Project 1: Predicting Catalog Demand

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

### **Key Decisions:**

1. What decisions needs to be made?

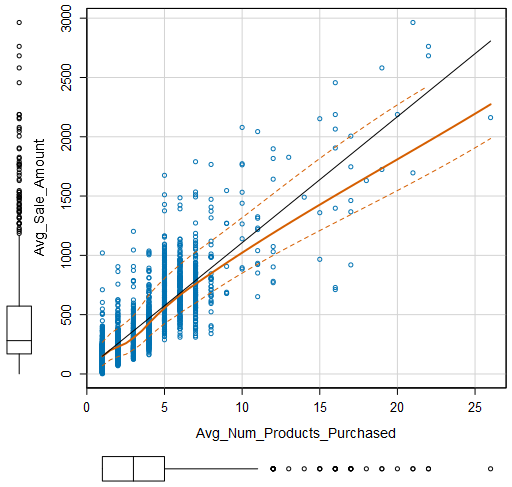
Determine if we should send out the catalog to potential customers or not. Since it costs $6.50 for printing and distributing a catalog, if the expected profit doesn’t exceed $10,000 then we should not send out them.

1. What data is needed to inform those decisions?

To predict the profit, we need customer purchase tendency historical data.

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

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.



I selected “City”, “Avg\_Num\_Products\_Purchased” and “Customer\_Segment” as the predictor variables in my model. To avoid over-fitting, I think low level information such as ZIP and address is not suitable for predictor variable. In addition, according to the scatter plot, it seems that there is a linear relationship between “num\_products\_purchased” and “sale\_amount.” On the contrary, there is no obvious relationship between “sale\_amount” and “#\_of\_years\_as\_customers” from scatter plot' s view.

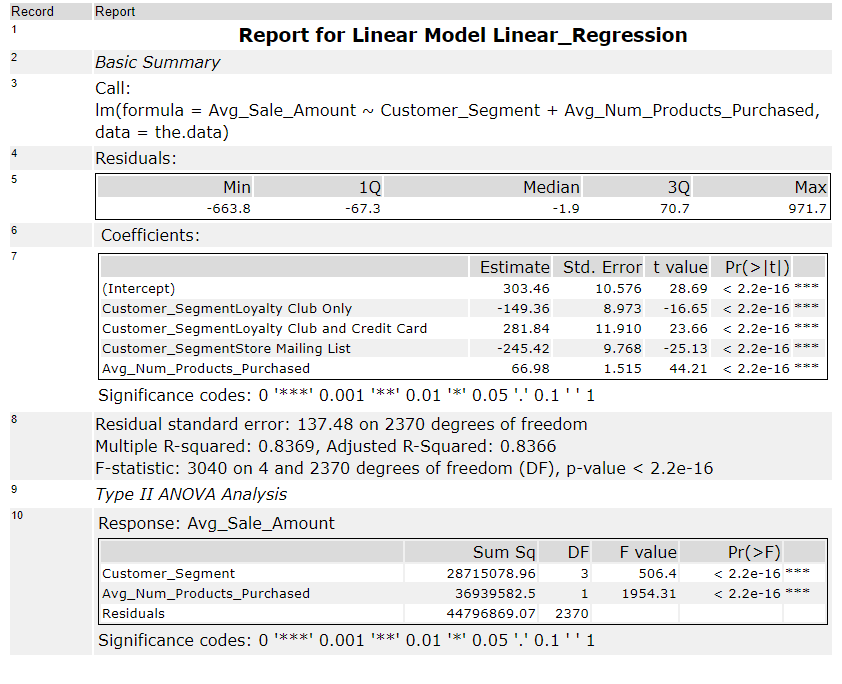
2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

In this model, the adjusted R-Squared value is 0.8366, not too bad.

The p-value of all predictor variables are less than 2.2e-16, indicating statistical significance.

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)

Avg\_Sale\_Amount = 303.46 + 66.98 \* Avg\_Num\_Products\_Purchased -149.36 (Loyalty Club Only) + 281.84 (Loyalty Club and Credit Card) – 245.42 (Store Mailing List)



## **Step 3: Presentation/Visualization**

1. What is your recommendation? Should the company send the catalog to these 250 customers?

Yes, based on the prediction, we should send out these catalogs.

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

I calculated the predicted sale amount through a linear regression model based on existing customer historical data. Then, substrate it with cost of sending out catalogs. Finally, multiply it with 50% margin. The result indicates that expected profit exceeds our target. Thus, we should send out the catalogs for profit.

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

Based on the calculation mentioned above. The expected profit is $22800.