# Key Decisions:

Answer these questions

1. What decisions needs to be made?
   a. Whether it's worth the cost and justification to send out these catalogs. Profits need to exceed $10,000 to justify this decision due to cost of printing and distributing catalogs, gross margins of products, and human resource cost. Along with profitability this marketing strategy can potentially turn new customers into long-term loyal customers which can affect the company's business down the road.

2. What data is needed to inform those decisions?
   a. Cost of printing and distributing is $6.50 per catalog
   b. Average gross margin on products sold through catalog is 50%
   c. Existing customer data like average sale, customer segment, average number of products purchased and # of years as customers.
   d. 250 new customer mailing list with the above variables and score yes/no probabilities.
   e. Linear regression model data with statistically significant p-values, adjusted r-squared and coefficient estimates.
   f. Score tool predictability model

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.
   a. Before I setup my linear regression model, I tested predictor variables along with my target variable using scatterplot. Since we're trying to predict profits, I set average sale amount as Y field and average number of products purchased as X field. Here's the result:

rplot of Avg_Num_Products_Purchased versus Avg_Sale
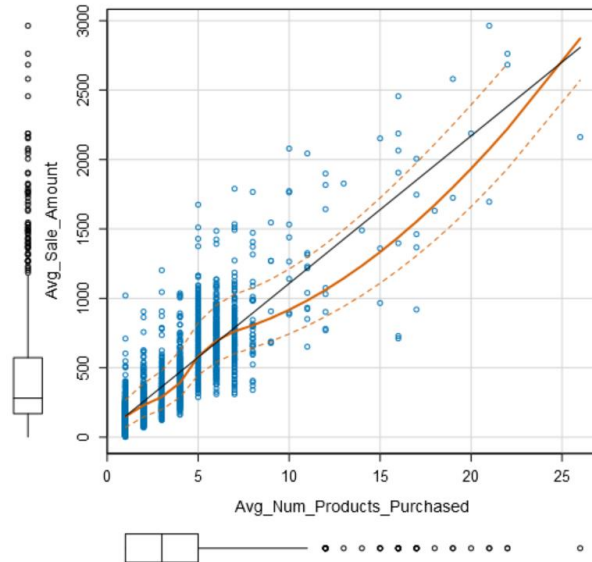
The scatterplot shows a linear relationship between the X and Y variable. As the number of products purchased increases so does the average sale amount. It's not a great linear relationship but nonetheless a good indication that it's useful data.

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.
   a. For my linear regression model I set average sale amount as target variable and (Customer segment and average number of products purchased) predictor variables. Running the model adjusted R-squared is 0.83 which indicates high variation explained in the target variable by the variation in the predictor variable. For p-values customer segment and average number of products purchased are well below 0.05 indicating strong relationship between it and the target variable. Also they have 3 stars next p-value indicating strong statistical significance.

	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

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)

Multi linear regression equation is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\ldots$

Therefore it's $Y = 303.46 - 149.36$(customer_loyaltyclub) $+ 281.84$(customer_loyalty+creditcard) $- 245.42$(customer_storemailinglist) $+ 0$(if_creditcard_only) $+ 66.98$(avg_num_products_purchased)

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

      a. Yes the company should send out catalogs to the new 250 customers.

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

      After building linear regression model, I plugged the model and mailinglist.xlsx to score tool. This gives us how much customers will purchase before factoring in score probability and cost. Then I plugged in scatter plot for score and average number of products purchased which showed confidence in my current data analysis. I use formula tool [score]*[score_yes] to factor in their probability to purchase. Then I summarized the new score and used formula tool again to deduct gross margin and cost per catalog [sum_score]*.5-(250*6.5)

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

Expected profit = $21,987.44