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?

Ans: A company that manufactures and sells high-end home goods, wants to determine if they send their printed this year's catalog to the 250 new customers will be profitable or not. That is, if the expected profit from these new customers exceed $10,000, they will send the catalog. Otherwise, they will not send it.

2. What data is needed to inform those decisions?

Ans: Two files are given.

1. **p1-customers.xlsx -** This dataset contains historical data about 2,300 customers.
   - This data set is used to build the linear regression model to predict the expected revenue from the new 250 customers.
2. **p1-mailinglist.xlsx -** This dataset contains information of 250 new customers.
   - This is the list of customers that the company wants to send catalog. So that decision will be profitable or not, we have to find.
   - The total profit has to be calculated. For that we need the average gross margin (given 50%), the cost of printing and sending one catalog (given $ 6.5) and the probability that the customer WILL respond to the catalog and make a purchase (given Score_Yes).

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)

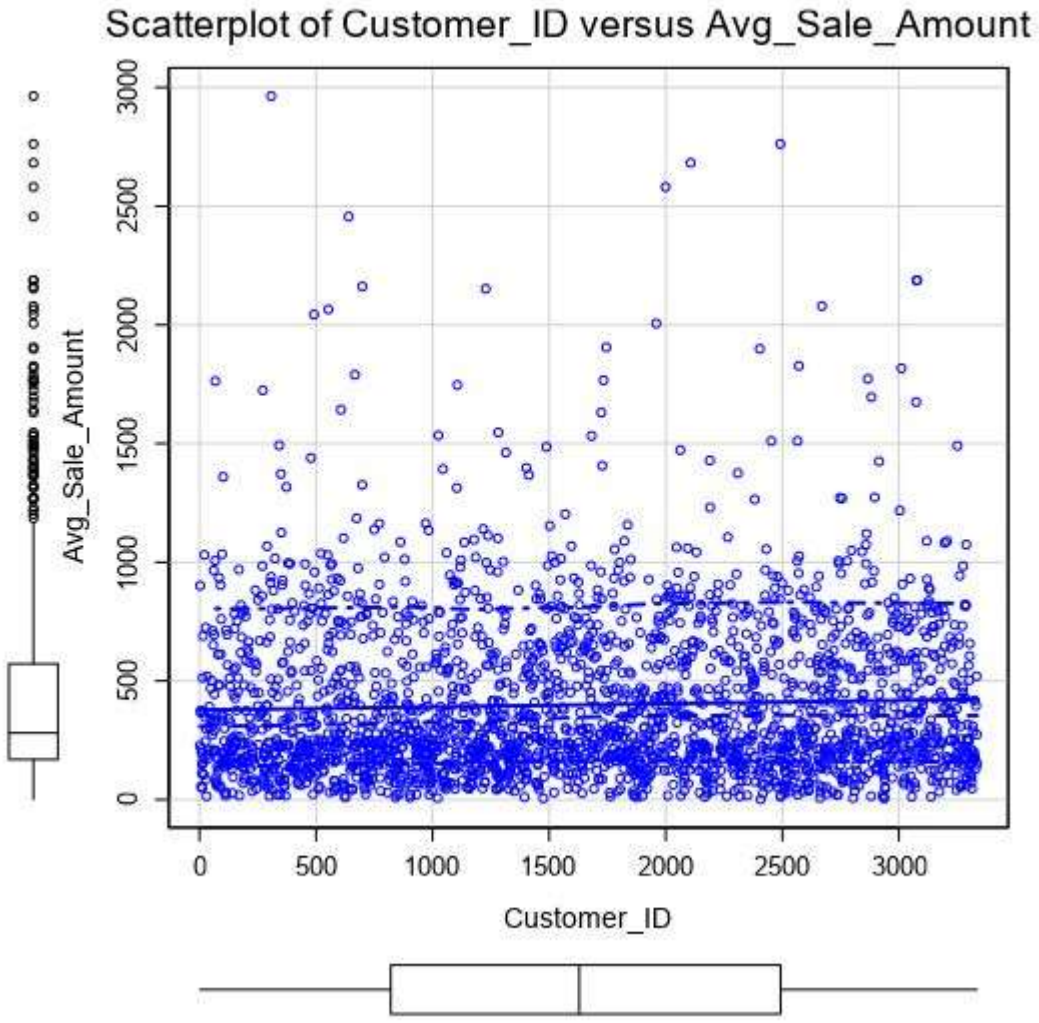
Important: Use the p1-customers.xlsx to train your linear model.

