Predicting Catalog Demand

Step 1: Business and Data Understanding

The management of the company I work with, needs to decide if we should send out printed catalogues to the 250 new customers they have on their mailing list or we they shouldn't. We also need to decide if the resources we have is worth investing into printing and mailing the catalogues, we would make this decision based on the profit we expect to get from this move and which thankfully the management have put a threshold to.

Key Decisions:

Answer these questions

1. What decisions needs to be made?
    - We need to decide if we should put in resources to print and mail catalogues to our new customers. We would do this by checking if our expected profit from them will exceed $10,000.

2. What data is needed to inform those decisions?
    - Data containing information on old customers which responded to catalogues that were sent to them. Data such as the average number of products purchased by each customer, their customer segments etc. will be used to train a regression model to make predictions.
    - In our training data, variables such as average number of products purchased and any other variable which has a linear relationship with our target variable or a significant p-value will be relevant to getting our prediction. The probability that a customer will actually buy our products will help us determine the expected revenue and hence the expected profit.

Step 2: Analysis, Modeling, and Validation

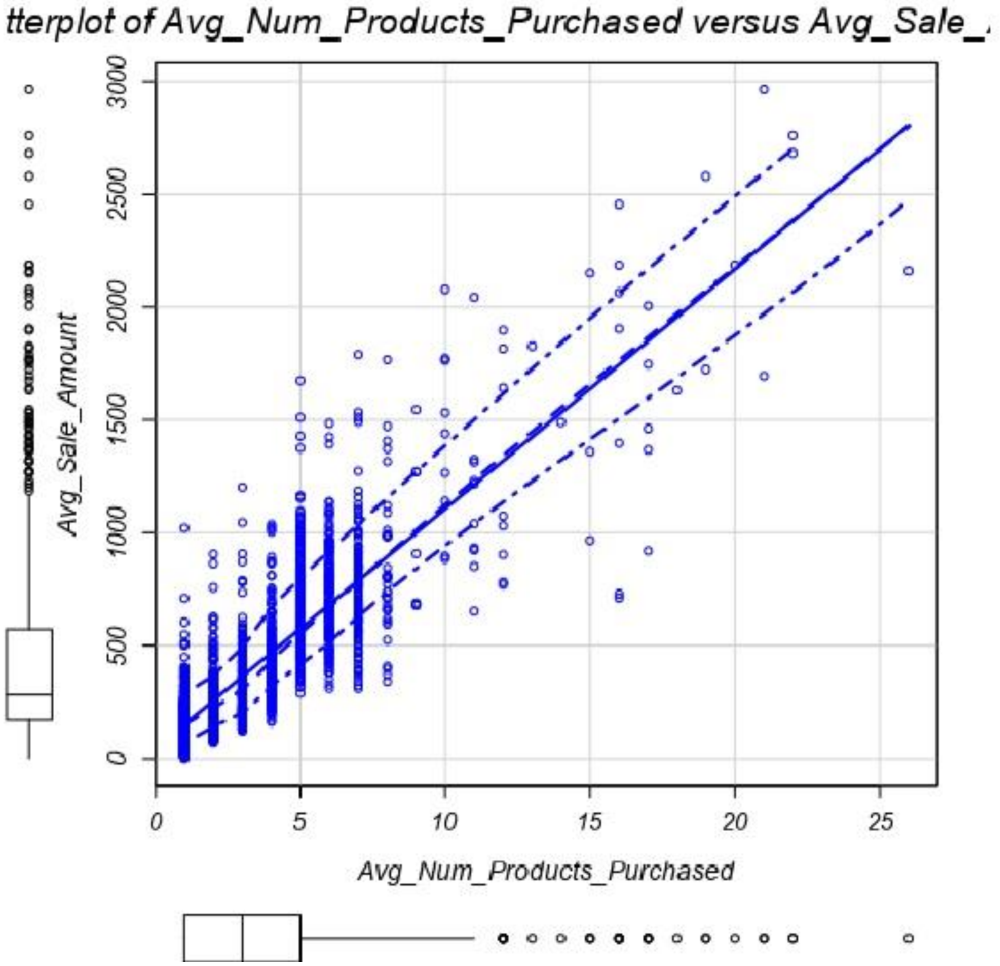
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

