<u>Project 2: Predicting Catalog Demand</u>

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?

If total expected profit calculated from sending a catalog to 250 new customers is equal or greater than $10,000, then send them out.

2. What data is needed to inform those decisions?

2375 customers data consist of average sale amount, customer segment, average number of products purchased, year number as customer, etc.

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.

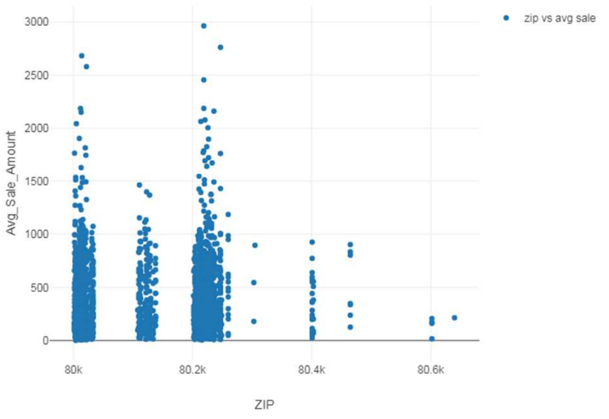
At the minimum, answer these questions:

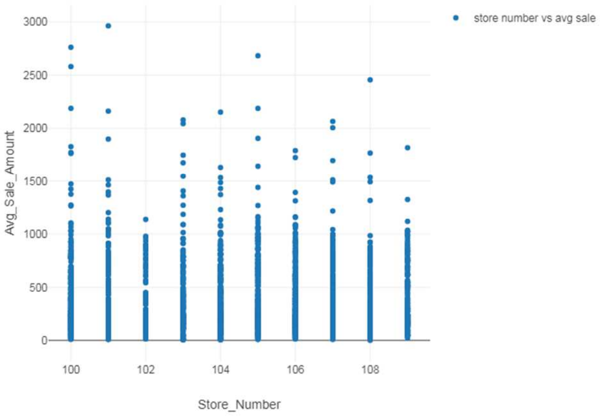
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

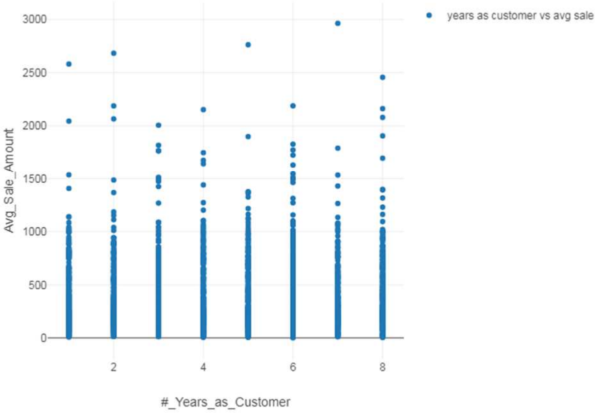
Customer_ID, Name, Address, City, State, and ZIP are not predictors since they are unique observations.

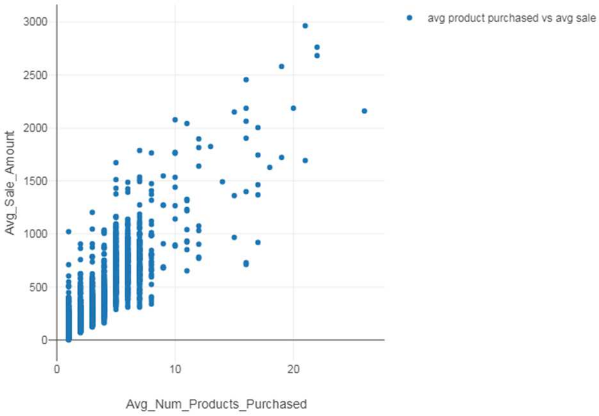
Store_Number and Responded_to_Last_Catalog are categorical predictors.

