Step 1: Business and Data Understanding

The key decision this company is trying to make is whether to send this year's catalog out in coming months to 250 new customers from their mailing list. Management does not want to send the catalog unless they can expect more than $10,000 in profit contribution.

Key Decisions:

1. What decisions needs to be made?

- Whether the company should send this year's catalog out
- Whether the company can expect the profit contribution from new catalog sales exceeds $10,000.

2. What data is needed to inform those decisions?

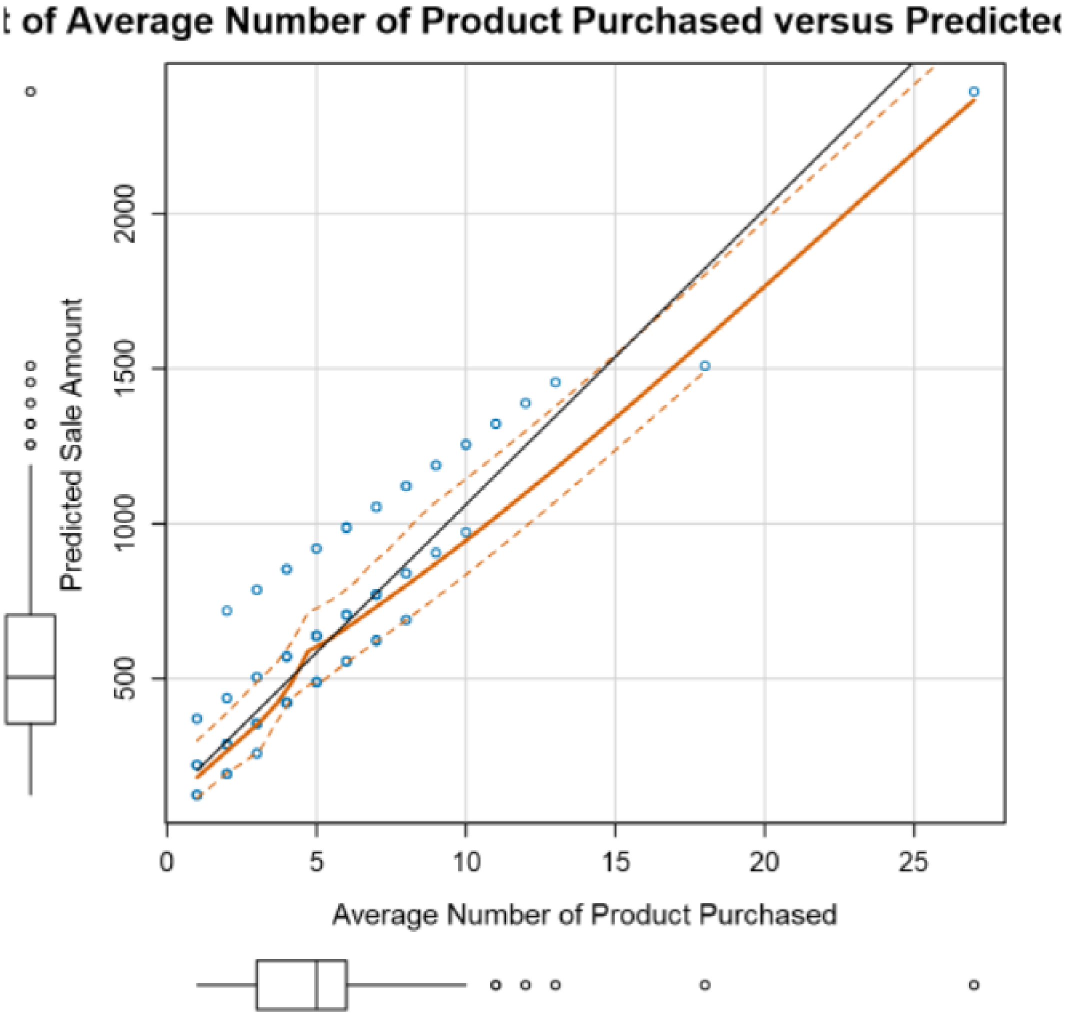
- Historical/Past sales data to predict sales from 250 customers
- Chance (probability) of buying something from this company for all 250 customers
- Cost of printing and distributing - $6.50 per catalog.
- Average gross margin on all products sold through the catalog - 50%

Step 2: Analysis, Modeling, and Validation

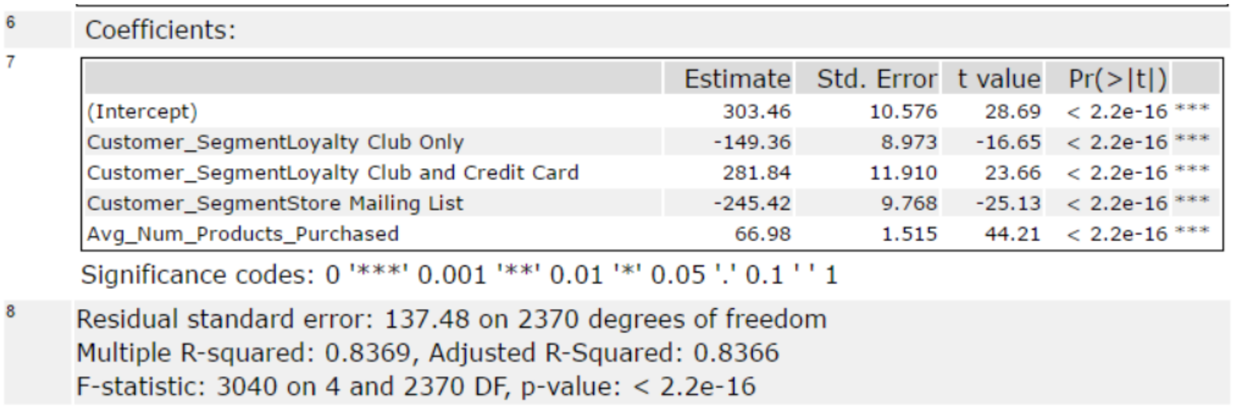
1. How and why did you select the in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. You must include scatterplots in your answer.

