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?

Predicting the estimated profit if a catalog was sent to new customers and then on the basis of profit, decide whether the catalog should be sent or not?

2. What data is needed to inform those decisions?

We need to calculate the average number of sale amount per customer. We can then get expected revenue by multiplying the average number of sale with the score value. Once we have this, we can find profit by multiplying average gross margin (0.5) and then subtracting 6.50 (cost of printing and distributing per catalog).
Data about the sales occurred last year when company sent out its first print catalog.
Probability that a new customer will buy a catalog and purchase items?
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)

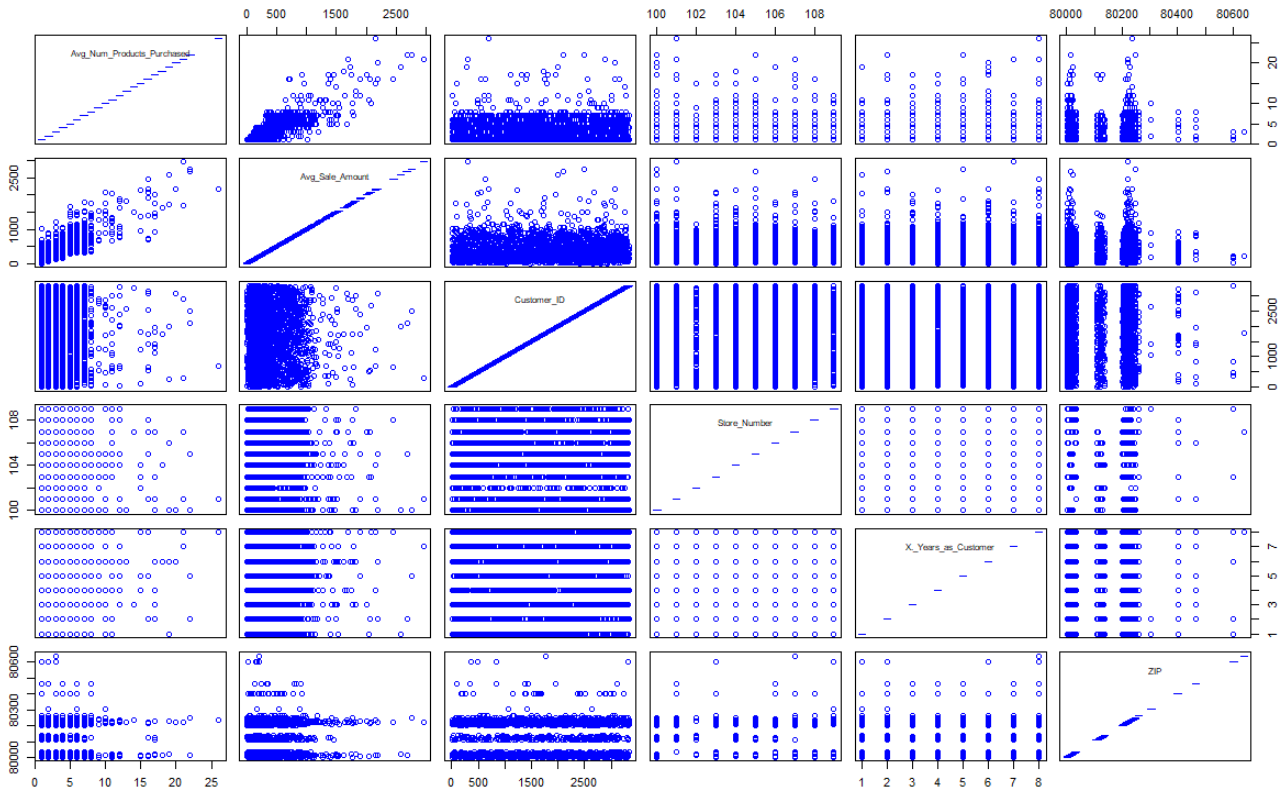
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

