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?

Decision is to whether or not to print and ship catalogs to 250 new customers.

Criteria is whether the predicted profits from the 250 new customers exceed $10,000. This is after considering the cost of printing and distributing each catalog at $6.50 as well as the average 50% gross margin from each product sold.

2. What data is needed to inform those decisions?

Need to predict how much sales can be made from each new customer based on historical customer data such as previous sales amount, number of products sold, segment or type of customer (loyalty card, credit card, mailing list), number of years as customer and geospatial location.

## **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.***

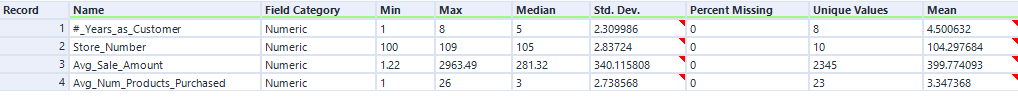
*At the minimum, answer these questions:*

1. How and why did you select the predictor variables in your model?

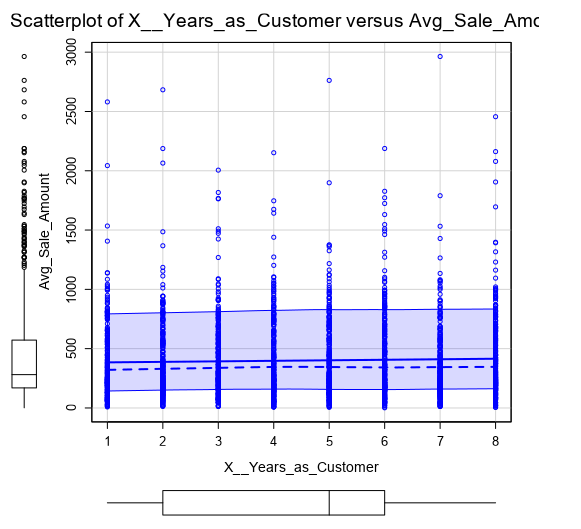
Data exploration is first performed to understand the distribution of the variables and their relationship with the target, ‘Avg\_Sale\_Amount’. Unique variables, ‘Name’, ‘Customer\_ID’ and ‘Address’ are dropped, as is ‘City’.

An initial linear regression model was fitted using the remaining variables, ‘Customer\_Segment’, ‘ZIP’, ‘Store Number’, ‘Avg\_Num\_Products\_Purchased’, ‘#\_Years\_as\_Customer’ against the target of ‘Avg\_Sale’Amount’ to calculate the strength of each of their relationship through their p values.

Numeric variables includes ‘#\_Years\_as\_Customer’, ‘Avg\_Sale\_Amount’, ‘Store\_Number’ and ‘Avg\_Num\_Products\_Purchased’. The overall statistics are shown below.

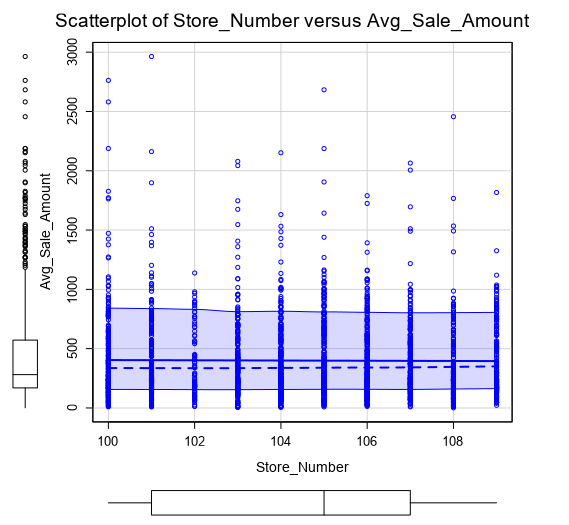


The scatterplot between ‘#\_Years\_as\_Customer’ and ‘Avg\_Sale\_Amount’ shows a rather poor relationship. This is supported by a p value of 0.05825 which indicates a weak correlation according to significance codes.

