Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made.

Key Decisions:

1. What decisions needs to be made?

Background: Last year the company sent out its first print catalog and is preparing to send out this year's catalog. The company has 250 new customers from their mailing list to whom they want to send the catalog. Predict the expected profit from these 250 new customers. Management doesn't want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Decision: Do we send the catalog to the new customers?
Information: We need the "expected profit" from the 250 new customers to make the decision.

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

To predict the expected profit from the 250 new customers, we must use predictive analysis to help us obtain the data we need. We need to calculate the profit from these 250 new customers. What historical information we can use to make this decision? From the historical data, we see that we have customer personal information, average sale amount, where they made the purchase, whether they responded to the last catalog, average number of purchases made, and customer segment information. One could use this historical information to make a prediction.

Step 2: Analysis, Modeling, and Validation

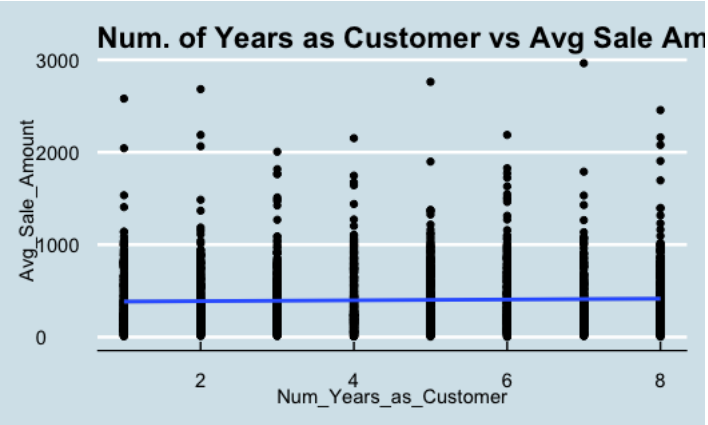
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged.

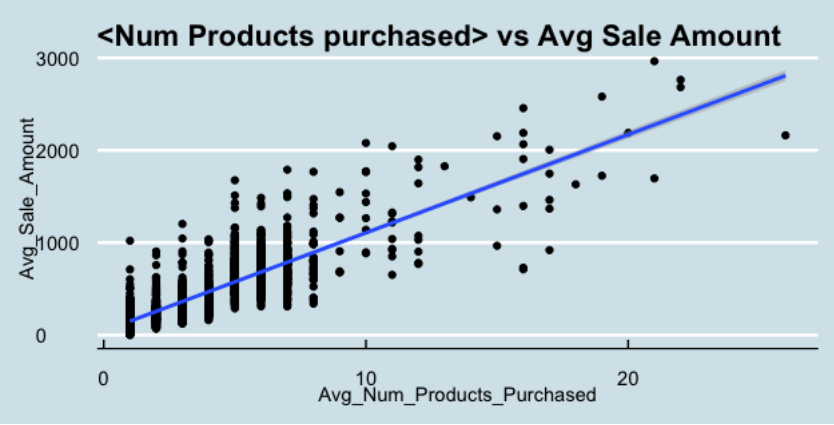
Important: Use the p1-customers.xlsx to train your linear model.

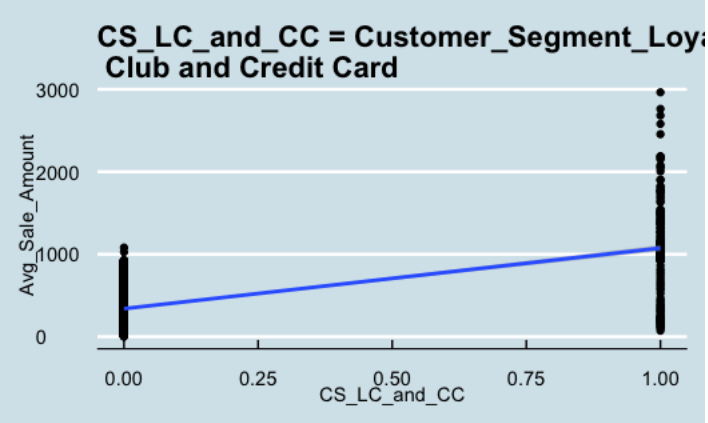
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

