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?

- Whether or not to send the catalog to new customers.

2. What data is needed to inform those decisions?

- The estimated profit from new customers, which required the shared data feilds of the current customers dataset and new customers dataset

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)

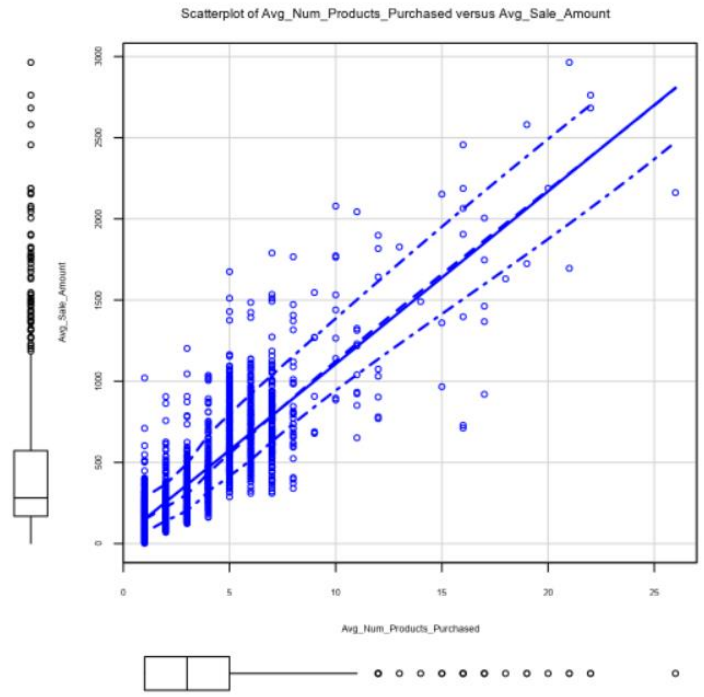
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

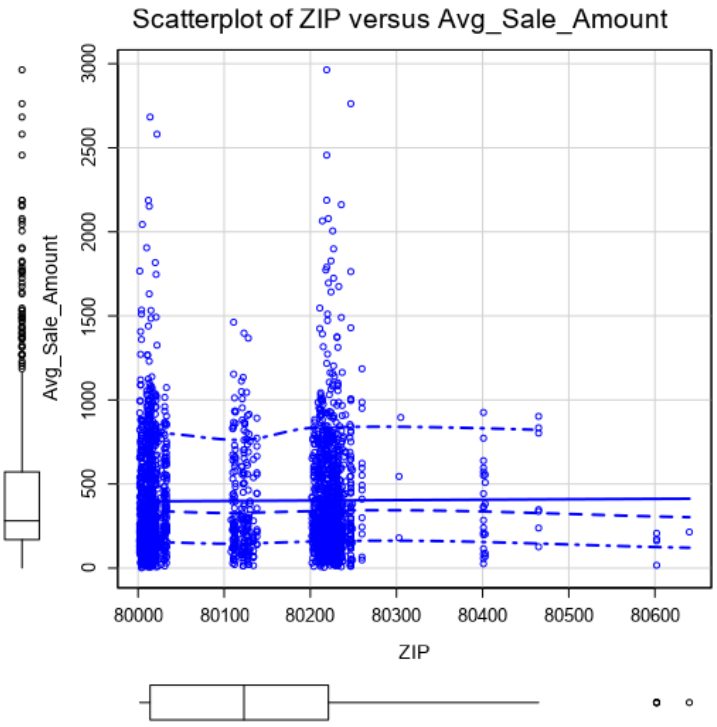
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

The scatterplots of other numeric variables vs. Avg_Sale_Amount (Revenue):

