Project 2: Predicting Catalog Demand

Step 1: Business and Data Understanding

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

Key Decisions:

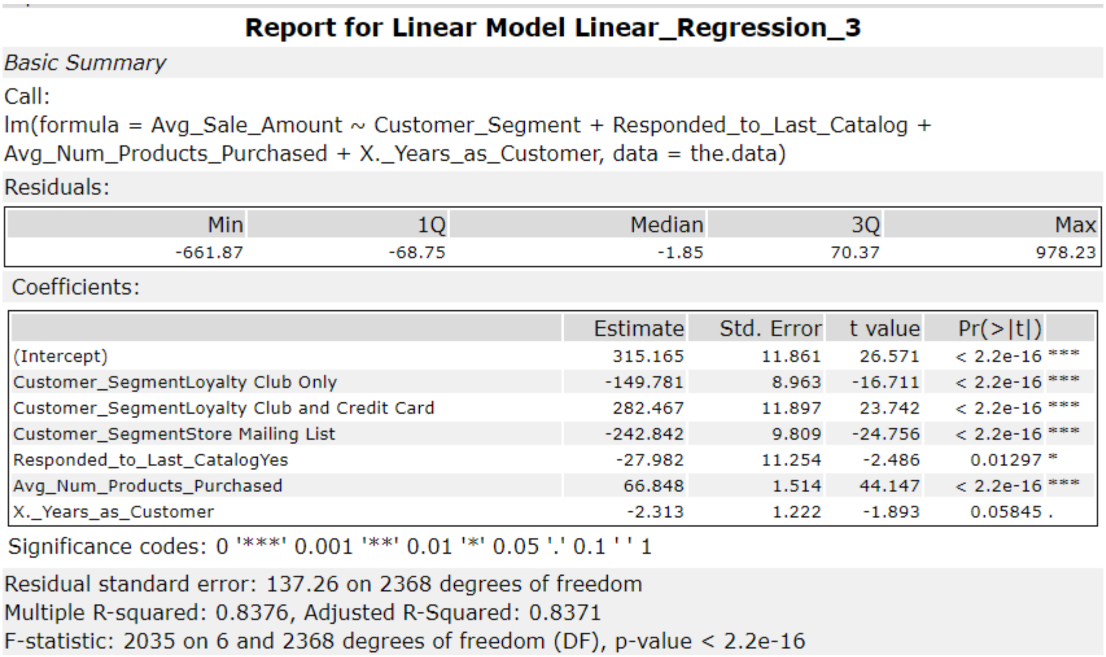
Answer these questions

1. **What decisions needs to be made?**
   The decision needs to be made in this case is to whether the store need to send the catalog to 250 new customers, depends on the predicted profit that they can make on these new customers.

2. **What data is needed to inform those decisions?**
   We need the data with to predict the sales, which can be trained from current customer data. We will need the features like customer segment, store number, responded to last catalog or not, years as customer, margin and cost of catalog, these numerical and factor variables.

Step 2: Analysis, Modeling, and Validation

1. **How and why did you select the predictor variables in your model?**
   I quickly build a model to test the significance of each variable, check if it is important in building regression model. For some variables like customer personal information which is not important at all.



Report for Linear Model Linear_Regression_3

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-661.87	-68.75	-1.85	70.37	978.23

Coefficients:

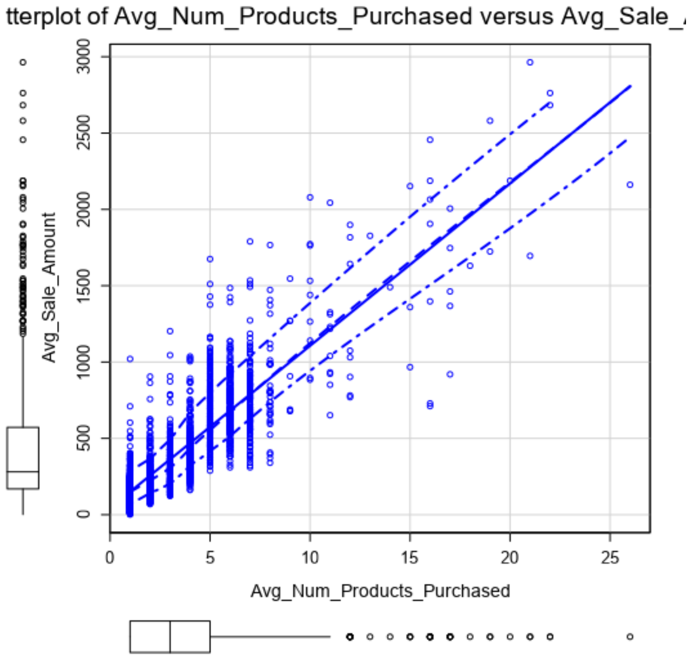
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	315.165	11.861	26.571	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.781	8.963	-16.711	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.467	11.897	23.742	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.842	9.809	-24.756	< 2.2e-16	***
Responded_to_Last_CatalogYes	-27.982	11.254	-2.486	0.01297	*
Avg_Num_Products_Purchased	66.848	1.514	44.147	< 2.2e-16	***
X._Years_as_Customer	-2.313	1.222	-1.893	0.05845	.