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.

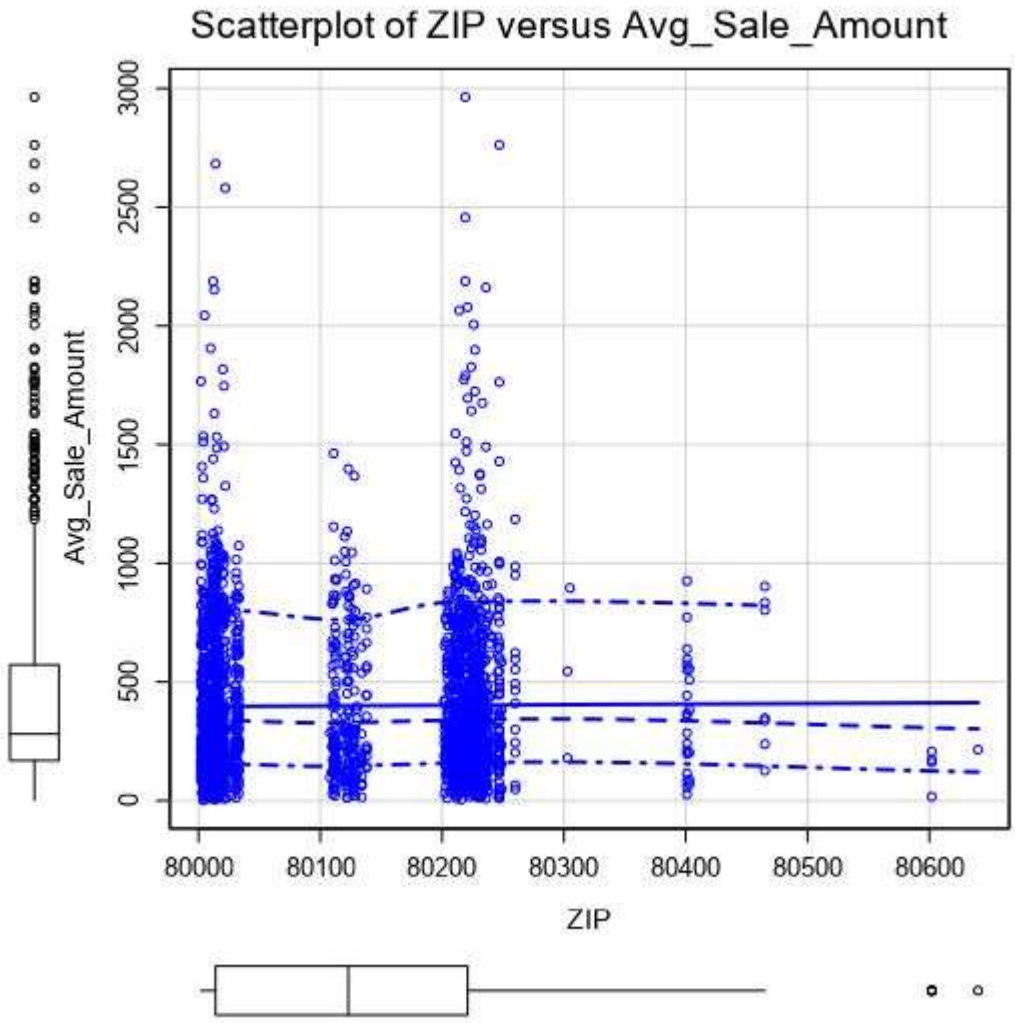
Ans: Step 1: Responded_to_Last_Catalog variable cannot be a predictor variable, because it is not there in the **p1-mailinglist.xlsx** dataset.