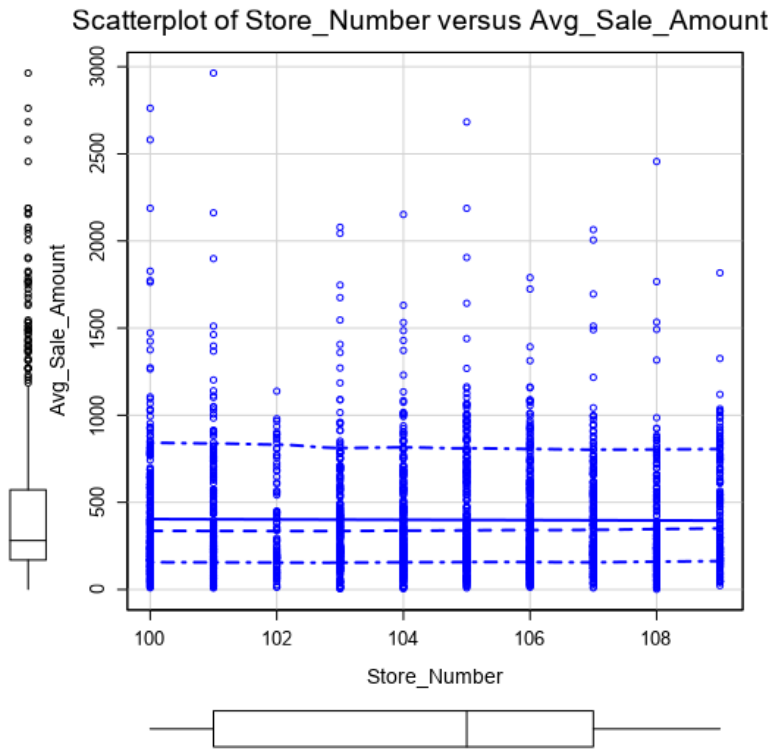
- ***Avg_Num_Products_Purchas****ed vs.* ***Avg_Sale_Amount****:*



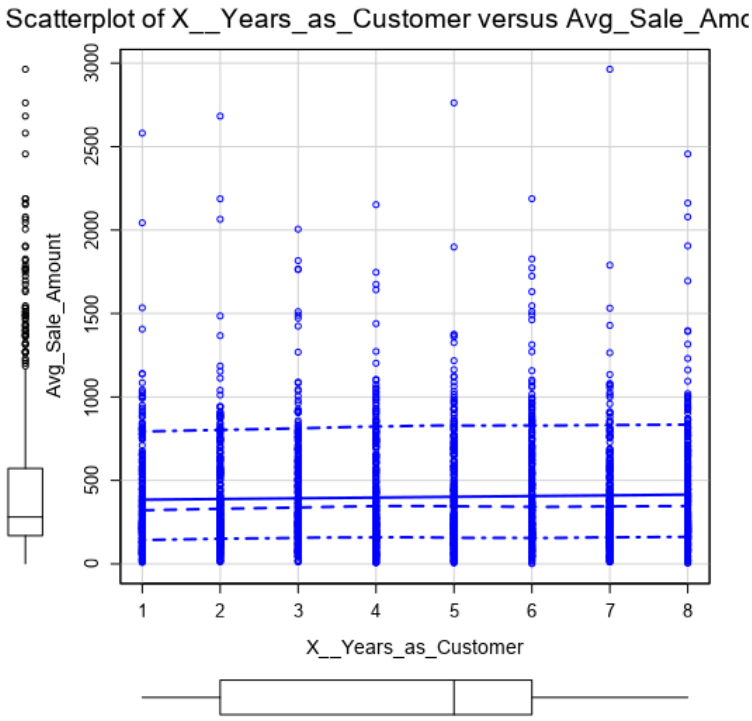
- ***Zip*** *vs.* ***Avg_Sale_Amount****:*



- ***Store_Number*** *vs.* ***Avg_Sale_Amount****:*



- ***X_Years_as_Customer*** *vs.* ***Avg_Sale_Amount****:*



*Base on the scatterplots above, there's no significant correlation between **Zip/Store_Number/ X_Years_as_Customer** and **Avg_Sale_Amount**, whereas the relationship between **Avg_Num_Products_Purchased and Avg_Sale_Amount** shows a strong positve correlation.*

*Addtionally, the table below depicts the correlation between all revelant variables and the target variables (**Avg_Sale_Amount**), which indicates that the **Customer_Segment** and **Avg_Num_Products_Purchased** are significant predictor variables (with p-values < 0.05).*

