Project 1: Predicting Catalog Demand

## **Step 1: Business and Data Understanding**

### **Key Decisions:**

1. What decisions needs to be made?

The decision that need to made for the problem in the mail-order catalog business is, “how much does the company sell a catalog on an average per customer and what are the factors that are effecting the sale. Moreover, the company is required to generate profit from 250 new customers such that it exceeds $10000 and to expand business to new clients by sending new catalog.

1. What data is needed to inform those decisions?

Information that is needed to inform the decisions of the company:

* Past customer data including average amount of sale, average number of product purchases from previous catalog, ways by which customer purchase the product (segmentation).
* Average amount spent by customers for the products bought from the company
* Data of new customers to whom the company want to send the new catalog including probability of sure purchases, associated costs like printing and distribution cost
* The average gross margin on all products sold through the catalog
* Average profit or amount that the company made to set it as a bench mark to achieve more profit for next year.

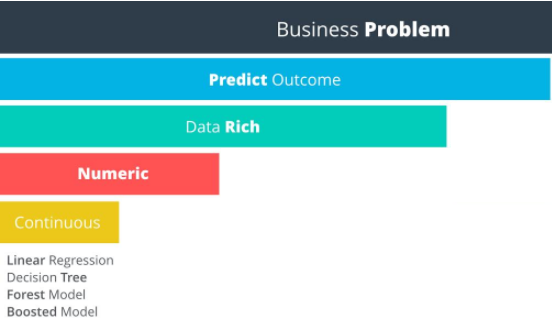
We need to calculate the average sale amount (Score) of the catalogue per customer and will calculate sale amount (Score) for 250 customers with their probability to purchase (Score\_Yes) next year. Then we will calculate the expected revenue from the 250 customers by average gross margin of 50% on all the products sold through the catalog. Lastly, we will calculate the profit (revenue-cost) or predict amount of money the company can expect from sending catalog to new customers.

I will use the Methodology Map flowchart to help us determine the type of analysis we should use to provide the exact information needed to inform the decision. We want to predict the Total profit for next year for sending the catalog to new customers.

## **Step 2: Analysis, Modeling, and Validation**

*At the minimum, answer these questions:*

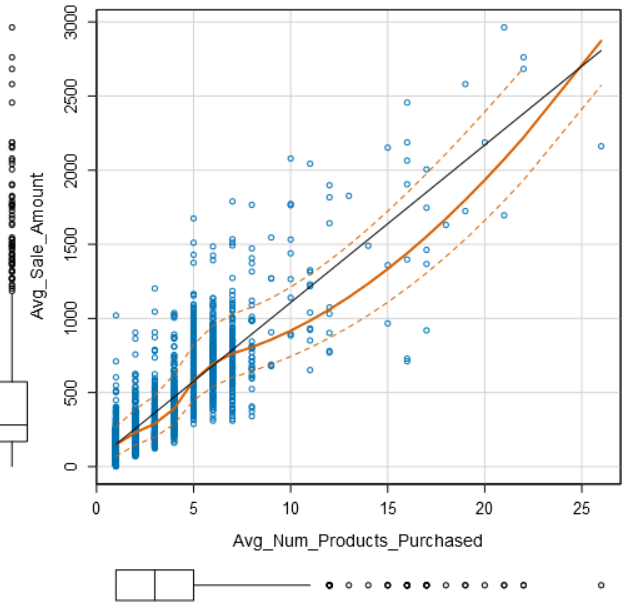
1. How and why did you select the [predictor variables](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48)  in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable.



From the historical data provided, it can easily be viewed that the outcome will be a numeric data. Example, the outcome for this scenario is the amount of profit the company will earn next year will be numeric. Hence, the Target variable is numeric.   
In the given problem, the Target variable is the “Average sale amount” as it is continuous and takes the value from all the factors effecting it. Hence, we can use continuous models such as linear regression model.

