Project 1: Predicting Catalog Demand

Dung Nguyen

## **Step 1: Business and Data Understanding**

1. *What decisions needs to be made?*

* Whether or not sending out catalog to specific customers

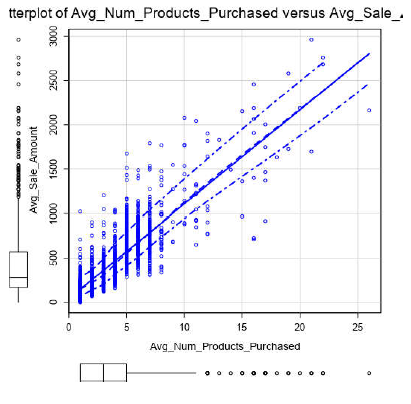
1. *What data is needed to inform those decisions?*

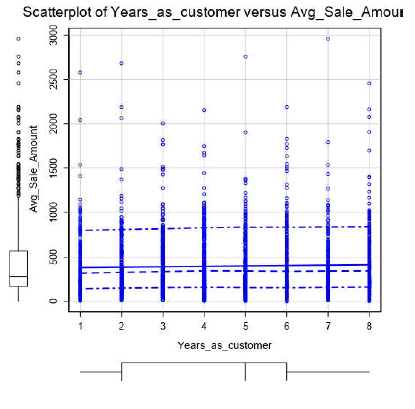
* Information needed: how much profit the company can expect from sending a catalog to these customers.
* Specifically, these data is needed:
  + Outcome of expected profit: Average sale amount (Avg\_Sale\_Amount)
  + Potential predictors such as Average number of product purchased, Years as customer, and Customer segment.
* Predictive analysis to help us obtain the data we need. According to Methodology Map, I choose linear regression because the data is rich and outcome is numeric and continuous.

## **Step 2: Analysis, Modeling, and Validation**

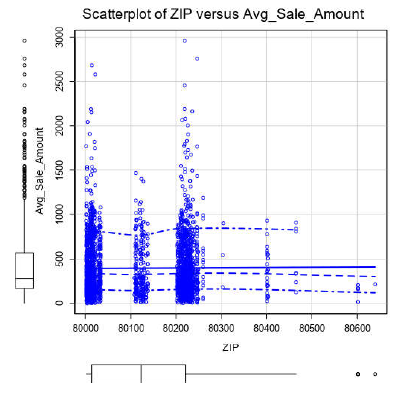
*1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer back to the “Multiple Linear Regression with Excel” lesson to help you explore your data and use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

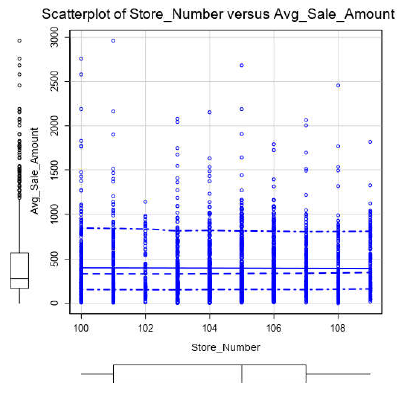
* For continuous variables: only Avg\_Num\_Product\_Purchased and #years\_As\_Customer showed some correlation with Avg\_Sale\_Amount



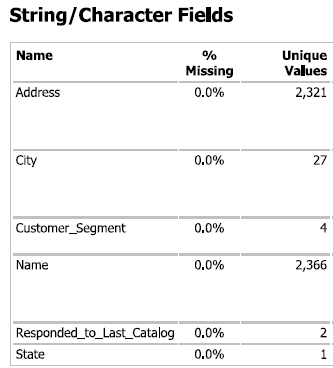


* Some variables such as ZIP and Store\_Number have no meaning in predicting Avg\_Sale\_Amount:



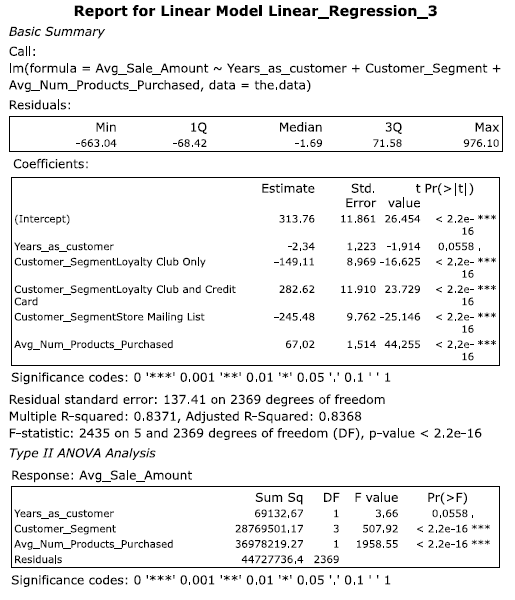


* Furthermore, non-numeric variables were inspected, only Customer\_Segment has a reasonable number of unique values to be a categorical predictor:

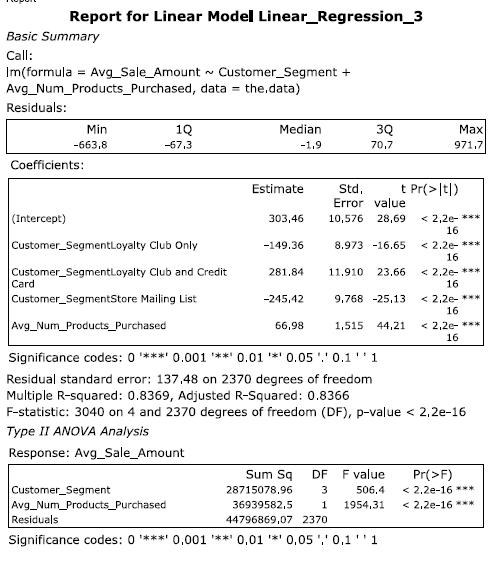


*2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.*

* Firstly I included 3 variables in the linear regression model: Years\_as\_customer, Customer\_Segment, and Avg\_Num\_Products\_Purchased. In the model, Years\_as\_customer has p-value > 0.05, so I excluded it. R and R adjust do not increase much in the new model. My final model includes Customer\_Segment and Avg\_Num\_Products as predictors.



First model



Second model (Years\_as\_customer excluded)

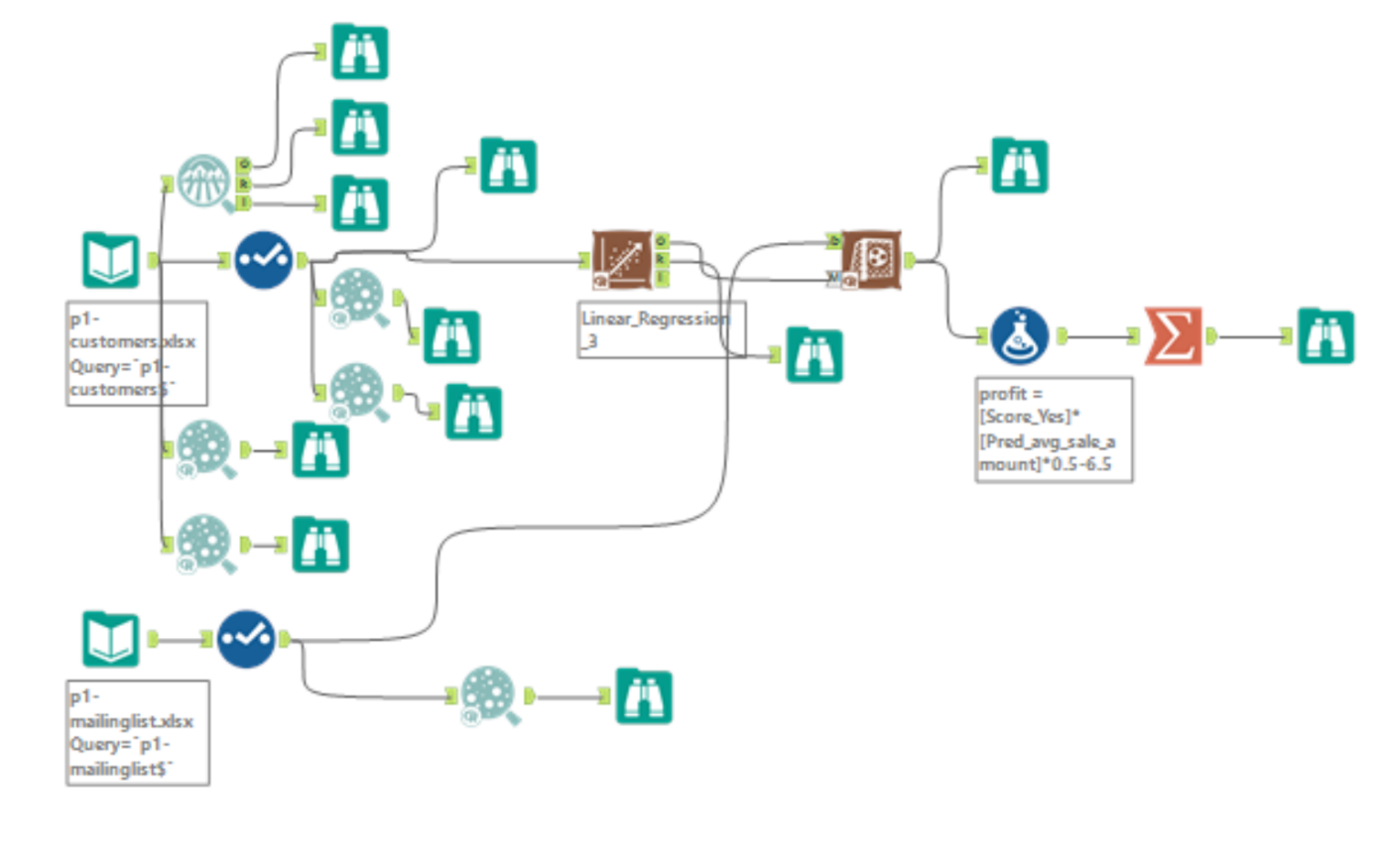
*3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)*

* Y = 303.46 + 66.98 \*Avg\_num\_products\_purchased + 0 (If Customer\_segment: Credit Card Only) – 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)

## **Step 3: Presentation/Visualization**

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

* I predicted the Avg\_sale\_amount based one the data mailinglist.xlsx. After that I multiply the predicted values with 0.5 and minus 6.5. Finally I summed up.



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

* $ 21,987.44, higher than $10,000

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

* My recommendation is the company should send the catalog to those 250 customers.