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 need to be made?*

The company has 250 new customers from their mailing list that they could send the catalogs to. They want to estimate how much profit they can expect by sending the catalogs to these customers based on historical data. The predicted profit will help to take the decision of sending or not the catalogs to the customers.

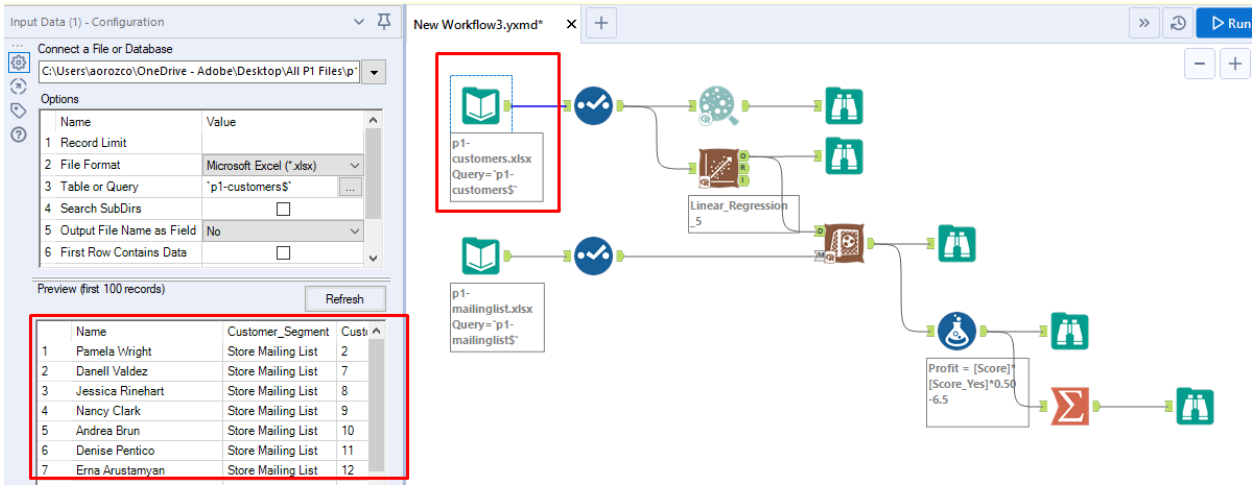
2. *What data is needed to inform those decisions?*

Historical data of customers who received the catalogs and their sales.
Current data of the new customers who potentially would receive the catalogs to predict the sales.

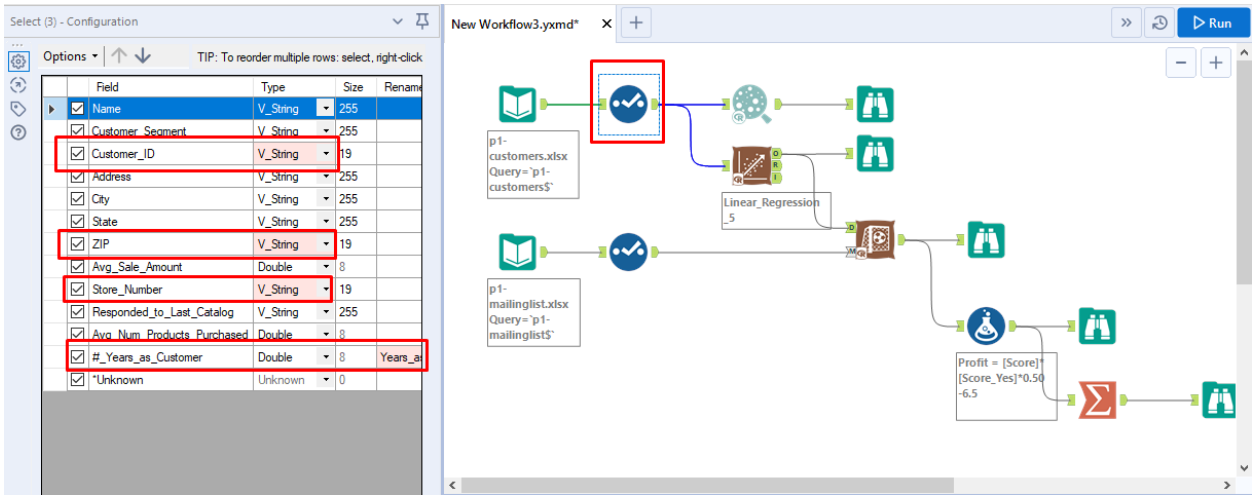
Step 2: Analysis, Modeling, and Validation

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.