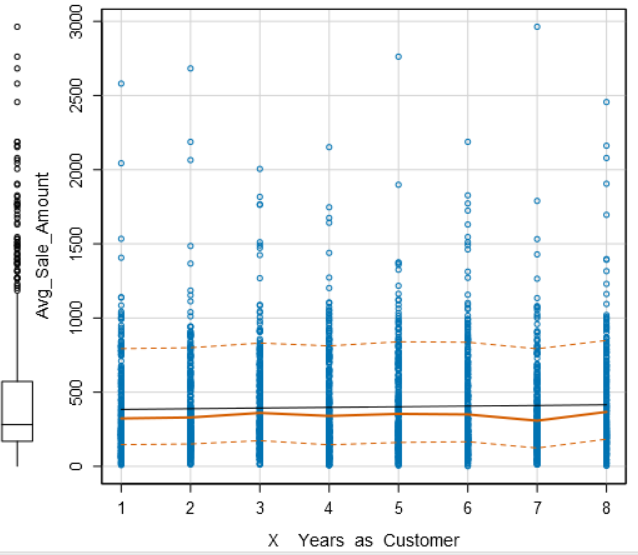
Numerical Variable:

Scatter plot of Avg\_Num\_Products\_Purches Vs Avg\_Sale\_Amount



There is a positive linear relationship between target variable and avg\_num\_product \_purchased, i.e. as the average purchases increases, average sale will also increase. Hence, Avg\_num\_product\_purchased is chosen as the predictor variable

Scatter plot of other variables Vs the target variable



There is no relationship between the target variables and other given variables

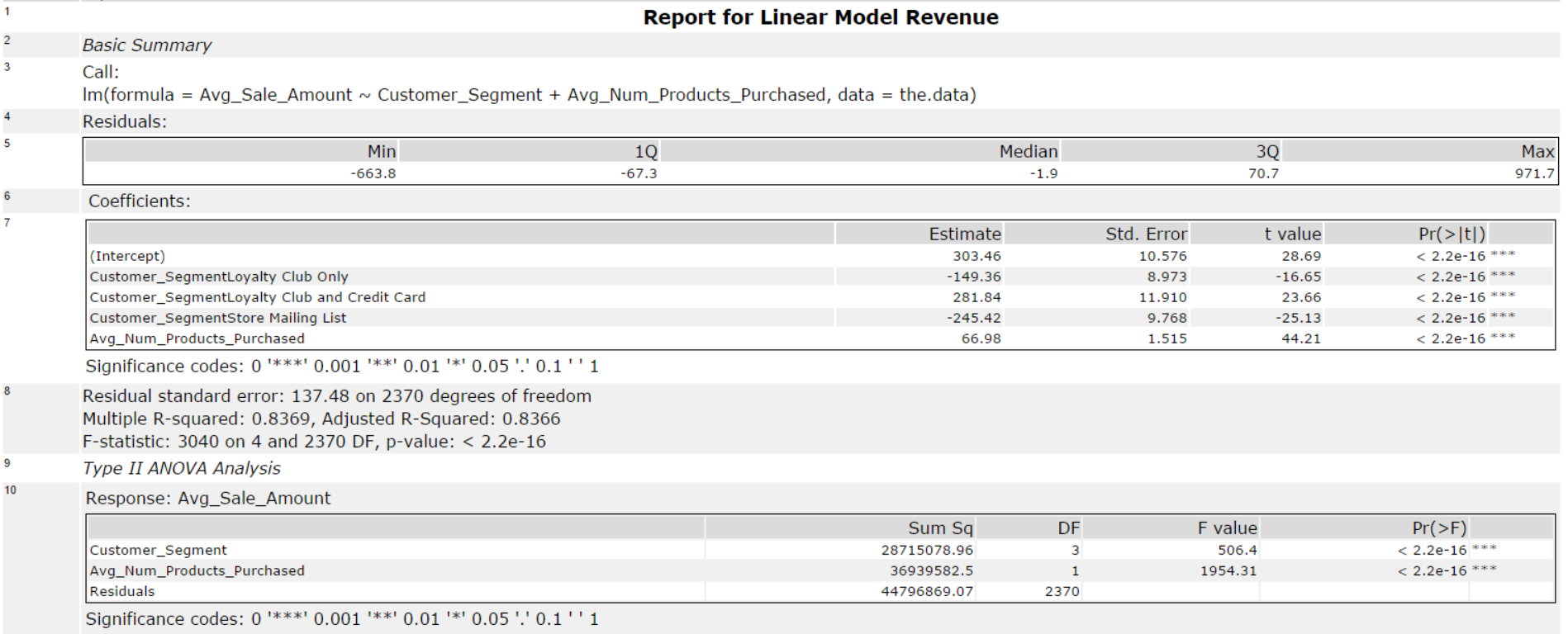
Categorical Variable:

The best way to check for a linear relationship is to run the categorical variables through the regression model, some of the categorical variable are Store no., customer segmentation.