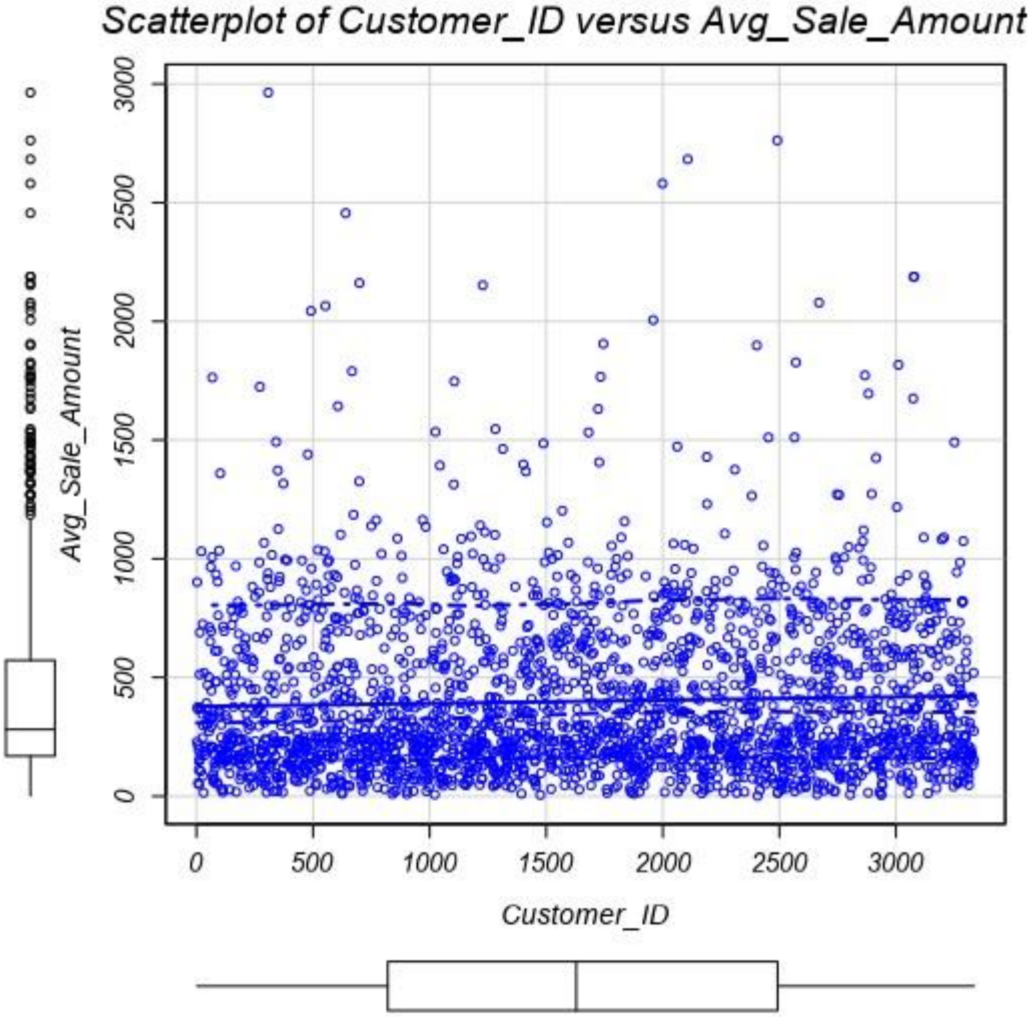
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

