# 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?*

**Answer:** Does it worth to pay $6.50 for each catalog and send them to 250 of new customers or not? If it's a profitable plan and by decided threshold the profit is more than $10,000, the decision will be publishing and distributing the catalog to new customers and if it's not a profitable act, the decision will be forgetting about the catalog advertisement and find another way for improving sales.

2. *What data is needed to inform those decisions?*

**Answer:** We need a dataset that contains some of the customers' characteristic and buying behavior. Like different groups of customers and how much each of them bought from the company's product on average. Based on these information we will build a model to predict how much product each of new customers would buy from the company.

## 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 (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.*

**Answer:** Between all available variables in the data set, I consider these fields to have a relationship with "Average Sales Amount(Avg_Sales_Amount)":

- Customer Segment
- Store Number

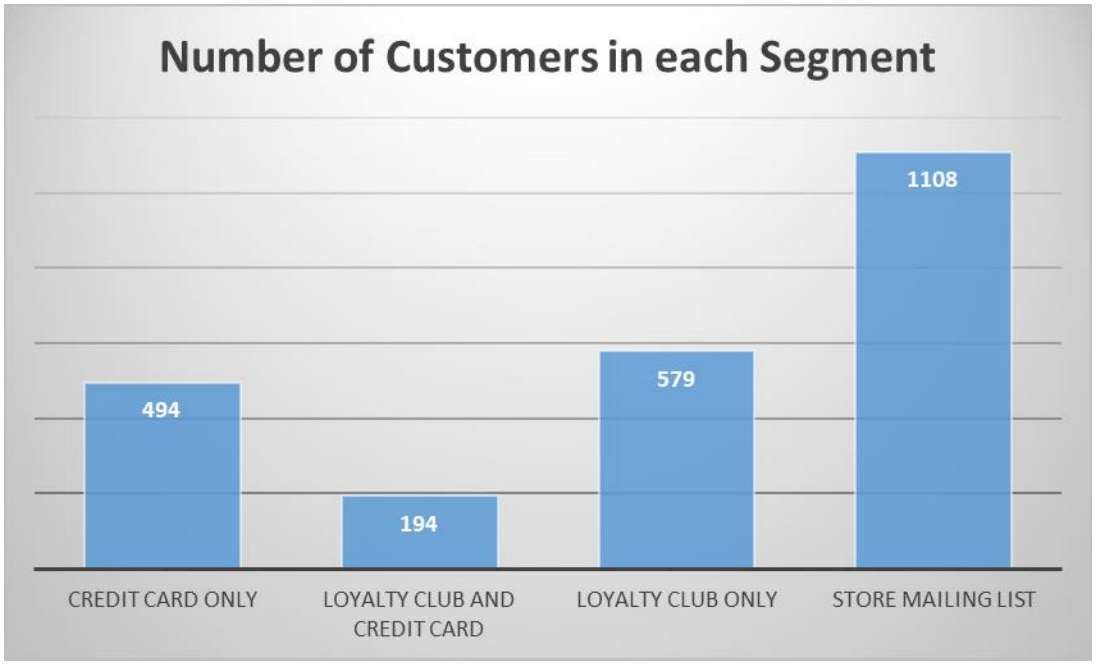
- Average Number of Products Purchased
- Number of Years as Customer

Customer segment could be a potential predictor for sales amount because it shows the customer category and group.
Store number could be another potential predictor because each store has its own identity and maybe some store attract more customers base on the customer service they provide and the behavior they show to the customers.
Average Number of Products that a customer purchased previously shows how engaged a customer with the products of the company and it could be a predictor for future purchases.
Number of years as customer shows the loyalty and degree of interest each customer has to the products and could be a predictor for amount of purchase they will have.