Step 2: Next, I took the numeric variables one by one, and examine their relationship with the Target variable **Avg_Sale_Amount.** Since these are Numeric variables, their relationship with the target variable can be found using SCATTER PLOT. So I plotted them one by one and arrived a conclusion of selectin as a Predictor variable.
First, I tried a scatterplot of Customer_ID vs Avg_Sale_Amount.



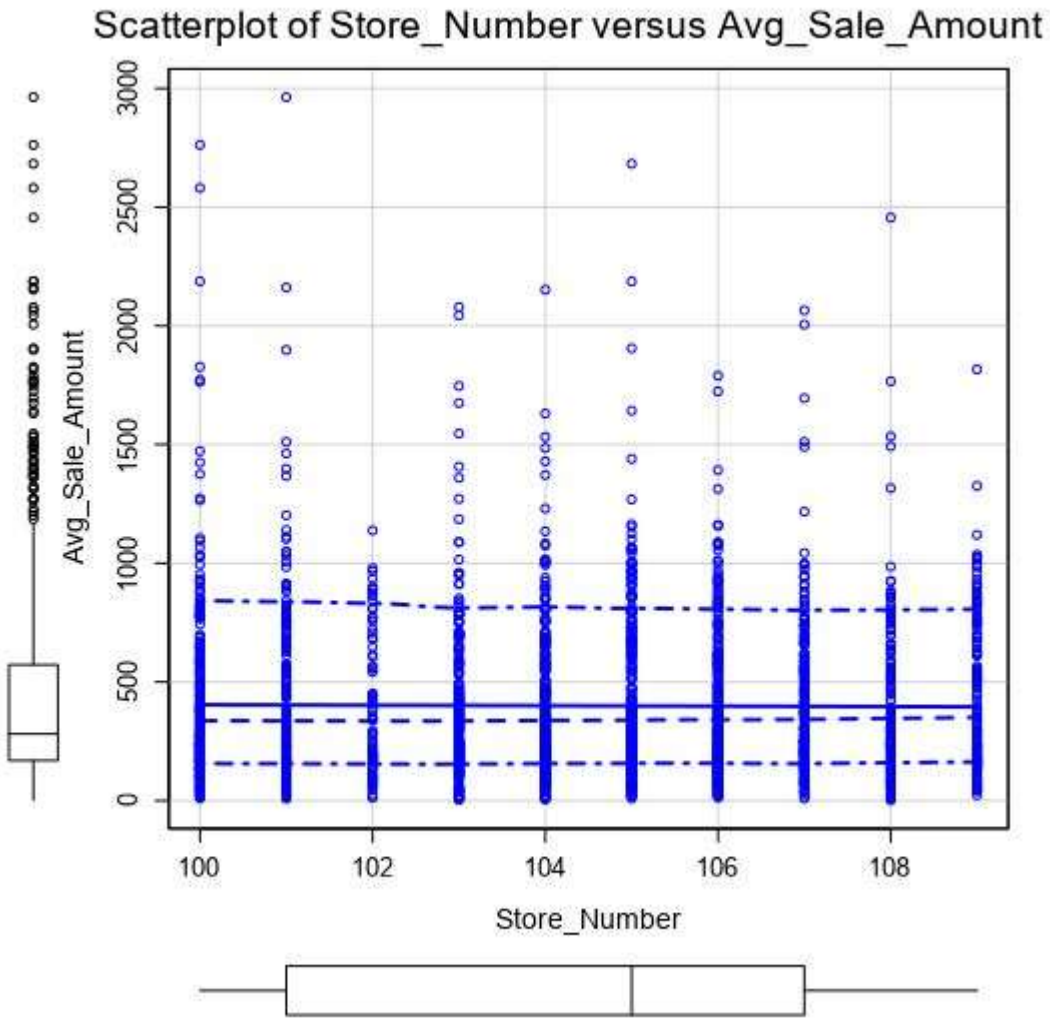
The above Scatterplot clearly shows that there is **NO Linear Relationship** between these 2 variables. So I decided to exclude Customer_ID from my linear model.