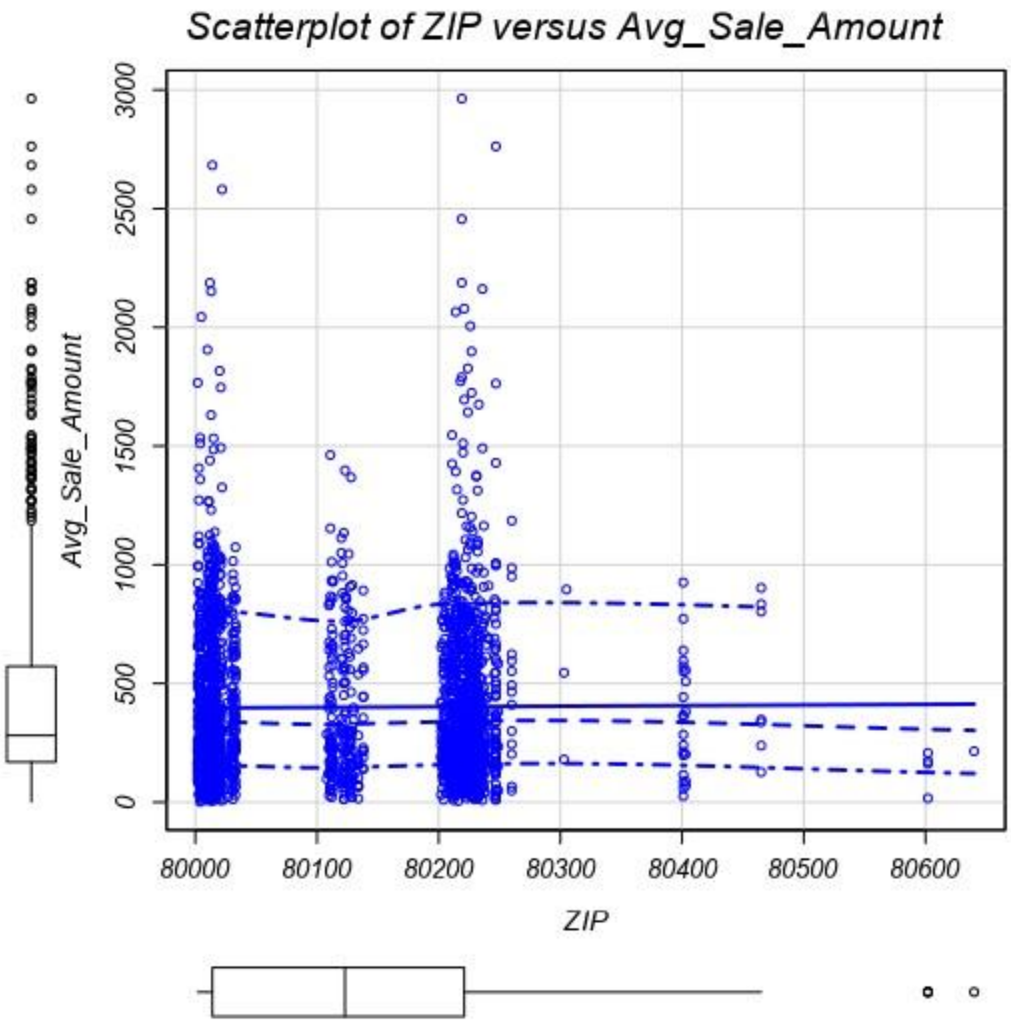
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

Ans: I selected some of my predictor variables by first checking the relationship between the continuous variables and the variable I intend to predict using a scatterplot. If I notice a linear relationship between the variable and the predicted or target variable, it gets selected as a predictor variable. Other predictor variables which I couldn't plot on the scatter plot where selected based on experiment and common sense, that is If I think a particular variable might contribute to the behavior of the variable I want to predict, I would include it and run my model, then I would check it's p-value, if it's p-value is less than 0.05 I would keep it, if it is 0.05 or more, I would drop it. Please find below, the scatterplots for all my numeric predictor variables.

