Project 1: Predicting Catalog Demand

Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project>

## Step 1: Business and Data Understanding

*Provide an explanation of the key decisions that need to be made. (500 word limit)*

The main decisions that need to be made in this project is to determine the amount of profit expected by the company by sending catalogs to 250 new customers and the catalogs can only be send if the profit expectation exceeds $10,000. The company already have dataset of 2,300 customers that catalog has been sent to in the past. We will have to build the model with this dataset, with Customer Segment and Average Number of Products Purchased serving as good predictor variables. The dataset already has average amount of sales which will be our target training variable that will help predict the sales in the future. In a nutshell, the key decision will be to know if the company can send out its catalog to the new customers or not.

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

The company need to decide if the company can send out its catalog to new customers or not by predicting the amount of money that being expected to be made from them (new customers) after sending out catalog to them.

1. What data is needed to inform those decisions?

The company already have dataset of 2,300 customers that catalog has been sent to in the past. We will have to build the model with the dataset of 2,300 customers provided by the company which the past record of customers average amount of sale per customer. We will use that to build a model and then run prediction for the new customer dataset to make an informed decision. We will also be needing 50% average gross margin (price - cost) on all products sold through the catalog and the costs of printing and distributing of all the catalogs to make an informed decision.

## Step 2: Analysis, Modeling, and Validation

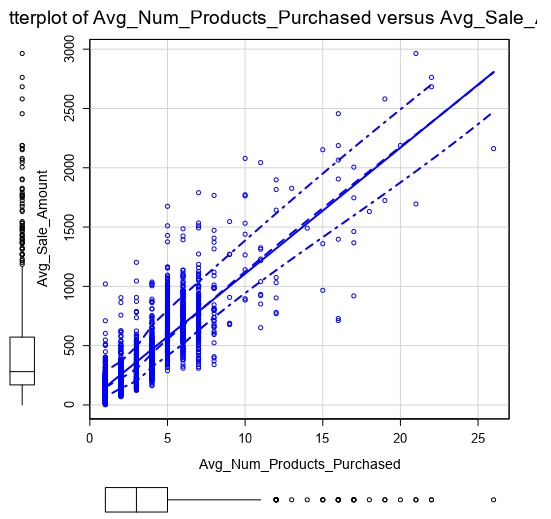
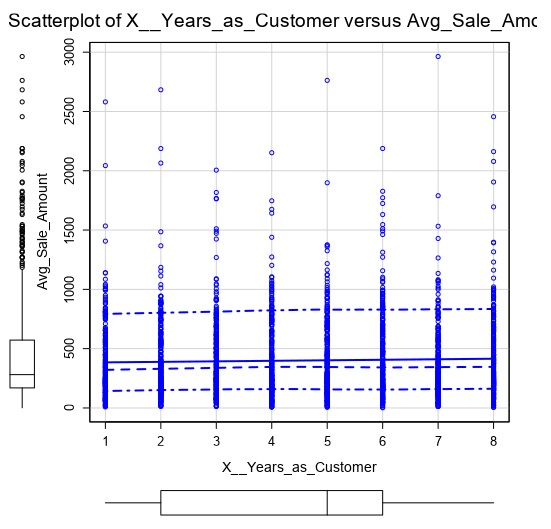
*Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)*

***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)](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) 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](https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/modules/bf705147-0d7c-4492-842a-698a6410a8a3/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

The predictor variable (Customer Segment and Average Number of Products Purchased) I used have higher correlation with the target variables while the remain variable don’t have significate relationship with the target variable. Also, the continues predictor variable, Avg Num Products Purchased, has a p-value of 2.2 \* 10^-16 which is much statistically significate for predictor variable (i.e p-value <= 0.05).

Visual correlation between predictor Visual correlation between predictor

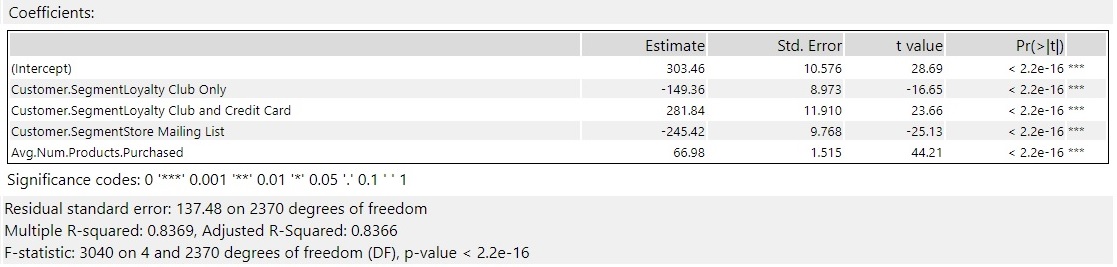
variable, Avg Num Products Purchased, variable, Num. of years as Customer,

and the target variable, Avg. Sale Amount. and the target variable, Avg. Sale

(p-value <= 0.05) Amount. (p-value >= 0.05)

1. 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.

The model we are trying to build is for predicting continues number, and also 84% of the variables (training dataset) used was explained by the model (Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366). With each variable used having p-value of below 0.05 (Customer Segment and Average Number of Products Purchased with both p-value of 2.2 \* 10^-16), we be 95% confidence there are good fit for our model.



1. 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)

*Y = 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)*

*Credit Card Only is the 0 coefficient for the Customer Segment categorical variable.*

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.

## Step 3: Presentation/Visualization

*Use your model results to provide a recommendation. (500 word limit)*

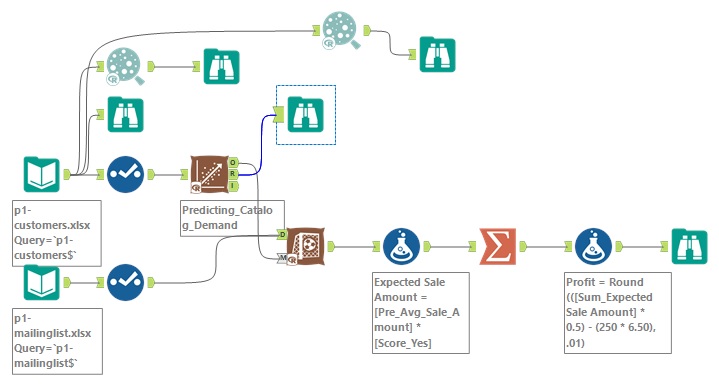
*At the minimum, answer these questions:*

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

Yes, the management was to send out the catalog only when the expected profit contribution exceeds $10,000. The profit expectation we have is $21,987.44, which is significate enough to send out the catalog to the 250 new customers.

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

By first predicting the Pre\_Avg\_Sale\_Amount which is then multiple by Score\_Yes (which is the probability the customer will respond to the catalog and make a purchase). The multiplication gives us the Expected Sale Amount variable, then sum all the Expected Sale Amount variable and then multiple it with 50% (average gross margin (price - cost) on all products sold through the catalog). Then subtract the costs of printing and distributing of all the catalogs (250 \* $6.50) from the outcome.



Model Workflow

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

The expected profit is $21,987.44

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/186/view) here. Reviewers will use this rubric to grade your project.