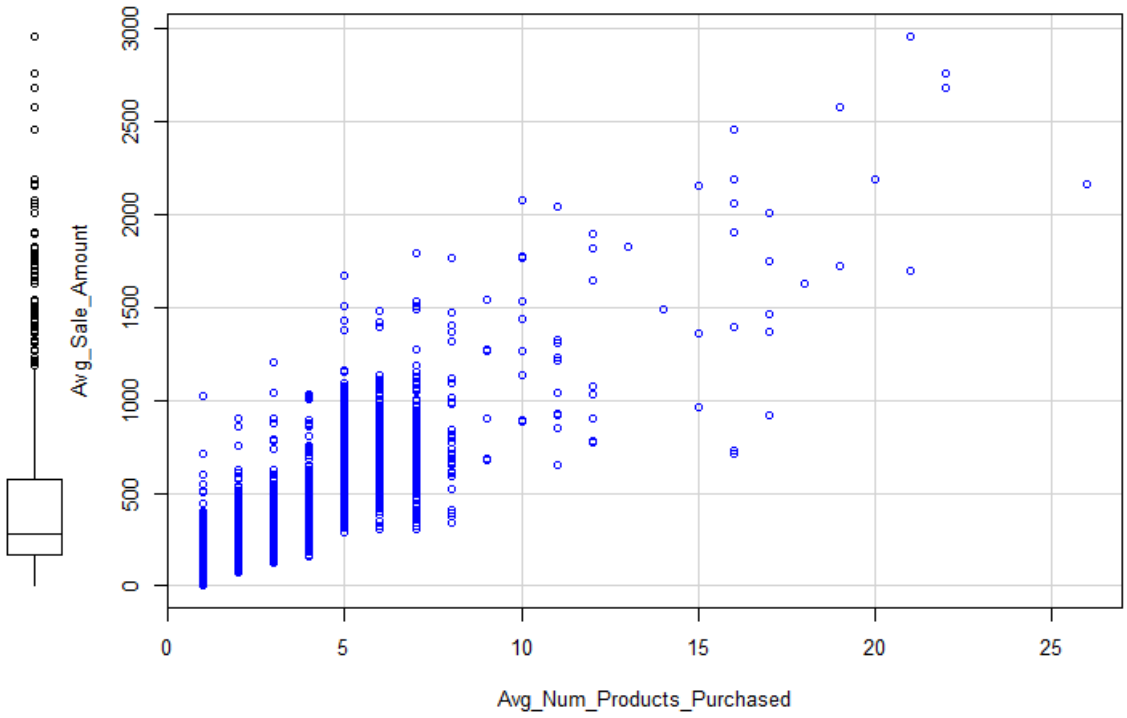
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel"

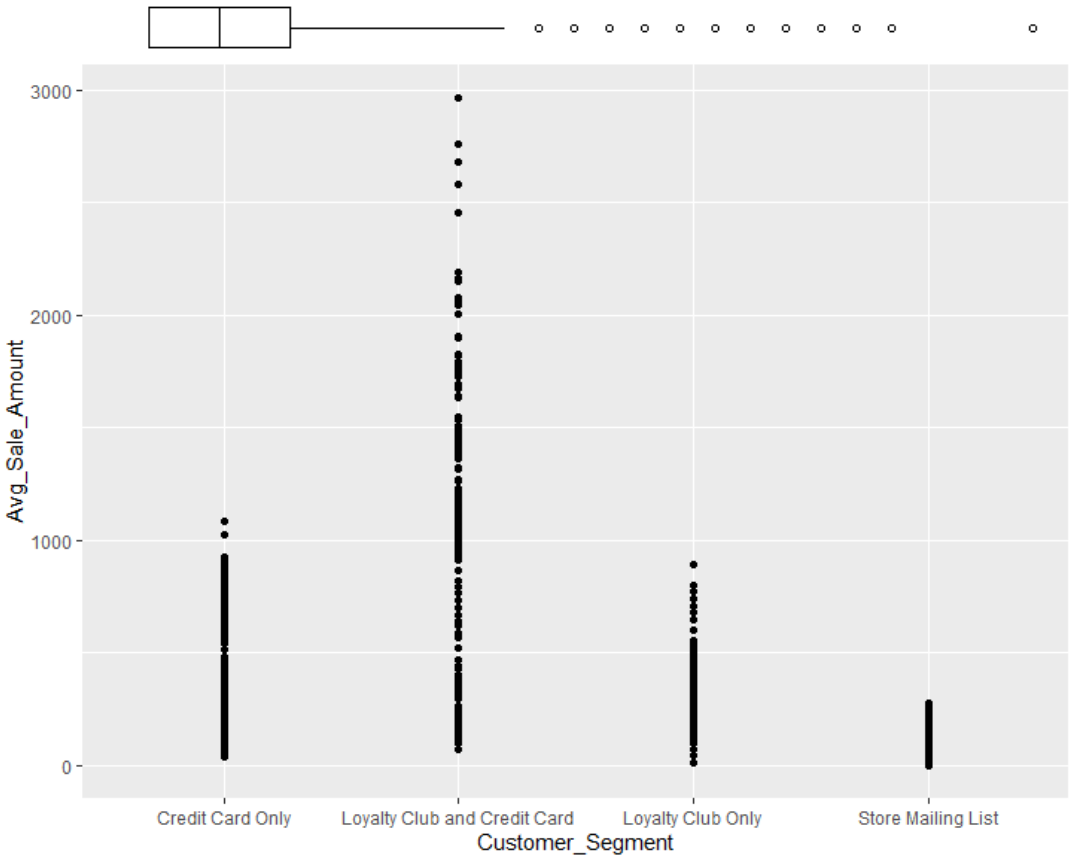
lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

Let us see the scatter plot matrix for the variables to understand the relationships.



