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?**
   We need to decide whether to send a catalog to a prospective consumer or not depending upon the expected profit.
2. **What data is needed to inform those decisions?**
   We need to following data values to be able to take the aforementioned decisions -
   - The data regarding the cost i.e 6.50 and the profit margin i.e. 50% is very important and fundamental to the analysis.
   - Data about consumer's past purchase value is needed so as to make a model for predicting the average sales.
   - Following which data for the 250 consumer's parameter values are needed so that we can fit that into the model so as to predict the average sales for these consumers.
   - The probability of sales, so as to calculate the expected profit.
   - We need data that pertains to intrinsic details about the consumers purchases like the average number of products purchased because these are parameters that impact the target variable in our study.
   - We need data about the various consumers segments so as to understand better which consumer segment is expected to add the most to the profits.
   - Lastly the threshold of 10,000 is important so that we can make a comparison.

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.

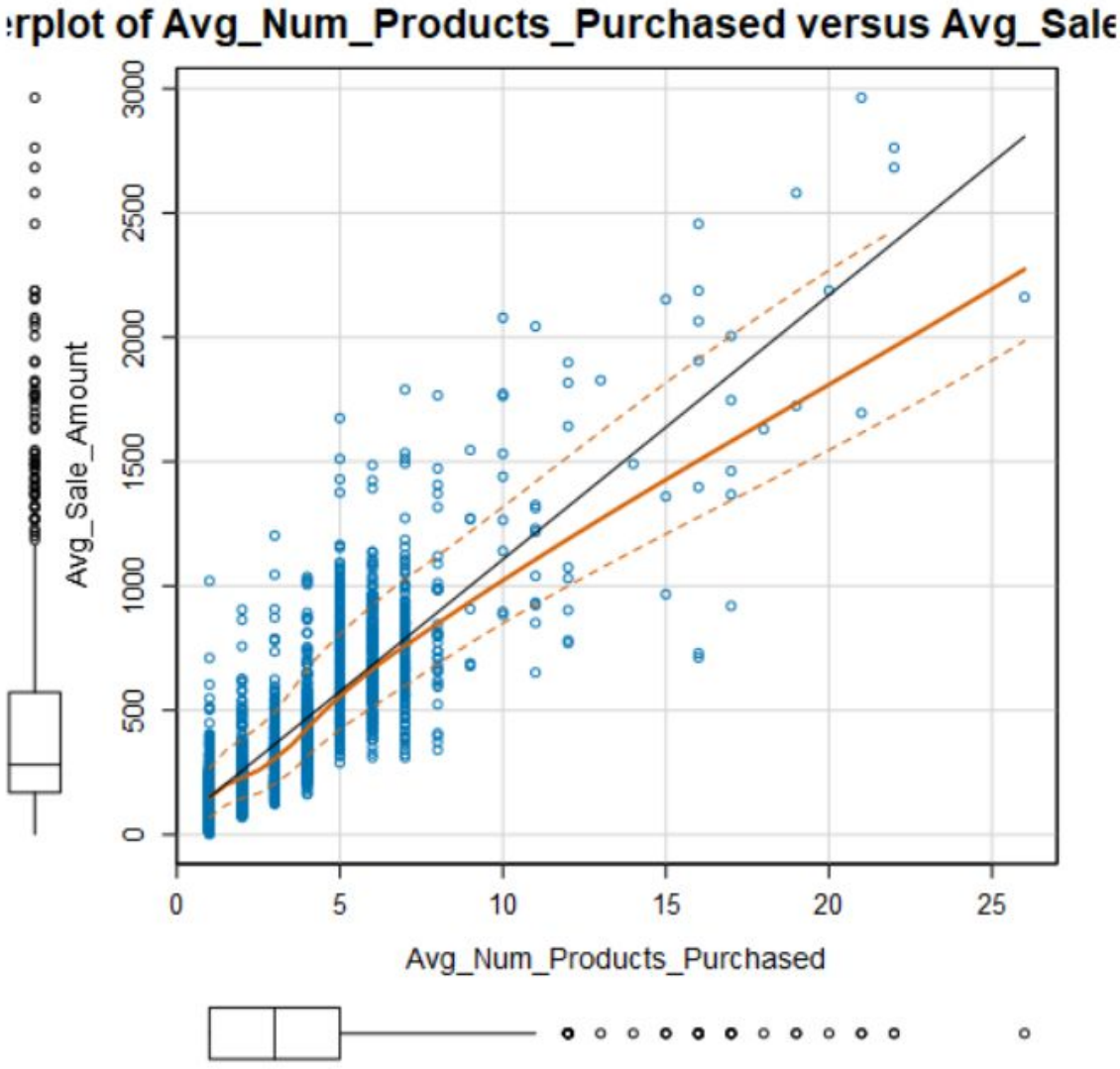
At the minimum, answer these questions:

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.