Step 3: Next I tried scatter plot of Zip vs Avg_Sale_Amount.



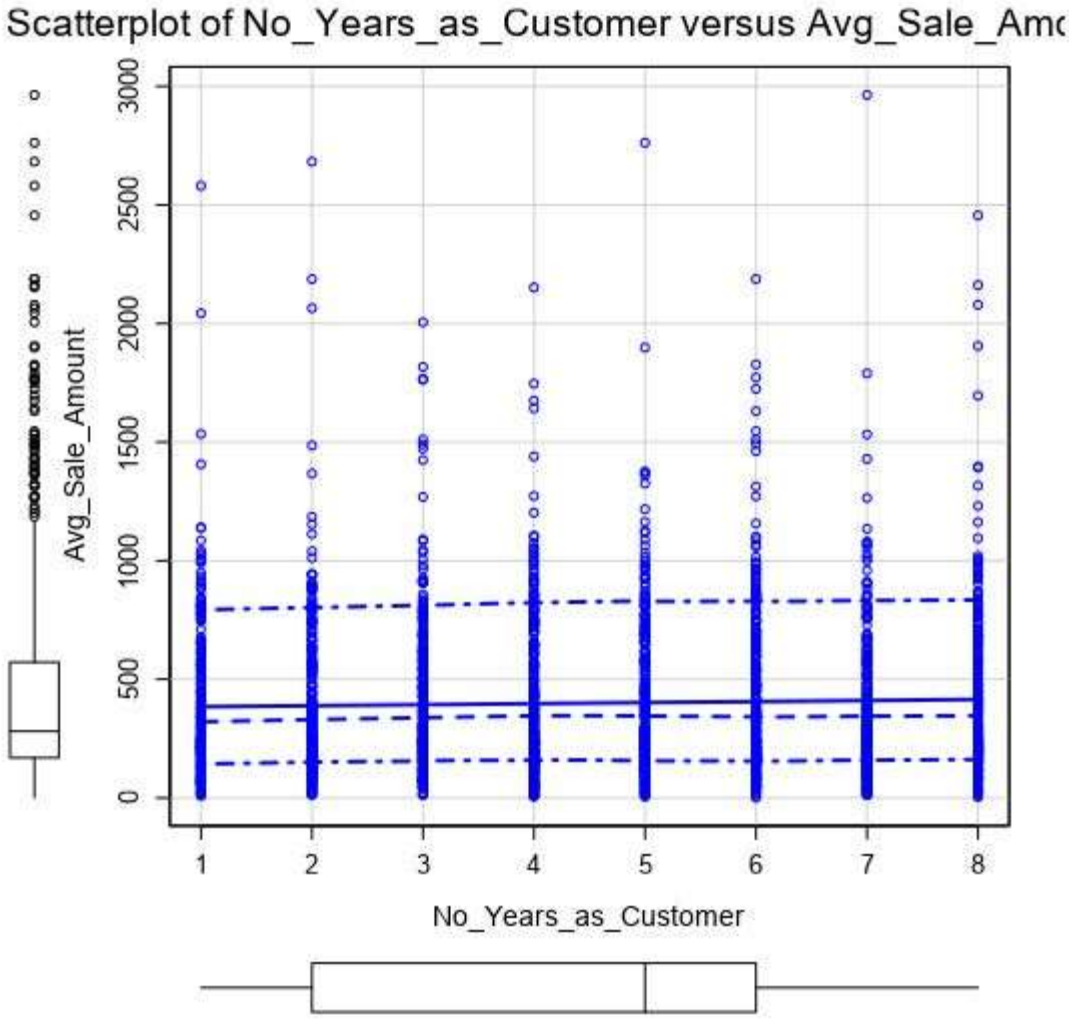
The above Scatterplot clearly shows that there is **NO Linear Relationship** between these 2 variables. So I decided to exclude Zip from my linear model.

