MAKE A COPY

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

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

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Determine how much profit the company can expect from sending a catalog to these customers. then decided whether the catalog should be sent or not?

2. What data is needed to inform those decisions?

- Data about the sales occurred last year when company sent out its first print catalog. (Given)
- Probability that a new customer will buy a catalog and purchase items? (Given)
- Information about current customers, shopping behavior, location etc. (Given)
- Profit Margin (Given 50%)
- Cost structure (Cost for catalog is given)
- Since, we have the past data about sales, we can predict the sales for current year. And then multiplying the sales by probability that a new customer will respond to a catalog and make a purchase (Score_Yes), we get the sales for current year. On the basis of which profit can be calculated and then a decision could be taken if the catalog should be sent or not.

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. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

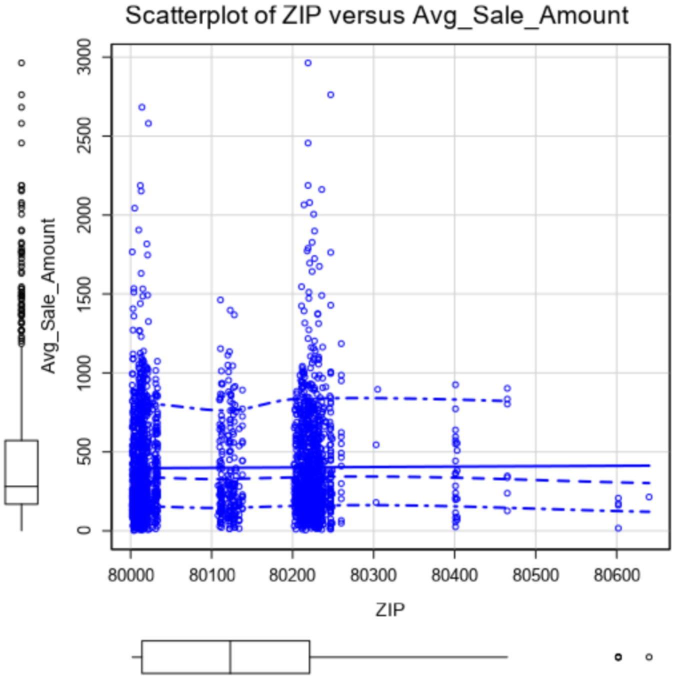
At the minimum, answer these questions:

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.

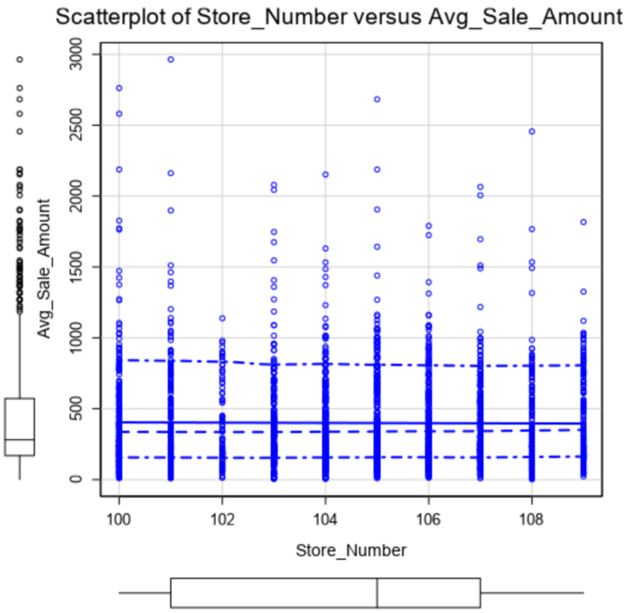
in the variables, just a few variables lead to be a good fit to predict the average sales for the current year. and some variables aren't important such as : ***Name***, ***Customer_Id***, ***Address***, ***State***.

Analyzed the scatterplot between numerical predictors and target variable.