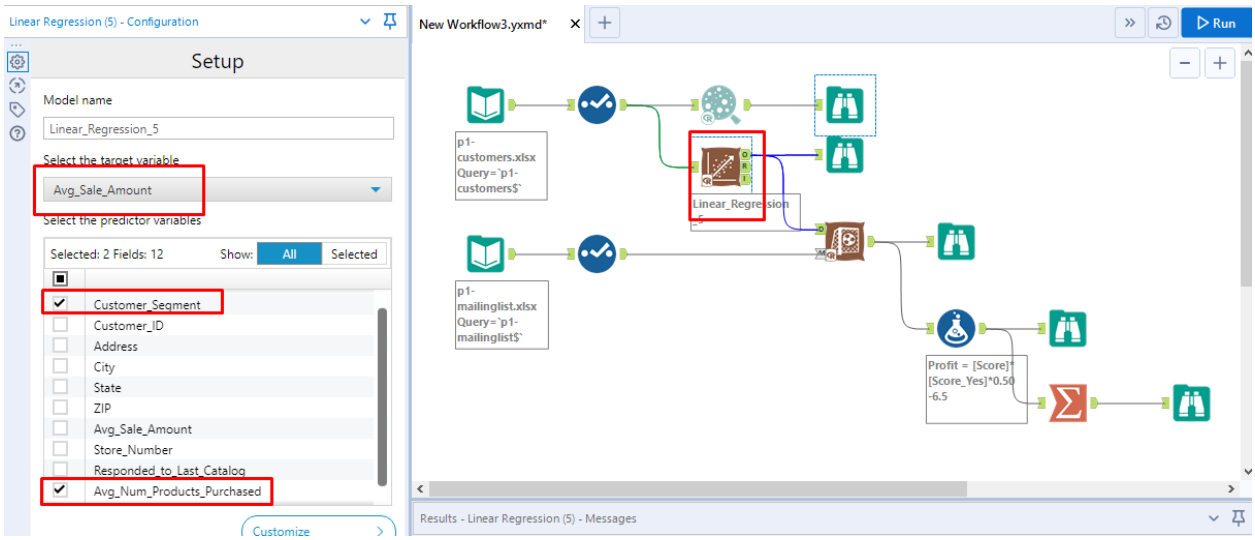
1. Input the data historical data into Alteryx by Input Data tool



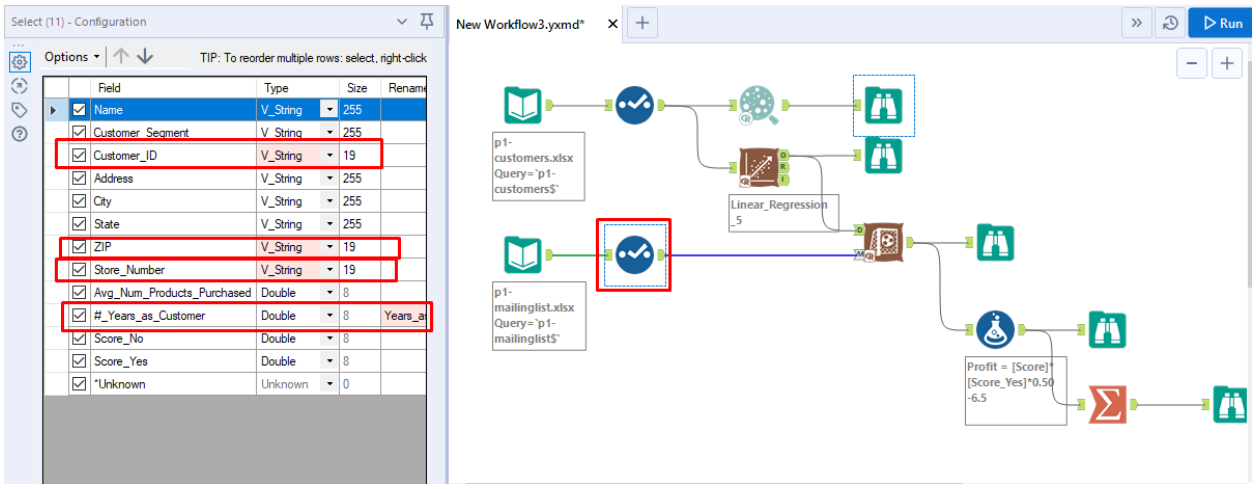
2. In customers data verify and ensure the correct data types in the data source by Select Tool, considering that there are Customer_ID, ZIP, Store_Number are numbers but this should be process as strings due to the nature of the data, Also due to some issues in the name it is convenient to rename #_Years_as_customer as Years_as_customer



