PREDICTIVE ANALYTICS: PREDICTING CATALOG DEMAND

A company that manufactures and sells high-end home goods is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to. This project determines how much profit the company can expect from sending a catalog to these customers.

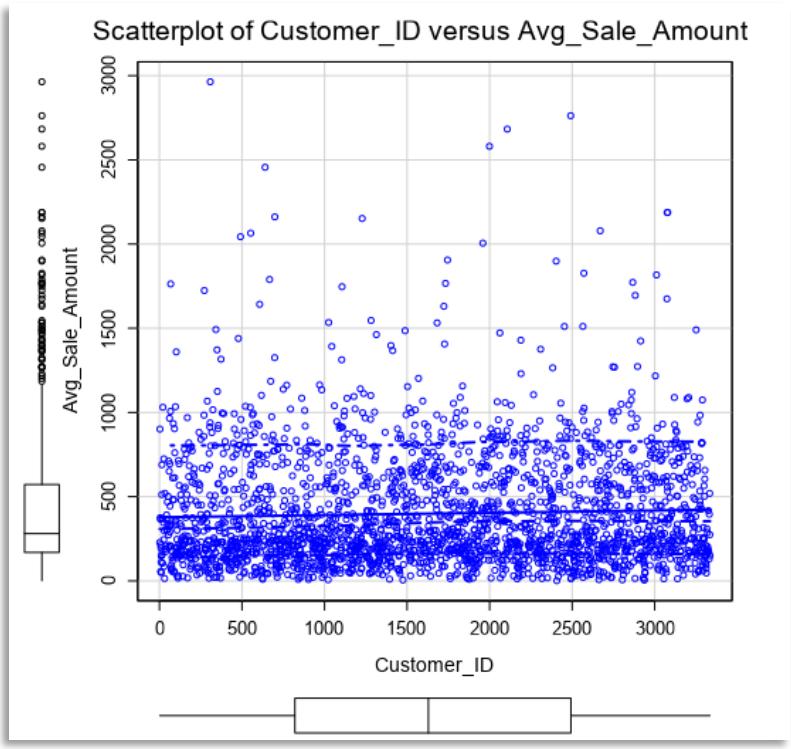
STEP 1: BUSINESS AND DATA UNDERSTANDING

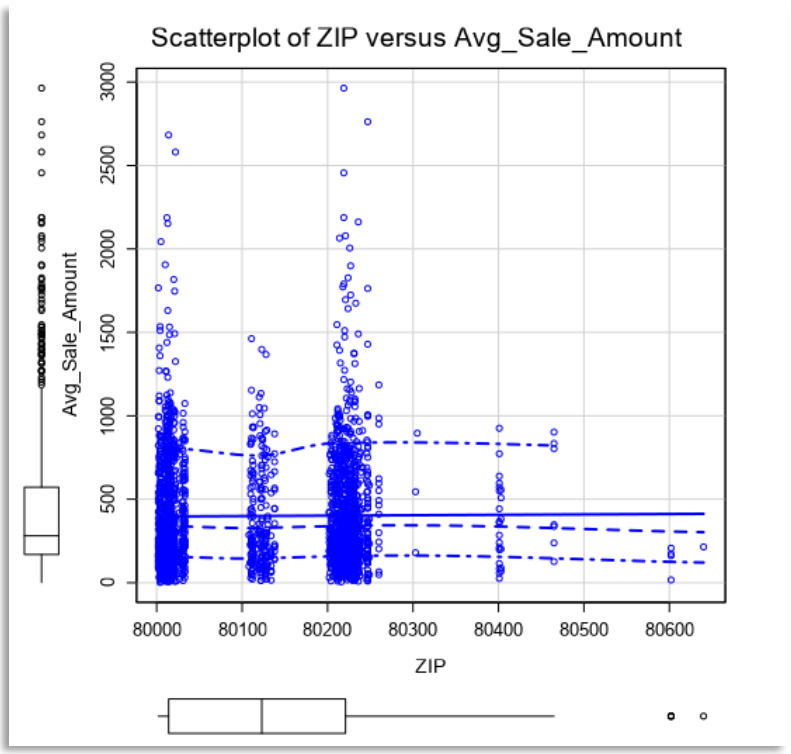
1. What decisions needs to be made?
    a. The decision that needs to be made is whether to send the catalogs out to the 250 new customers or not to send them. To make this decision, we will need to calculate the predicted profit. If it is greater than $10,000, the decision will be to send the catalogs. If it is not greater than $10,000, the decision will be not to send the catalogs.
2. What data is needed to inform those decisions?
    a. Average cost of printing and distributing
        i. $6.50/catalog
    b. Average gross margin per item sold through the catalog
        i. 50%
    c. Probability that a person will make purchases
        i. score_yes calculation
    d. Predicted Revenue
        i. $ worth of products * % likelihood of purchase
    e. Predicted Profit
        i. revenue * gross margin .5 – COGS 6.5