Step 4: Next I tried a scatterplot of Store_Number vs Avg_Sale_Amount.



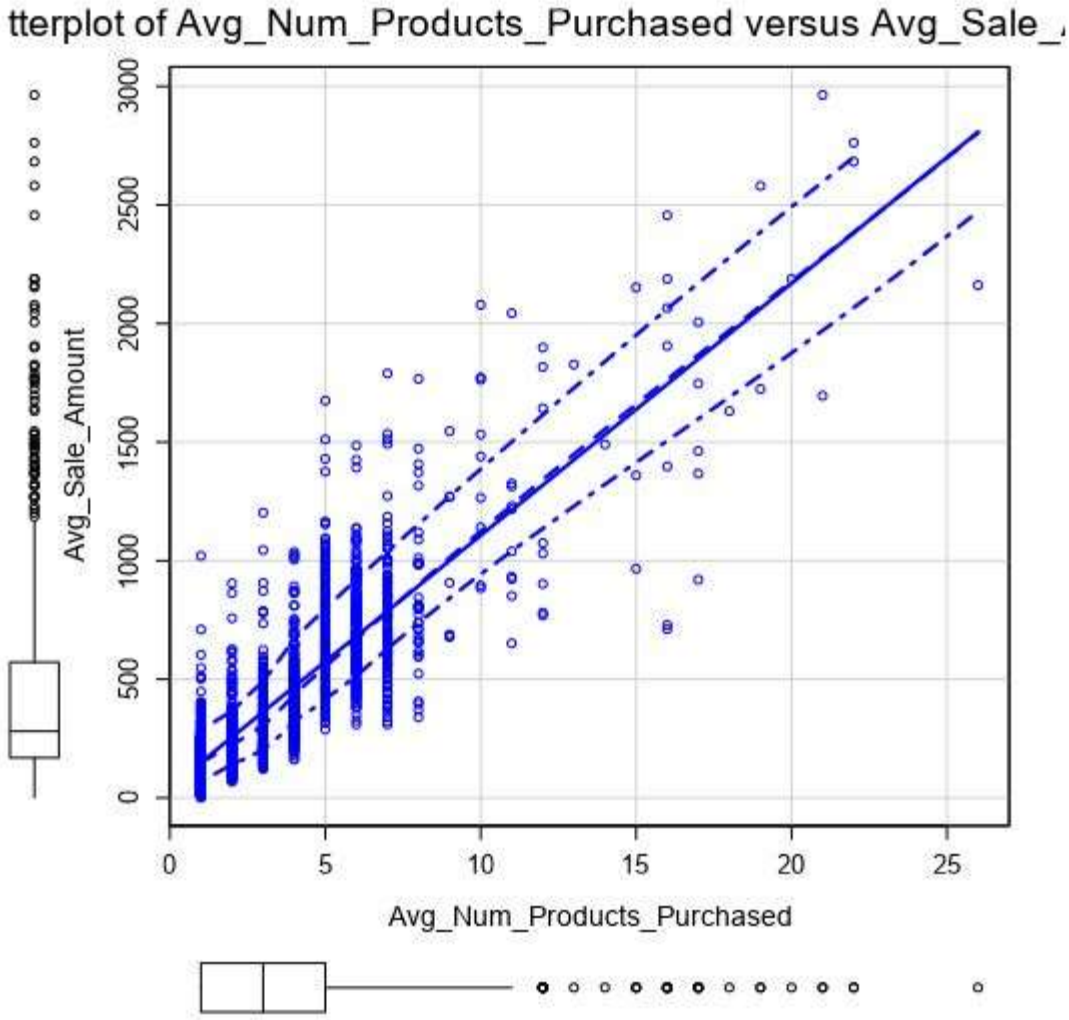
The above Scatterplot clearly shows that there is **NO Linear Relationship** between these 2 variables. So I decided to exclude Store_Number from my linear model.