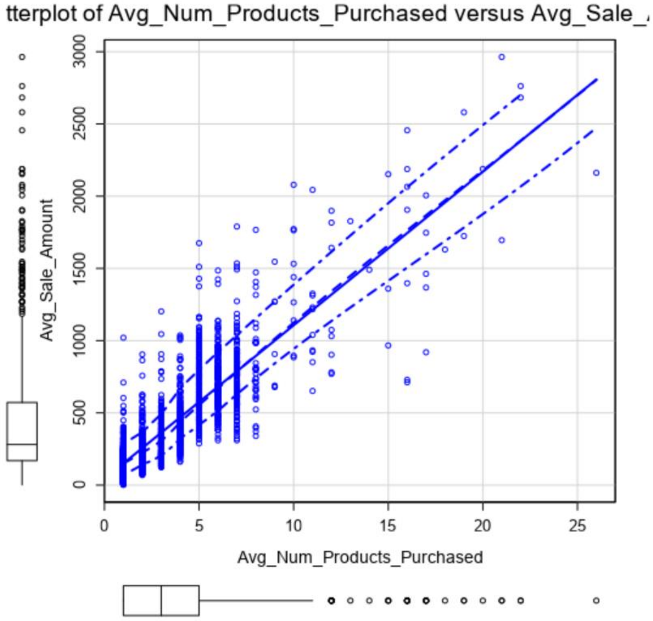
- **Avg_Sales – ZIP (Not linear)**



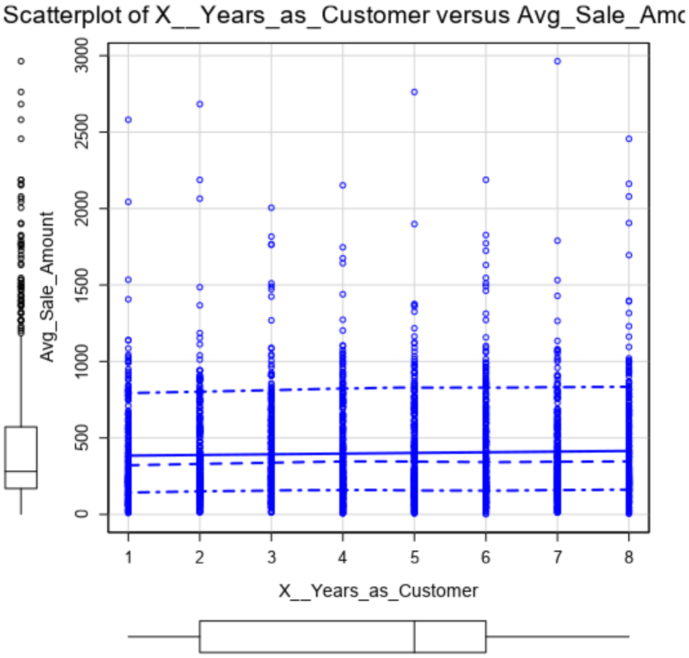
- **Avg_Sales – Store_No (Not linear)**



- **Avg_Sales – Avg_Number_Of_Products_Sold (Not so strong Linear relationship)**



- **Avg_Sales - #_Of_Years_As_Customer (Not Linear)**

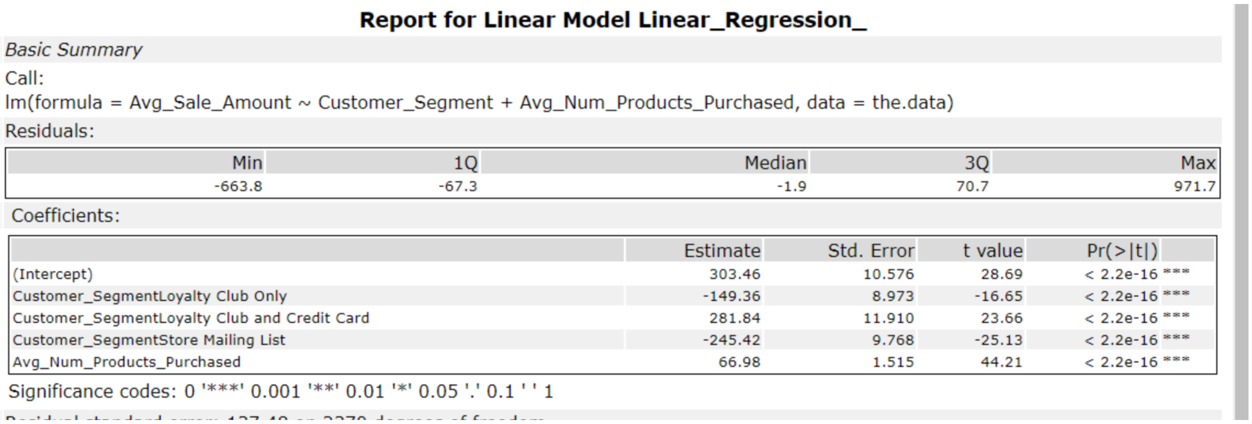


Avg_Number_Of_Products_Sold seems linearly related with target variable.