The store number variable is not considered significant as Multiple R-squared is 0.63 which is low.

the customer segment coefficient is significant as Multiple R-squared is 0.83 which is relatively high.

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.



Statistical significance:

The p-value for Customer segments and Avg\_Num\_Products\_Purchased is <2.2e-16, which implies lowest value. P-value => 0.05 <.00…

Therefore, there is a relationship between the target variable and the predictor and are significant.

R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. The higher the r-squared, the higher the explanatory power of the model.  
The R-Squared value and the Adjusted R-Squared value is 0.8366 which is high and implies good-fit for 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)

**Important: The regression equation should be in the form:**

*Y = Intercept + b1 \* Variable\_1 + b2 \* Variable\_2 + b3 \* Variable\_3……*

Let Y be Avg\_Sale\_Amount  
b1= Customer\_SegementLoyalty Club Only  
b2= Customer\_SegementLoyalty Club and Credit Card

b3= Customer\_SegementStore Mailing List

b4= Avg\_Num\_Products\_Purchased

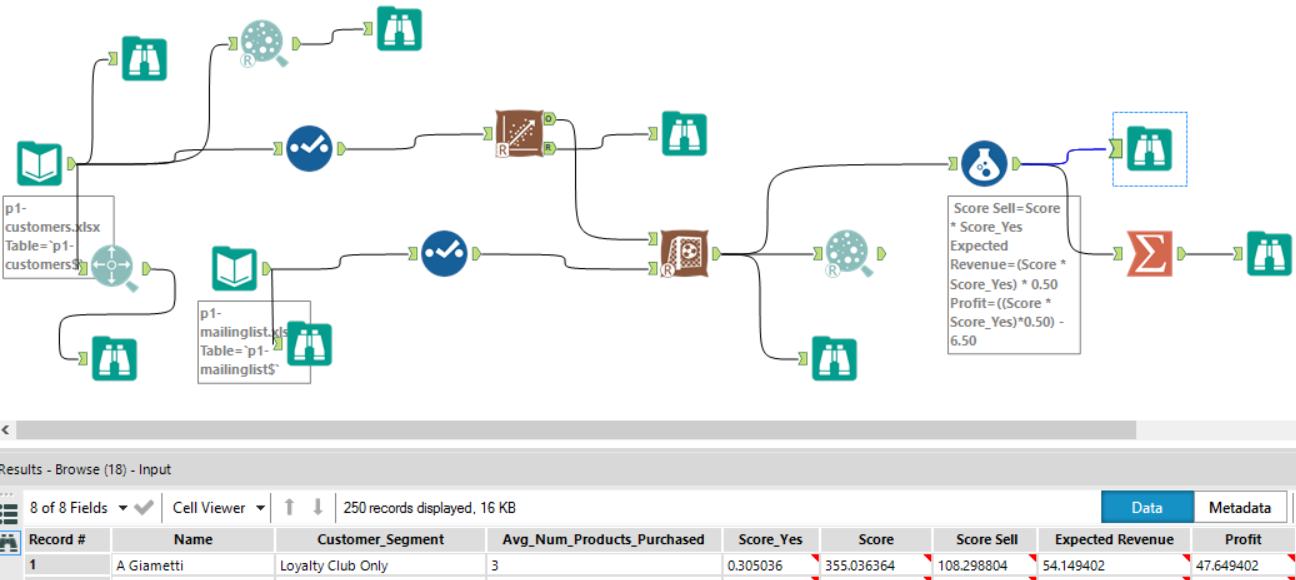
Y = 303.46 – 149.39 \* Customer\_SegementLoyalty Club Only + 281.84 \* Customer\_SegementLoyalty Club and Credit Card – 245.42 \* Customer\_SegementStore Mailing List + 66.98 \* Avg\_Num\_Products\_Purchased + 0 \* (If credit card only)

## **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?



I recommend that the company should send the catalog to all 250 customers as the profit exceeds $10000

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

* I have calculated the average sale amount (Score) of the catalog per customer = 355.03
* Then calculate sale amount (Score) for 250 customers with their probability to purchase (Score\_Yes) next year = 108.29
* Calculate Expected Revenue by multiplying Average Gross Margin of 50%
* Calculated the Profit = Expected Revenue – Cost (i.e. 6.50 per catalog )  
  Then Calculated the Total profit by summing the profit from all 250 customers.

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

The Expected Total Profit from the new catalog from 250 customers is $21987.43