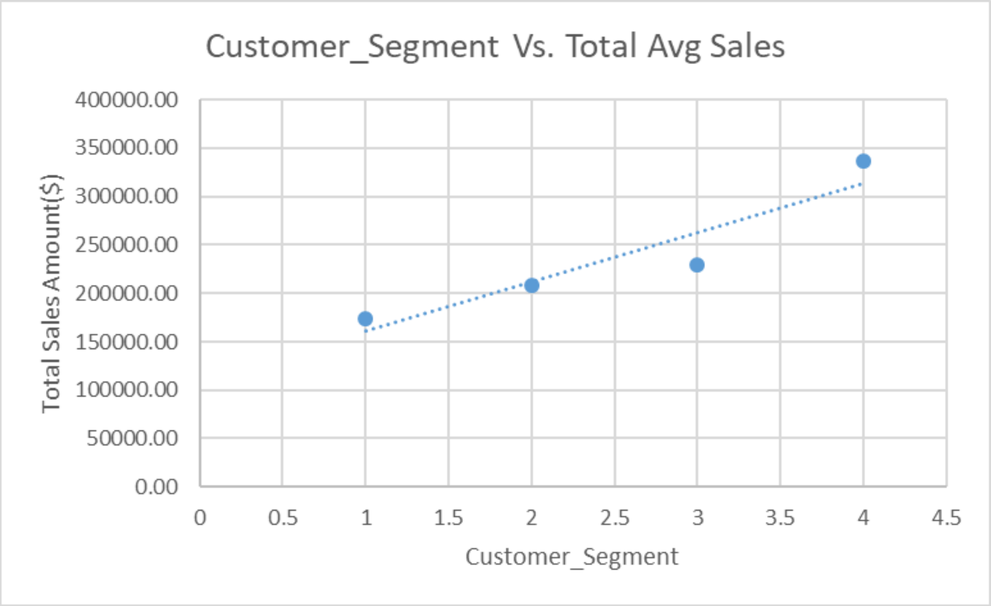
These are common sense and not enough. For choosing which variables are real predictors I will look at scatter plot for each of these variables and Average Sales Amount and I will calculate the correlation coefficient between each pair and if this parameter is more than 0.80 I will choose that variable as a predictor for sales amount.

Here are scatter plots:

Customer Segment:



As it has shown in the above bar chart, there is a clear difference number of customers in 4 different customer segments. Now I want to see what is the difference in total sales amount for each customer segment. So I aggregate the Avg_Sales_Amount on customer segments and if I represent each segment with a number the result will be this scatter plot:

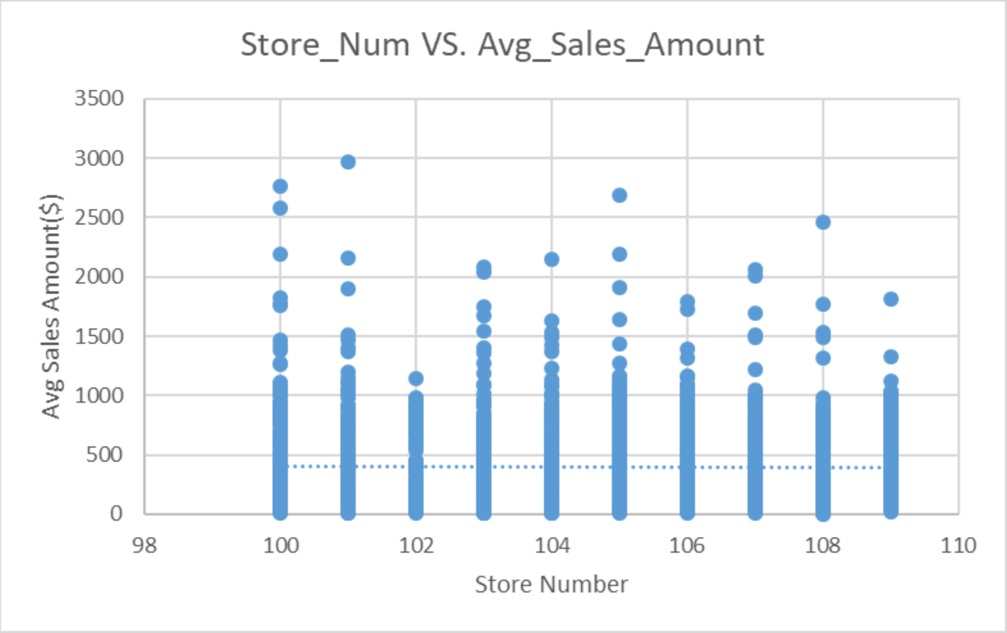


1: Store Mailing List
2: Loyalty Club and Credit Card
3: Loyalty Club Only
4: Credit Card Only

I am concluding that the Customer_Segment should be a good predictor for Avg_Sales_Amount and I will keep this categorical variable.

