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?

The Company wanted to know how much the expected profit from sending the catalog to 250 new customers from their mailing list. The situation required predicting sales from the new mailing list, and calculated profit based on the predicted sales. And the manager would decide to send out the catalog if the profit exceeds 10,000, otherwise, the manager would not send out the catalog.

2. What data is needed to inform those decisions?

For the problem, we need two different datasets. First part is the historical data of the customer sales, including Customer ID, City, State, Address, Zip, Customer Segment, Store Number, Number of Years As Customers, Average number of Products Purchased and the Average Sales Amount. We need to build our prediction model based on these data. Second part is the customer information of the 250 new customers from the mailing list, containing the same categories as the first part.

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.

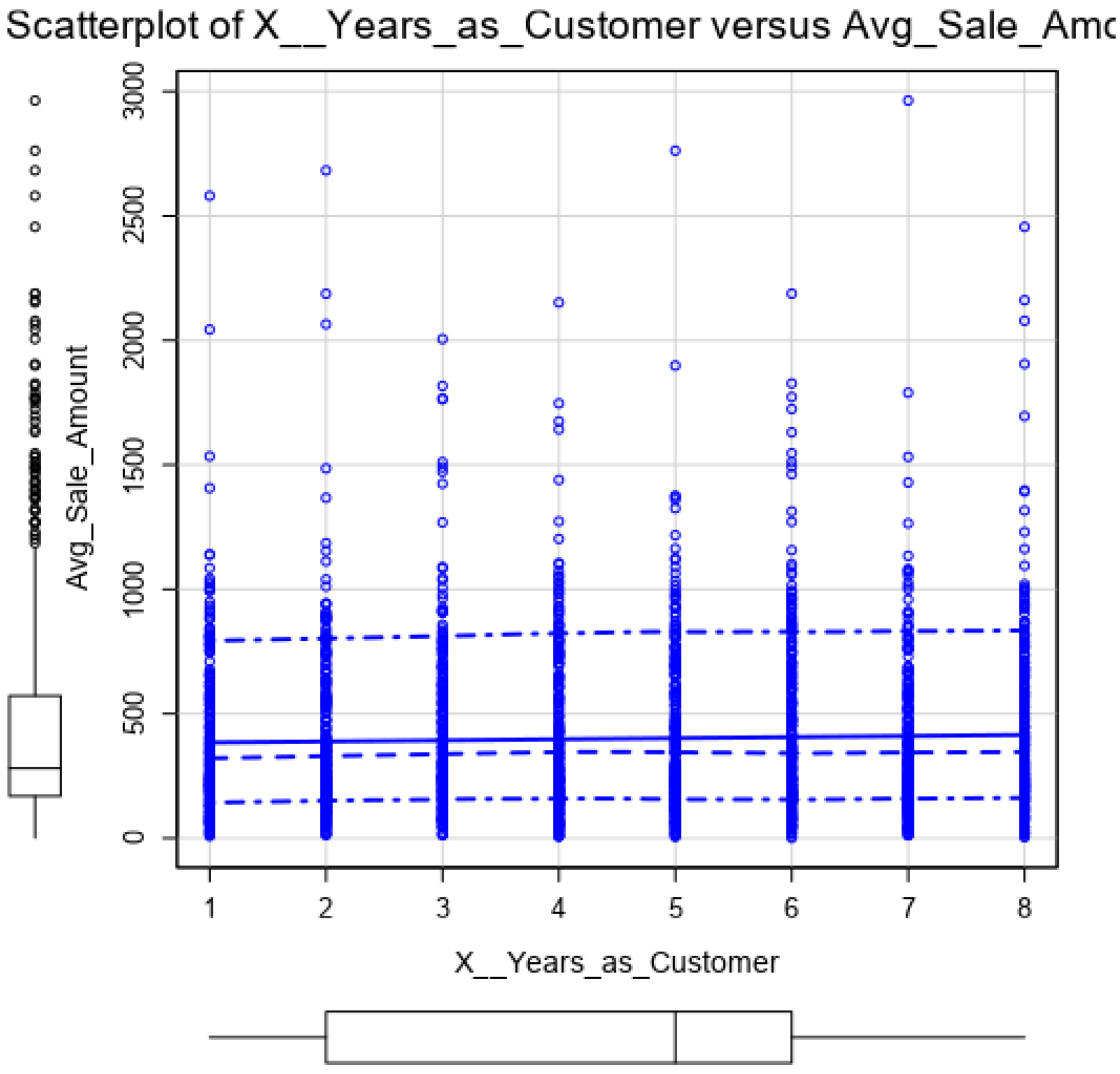
At the minimum, answer these questions:

