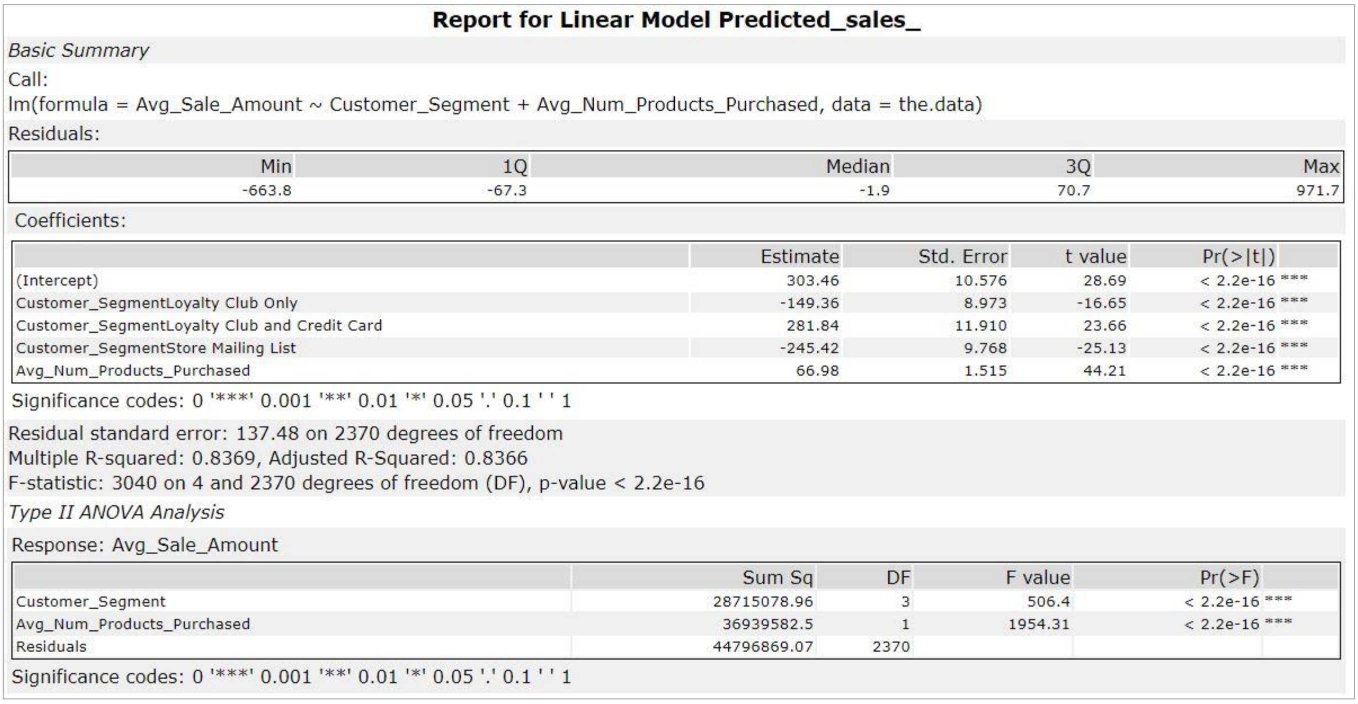
Step 2: Analysis, Modeling, and Validation

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?**

The target variable for the analysis is Avg_Sale_Amount .
And the predictor variables selected for the model are Customer_Segments and Avg_Num_Products_Purchased .
The reason for we selected this two variables as predictor variables is because their p-value less than 0.05 which that mean these two variables are statistically significant.