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

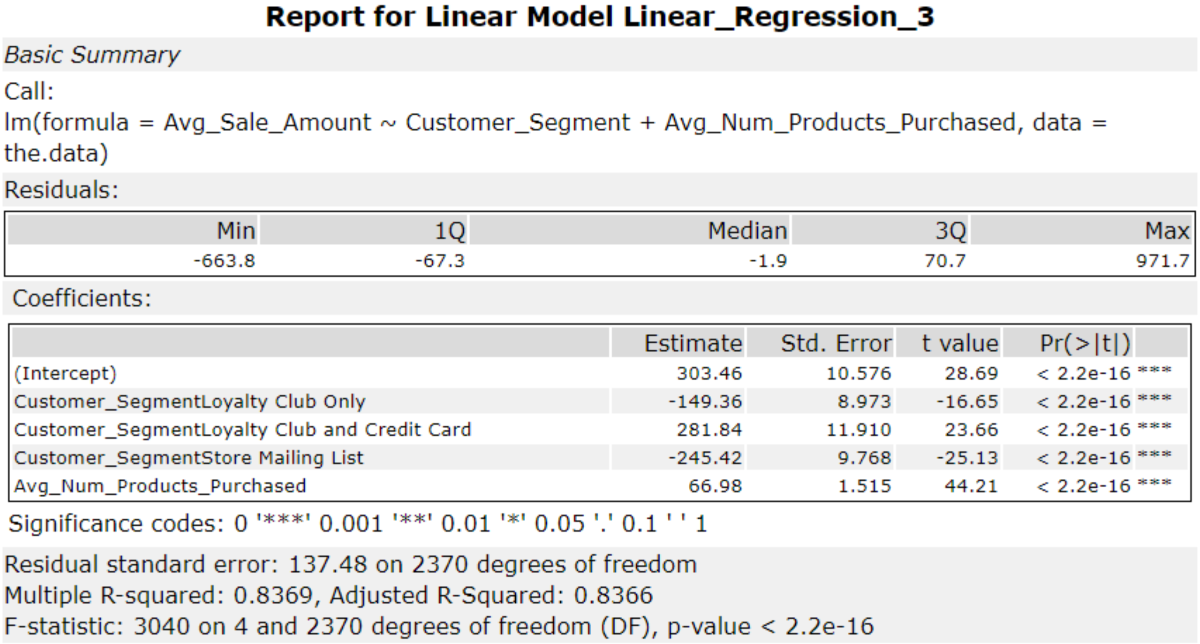
Residual standard error: 137.26 on 2368 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8371
F-statistic: 2035 on 6 and 2368 degrees of freedom (DF), p-value < 2.2e-16

As one can see from P_value<0.05 of coefficient variables, customer_segment factor variable is very important, same for avg_num_products_purchased, respond_to_last_catalogYes, and years_as_customer is less significant, but still important.

I will also show scatter plots between target and continuous variable.



2. **Explain why you believe your linear model is a good model.**
Due to the fact that predicting data do not have respond_to_last_catalogYes, we will not use this column to build the 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:

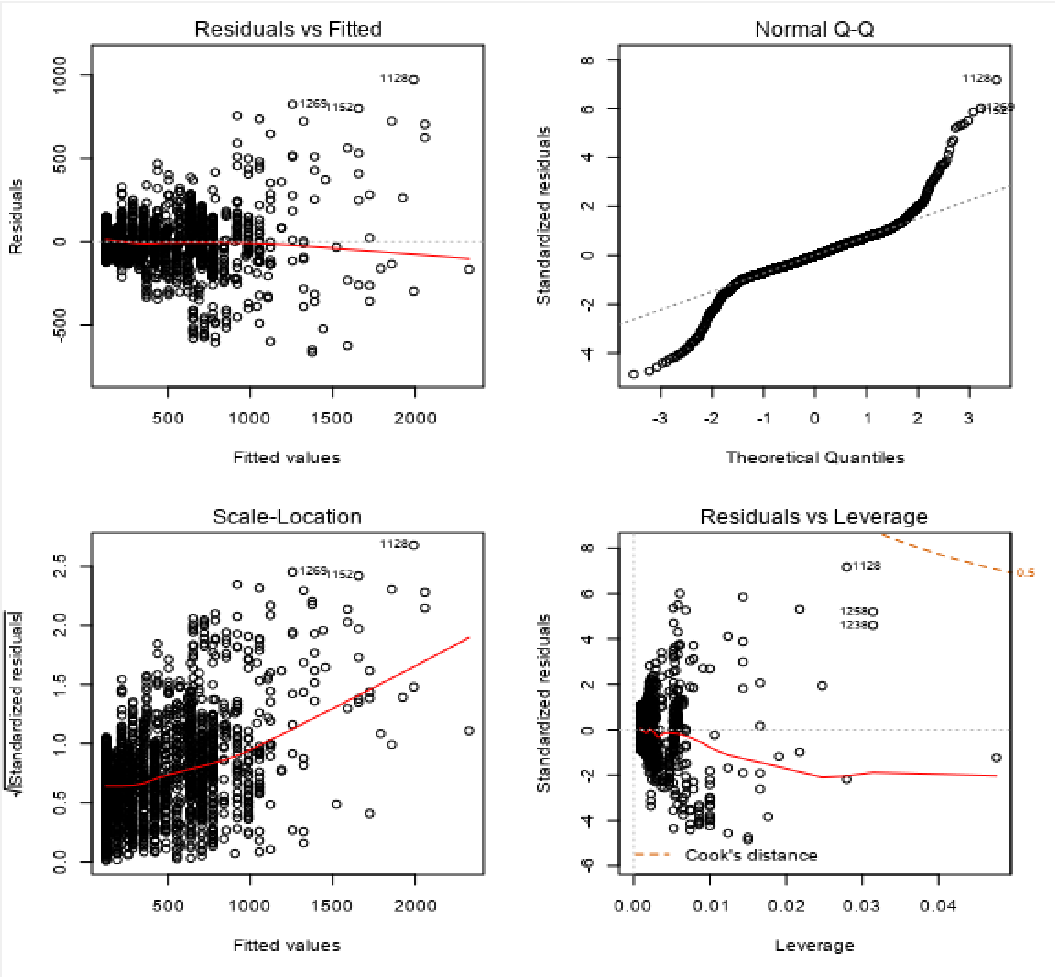
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
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