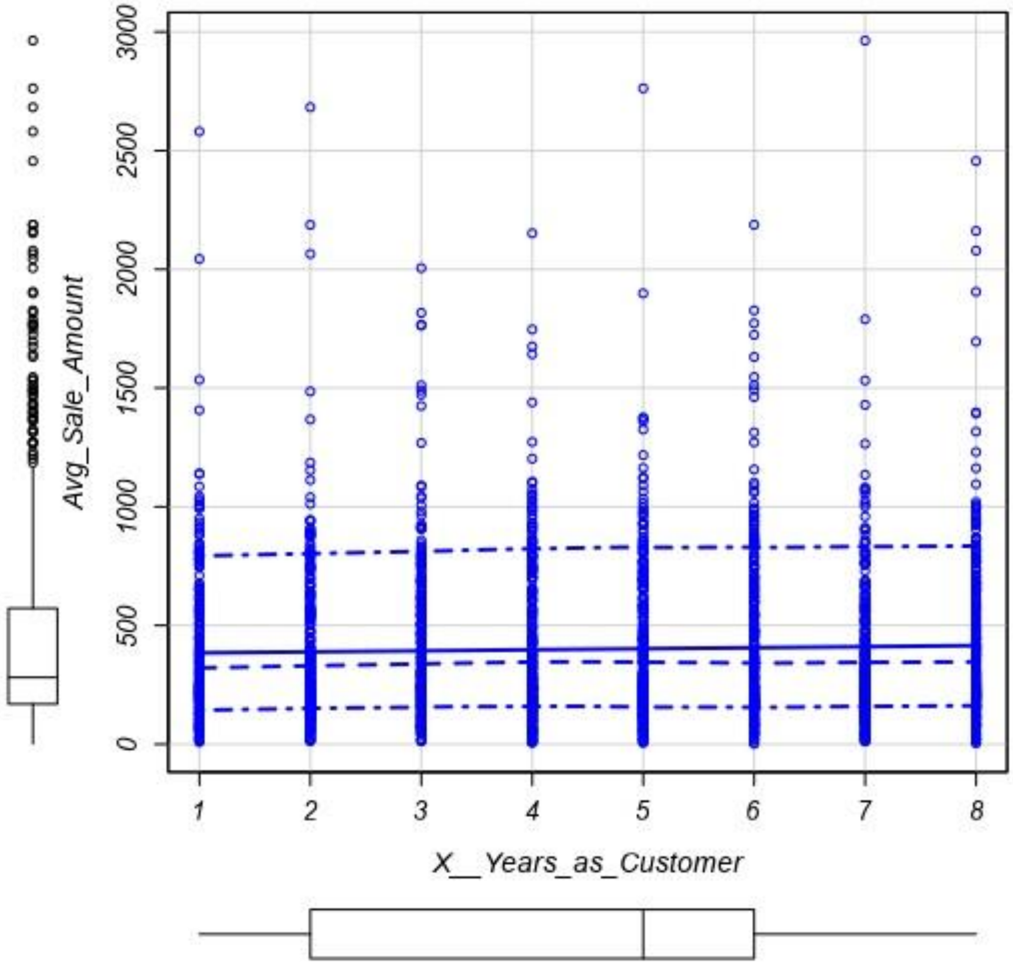

tterplot of Avg_Num_Products_Purchased versus Avg_Sale_
Avg_Sale_Amount
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
25
Avg_Num_Products_Purchased