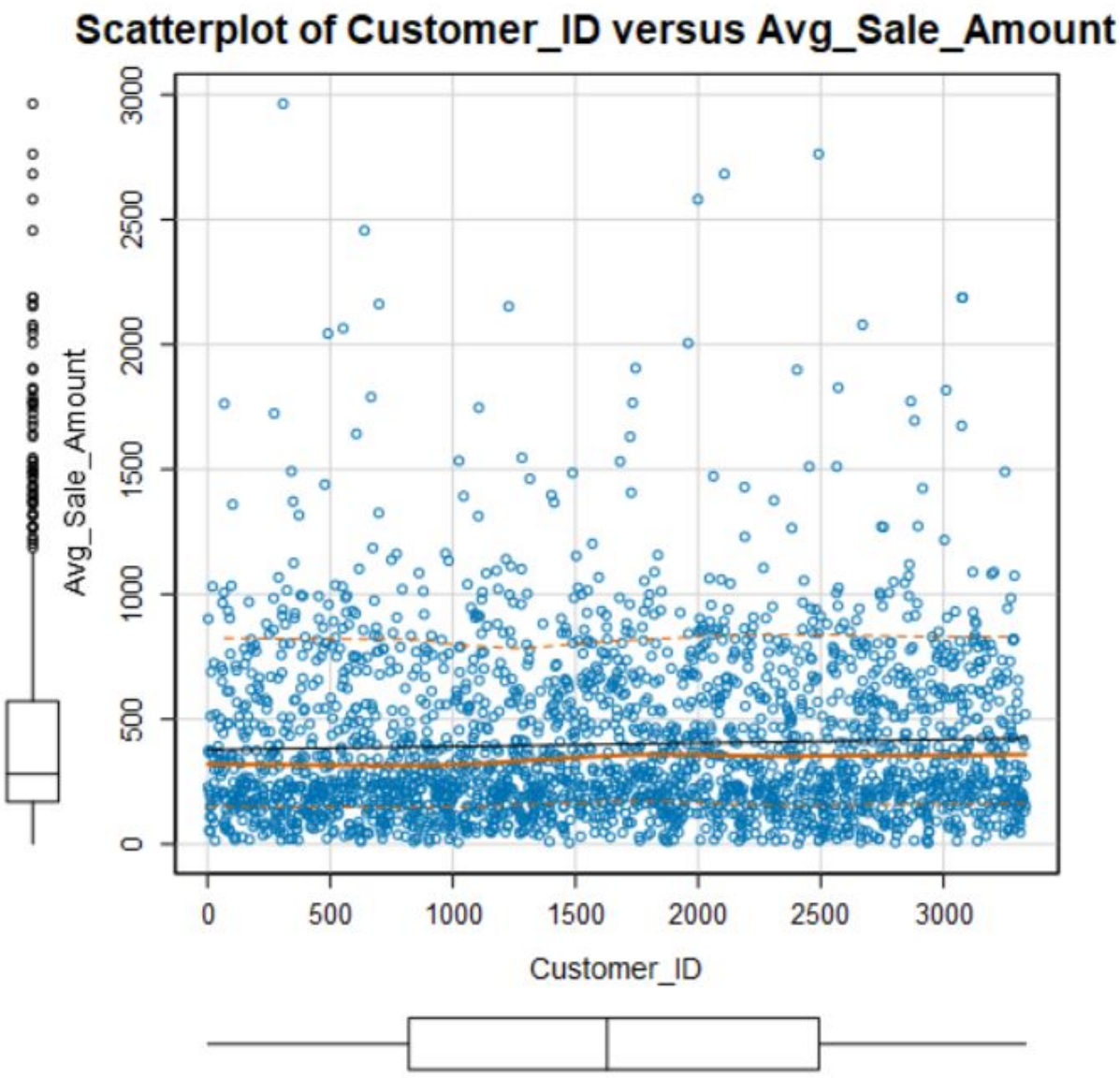
Predictor Variables chosen - Customer Segment and Average number of products purchased.