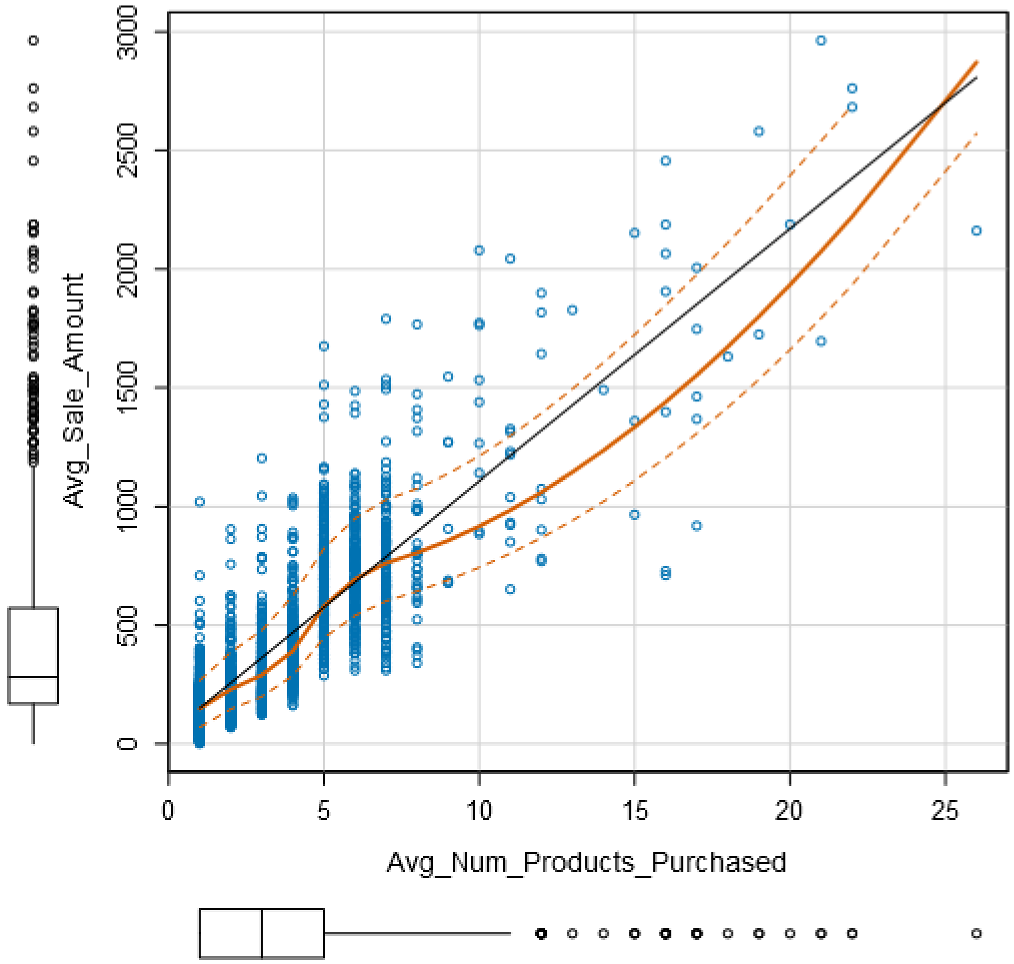
Store_Number:

Now I am looking at potential relationship between Store_Number and Avg_Sales_Amount.