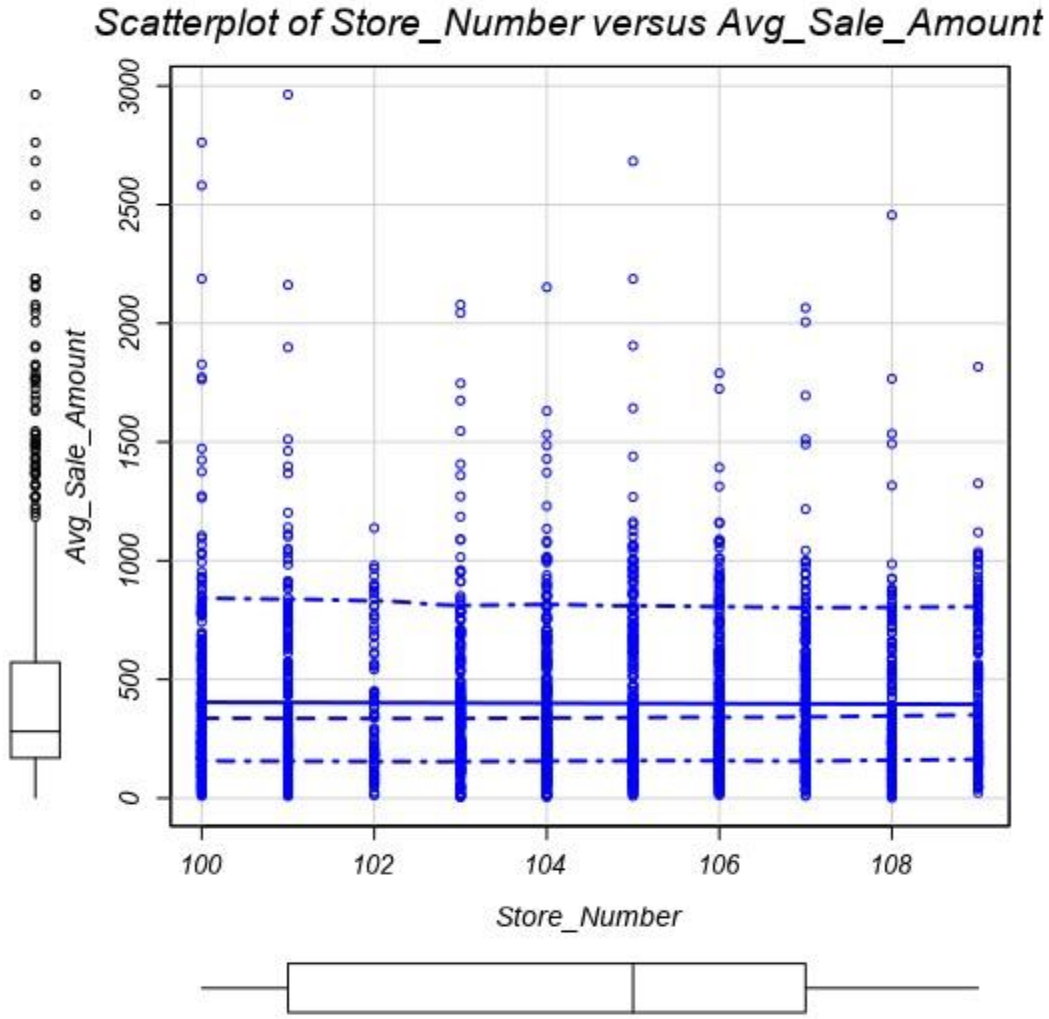


Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
Customer_ID
0
500
1000
1500
2000
2500
3000
0
500
1000
1500
2000
2500
3000