Report for Linear Model Predicted_sales_

Basic Summary

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

Residuals:

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

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

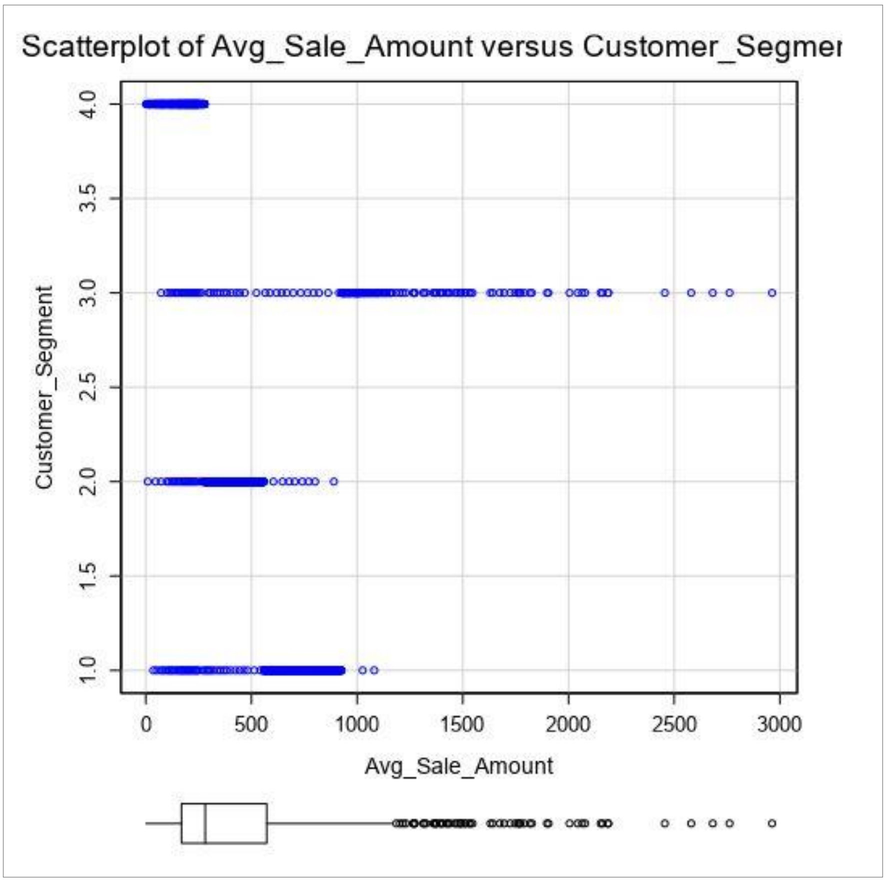
Type II ANOVA Analysis

Response: Avg_Sale_Amount

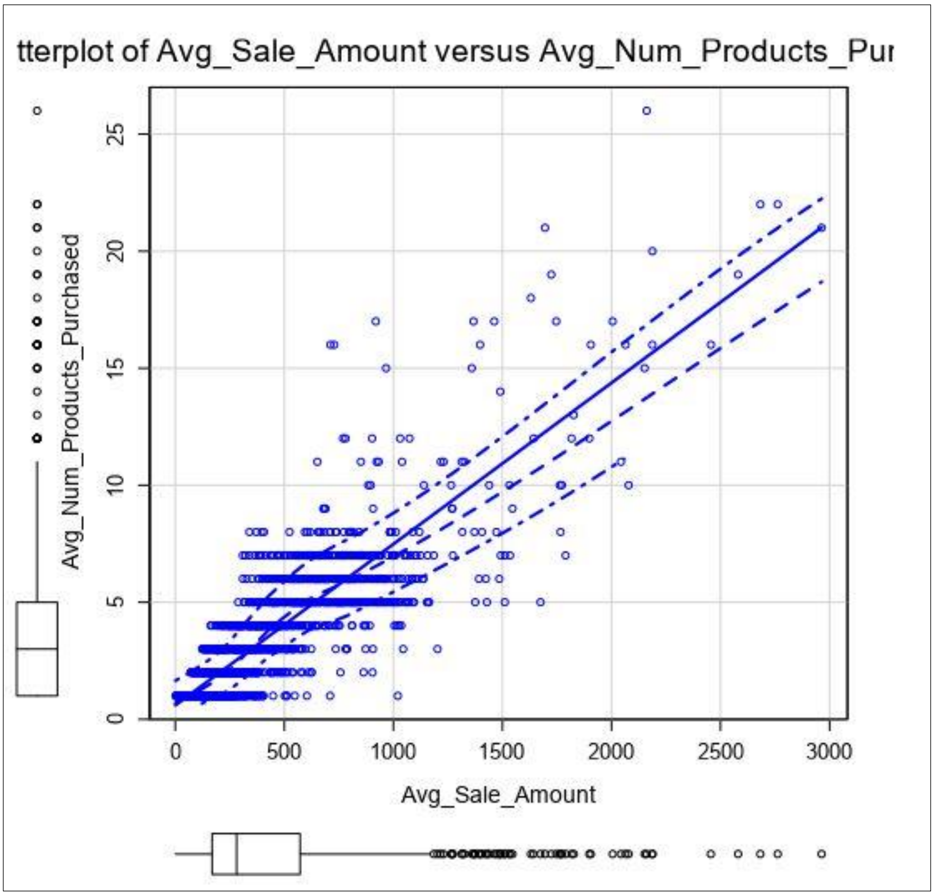
	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