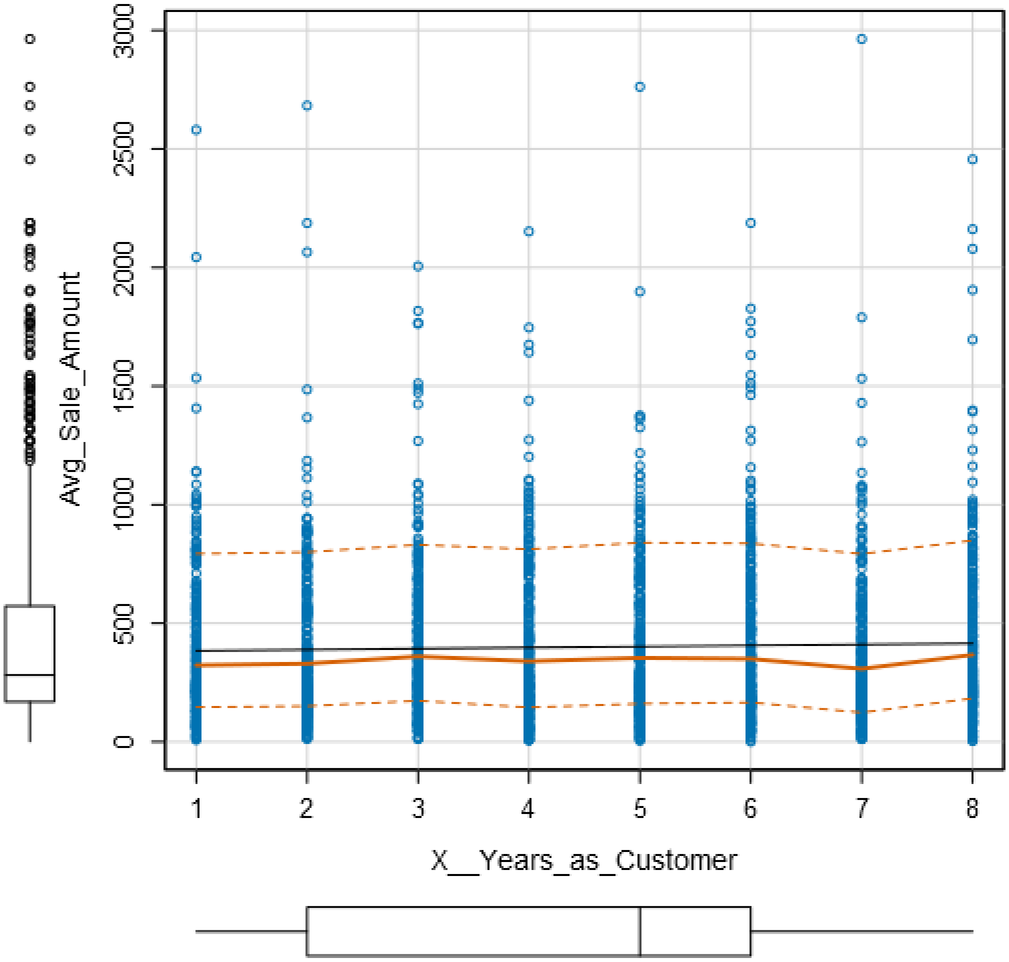
As it has been shown in the above plot, there is not a strong relationship between these two variables. Per the trend line there is no relationship and I don’t consider Store_Number as one of the predictors.

<u>Average Number of Products Purchased:</u>



Base on the above scatter plot there is a clear linear relationship between these two variables. The correlation coefficient for these two variables is: 0.86; So I will keep Avg_Num_Products_Purchased as one of the predictors.

Number of Years as Customer:



Base on this scatter plot there is no strong relationship between Number of years as customer and Avg_Sales_Amount. The correlation coefficient between these two variables is : 0.03 and it's very weak correlation. So I don't consider this variable as a predictor in my model.

In summary I choose two predictors:
One categorical variable: **Customer_Segment**
One numeric variable: **Avg_Num_Products_Purchased**

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.

**Answer:** This linear model is a good model because Multiple R-Square and Adjusted R-Square is 0.84 and in most cases if the R-Square is more than 0.70 the model is a good one. On the other hand for all selected predictors the p-value is very close to zero: < 2.2e-16; It means the probability of getting the results base on chance and like random numbers is almost zero.

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……*

**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.

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

**Answer:** I am recommending to send the catalog to these 250 customers. Because base on the linear model we implemented, there is a chance of achieving $21,987 net profit by sending these catalogs out and this profit is more than the $10,000 threshold; So the company should send catalog to customers.

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

**Answer:** We can use the linear model and plug the 250 new customers data and calculate the Avg_Sales_Amount for each of them. Then sum up the total Avg_Sales_Amount for all 250 customers. Base of information we have we know %50 of this total amount should be counted as profit. And by subtracting the cost of catalogs preparation ( 250 * 6.5) from that gross profit we will find out the net profit as : $21,987.44. So it worth to go for the catalog project and send them out.

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

**Answer:** The expected net profit is: $21,987.44;