# Predicting Catalogue Demand

## The Business Problem

Our company manufactures and sells high-end home goods. Last year the company sent out its first print catalogue, and is preparing to send out this year's catalogue in the coming months. The company has 250 new customers from their mailing list that they want to send the catalogue to.

Now, we need to determine how much profit the company can expect from sending a catalogue to these customers.

I will try to predict the expected profit from these 250 new customers. Management does not want to send the catalogue out to these new customers unless the expected profit contribution exceeds $10,000.

## Details

- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- We will multiply the revenue by the gross margin first before we subtract out the $6.50 cost when calculating the profit.

## Business and Data Understanding

### Key Decisions:

The company has 250 new customers we want to calculate the expected profit assuming that the new 250 customers will respond to the catalog received by mail.
If the expected profit excceds $10,000 then company will send the catalog to these new customers.

We have a dataset of 2,376 customer with details on the *customer segment, average number of products purchased, City, number of years as customer*, respond to last catalog and other information.
I will look for correlation between target variable which is the *Average sale amount* and each of the fields as predictor variables using linear regression. Then I will perform a multiple linear regression using selected fields with a high statistical significance (low p-value <=0.05) and high R-squared value.

To get the expected revenue, I will apply the multiple linear regression formula on the mailing list dataset. The I will use the result to calculate the expected profit

# Analysis, Modeling, and Validation

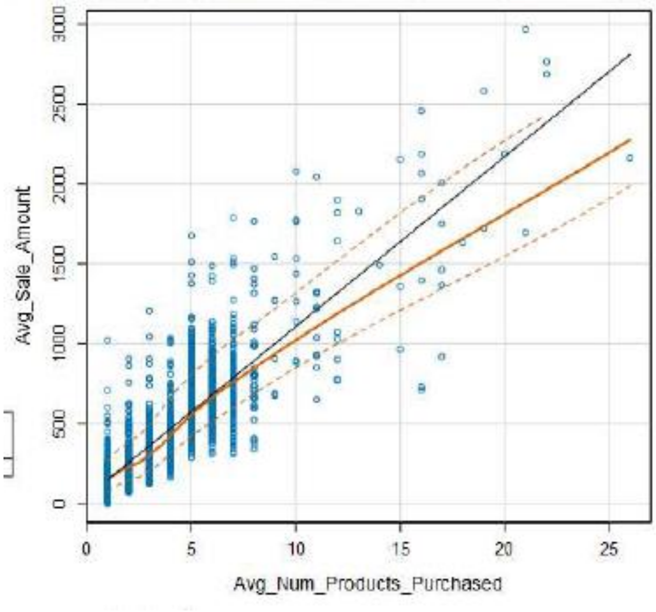
The *customers* dataset has multiple fields that may have an effect on the average sale amount.
Using multiple linear regression with these fields as predictor variable will give a clear picture on which of the predictor variables has a correlation with the average sale amount which means that we can find out which data we can include in our formula to calculate the expected revenue using the result of 2,376 customers.

Before I perform the multiple regression, I will investigate 2 continuous variables (average number of products purchased, number of years as customer) and apply a linear regression on each of the variables:

### Average number of products purchased vs Average sale amount

**R-squared** value of 0.7323 is a good fit to the data
**p-value** is 2.2e$^{-16}$ indicates a high statistical significance

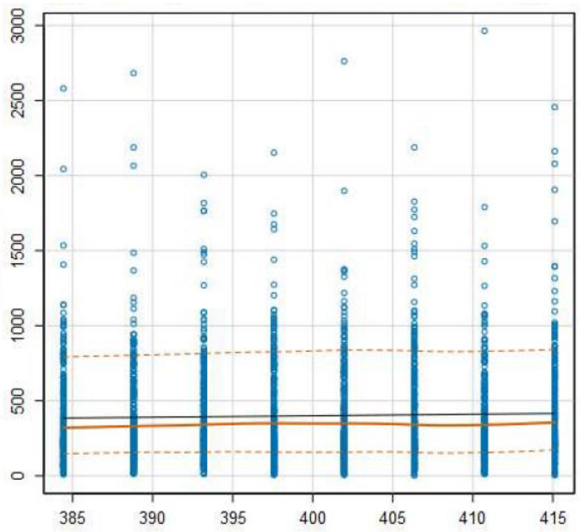
the **scatter plot** shows a strong positive correlation between *average number of products purchased items* ***vs*** *Average sale amount*



### Number of years as customer vs Average sale amount

**R-squared** value of 0.000887 is very low and the model doesn't explain the variability of the data
**p-value** is 0.1468 indicates a low statistical significance

The **scatter plot** shows no relationship between *number of years as customer* ***vs*** *Average sale amount. So, I will not include it in the multiple regression formula.*



Additional to the continuous variables, there are **3 categorial** variables (customer segment, city and respond to last catalog).
After performing linear regression on three variables the resulted values are as follows:

| Predictor variable | R-squared | p-value | Decision |
|---|---|---|---|
| **Customer Segment** | 0.7024 | $<2.2e^{-16}$ | Include the variable in multiple linear regression. It has high statistical significance (low p-value) and high R-squared value |
| **Respond to last catalog** | 0.0397 | $<2.2e^{-16}$ | Exclude. Although the p-value indicates high statistical significance but the R-squared is almost 0 |
| **City** | 0.008008 | 0.8374 | Exclude. (low statistical significance) |