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

The relationship between Avg_Sale_Amount and Customer_Segments represent by scatterplot :

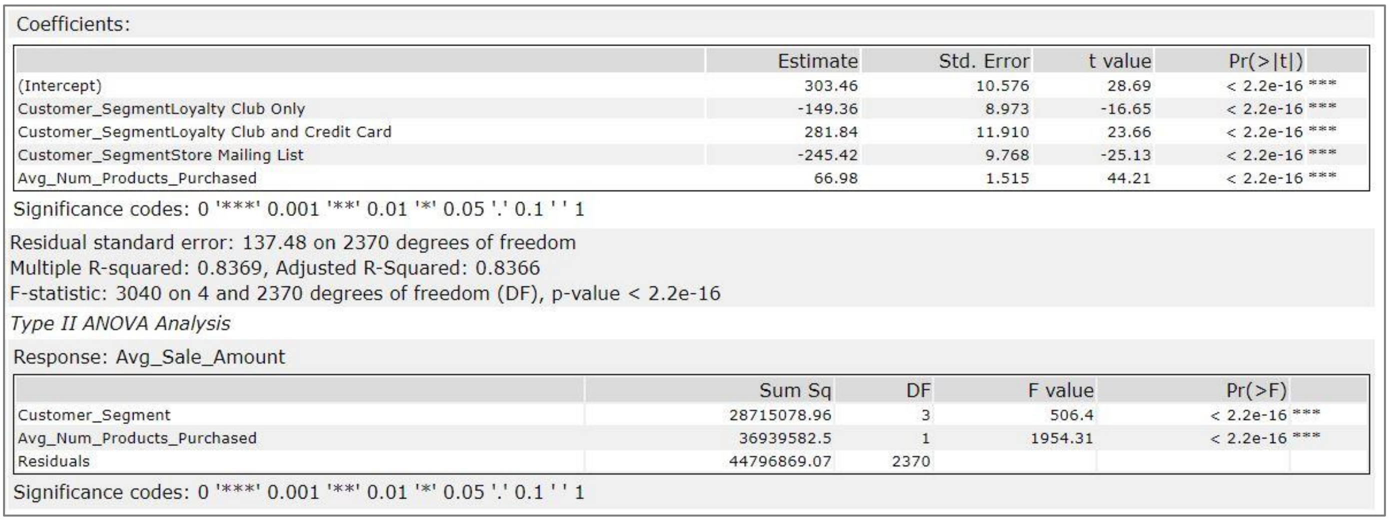


The relationship between Avg_Sale_Amount and Avg_Num_Products_Purchased represent by scatterplot :



2. Explain why you believe your linear model is a good model.

As shown below :
- The Customer_Segment and Avg_Num_Products_Purchased have p-values less than 0.05.
- The Adjusted R Squared value isequal 0.8366 which is quite a large value.
This mean that our model is a good model because p-values and R-Squared value is statistically significan



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 form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

Avg_Sales_Amount = 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) + (66.98 * Avg_Num_Products_Purchased)

Step 3: Presentation/Visualization

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 these catalogues to these 250 customers.