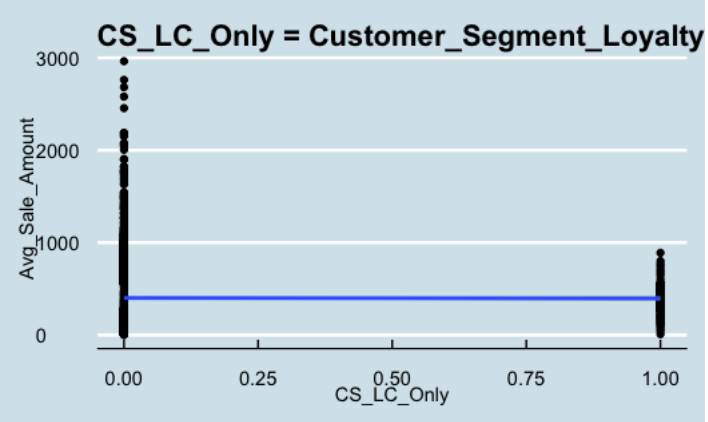
Let us do Exploratory Data Analysis on the p1-customers.xlsx. The relevant variables are, a) Avg_Sale_amount, b) #_Years_as_Customer (renamed as Num_Years_as_Customer), c) Avg_Num_Products_Purchased, d) Customer_Segment (What group the customer belong to).

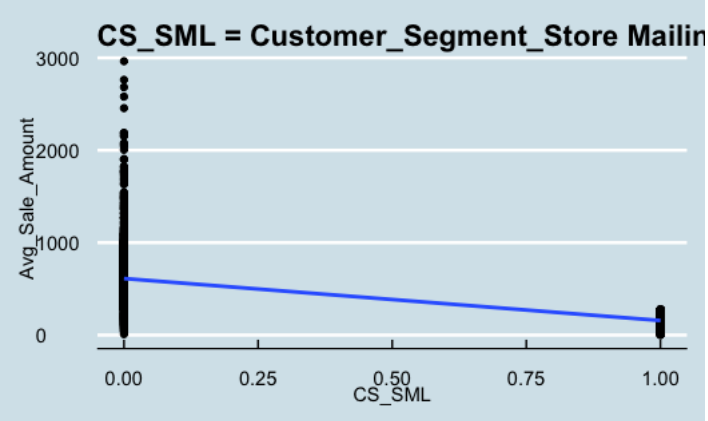
Please note that the relevant variables are the ones that are either present in both the training and test datasets and the ones that are meaningful to be used as features. Here are some of the features plotted against the outcome variable, Avg_Sale_amount











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.

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:

```
Call:
lm(formula = Avg_Sale_Amount ~ Avg_Num_Products_Purchased + CS_LC_and_CC +
    CS_LC_Only + CS_SML, data = p1_customers)

Residuals:
    Min      1Q  Median      3Q     Max
-663.77  -67.31   -1.90   70.69  971.69

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                303.463     10.576   28.69   <2e-16 ***
Avg_Num_Products_Purchased  66.976      1.515   44.21   <2e-16 ***
CS_LC_and_CC               281.839     11.910   23.66   <2e-16 ***
CS_LC_Only                -149.356      8.973  -16.64   <2e-16 ***
CS_SML                    -245.418      9.768  -25.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137.5 on 2370 degrees of freedom
Multiple R-squared:  0.8369,    Adjusted R-squared:  0.8366
F-statistic:  3040 on 4 and 2370 DF,  p-value: < 2.2e-16
```

Avg_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased + 281.84 * CS_LC_and_CC - 149.36 * CS_LC_Only – 245.42 * CS_SML