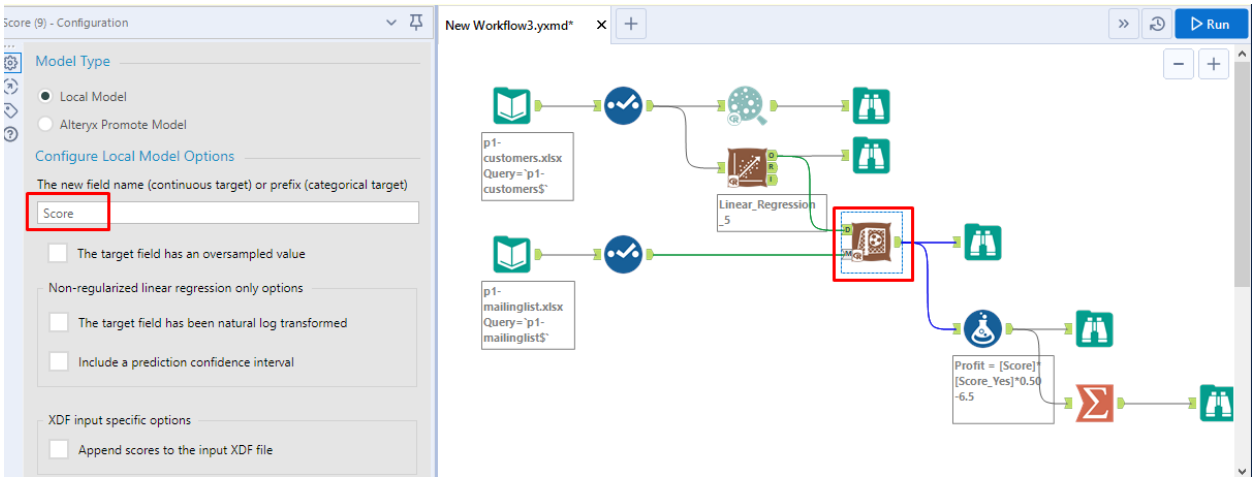
3. Then using the Linear_regression tool, selecting Avg_Sale_Amount as target variable because we are looking to get the predicted sales, and adding as predictor variables Customer_Segment and Avg_Num_Product purchase.