Now we can see that all the coefficients P_value<0.05, which proves the significance of variable. Adjusted R^2 has 0.8366 which also confirm the linear model is a good fit. Also, from residual plot, we can see that they are more like normal distribution now.

Basic Diagnostic Plots



3. **What is the best linear regression equation based on the available data?**

Avg_Sale_Amount = 303.46 + 0 x (If Customer is Credit Card Only) – 149.36 x (If Customer is Loyalty Club Only) + 281.84 x (If Customer is Loyalty Club and Credit Card) – 245.42 x (If customer is Store Mailing List Only) + 66.98 x (Avg_Num_Products_Purchased)

Step 3: Presentation/Visualization

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

Based on my calculations, the company should send the catalog to 250 customers.

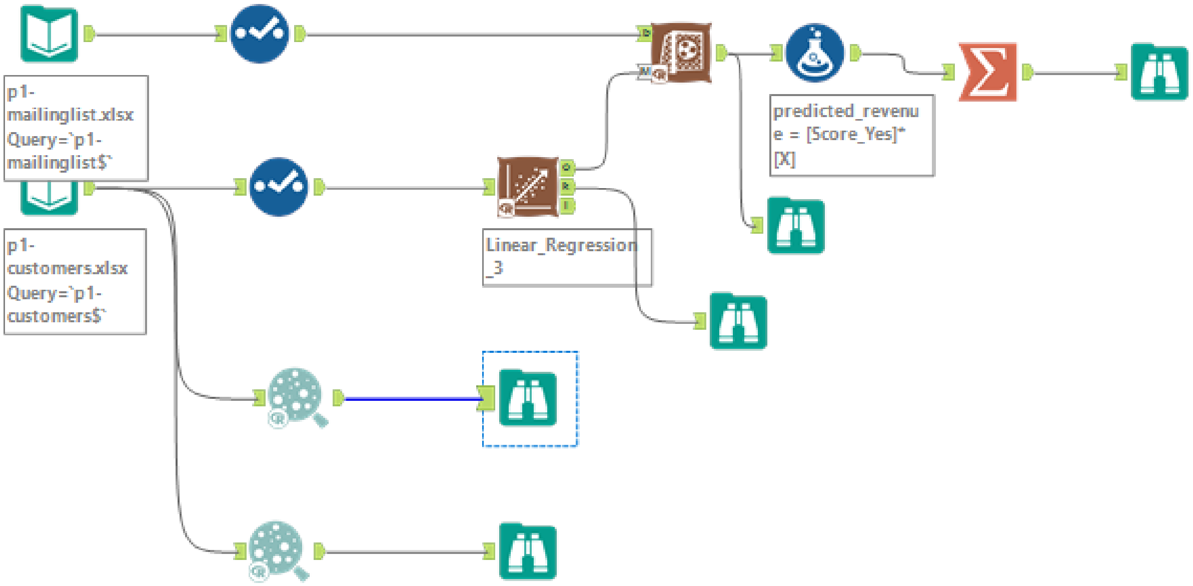
3. **How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)**

I first calculated the predicted revenue for each of 250 customers from the linear regression model, then for predicted revenue time the variable [score_Yes] (probability that the customer will buy the product), then sum them to get the total predicted revenue. Then use this value time margin subtract the cost of catalog, to get the predicted profit.

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

The total predicted revenue is $47224.87, consider 50% margin and cost of catalog,
Predicted profit = $47224.87 * 50%(margin) - 250 * $6.5 (each catalog cost)
= $23612.44 - $1625
= **$21987.44**

Attach the workflow here.


p1-mailinglist.xlsx
Query=`p1-mailinglist$`
p1-customers.xlsx
Query=`p1-customers$`
Linear_Regression_3
predicted_revenue = [Score_Yes]*[X]