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.379e+03	2.149e+03	-0.6416	0.52118
Customer_SegmentLoyalty Club Only	-1.497e+02	8.980e+00	-16.6659	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	2.824e+02	1.193e+01	23.6659	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-2.459e+02	9.774e+00	-25.1627	< 2.2e-16 ***
Customer_ID	-1.373e-03	2.941e-03	-0.4669	0.64063
ZIP	2.248e-02	2.660e-02	0.8451	0.39814
Store_Number	-1.011e+00	1.007e+00	-1.0042	0.31539
Avg_Num_Products_Purchased	6.700e+01	1.517e+00	44.1582	< 2.2e-16 ***
X._Years_as_Customer	-2.345e+00	1.223e+00	-1.9167	0.0554 .

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 model is a good model due to the significant predictor variables (with their p-values above 0.05), and the adjusted R-squared value is 0.8366, which suggests the model is highly predictive.

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 – 149.36 (If Type: Loyalty Club Only) + 281.84 (If Type: Loyalty Club and Credit Card) – 245.42 (If Type: Mailing List) + 0 (If Type: Credit Card Only)*

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?

As the estimated total profit from new customers is 21,987.44 USD, exceed 10,000 minimum expectation, we should recommend to send the catalog to the new customers.

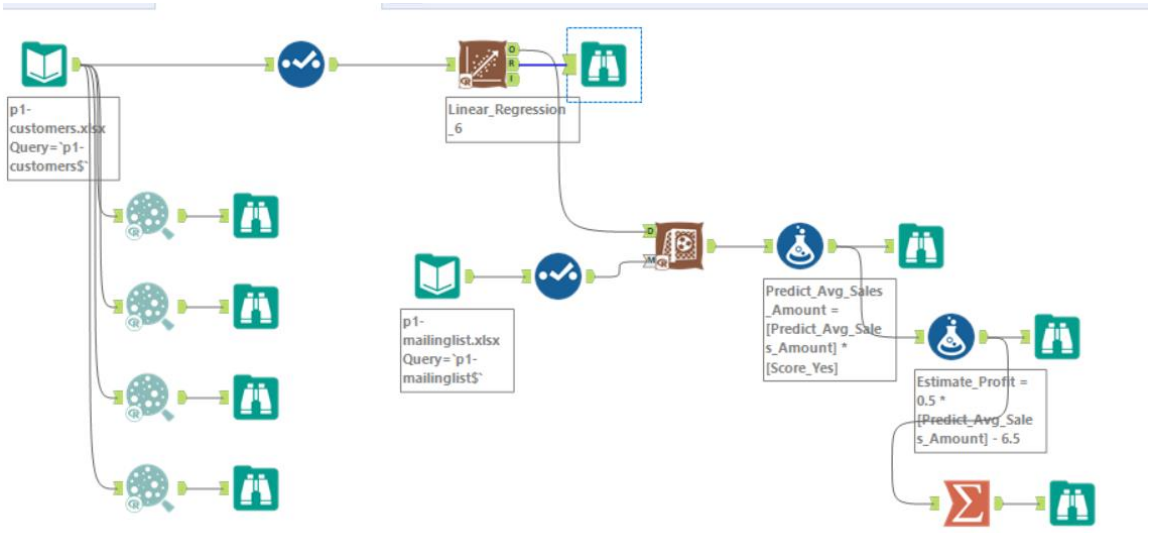
Record	Sum_Estimate_Profit
1	21987.435696

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

The prediction analysis process:

- *After choosing the appropriate significant predictor variables from the scatterplots and the initial model, I used the 'Score' tool to predict the estimated revevnue(**Predict_Avg_Sales_Amount**) with the data from the new customer dataset 'p1-mailinglist.xls'.*
- ***Predict_Avg_Sales_Amount** is calculated by multiply the probability of product purchasement to the intitial predicted revenue: **Predict_Avg_Sales_Amount = Predict_Avg_Sales_Amount * Score_Yes***
- *Then calculate the estimate profit: **Estimate_Profit = 0.5 * Predict_Avg_Sales_Amount - 6.5**, before come up with the total estimated revenue, which is 21,987.44 USD*

Per the Alteryx workflow below:



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

The estimated total profit from new customers is 21,987.44 USD