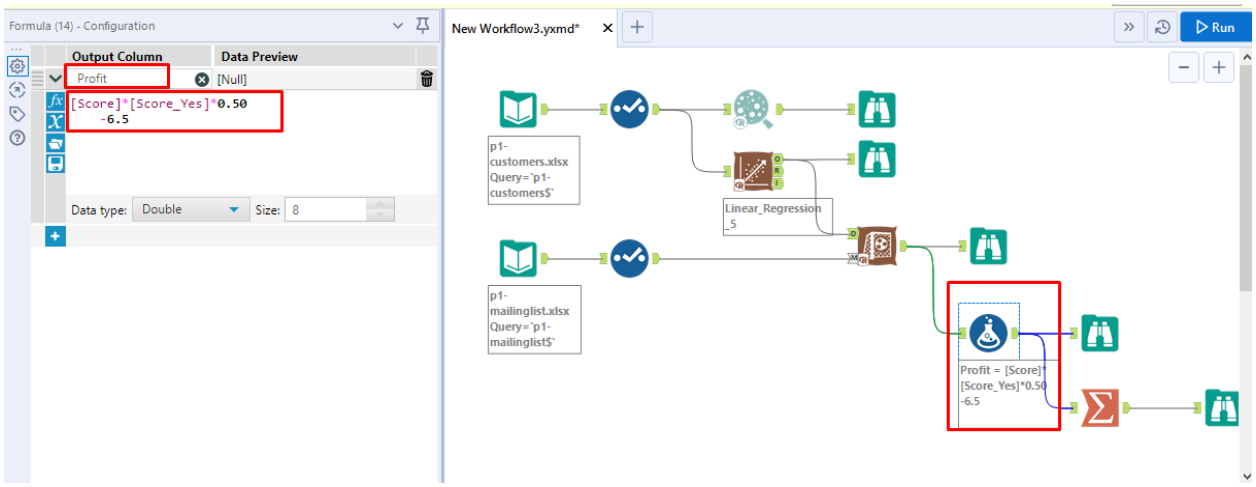
4. In the mailing list verify and ensure the correct data types in the data source by Select Tool, considering that there are Customer_ID, ZIP, Store_Number are numbers but this should be process as strings due to the nature of the data, Also due to some issues in the name it is convenient to rename #_Years_as_customer as Years_as_customer



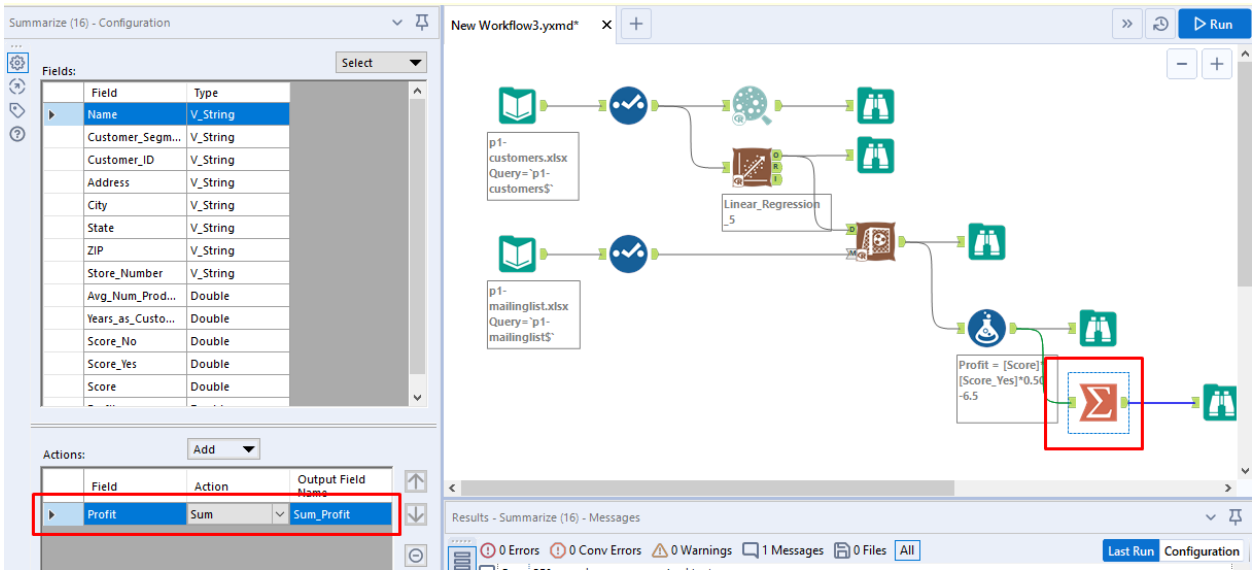
5. Using the Score tool to apply the linear regression to the mailing list data, in this case setting name of the new field as ‘Score’.