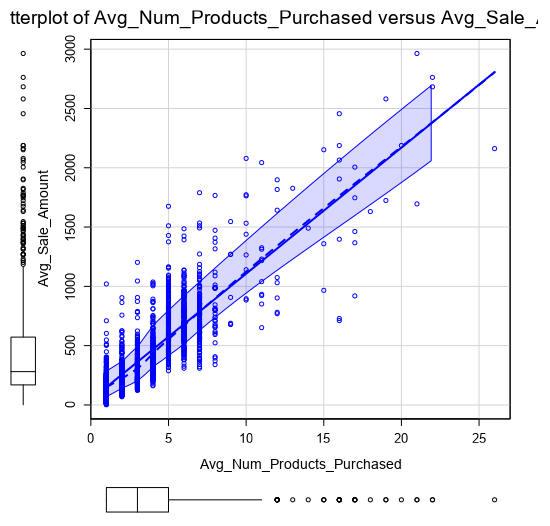




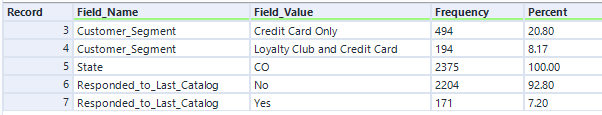
Similarly, for ‘Store\_Number’, the scatterplot with ‘Avg\_Sale\_Amount’ also shows a rather poor relationship and is supported by a p value of 0.32036 which indicates a weak correlation as it is greater 0.05.



For ‘Avg\_Num\_Products\_Purchased’, the relationship with ‘Avg\_Sale\_Amount’ is strongly linear based on the scatterplot and a very low p value of <2.2e-16.



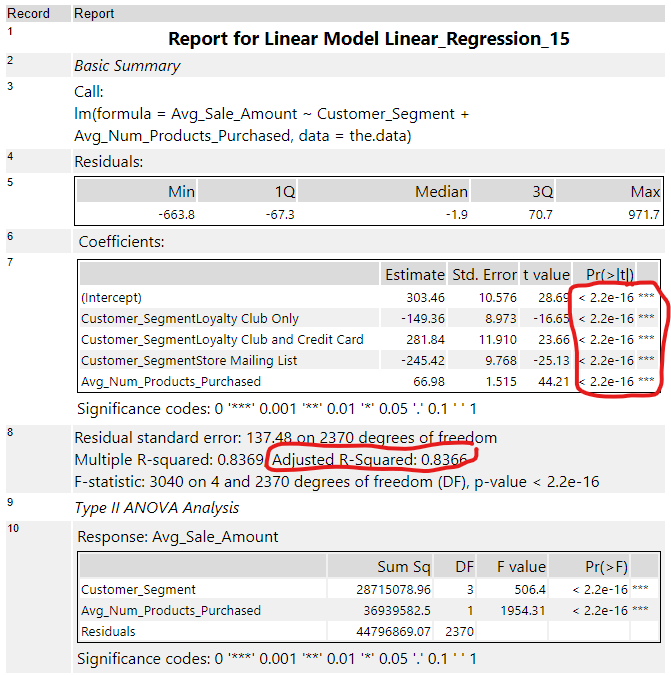
For non-numeric or categorical variables, ‘Customer\_Segment’, ‘State’ and ’Responded\_to\_Last\_catalog’, the overall statistics are shown below. Since all previous data are from the same state, CO, this variable is not useful and is dropped. ‘Responded\_to\_Last\_Catalog’ is also dropped since it is not available in the dataset for the new customers.



The final variables, which have strong correlations with ‘Avg\_Sale\_Amount’, selected are ‘Customer\_Segment’, ‘Avg\_Num\_Products\_Purchased’.

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

Fitting ‘Customer\_Segment’ and ‘Avg\_Num\_Products\_Purchased’ to ‘Avg\_Sale\_Amount’ with a multilinear regression model resulted in a strong correlated model with a high adjusted R2 of 0.8366 and very small p values of < 2.2e-16 for each coefficient as shown in summary below.



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)

Ayg\_Sale\_Amount = 303.46

– 149.36 \* Customer\_Segment(Loyalty Club)

+ 281.84 \* Customer\_Segment(Loyalty Club and Credit Card)

– 245.42 \* Customer\_Segment(Mailing List)

+ 66.98 \* Avg\_Num\_Products\_Purchased

## **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?

It is recommended to print and ship the catalogs to the 250 new customers as the expected profit is predicted to be greater than $10,000.

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

The linear regression model consists of 2 variables, one categorical (Customer\_Segment) and the other is numerical (Avg\_Num\_Producted\_Purchased).

For each new customer, the linear regression model was applied to predict their sales amount, which was then multiplied with the probability of sales (‘Score\_Yes’) to obtain the expected sales.

The 50% gross margin was then applied to the probable sales to obtain the gross. The cost due to printing and shipment for each catalog was offset by subtracting the $6.50 to get the profit for each new customer. Summing up the net profit for all 250 customer gives the total net profit which is greater than the $10,000 threshold.

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

Expected profit for 250 new customers is $21,987.44.