Project 1: Predicting Catalog Demand

Dung Nguyen

Step 1: Business and Data Understanding

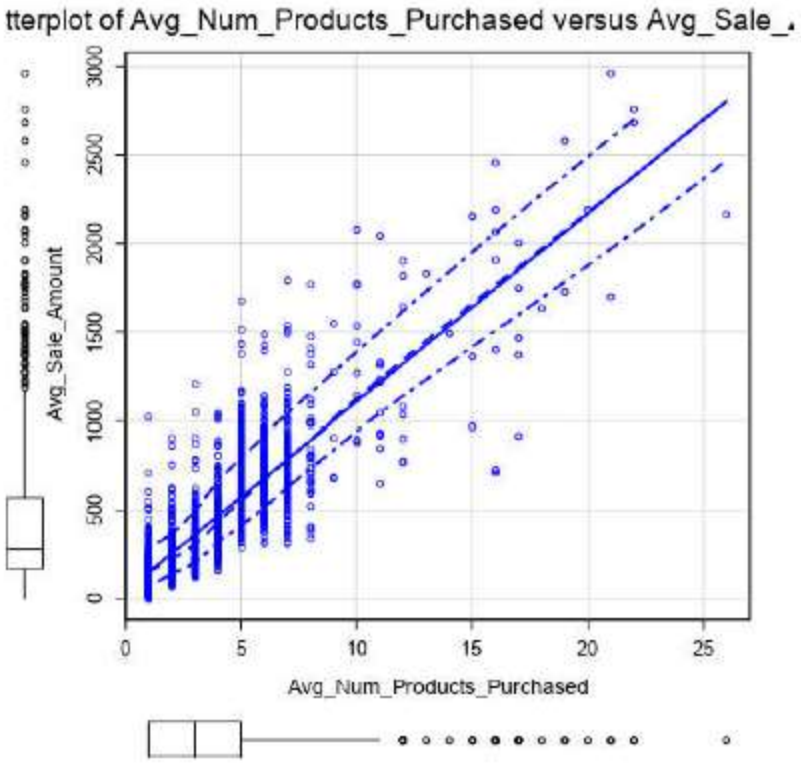
1. *What decisions needs to be made?*
    - Whether or not sending out catalog to specific customers

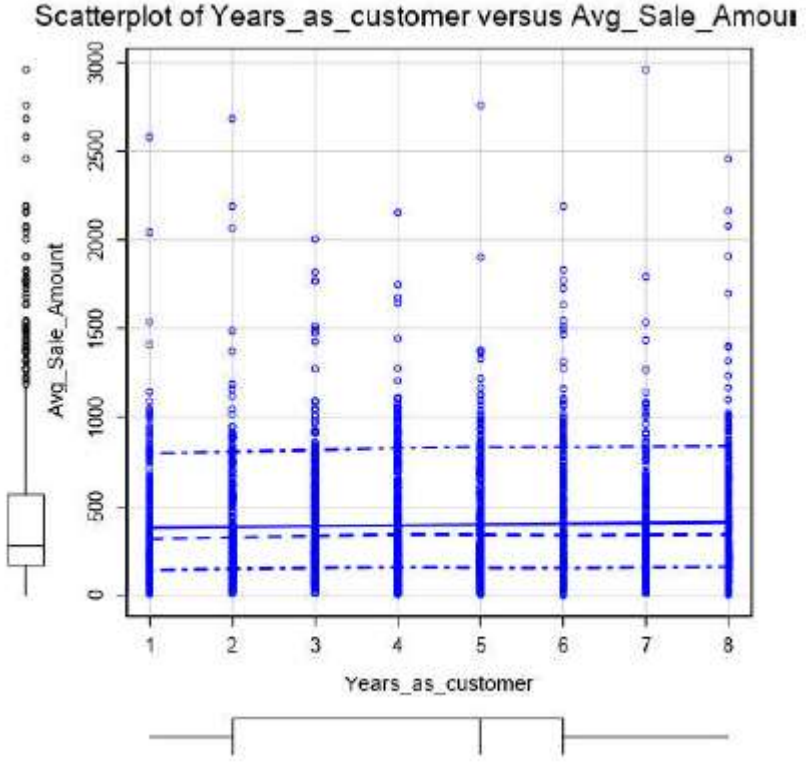
2. *What data is needed to inform those decisions?*
    - Information needed: how much profit the company can expect from sending a catalog to these customers.
    - Specifically, these data is needed:
        - Outcome of expected profit: Average sale amount (Avg_Sale_Amount)
        - Potential predictors 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

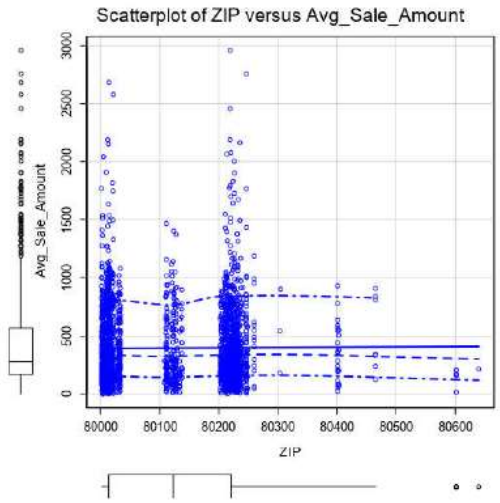
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