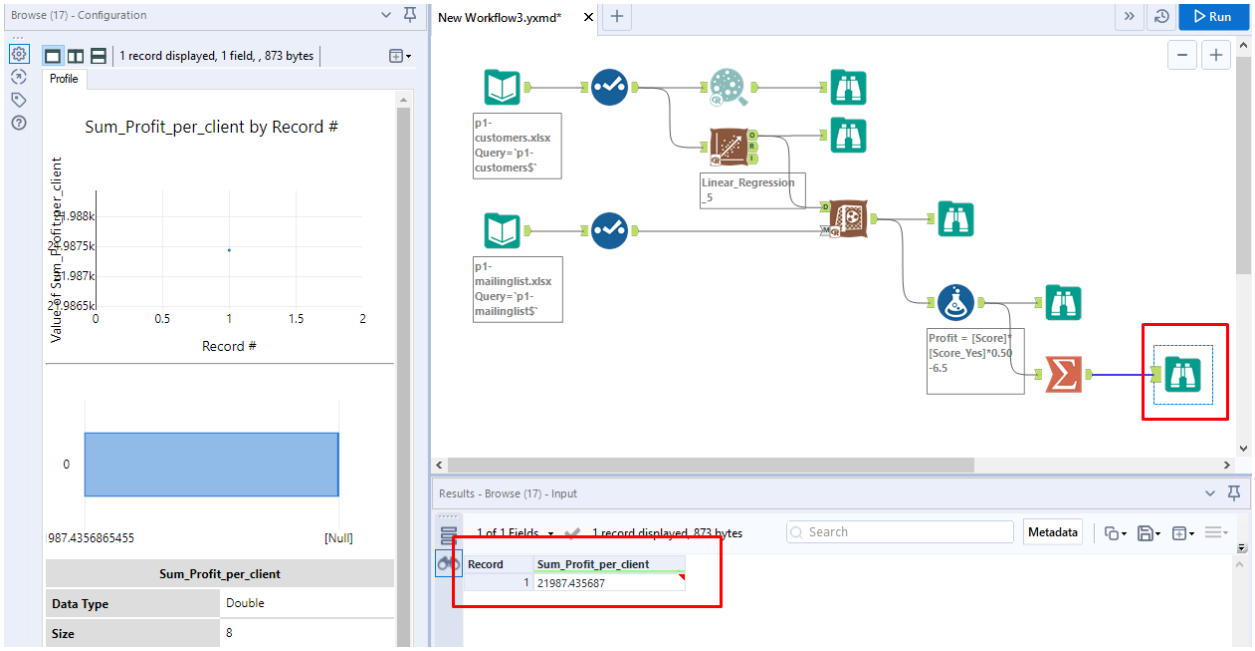
6. Using the Formula tool is possible to calculate the predicted profit per client, multiplying the Score from the Score to Score_Yes that is the probability that the customer will respond to the catalog and make a purchase, furthermore, to considering the costs of printing and distributing is $6.50 per catalog and the average gross margin (price - cost) on all products sold through the catalog is 50%. It is possible to apply a formula of Score* Score_Yes*0.50-6.50 that will return the predicted profit per customer.



7. In order to get the total of the predicted profit for sending the catalogs to the 250 customers in the mailing list, it is possible to use the Summarize tool and select the field profit that is coming from the Formula tool.



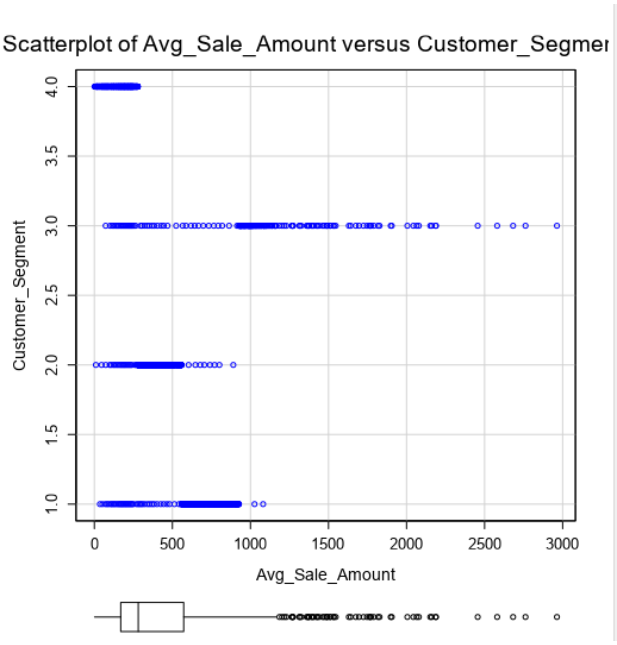
8. Using the Browse tool it is possible to see the sum of predicted profit from customer list as total of $2,1987.44



