Project 1: Predicting Catalog Demand

# Step 1: Business and Data Understanding

## Key Decisions:

1. What decisions need to be made?

Ans:- Whether to print and send catalogs to the new 250 customers or not.

2. What data is needed to inform those decisions?

Ans:- The company will print catalog only if the expected profit exceeds $10,000. To find out the profit, one needs revenues, probability of ordering and costs. The profit is revenues times the probability of customer ordering less the costs of printing and sending out. Cost of printing and probability of customer ordering are provided. As for the revenues, they are dependent on two factors found by multiple linear regression namely customer segment and number of products ordered.

# Step 2: Analysis, Modeling, and Validation

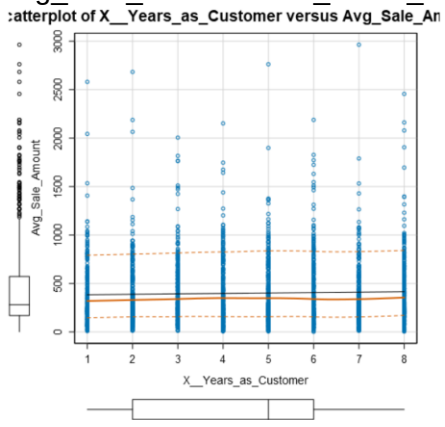
1. How and why did you select the predictor variables (see supplementary text) in your model?
2. Explain why you believe your linear model is a good model.

Ans:- For the numeric variables, I observed scatter plots of the variables vs Avg_sale_amount (revenues). For the non-numeric variables, I used linear regression and selected the variable with p <= 0.05 – selected numerical variable avg_num_products was in the model as well. The only one to fit the condition of p <= 0.05 was customer_segment with a very significant p-value of 0. p-value for the numeric variable avg_num_products was 0 i.e. very significant as well.

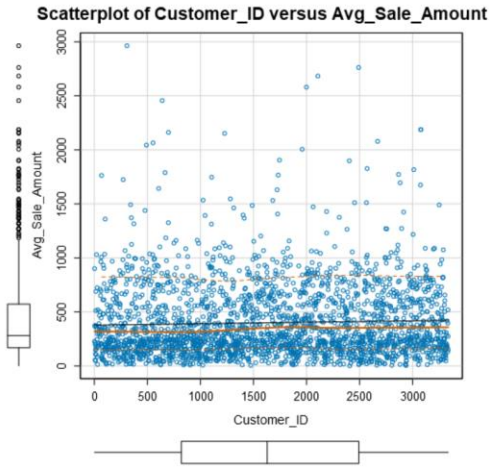
I think that the linear model is good because I chose variables with p <= 0.05 and that the high adjusted r-squared value is 0.8366 which is a very high value and suggests that the model is a good fit.

Below are the scatter plots:

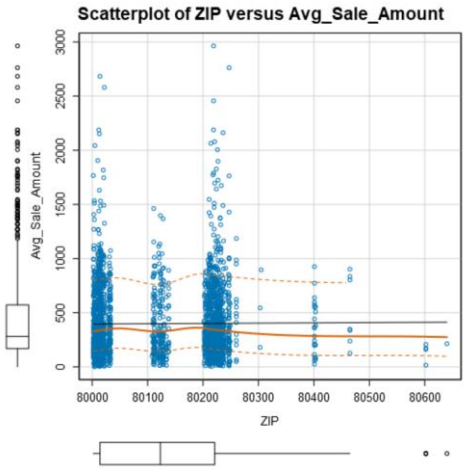
1. Avg_Sale_Amount vs #_Years_as_customer



2. Avg_sale_amount vs Customer_id

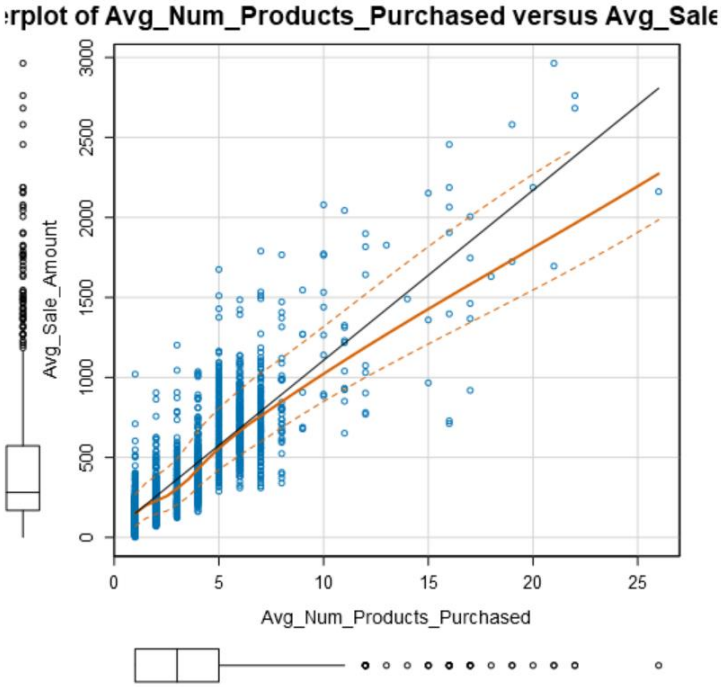



3. Avg_sale_amount vs ZIP




4. Avg_sale_amount vs store number

5. Avg_sale_amount vs Avg_Num_products. This was the only graph where any trend between the predictor and predicted variable was observed through the scatter plot. The two variables seem to be directly proportional i.e. increase in number of products brought seems to increase the sales amount.




3. What is the best linear regression equation based on the available data?

**Avg_Sale_Amount = 303.46 + 0.00 * Customer_SegmentCredit Card Only**
**- 149.36 * Customer_SegmentLoyalty Club Only**
**+ 281.84 * Customer_SegmentLoyalty Club and Credit Card**
**- 245.42 * Customer_SegmentStore Mailing List**
**+ 66.98 * Avg_Num_Products_Purchased**

# Step 3: Presentation/Visualization

1. What is your recommendation? Should the company send the catalog to these 250 customers?
2. How did you come up with your recommendation?
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

Ans:- Yes, the company should send catalog to the 250 new customers.

Ans: To come up with the recommendation, I used the management's directive which stated that it would make sense to print and send catalogs if the expected profit from these 250 customers would be greater than $10,000.

To calculate the profit, I needed revenues i.e. avg_sale_amount which I got from running linear regression using variables customer_segment and avg_num_products. Additionally, I also had the data on the probablitly

of customer ordering and not ordering. I used the probability of customer ordering with revenues expected less the cost of printing and sending to find the profit per customer. Profit_per_customer = revs * yes_probability – $6.50. I summed up for all the 250 clients and the amount came out to be $21,987.44. Since expected profit $21,987.44 is greater than $10,000 I recommed that the company should print and send catalogs to the new 250 customers.