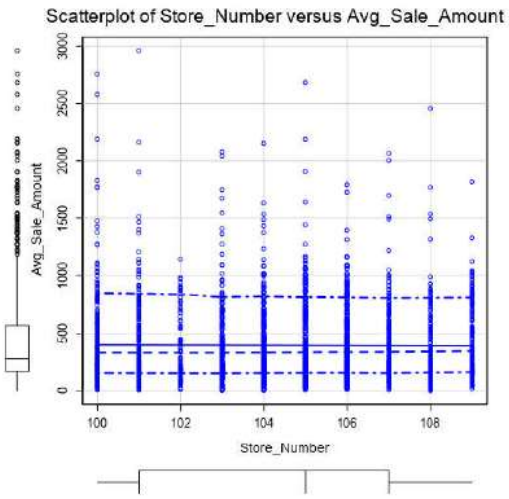
- For continuous variables: only Avg_Num_Product_Purchased and #years_As_Customer showed some correlation with Avg_Sale_Amount


tterplot of Avg_Num_Products_Purchased versus Avg_Sale_
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
Avg_Num_Products_Purchased


Scatterplot of Years_as_customer versus Avg_Sale_Amou
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
Years_as_customer

- 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:

Name	% Missing	Unique Values
Address	0.0%	2,321
City	0.0%	27
Customer_Segment	0.0%	4
Name	0.0%	2,366
Responded_to_Last_Catalog	0.0%	2
State	0.0%	1

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.

Report for Linear Model Linear_Regression_3

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Years_as_customer + Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.04	-68.42	-1.69	71.58	976.10

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	313.76	11.861	26.454	< 2.2e-16	***
Years_as_customer	-2,34	1,223	-1,914	0,0558	,
Customer_SegmentLoyalty Club Only	-149,11	8,969	-16,625	< 2,2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.62	11.910	23.729	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.48	9.762	-25.146	< 2.2e-16	***
Avg_Num_Products_Purchased	67,02	1,514	44,255	< 2,2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.41 on 2369 degrees of freedom
Multiple R-squared: 0.8371, Adjusted R-Squared: 0.8368
F-statistic: 2435 on 5 and 2369 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Years_as_customer	69132,67	1	3,66	0,0558	,
Customer_Segment	28769501,17	3	507,92	< 2,2e-16	***
Avg_Num_Products_Purchased	36978219.27	1	1958.55	< 2.2e-16	***
Residuals	44727736,4	2369			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

First model

Report for Linear Model Linear_Regression_3

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment +
Avg_Num_Products_Purchased, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663,8	-67,3	-1,9	70,7	971,7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	303,46	10,576	28,69	< 2,2e-16	***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245,42	9,768	-25,13	< 2,2e-16	***
Avg_Num_Products_Purchased	66,98	1,515	44,21	< 2,2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137,48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2,2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28715078,96	3	506,4	< 2,2e-16	***
Avg_Num_Products_Purchased	36939582,5	1	1954,31	< 2,2e-16	***
Residuals	44796869,07	2370			

Significance codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Second model (Years_as_customer excluded)

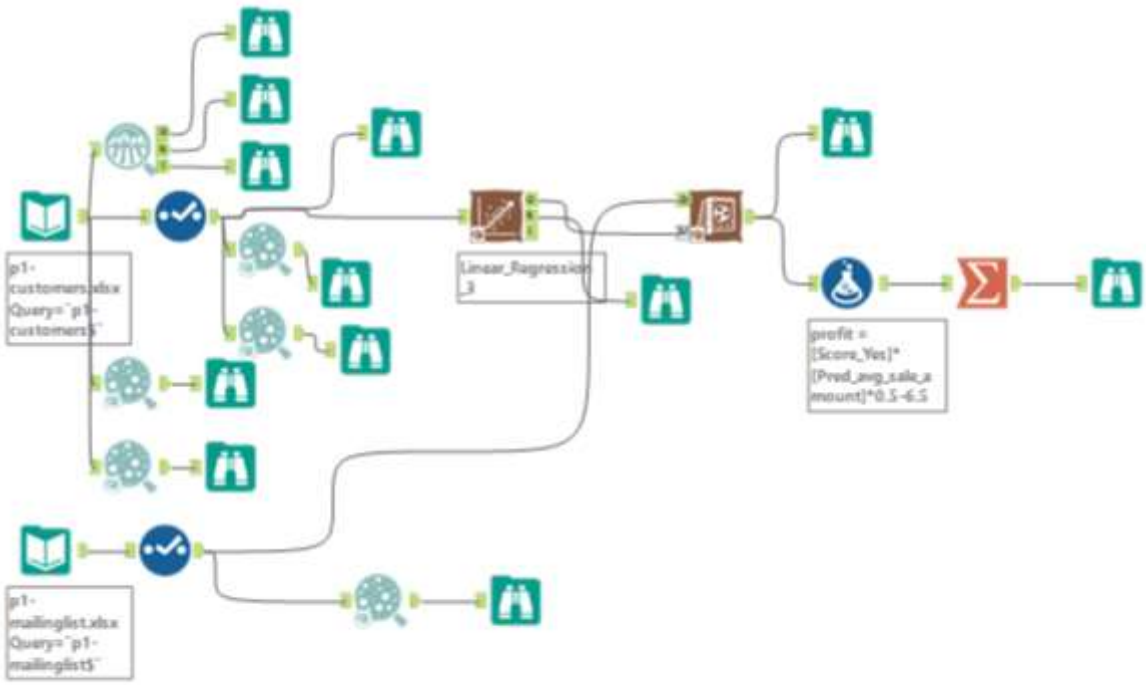
3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

- 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.



2. *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

3. *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.