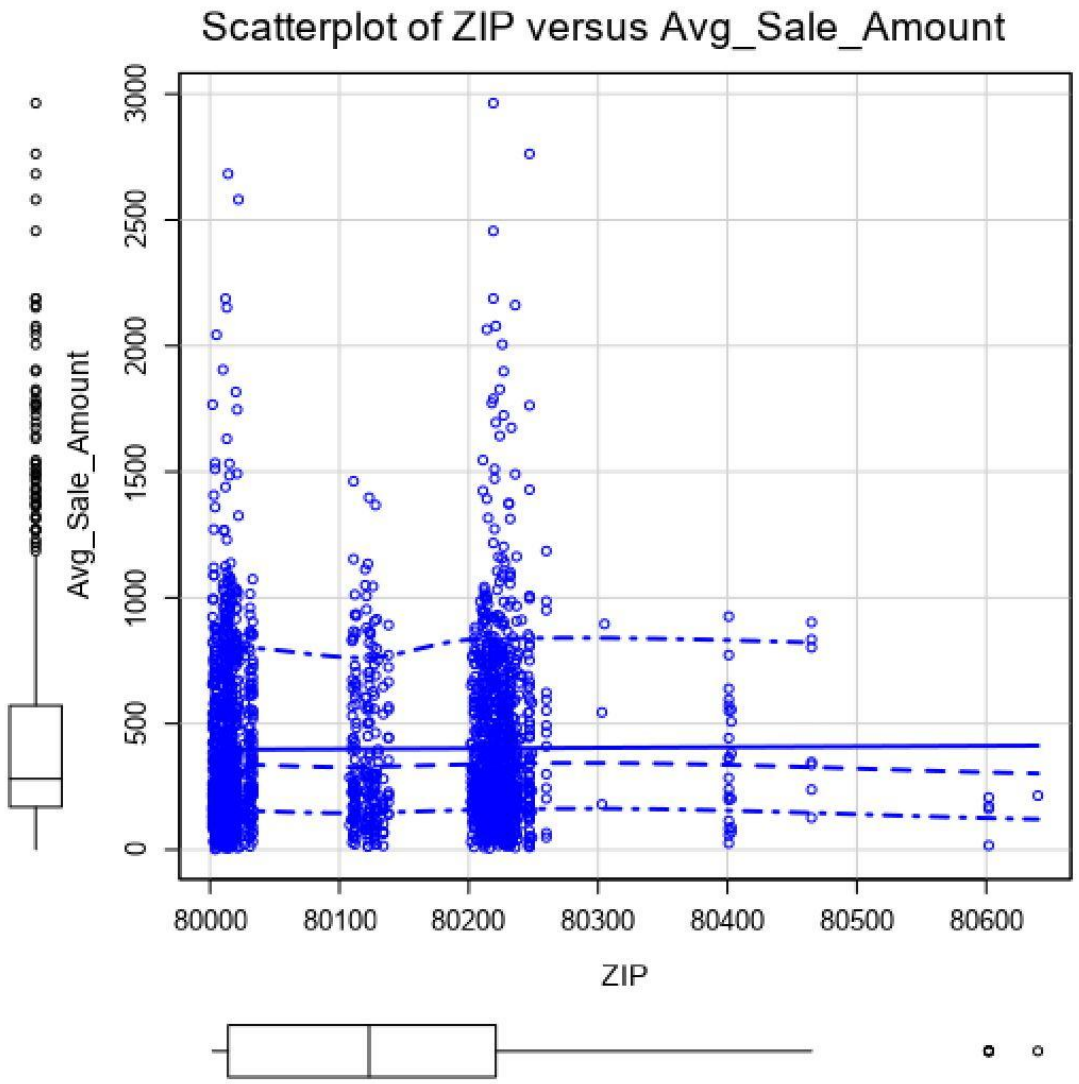
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

After examining relationships between Averages Sales Amount and other variables, I chose Customer Segment and Average Number of Products Purchased as my predictor variables.