Step 5: Next I tried a scatterplot of No_Years_as_Customer vs Avg_Sale_Amount.



The above Scatterplot clearly shows that there is **NO Linear Relationship** between these 2 variables. So I decided to exclude No_Years_as_Customer from my linear model.

Step 6: Next I tried a scatterplot of Avg_Num_Products_Purchased vs Avg_Sale_Amount.



The above scatterplot clearly depicts the Linear Relationship between Avg_Num_Products_Purchased and Avg_Sale_Amount. So I decided to **include** Avg_Num_Products_Purchased for my linear model.
From the above scatterplots, I came to a conclusion that **Avg_Num_Products_Purchased** is the **ONLY numerical variable** that I am going to use as a **PREDICTOR** variable for building my model.

Step 7: Categorical variables cannot be used in scatter plots. So first I've used all the categorical variables for building the model, and check their significance from the report. If their statistical significance value is < 0.05, I decided to use it as a Predictor. Otherwise I planned to exclude from my model.

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Customer_ID + ZIP + Store_Number + Avg_Num_Products_Purchased + No_Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-667.40	-67.94	-2.06	71.85	969.04

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.379e+03	2.149e+03	-0.6416	0.52118	
Customer_SegmentLoyalty Club Only	-1.497e+02	8.980e+00	-16.6659	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.824e+02	1.193e+01	23.6659	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.459e+02	9.774e+00	-25.1627	< 2.2e-16	***
Customer_ID	-1.373e-03	2.941e-03	-0.4669	0.64063	
ZIP	2.248e-02	2.660e-02	0.8451	0.39814	
Store_Number	-1.011e+00	1.007e+00	-1.0042	0.31539	
Avg_Num_Products_Purchased	6.700e+01	1.517e+00	44.1582	< 2.2e-16	***
No_Years_as_Customer	-2.345e+00	1.223e+00	-1.9167	0.0554	.

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

Residual standard error: 137.43 on 2366 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8367
F-statistic: 1522 on 8 and 2366 degrees of freedom (DF), p-value < 2.2e-16

The above is the report of the Linear Regression Model.

- The Customer_ID has significance value 0.64063, which is greater than 0.05, so I decided to exclude that variable.
- Similarly the significance of ZIP is 0.39814 which is greater than 0.05, I decided to exclude that variable.
- Also, the significance of Store_Number is 0.31539 which is greater than 0.05, I decided to exclude that variable.

Step 8: Now I used Customer_Segment and No_Years_as_Customer as my predictor variables and rerun my model. The result is as below:

Report

Report for Linear Model Catalog_Demand

Basic Summary

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-663.04	-68.42	-1.69	71.58	976.10

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	313.76	11.861	26.454	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.11	8.969	-16.625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.62	11.910	23.729	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.48	9.762	-25.146	< 2.2e-16	***
Avg_Num_Products_Purchased	67.02	1.514	44.255	< 2.2e-16	***
No_Years_as_Customer	-2.34	1.223	-1.914	0.0558	.

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

Residual standard error: 137.41 on 2369 degrees of freedom

Multiple R-squared: 0.8371, Adjusted R-Squared: 0.8368

F-statistic: 2435 on 5 and 2369 degrees of freedom (DF), p-value < 2.2e-16

The above result shows the significance of No_Years_as_Customer is) 0.0558 ≈ 0.06 which is greater than 0.05. So I've decided to exclude this variable also.
Hence, the **ONLY Categorical Variable** I've decided to use as a **PREDICTOR** variable for my linear model **is Customer_Segment**.
Finally, Customer_Segment and Avg_Num_Products_Purchased are the two predictor variables I am using to build my linear model because only these 2 variables are statistically significant.

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.

Ans: I believe that my linear model is a good model. Because the 2 variables I've used to build my model has high statistical significance.

1. Customer_Segment has p-value < $2.2e^{-16}$, which is much less than 0.05. So this shows it is highly statistically significant.