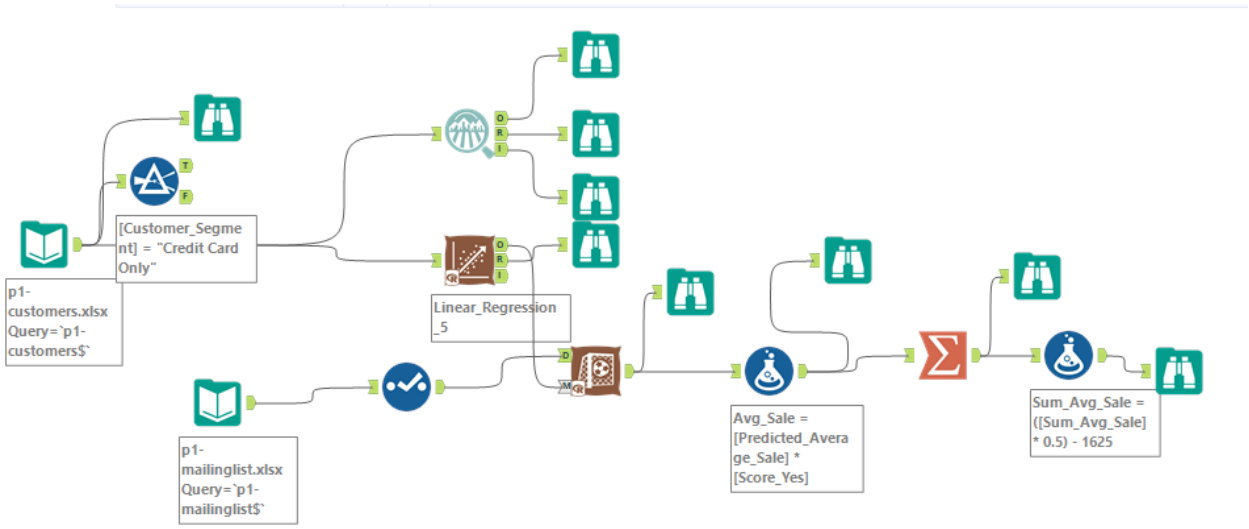
Note: CS_LC_and_CC: Customer_SegmentLoyalty Club and Credit Card
CS_LC_Only: Customer_SegmentLoyalty_Club_Only
CS_SML: Customer_SegmentStore Mailing List

Note the base case is Credit Card Only.

Step 3: Presentation/Visualization

Use your model results to provide a recommendation.

Here is the model pipeline in Alteryx



For the rest of the project, we will be using R/Rstudio. Here is a linear regression model that built using Avg_Num_Products_Purchased, CS_LC_and_CC, CS_LC_Only, CS_SML.

```
Call:
lm(formula = Avg_Sale_Amount ~ Avg_Num_Products_Purchased + CS_LC_and_CC +
    CS_LC_Only + CS_SML, data = p1_customers)

Residuals:
    Min      1Q  Median      3Q     Max
-663.77  -67.31   -1.90   70.69  971.69

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                303.463     10.576   28.69   <2e-16 ***
Avg_Num_Products_Purchased  66.976      1.515   44.21   <2e-16 ***
CS_LC_and_CC               281.839     11.910   23.66   <2e-16 ***
CS_LC_Only                -149.356      8.973  -16.64   <2e-16 ***
CS_SML                    -245.418      9.768  -25.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137.5 on 2370 degrees of freedom
Multiple R-squared:  0.8369,    Adjusted R-squared:  0.8366
F-statistic:  3040 on 4 and 2370 DF,  p-value: < 2.2e-16
```

Note that all the variables are statistically significant.

At the minimum, answer these questions:

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

The profit is $21,987.44. This is more than the expected profit contribution of $10,000. Based on the data analytics and modeling, the company should send the catalog to the 250 customers

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

   I used the company's historical data (P1-customers.xlsx) and used the features "Customer_Segment (categorical data)", Avg_Num_Products_Purchased to model the outcome variable, Avg_Sale_Amount. I used linear modeling approach for this problem. Based on the EDA, this is a valid assumption. The variables used for modeling had been chosen based on the following criteria: a) the availability of the variables in both the training and testing data set b) the statistical importance of the variable. For example, based on the second assumption, I decided to drop the feature, "#_Years_as_Customer".

   Business problem made the following assumptions:

   - Predict the expected profit from the 250 customers (p1-mailinglist.xlsx dataset).
   - Management wanted to send the catalog to the new customers only if the expected profit contribution exceeds $10,000
   - When calculating the profit, please remember the cost of printing/distributing is $6.50/catalog
   - The average gross margin on all products sold through the catalog is 50%

     Calculation:

   - Using the linear model, predict the average sales
   - Use the probability that the customer will respond to the catalog (var: Score_Yes) to calculate the % chance the customer will buy from the company (Score_Yes * Predicted_Average_Sale)
   - Sum this product from the previous step
   - Multiply by 0.5 for (50%; because the only 50% of the product will be sold through the catalog)
   - Finally subtract $6.50 * 250 = 1625 to get the result

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

   21987.435672. or ~ $21,987.44