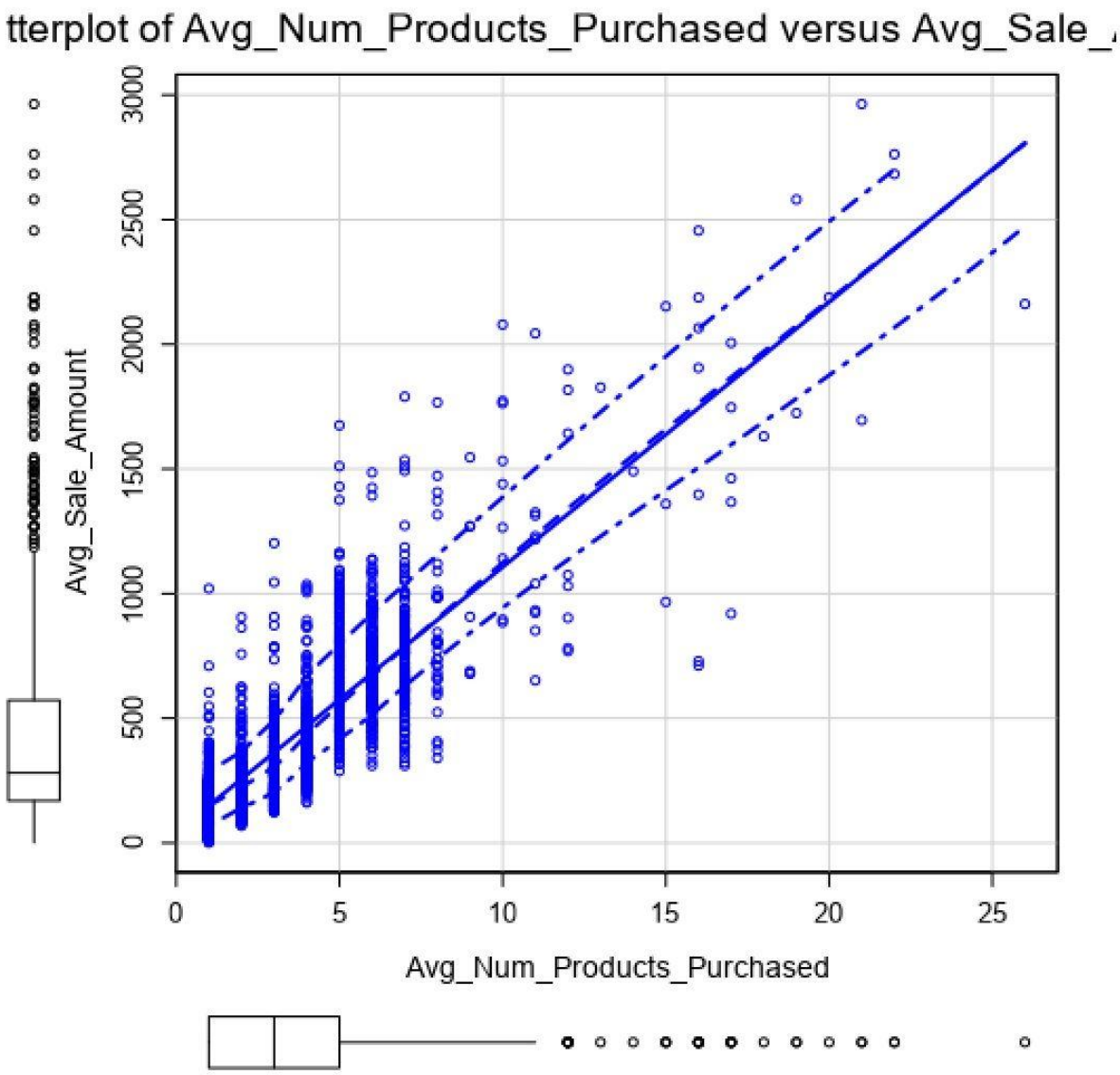
Explore data analysis:

Here I attached scatter plots to present relationships between sales and other variables, from the graphs, we can see that most of the relationships were not significant, such as relationships between Number of years as Customers and Sales Amount.





As we can see from the graph below, the relationship between Average Sales Amount and the Average Number of Products Purchased was 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.

Here is the Linear Regression Model Result:

Record	Report
1	**Report for Linear Model RegressionModel_1**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	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

8	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
9	*Type II ANOVA Analysis*
10	Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28715078.96	3	506.4	< 2.2e-16	***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16	***
Residuals	44796869.07	2370			

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

The result showed that for each variable, the P-value was significant (<0.05). The model's R-squared was 0.84 with adjusted R-squared of 0.84.
Both of the indicators showed our model result was significant. The low P-value means that this result was very unlikely due to random noise, and the result was replicable. And R-squared of 0.84 means that 84% of variance for Sales Amount can be explained by independent variables in this model. Thus we believe our linear regression model is appropriate.