2. Avg_Num_Products_Purchased has also p-value < $2.2e^{-16}$, which is much less than 0.05. So this shows it is highly statistically significant.

3. **The R-squared value** of my model is: **0.8369**, which shows that my model is highly **Strong Positive.**

Report for Linear Model Catalog_Demand

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)

Ans: The Linear Equation from the model I built is:

Avg_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased - 149.36 (If Customer_Segment: Loyalty Club Only) + 281.84 (If Customer_Segment is Loyalty Club and Credit Card) – 245.42 (If Customer_Segment is Store Mailing List) + 0 (If Customer_Segment is Credit Card Only)

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.

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?

Ans: The question asked is, if the revenue generated by sending the catalog to the new 250 customers exceeds $10,000, the company should send the catalog. Otherwise no. From my prediction, the net profit for the company by sending this year's catalog is **$ 21,987.44,** which is above $10,000.
Hence, 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)

Ans: My Linear regression model produces the following formula.

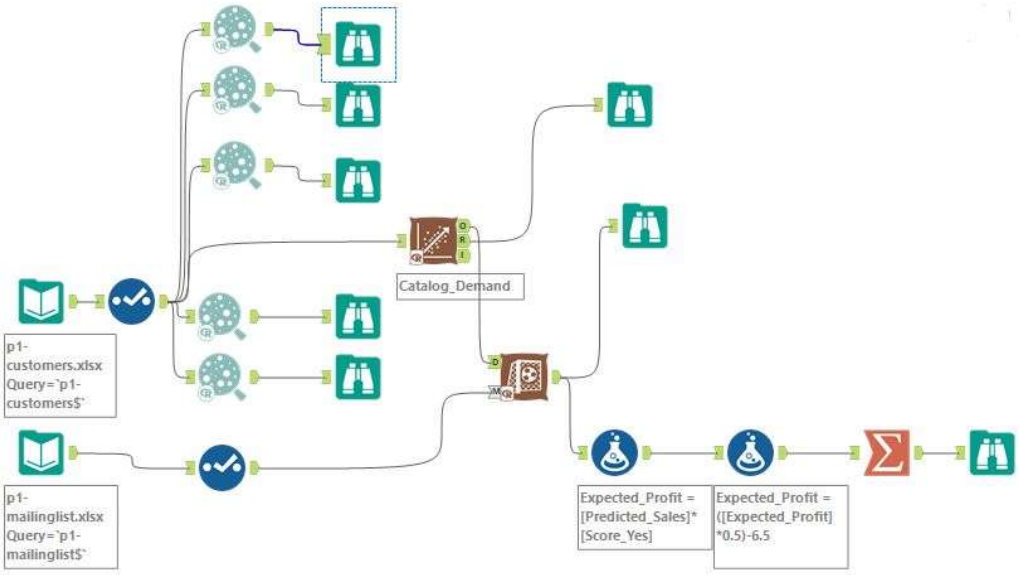
Avg_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased - 149.36 (If Customer_Segment: Loyalty Club Only) + 281.84 (If Customer_Segment is Loyalty Club and Credit Card) – 245.42 (If Customer_Segment is Store Mailing List) + 0 (If Customer_Segment is Credit Card Only).

- Apply the model to the mailinglist data set to get the Predicted_Sales.
- Then multiply this Predicted_Sales by Score_Yes (which is the probability to buy the catalog) for each customer which gives the Expected_Profit.
- Next, multiply Expected_Profit by 50% (0.5) which is the gross margin, then subtract the catalog cost ($6.5).
- Finally Sum up the Expected_Profit to get the total profit, which is calculated by our model : **$ 21,987.44.**

Below are the formulas I used:

Expected_Profit = Predicted_Sales * Score_Yes *
Predicted_Avg_Sales = (Expected_Profit * 0.5) – 6.5
Total Expected Profit = SUM (Expected_Profit).

I have used Alteryx software for building my model. The following is the Alteryx workflow of my Project:



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

Ans: The expected profit from my model prediction is: **$ 21,987.44.**

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.