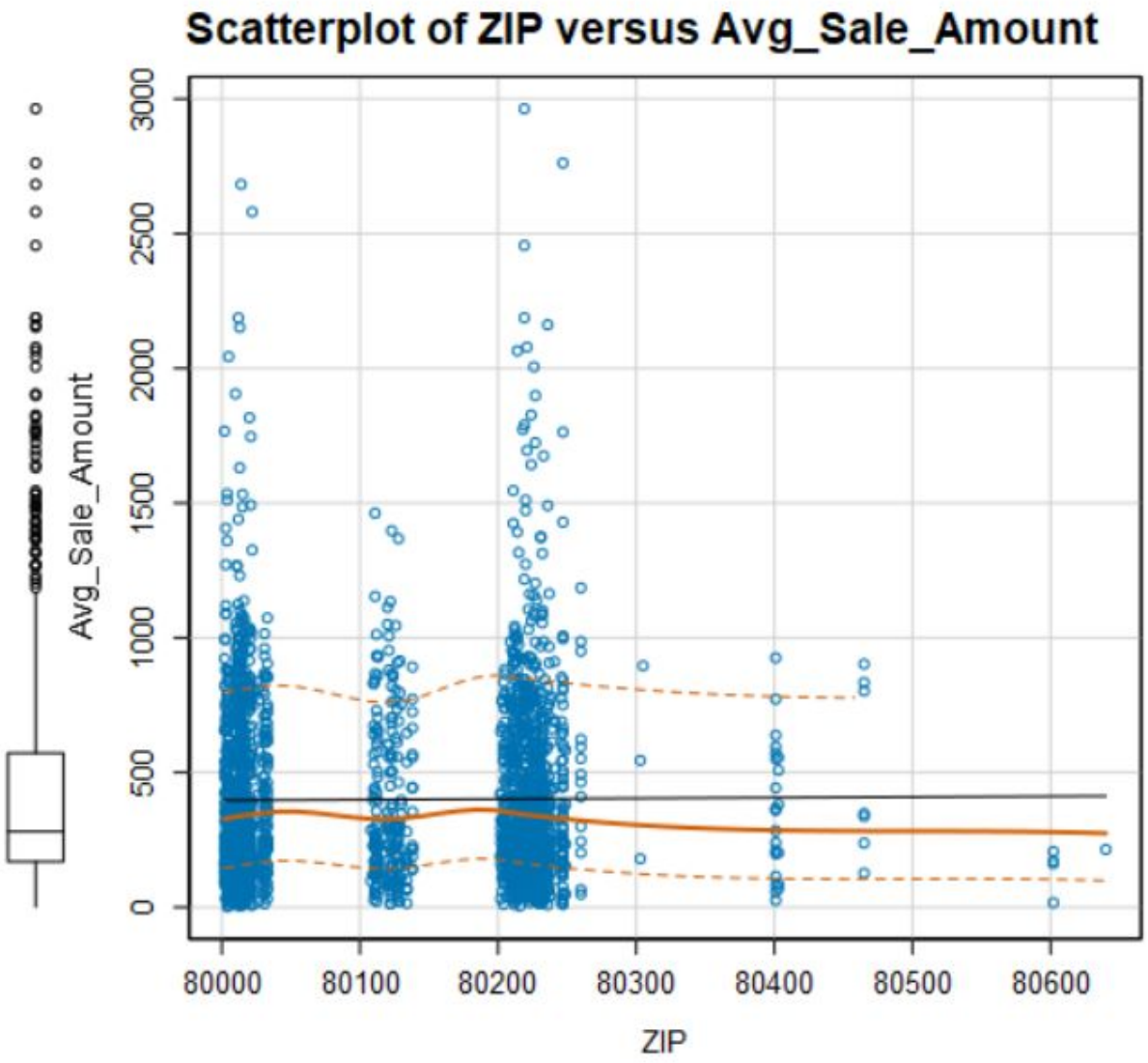
Target Variable chosen - Average sales

I have chosen the aforementioned predictor variables because apart from these two, the parameter that the consumer responded to last year's catalogue and the addresses are not important in the prediction of the average sale. It is clear from the scatter plot of average number of products purchased vs the average amount of sale that there exists a positive relationship among the two variables. When the number of products purchased increases the value of the average expected sales also increases.



It can be seen from these scatter plots that variables like Customer ID, ZIP, and store number do not have any relationship with avg_sale, therefore these parameters are not used in the estimation process.


Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
Customer_ID


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

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 linear model is a good model in this case because the data is showing a linear trend, when we plot the regression results we see that the data has a positive relationship. Moreover by using a linear regression model, we are getting the following results -

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

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

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

This result shows that the coefficients are statistically significant due to the very low p-values, and the R-squared values are high. This shows why using linear regression is a good choice for the given data.

A low p-value suggests that the predicted coefficients are statistically significant, implying that they are not zero. This means that the estimated parameters i.e. the predictor variables effectively have an impact on the target variable.

A high R-squared value suggests “Goodness-of-fit”, since the value is 0.84 this means that our model predicts 84% of the change in the target variable due to the predictor variables. Also this value is above 0.7, beyond which a model is considered a strong model.

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 x (If Type: Loyalty Club Only) + 281.84 x (If Type: Loyalty Club and Credit Card) – 245.42 x (If Type: Store Mailing List)
+ 0 x (If Type: Cash Only) + 66.98 x (Avg_Num_Products_Purchased)

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?**

Since the expected profit exceeds $10,000, as the value os coming out to be 21987.44, the company must send out the catalogs.

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

I used the predicted average sales value and multiplied that by the probability of sale so as to get to the expected sales. Next I summed up the expected sales and multiplied that by 0.5, because there is a 50% profit margin. Finally I subtracted the total cost i.e.250*6.5 from the profit margin to reach the final expected profit.

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

The expected profit is - 21987.44

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.