Used “***Avg_Number_Of_Products_Sold***”, “***Customer_Segment***” and “***City***” to build linear model. As per the model created “City” was not significant at all, so removed city and again created a model using only two predictors.

On the basis of p-values, “***Avg_Number_Of_Products_Sold***”, “***Customer_Segment***”, both are significant.

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.



Report for Linear Model Linear_Regression_

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

- For both the predictor variables, we used in our linear model creation, p-value (probability that the coefficient is going to be 0) is very less. Hence, both the predictors are significant in deciding the target variable.
- This model is strong since the R-value is very high (0.8366)
- R Square equals 1- SSE/SST
- SSE – sum of squares of residuals (actual value – predicted value) using Predictive Model
- SST – sum of squares of residuals (actual value – predicted value) using Baseline Model
- SSE is always less than SST, hence R Square value lies in range 0-1. Higher the R Square means : SSE is very very small (Predicted values are very close to actual values) which leads to smaller (SSE/SST), hence higher R Square.

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)

Avg_Sales = 303 **–** 149.36**Loyalty_Club_Only** **+** 281.84**Loyalty_Club_And_Credit_Card** **–** 245.42**Store_Mailing_List** **+** 0**Credit_Card_Only** **+** 66.98 * **Avg_Number_Products_Purchased**

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

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

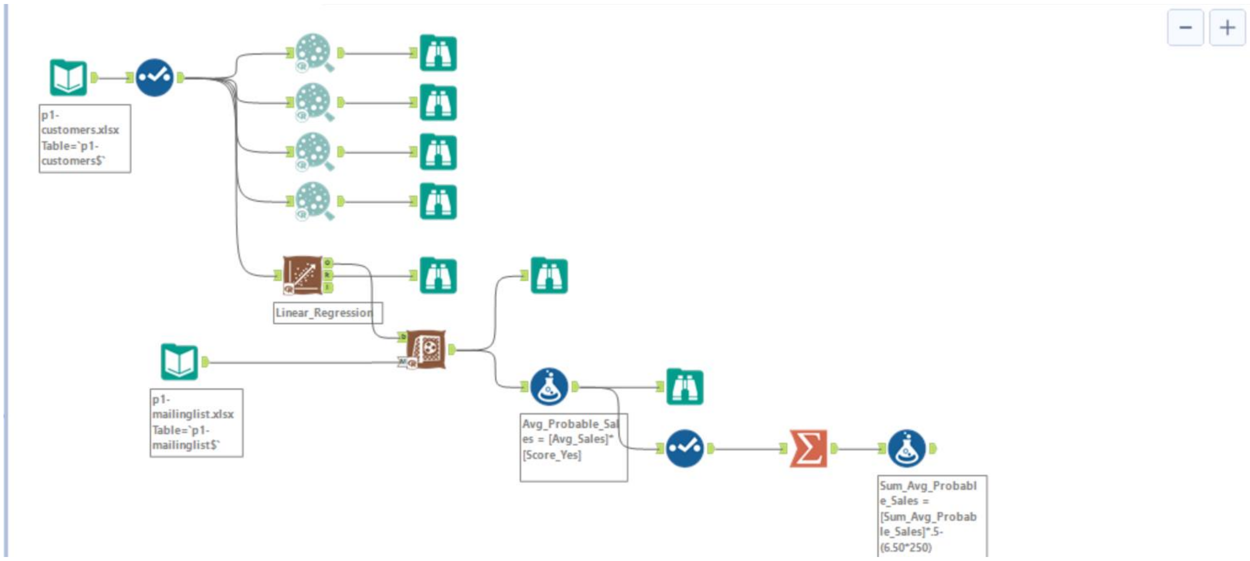
Yes, the company should send the catalog to the customer because we have condition if the profit exceeds $10000 and it actually exceeds as calculated using linear regression model.

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

- Calculated Avg_Sales using linear regression model. Score_Yes : Probability that a customer will respond to catalog and make a purchase.
- Created a new column (Avg_Probable_Sales = Avg_Sales * Score_Yes) Given profit margin is 50%, and cost for each catalog is $6.50, hence for all 250 customers
- Calculated the profit = Avg_Probable_Sales0.5-(6.50250)

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

Profit equals: $21987.43


−
+
p1-customers.xlsx Table=`p1-customers$`
Linear_Regression
p1-mailinglist.xlsx Table=`p1-mailinglist$`
Avg_Probable_Sales = [Avg_Sales]*[Score_Yes]
Sum_Avg_Probable_Sales = [Sum_Avg_Probable_Sales]*.5-(6.50*250)