The continuous predictors are Average_Num_Products_Purchased and #_Years_as_Customer but only the former have linear relationship with Avg_Sale_Amount as target.

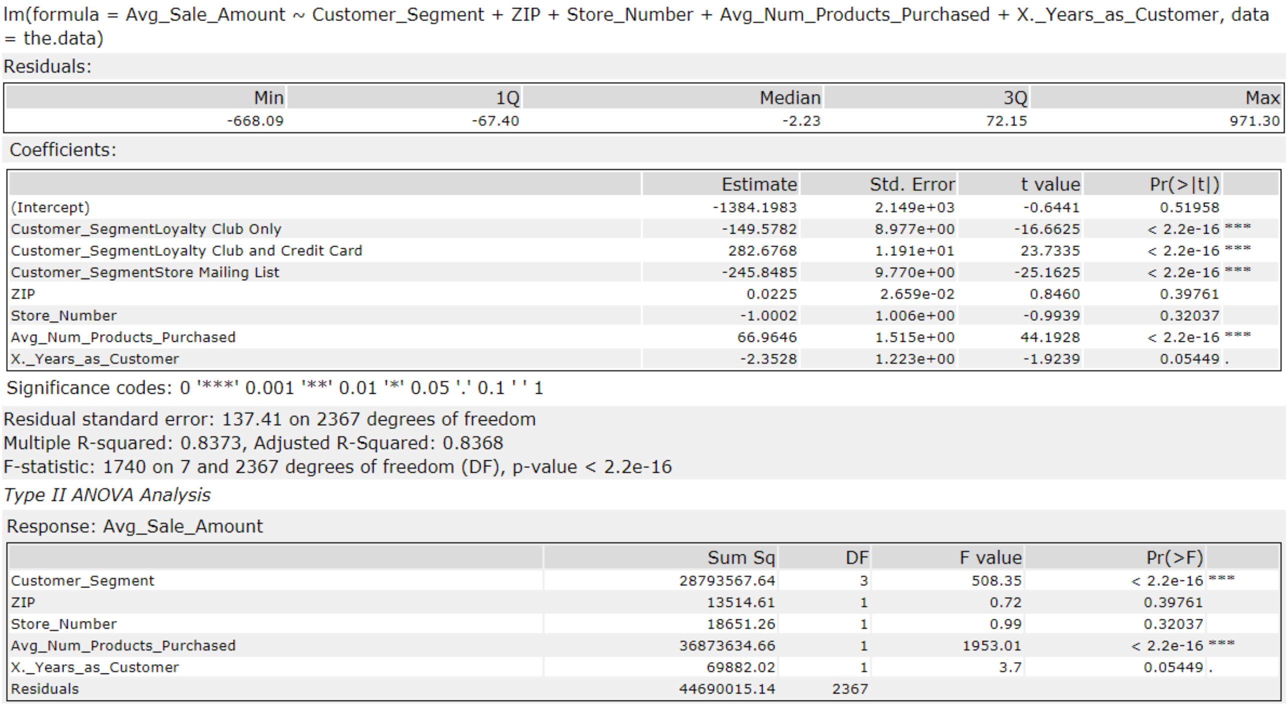








2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that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



lm(formula = Avg_Sale_Amount ~ Customer_Segment + ZIP + Store_Number + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-668.09	-67.40	-2.23	72.15	971.30

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958
Customer_SegmentLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.8485	9.770e+00	-25.1625	< 2.2e-16 ***
ZIP	0.0225	2.659e-02	0.8460	0.39761
Store_Number	-1.0002	1.006e+00	-0.9939	0.32037
Avg_Num_Products_Purchased	66.9646	1.515e+00	44.1928	< 2.2e-16 ***
X._Years_as_Customer	-2.3528	1.223e+00	-1.9239	0.05449 .