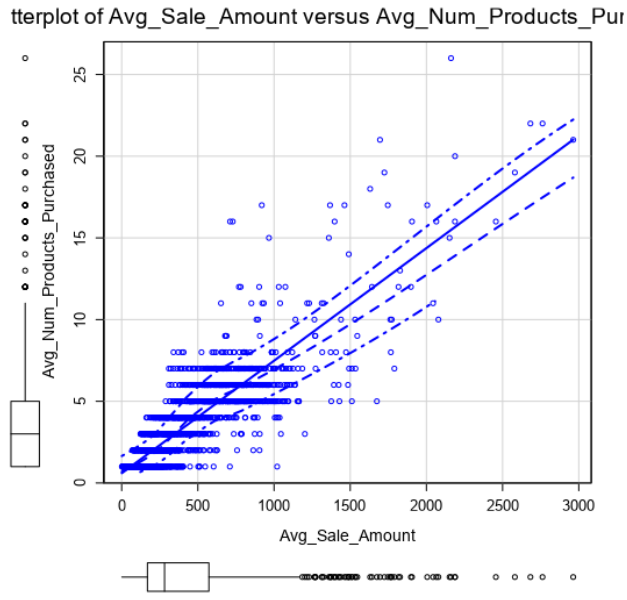
At the minimum, answer these questions:

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.

Considering that there are only four Customer Segments that are important for the type of customer, It is possible to use this as predictor variable, also by getting the scatterplot of the Avg_Sale_Amount vs Customer_Segment, it is possible to observe a trend on the sales high amount of sales for Loyalty Club and Credit Card, therefore it is an important factor and useful for the Linear Regression.



We could assume that the average number of product purchased directly connected with the sales, It is possible to use this as predictor variable, also by getting the scatterplot of the Avg_Sale_Amount vs Avg_Num_Product_Purchase, it is possible to observe a trend on the sales high amount of sales for average number of product purchase, therefore it is an important factor and useful for the Linear Regression.



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.

I believe my linear model is good, because it takes into consideration important factors such as Customer_Segmentation and Avg_Num_Products_Purchased to predict the profit.

We are looking for the lowest possible P-Values for the categorical variables, after trying and error I got to Customer_Segmentation and Avg_Num_Products, It is possible to observe in the report below how P-values have a value < 2.2e-16 this can be written as < .00000000000000022 This is very small number this is very significant variable for my model ensuring a more reliable output.

R-squared is a coefficient between 0 to 1, closer to 1 means almost all variance in the target variable are explained by the model, and a score equal or greater than 0.70 is considered a strong model. therefore when observe the report shows and shows a R-square of 0.836, it is possible to ensure that the model is strong to predict the profit.

Report for Linear Model Linear_Regression_5

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

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.*

The Linear Regression equation is
Y = 303.46 + - 149.36 * Loyality Club Only + 281.84 * Loyality Club Only and Credit Card + - 245.42 * Store Mailing List + 66.98 * Avg_Num_Products_Purchased

In the example of the first record in the mailing list 'A Giametti' it would be:
355.04 = 303.46 + - 149.36 * 1 + 281.84 * 0 + - 245.42 * 0 + 66.98 * 3

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 recommendation is that the company send out the catalog for the 250 customers.

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

Considering that the predicted profit from customer list delivery of catalogs is in a total of $2,1987.44 after the costs of printing and distributing is $6.50 per catalog and the average gross margin (price - cost) on all products sold through the catalog is 50%, the recommendation would be to proceed and deliver the catalogs to the customers.

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

The predicted profit is 21,987.44