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 - 149.36 * (If Type: Customer_Segment Loyalty Club Only)
+ 281.84* (If Type: Customer_Segment Loyalty Club and Credit Card)
- 245.42* (If Type: Customer_Segment Store Mailing List)
+ 66.98* Avg_Number_of_Products_Purchased

Step 3: Presentation/Visualization

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

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

I would recommend the company send the catalog to these 250 customers. Because based on our prediction, the profit from sending catalog would be 21,987.44, which is much larger than 10,000.Please find below of question 3 for detailed calculation.

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

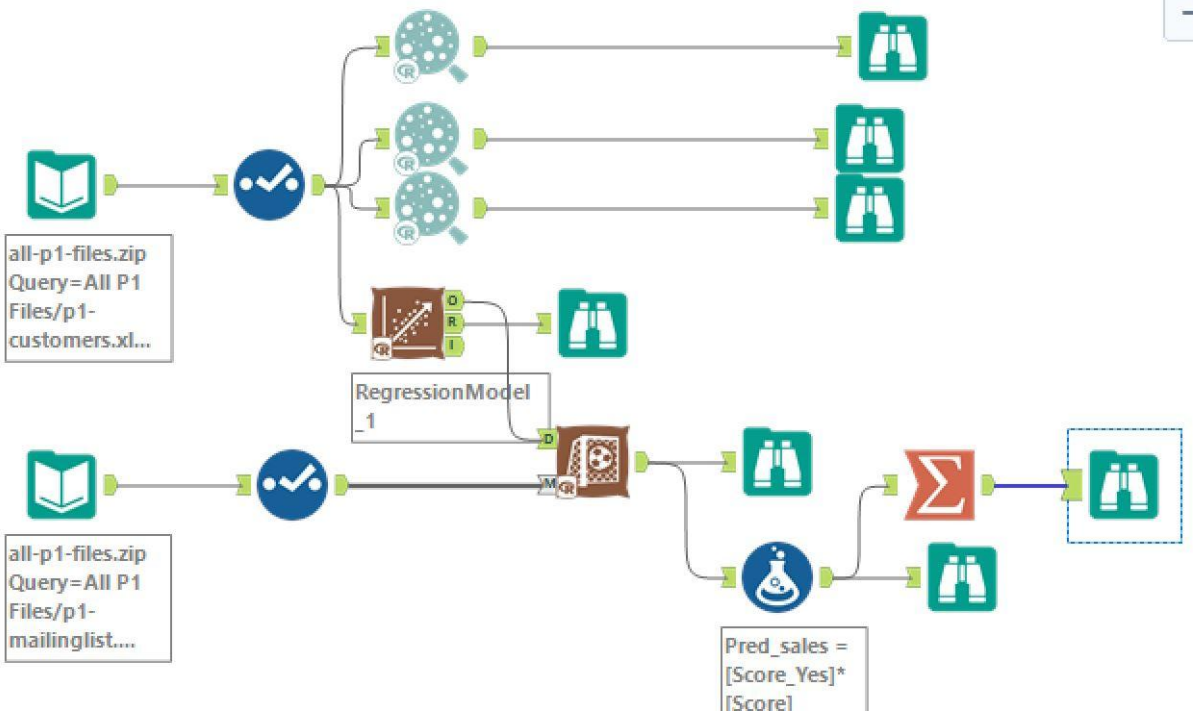
First, I implemented Alteryx to calculate predicted sales for each customer by the linear regression model (Score), multiplied by probability of buying catalog(Score of Yes) to get the potential sales for each customer. Then once I summed up all predicted sales, I have the total predicted sales.

Second, I calculated predicted profit using predicted sales, gross margin, and print cost for each catalog.
Profit = Sales * Gross Margin - Print Cost * Number of Mails

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

Profit Calculation:
(Based on Alteryx results of total sales: Sum of Predicted Sales
= Sum of (Predicted Sales for each customers * Probability of buying catalog)
= Sum of (Score *Score of Yes)
= 47224.87)
Profit
= Sales * Gross Margin - Print Cost * Number of Mails
= 47224.87 *50% - 6.5*250
= 21987.44

Additional
Alteryx workflow:


all-p1-files.zip
Query=All P1
Files/p1-
customers.xl...
RegressionModel
_1
all-p1-files.zip
Query=All P1
Files/p1-
mailinglist....
Pred_sales =
[Score_Yes]*
[Score]