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

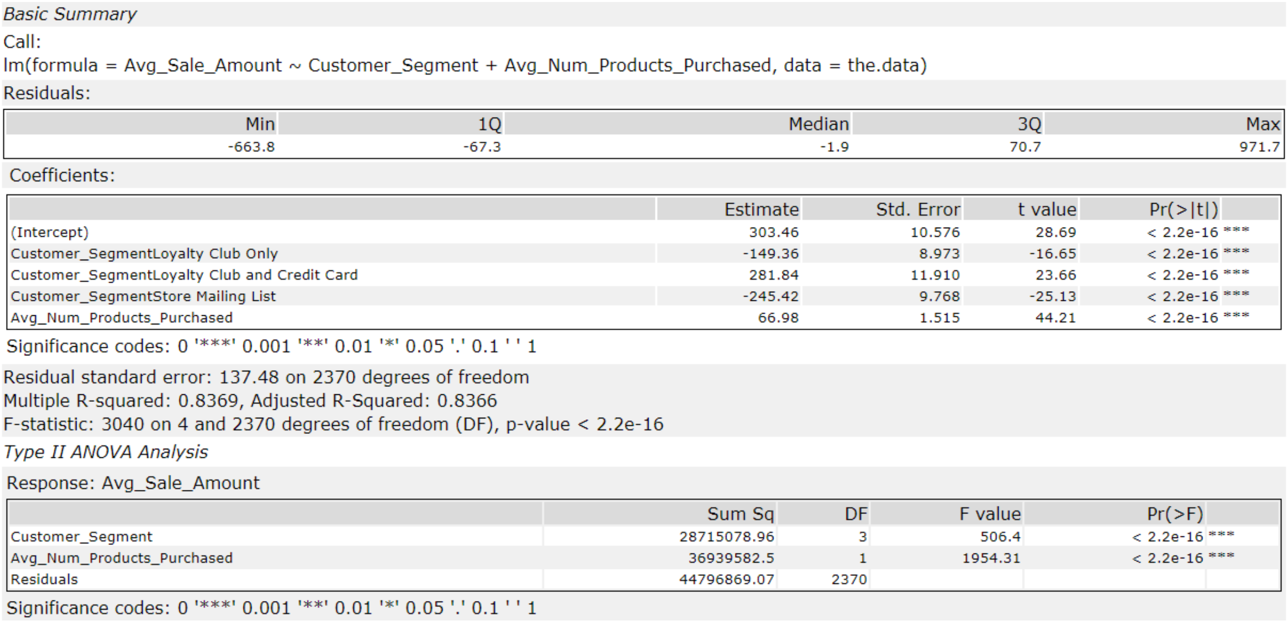
Residual standard error: 137.41 on 2367 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8368
F-statistic: 1740 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28793567.64	3	508.35	< 2.2e-16 ***
ZIP	13514.61	1	0.72	0.39761
Store_Number	18651.26	1	0.99	0.32037
Avg_Num_Products_Purchased	36873634.66	1	1953.01	< 2.2e-16 ***
X._Years_as_Customer	69882.02	1	3.7	0.05449 .
Residuals	44690015.14	2367		

ZIP, Store_Number and #_Years_as_Customer are not statistically significant predictors since their p-values > 0.05. From this first regression model, it is strong evidence that Customer_Segment and Average_Num_Products_Purchased are the statistically significant predictors.



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

Type II ANOVA Analysis

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

Residual standard error and adjusted R-Squared of this second model are not much different than that of first model but we are more confident with statistically significant predictors and intercept.

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 – 149.36 (If Customer_Segment: Loyalty Club) + 281.84 (If Customer_Segment: Loyalty Club and Credit Card) – 245.42 (If Customer_Segment: Store Mailing List) + 66.98 * Avg_Num_Products_Purchased

Important: The regression equation should be in the form:

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

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

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, they should.

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

The expected profit contribution from sending a catalog to them exceeds $10,000. It is calculated from multiply revenue by 50% of the gross margin first before subtract out the $6.50 cost.

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

$21,987.44