2. How did you come up with your recommendation?

I will explain the process in steps ::

1- calculated predicted_sales_amount using the linear regression and score tools (linear regression model) .

2- After that, I created a new column Predicted_Average_Sales = predicted_sales_amount * Score_Yes , by using formula tool .

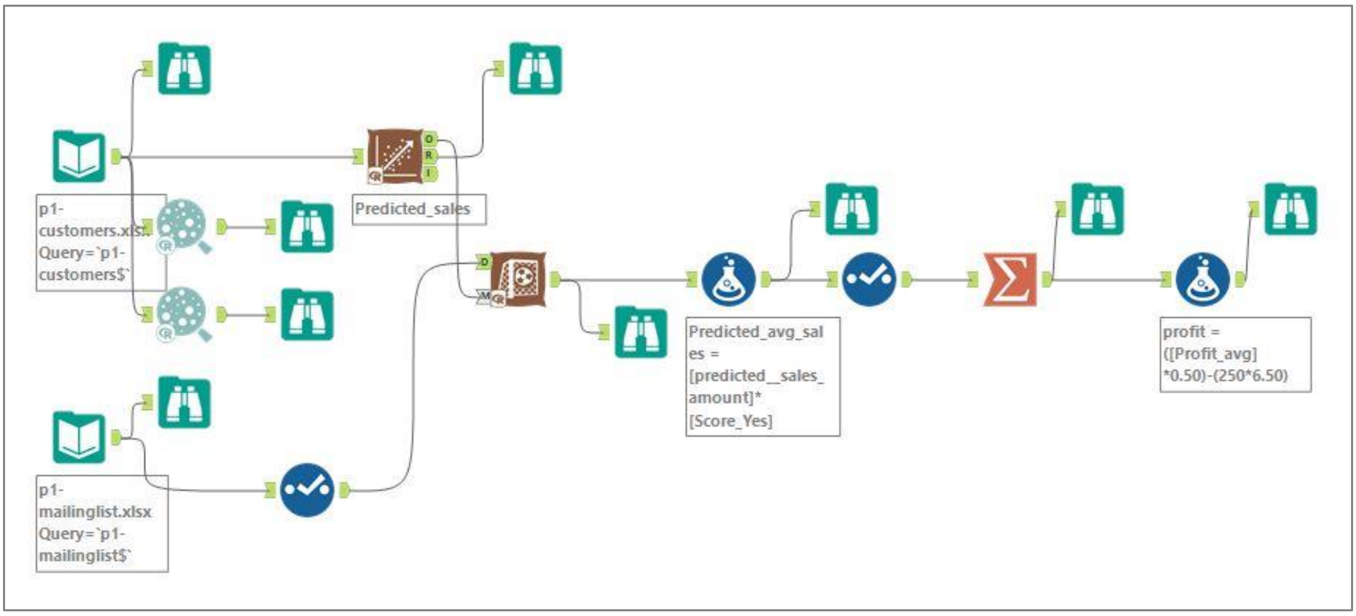
3- Then the profit is calculated with the given margin to be 50% and cost of each catalogue as $6.50, for all the 250 customers , sush as :

Profit = ([Profit_avg]*0.50)-(250*6.50)

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

Profit = ([Profit_avg]*0.50)-(250*6.50) = 21987.4356865455 $

Alteryx Workflow :


p1-
customers.xlsx
Query=`p1-
customers$`
Predicted_sales
O
R
I
D
M
p1-
mailinglist.xlsx
Query=`p1-
mailinglist$`
Predicted_avg_sal
es =
[predicted__sales_
amount]*
[Score_Yes]
profit =
([Profit_avg]
*0.50)-(250*6.50)