I have selected Customer segment and Average number of products purchased as the predictor variables. The reason why I have selected the customer segment is I wanted to make sure historical sales from the store mailing list (catalog) is weighed more for the regression model. For the average number of products purchased predictor variable, the predicted sale amount is determined by number of products purchased and is morelikely to be higher if the avarage number of product purchased is larger.

It is easy to see that the average number of products purchased predictor variable has a linear relationship with the target variable with the scatterplots graph below.



The customer segment is categorical variable, so scatterplots graph cannot be used to show a linear relationship. I confirm that the customer segrment predictor variable has a linear relationship with the target variable by checking the report for my model below. The coefficients are very significant having very small number 2.2e -16 (p-values <= 0.05) with relatively high multiple-R-squared (>= 0.70). Also the customer segment has 3-stars in the report which indicates the customer sequement predictor variable is significant.



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 DF, p-value: < 2.2e-16
```

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.

Looking at the report above (from Q1.), the coefficients for customer segment and average number of product purchased are very significant having very small number 2.2e -16 (p-values <= 0.05) with relatively high multiple-R-squared (>= 0.70, considered a good model).

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*Ave_Num_Products_Purchased
- 149.36*(If Loyalty_Club_Only)
+ 281.84*(If Loyalty_Club_And_Credit_Card)
- 245.42*(If Store_Mailing_List)

Step 3: Presentation/Visualization

I would recommend the company to send this year's catalog to these 250 customers. I have predicted the sales from the catalog to be $21987.44 which exceeds $10,000.

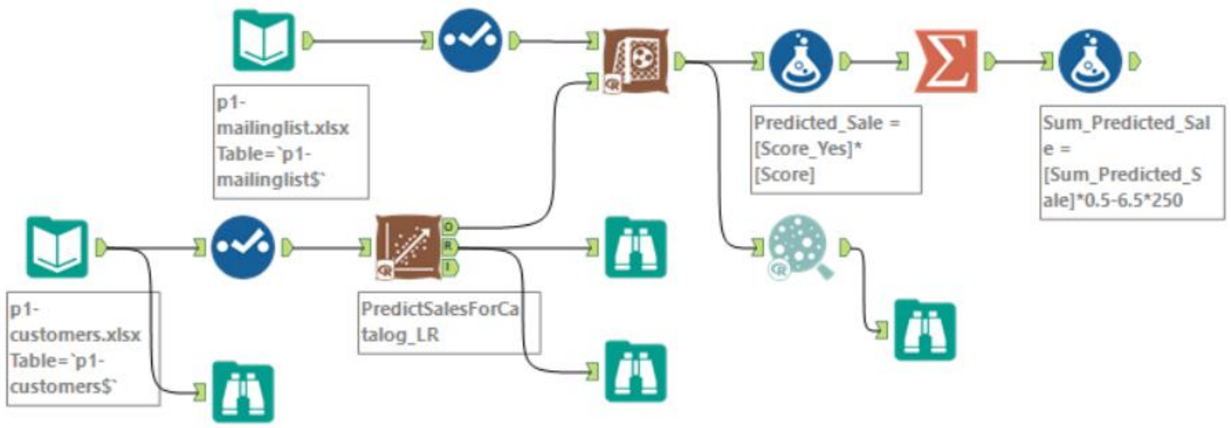
1. What is your recommendation? Should the company send the catalog to these 250 customers?

I would recommend the company to send this year's catalog to these 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 first took the historical/existing sales data and used customer segment and average number of products purchased as predictor variables to come up with the linear regression model. I looked at the report and made sure the model I came up with a good model. Next, I applied the model to 250 new customer data to come up with average sales amount (target) for each customer. I multiplied average sales amount by the score yes (probability) field in the data to come up with predicted sale amount for each customer. Then I used the following formula to come up with final number, the expected profit from the new catalog.

Expected profit = Sum of predicted sale amount from 250 customers * Average gross margin (50%) – Cost of print and distribution of catalog ($6.5) * 250


p1-mailinglist.xlsx Table=`p1-mailinglist$`
p1-customers.xlsx Table=`p1-customers$`
PredictSalesForCatalog_LR
O
R
I

Predicted_Sale = [Score_Yes]*[Score]

Sum_Predicted_Sale = [Sum_Predicted_Sale]*0.5-6.5*250

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

The expected profit from the new catalog is $21987.44.