STEP 2: ANALYSIS, MODELING, AND VALIDATION

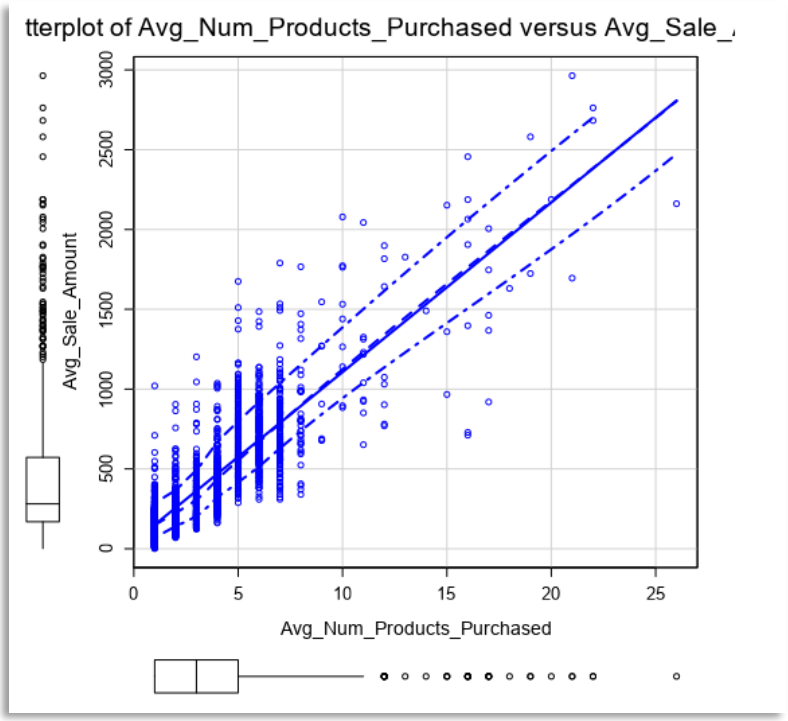
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model.

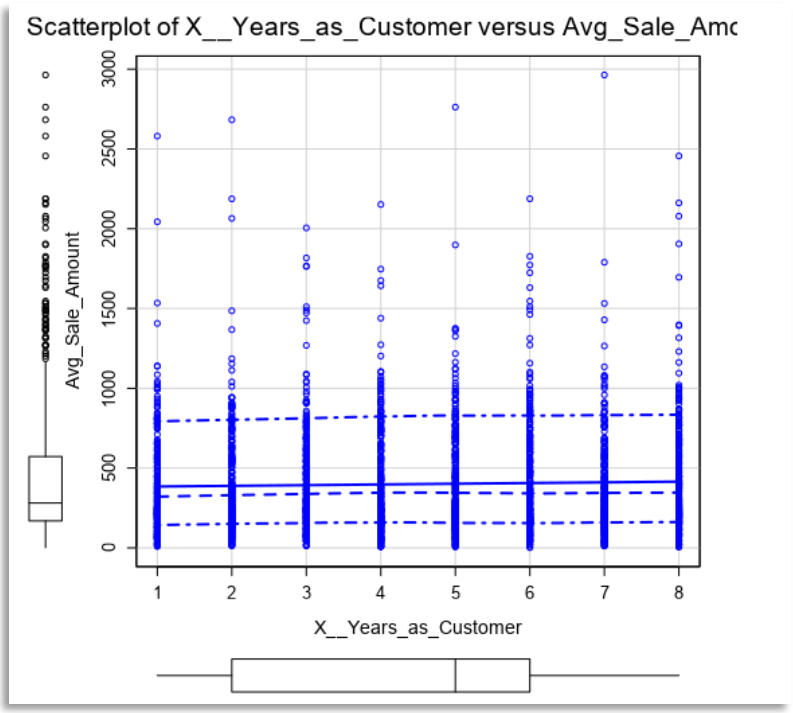
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.
    a. To explore relationships between the given data points, I have created scatter plots of all the numerical data using average sale amount as the target variable. I used average sale amount as the target variable because the business case is to predict profit, and this variable is an incremental part of price. Below are the outcomes of each predictor variable explored using average sale amount as the target.


Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
Customer_ID


Scatterplot of ZIP versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
80000
80100
80200
80300
80400
80500
80600
ZIP


Scatterplot of Store_Number versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
100
102
104
106
108
Store_Number


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



The only linear relationship present in the numeric data is average number of products purchased, therefore making it one of the predictor variables. The next predictor variable chosen would be categorical, so no scatter plots to show. But the only categorical variable that made sense to choose from was customer segment because it had no relationship to the numerical data (e.g. address and zip). So, after plugging in the linear regression test for customer segment and average number of products purchased for predictor variables as well as average sale amount for the target variable, the results came out like this:

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

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. This is a good model because for average number of products purchased, there is a linear relationship between the two variables. For the categorial variable, the P-values are less than .05, meaning that the probability for the results are significant and not random. Additionally, the adjusted r-squared value is high (.84), meaning that this is a strong model because the higher the score is to 1, the less variance there is with the linear relationship.

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)
    a. Using the regression results, the equation would look like this:
    Y = 303.46 - 149.36(Loyalty Club Only) + 281.84(Club and Credit) - 245.42(Store Mailing List) + 66.98 (Avg_Num_Products_Purchased) + 0(credit card only)

STEP 3: PRESENTATION/VISUALIZATION

Use your model results to provide a recommendation.

1. What is your recommendation? Should the company send the catalog to these 250 customers?
    a. My recommendation yes, send the catalog.
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
    a. The predicted profit surpasses the minimum value of $10,000. Predicted profit is $21,987. This was found by scoring the data from the regression results in order to calculate expected revenue (score*score_yes). Then, expected revenue was translated to profit with the equation (0.5*[Expected_Revenue])-6.5. The sum of those results equaled $21,987.
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
    a. $21,987