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

Scatterplot of X__Years_as_Customer versus Avg_Sale_Amc



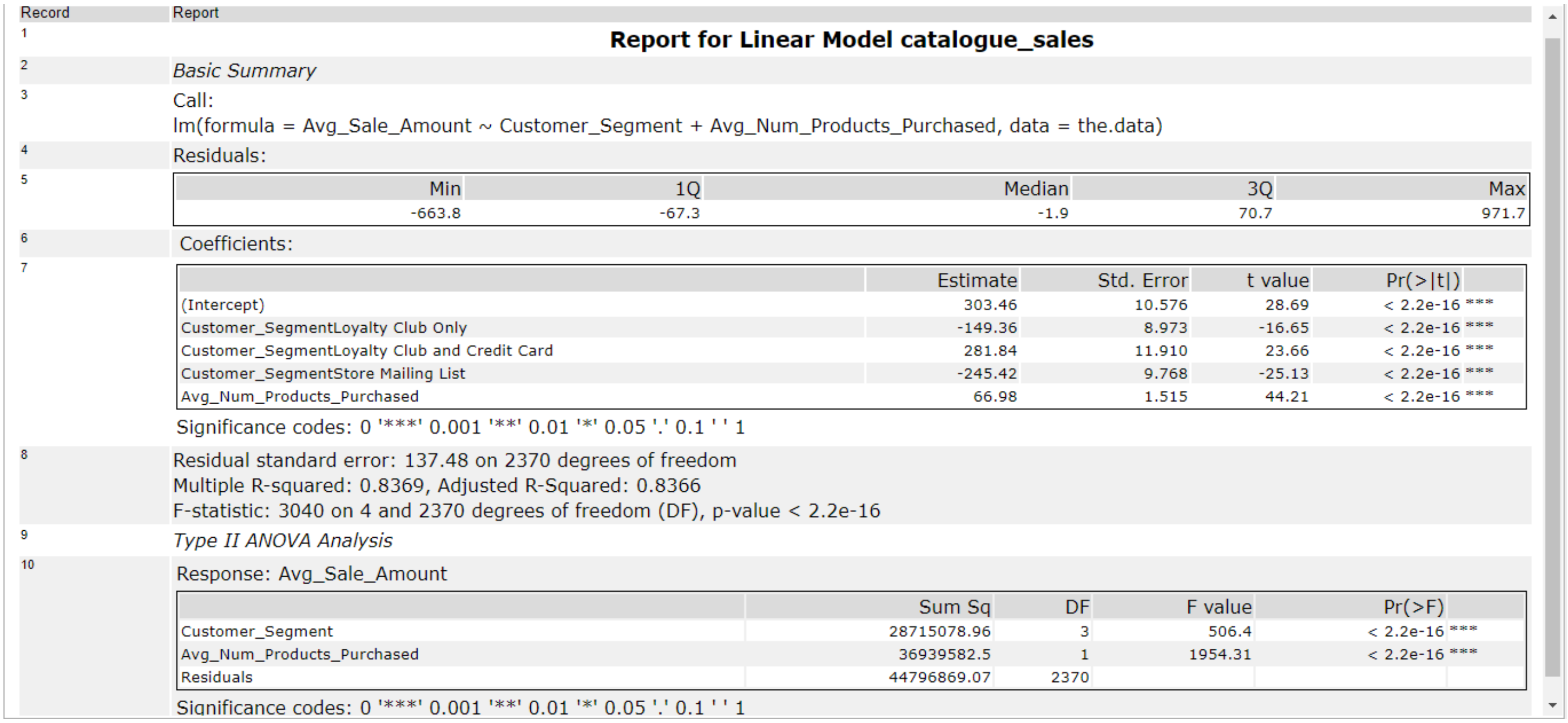


2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

Ans: I think my linear regression model is a good one because it has an R-squared value which is greater than 60%, to be precise it's R-squared and adjusted R-squared values are 83.69% and 83.66% respectively. This means 83% of the changes in our predicted variable

(Average sales amount) can be attributed to changes in the predictor variables that make up my linear regression model.

Each variable in my model is a good fit for the model because their p-values are all less than 0.05, this means the changes we observe in our predicted variable as a result of any of the predictor variables is statistically significant, that is the relationship between each of the predictor variables and the predicted variable is not due to chance.



Record Report

1 **Report for Linear Model catalogue_sales**

2 *Basic Summary*

3 Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 Coefficients:

7

	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

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)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

Ans:

Average_sale_amount

= 303.46 + 66.98*Avg_Num_Products_Purchased

– 149.36*Customer_Segment (Loyalty Club Only)

+ 281.84*Customer_Segment (Loyalty Club and Credit Card)

– 245.42*Customer_Segment (Store 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?

   ANS: Yes, the company should send out the catalogs to the 250 customers because the expected profit contribution exceeds the $10,000 threshold,

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

   ANS:
   Step 1: I built a linear regression model based on previous data, the model was trained to predict Average sale amount.

   Step 2: This model was used to predict the expected sale for each of our 250 new customers

   Step 3: Using the score_yes column, which gives us the probability of a customer actually buying our products, we got the expected revenue for each customer.

   Step 4: We summed up the expected revenue of all our 250 new customers, then multiplied our result by the average gross margin (50%) and subtracted the cost of printing and distributing all the catalogs, in order to get our expected profit. At the end of our procedure we had an expected profit which was larger than the threshold given by the management hence our recommendation above.

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

   ANS: Our expected profit is $21, 987