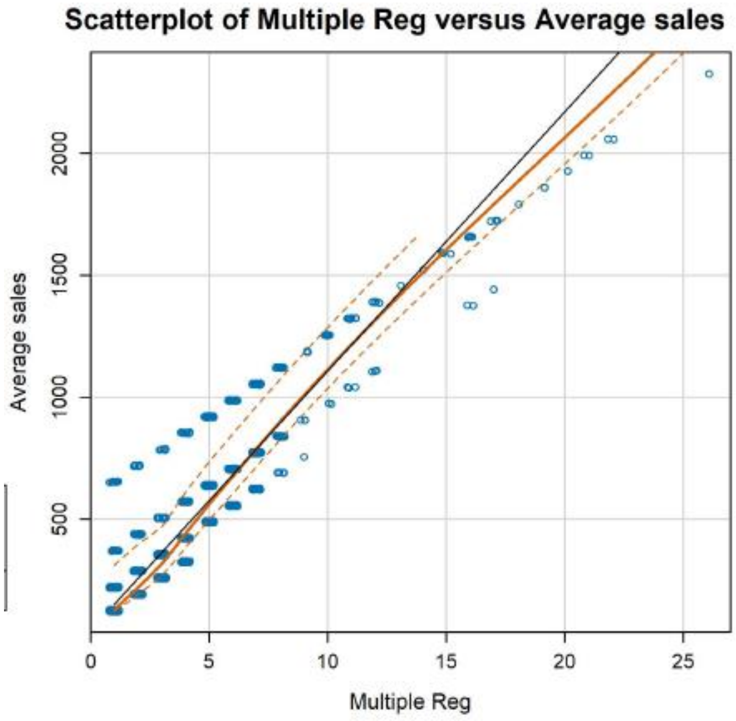
Based on the results in the above table I will include 1 categorial variable (Customer Segment) in the multiple regression model.

So, my multiple linear regression model will be based on 1 continuous variable (average number of products purchased) and 1 categorial variable (Customer Segment).

The formula resulted after performing the multiple linear regression is:

Avg_Sale_Amount = 303.46 + (66.89 *Avg_Num_Products_Purchased) – (149.36* Loyalty Club Only) + ( 281.84* Loyalty Club & credit card) – ( 245.42 * Store Mailing List)

It is a good model and highly predictive knowing that the **Adjusted R-squared** value is 0.8366 which indicated a good model fit to the customer data. And the **p-value** of < $2.2e^{-16}$ which is considered statistically significant.



The scatter plot for the multiple regression shows a positive relationship between average sales amount & the predictor variables mentioned in the formula above.

## Presentation/Visualization

### Recommendation

Based on the expected profit resulted of applying the multiple linear regression formula on the mailing list dataset which is greater than $10,000 I would recommend that the company will send the catalogs for the new 250 customers.

I came up with my recommendation after applying the following multiple regression formula on the mailing list dataset using the score tool in Alteryx to calculate the **Expected Revenue** :

Expected Revenue = 303.46 + (66.89 *Avg_Num_Products_Purchased) – (149.36* Loyalty Club Only) + ( 281.84* Loyalty Club & credit card) – ( 245.42 * Store Mailing List)

That created a new column **Expected Revenue.**

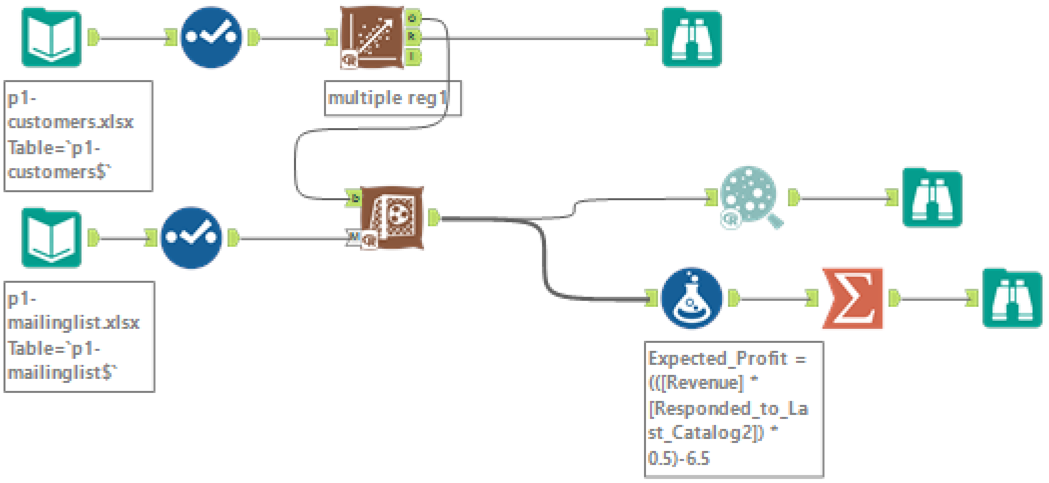
Then I have used the formula tool in alteryx to apply this formula expression to get the **Expected Profit :**

Expected Profit = (([Expected Revenue] * [Score_Yes]) * 0.5)- 6.5

**Where :**
**Score_yes** field indicates the probabilty that a customer will respond to the catalog
**0.5** is the Gross margin of %50
**6.5** is the cost of each catalog printing and distribution

### Conclusion

The expected profit from the new catalog (assuming the catalog is sent to these 250 customers) using the above expression is **$21,987.44**


p1-customers.xlsx
Table=`p1-customers$`
multiple reg1
p1-mailinglist.xlsx
Table=`p1-mailinglist$`
Expected_Profit = (([Revenue] * [Responded_to_Last_Catalog2]) * 0.5)-6.5