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?

This company need to predict whether worth to send a print catalogue to 250 new customers. To get this topic, first we need to calculate what the cost of goods is. 10000*2+6.5*250*2=$23250. So total sales should over than $23250 then we can send the flyers.

2. What data is needed to inform those decisions?

We must know average sale amount from customer table as target variable. Other variables may use for predictor variables. Then calculate total predict sale amount by multiplied by probability of customer in mailing list table.

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)

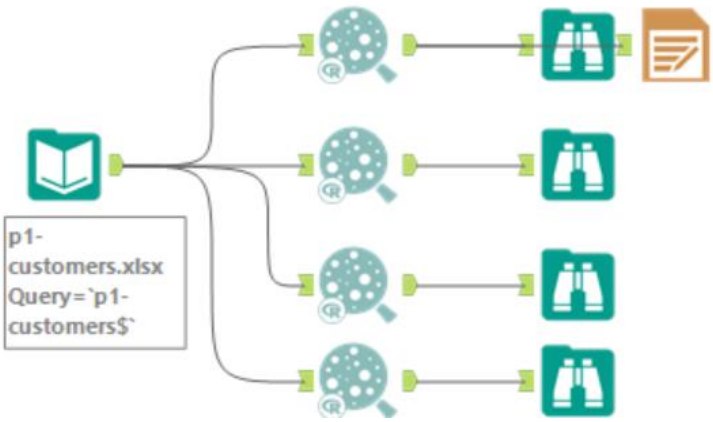
Important: Use the p1-customers.xlsx to train your linear model.

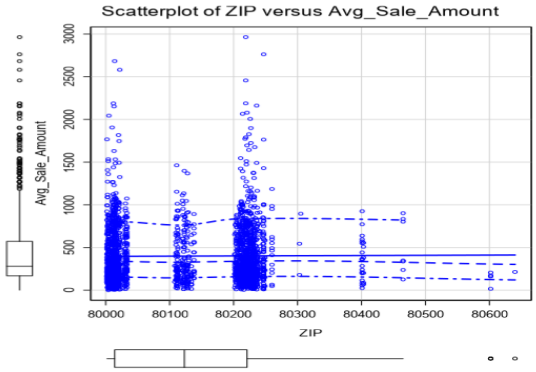
At the minimum, answer these questions:

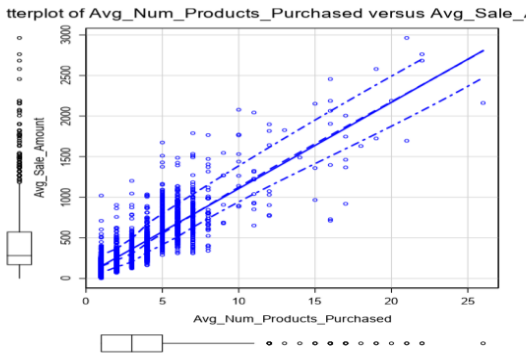
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

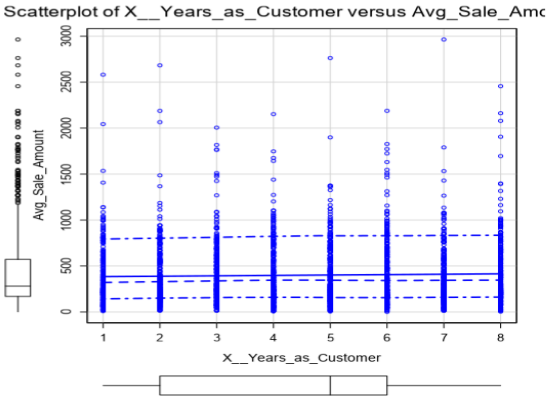
Because we have to predict sales amount, so average sales amount used as target variable.
First, we use correlation chart (Pearson) to check any liner correlation between each variable.
Now we know only variable (avg_num__products_purchased correated) number close 1. Others I don't think have liner correlation.

Second, for categorical variable, I only add variable (custom_Segment) because this one only has 4 segments, others have too many values.


p1-
customers.xlsx
Query=`p1-
customers$`


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

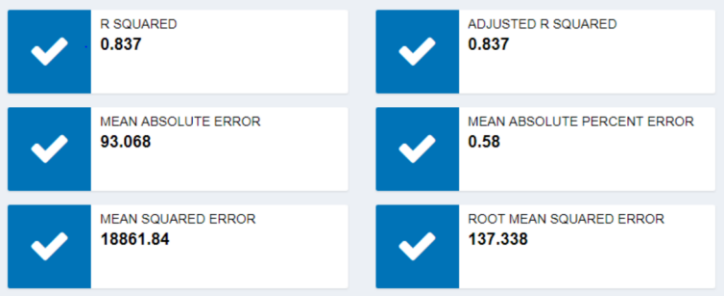

tterplot of Avg_Num_Products_Purchased versus Avg_Sale_
Avg_Sale_Amount
Avg_Num_Products_Purchased
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25


Scatterplot of X__Years_as_Customer versus Avg_Sale_Amc
Avg_Sale_Amount
X__Years_as_Customer
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8


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

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.

The adjust R-square is 0.837 which is a good score, so I will use variable (avg_Num_Products_Purchased) and variable (customer_segment) as predict variables.



Also, I use P-value to check the significance. Any value below 0.05 can be used for this liner regression. P value for Dummy variables for custom_Segment is way less than 0.05. now ready to use as predict variables.

Let see the results table.

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

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 avg_num__products_purchased correated –149.36 *customer_SegmentLoyalty club only +281.84* customer_SegmentLoyalty club and Credit Card -245.42* customer_SegmentStore Mailing List*

Step 3: Presentation/Visualization

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

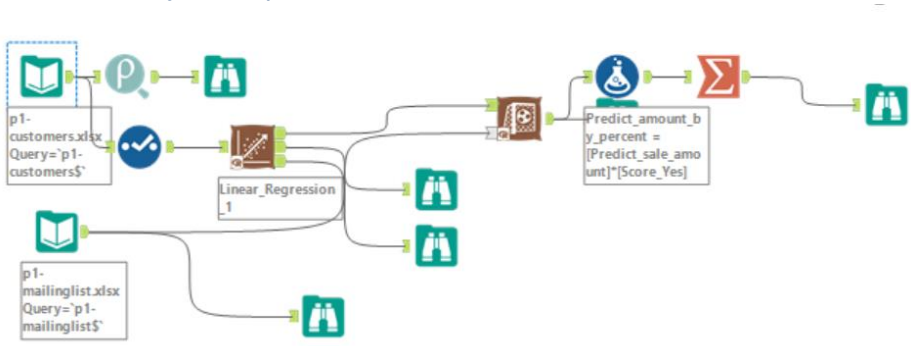
At the minimum, answer these questions:

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

Yes, I recommend manager to send catalogs to 250 new customers.

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

This is about my whole process.



At first, I used customers.xlsx to make a liner regression then connect score tool as input. Secondly, the score tools connect with mailinglist.xlsx as input and send output to get forecasting data (expect average purchase amount).
Thirdly, to use formula function to multiply by (Score_yes) field. Then choose summaries tool to sum the total expect value is $47224. This following table shows total_sum_predict_amount_by_precent.

Sum_Predict_amount_by_percent_total
47224.871373

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

The expect profit can 47224/2- 250*6.5=21987. So, I will suggest manger to send the catalogues.

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.