Out of the linear variables, only Avg_Number_Of_Products_purchased seems linearly related with target variable, Avg_sales_amount. Let explore the scatter plot.


Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
Avg_Num_Products_Purchased


Avg_Sale_Amount
3000
2000
1000
0
Credit Card Only
Loyalty Club and Credit Card
Loyalty Club Only
Store Mailing List
Customer_Segment

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.

```
 Call:
lm(formula = Avg_Sale_Amount ~ Avg_Num_Products_Purchased + Customer_Segment,
    data = customer)

Residuals:
    Min      1Q  Median      3Q     Max
-663.77  -67.31   -1.90   70.69  971.69

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                      303.463     10.576   28.69
Avg_Num_Products_Purchased                        66.976      1.515   44.21
Customer_Segment[T.Loyalty Club and Credit Card] 281.839     11.910   23.66
Customer_Segment[T.Loyalty Club Only]           -149.356      8.973  -16.64
Customer_Segment[T.Store Mailing List]          -245.418      9.768  -25.12
                                                Pr(>|t|)
(Intercept)                                       <2e-16 ***
Avg_Num_Products_Purchased                        <2e-16 ***
Customer_Segment[T.Loyalty Club and Credit Card]  <2e-16 ***
Customer_Segment[T.Loyalty Club Only]             <2e-16 ***
Customer_Segment[T.Store Mailing List]            <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.5 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

For both the predictor variables, we used in our linear model creation, p-val
ue (probability that the coefficient is going to be 0) is very less. Hence, b
oth the predictors are significant in deciding the target variable.

This model is strong since the R-value is very high (0.8369)
```

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_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased -149.36 (If Cus
tomer_Segment: Loyalty Club Only) + 281.84 (If mailCustomer_Segment is Loyalt
y Club and Credit Card) – 245.42 (If Customer_Segment is Store Mailing List)
+ 0 (If Customer_Segment is Credit Card Only)
```

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?
After conducting an analysis on the data, I would recommend that the company send out the catalog to these 250 customers as there is a profit of $ 21987.96 (greater than $10,000 profit) The calculated files attached.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
I have created a linear regression model, using the average sales amount as the target variable and using 2 predictors variable. The average number of products bought per customer and customer segment (this is a categorical variable), hence the equation is explained in step 2. Then a set of if elif statements in R to get the predicted value. I can use the coefficients provided to predict expected revenue and expected profit (revenue*0.5 -6.5).
The code sets as follows.

```
mail_list_p <- mutate(mail_list, predicted_sales= case_when(
  mail_list_p$Customer_Segment=='Loyalty Club and Credit Card' ~ 303.46 + 66.
98*mail_list_p$Avg_Num_Products_Purchased + 281.84,
  mail_list_p$Customer_Segment=='Loyalty Club Only' ~303.46 + 66.98 *mail_lis
t_p$Avg_Num_Products_Purchased-149.36,
  mail_list_p$Customer_Segment=='Store Mailing List' ~303.46 + 66.98*mail_lis
t_p$Avg_Num_Products_Purchased-245.42,
  mail_list_p$Customer_Segment== 'Credit Card Only' ~ 303.46 + 66.98*mail_lis
t_p$Avg_Num_Products_Purchased,
))


View(mail_list_p)


mail_list_profit <- mutate(mail_list_p, predicted_rev = mail_list_p$predicted
_sales*mail_list_p$Score_Yes)
mail_list_profit <- mutate(mail_list_profit, predicted_profit = mail_list_pro
fit$predicted_rev*0.5 -6.5)
View(mail_list_profit)
total_profit <- sum(mail_list_profit$predicted_profit)
total_profit
write.csv(mail_list_profit, file = "D:/Predictive Analytics/Udacity/Predictiv
e Analytics for Business/Project 2/mail_list_profit.csv")
```

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

Expected profit, $21987.96; from sum of all 250 customers.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.