MAKE A COPY

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?

If sending out a catalog to the 250 new customer's results in a profit of more than $10,000 It will be done
If sending out a catalog to the 250 new customer's results in a profit of less than $10,000 It will NOT be done
What predictor variables to select

2. What data is needed to inform those decisions?

We are given 2 datasets. 1 has a list of attributes of 2375 customers that have ordered products from our catalogue in the past, including how much $ on average they have spent
The other is a list of attributes of our 250 new customers.

Need data from current customers: Avg Sales, Customer segment, Avg number products purchased

Need data from future customers: Will respond to the catalog and make a purchase, Score_Yes & Score_No.

We also need to know the margin (50%) and cost of the catalogs ($6.50)

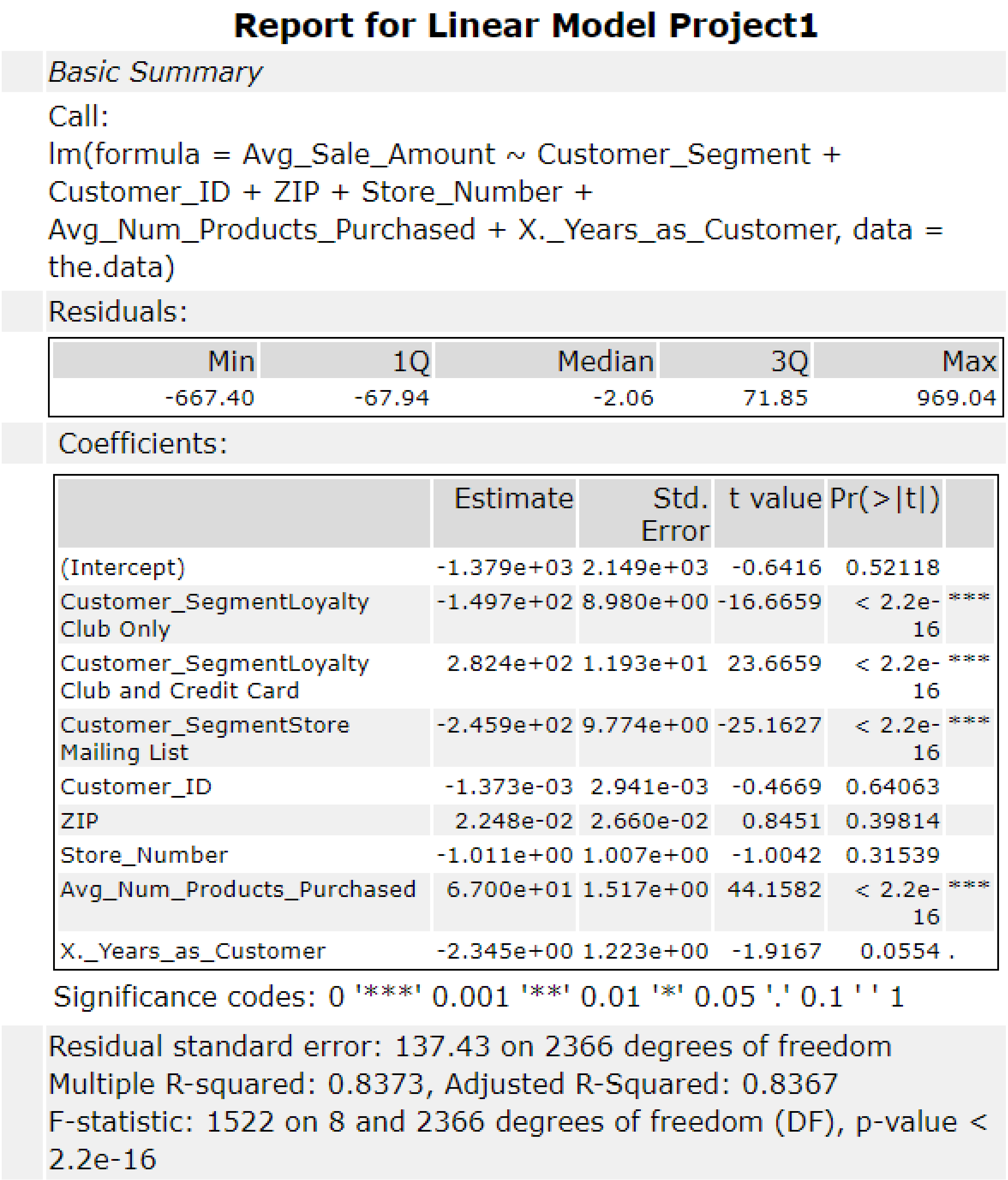
Step 2: Analysis, Modeling, and Validation

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.



Report for Linear Model Project1

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-667.40	-67.94	-2.06	71.85	969.04

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	.

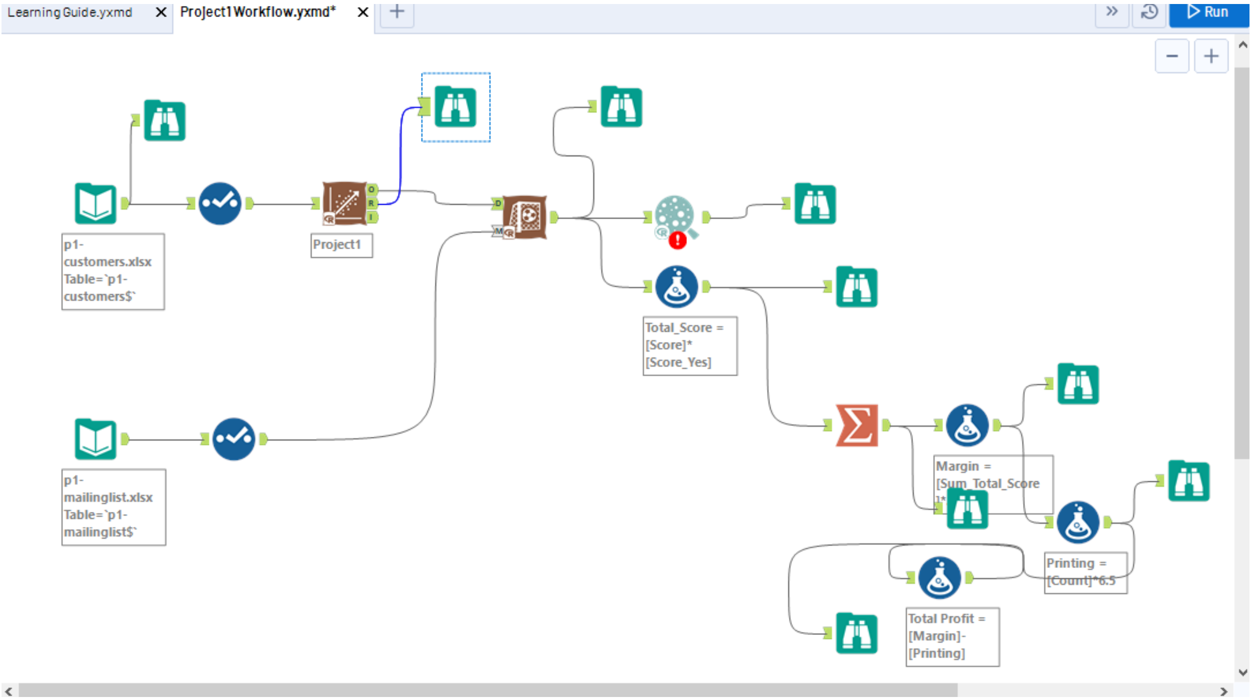
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

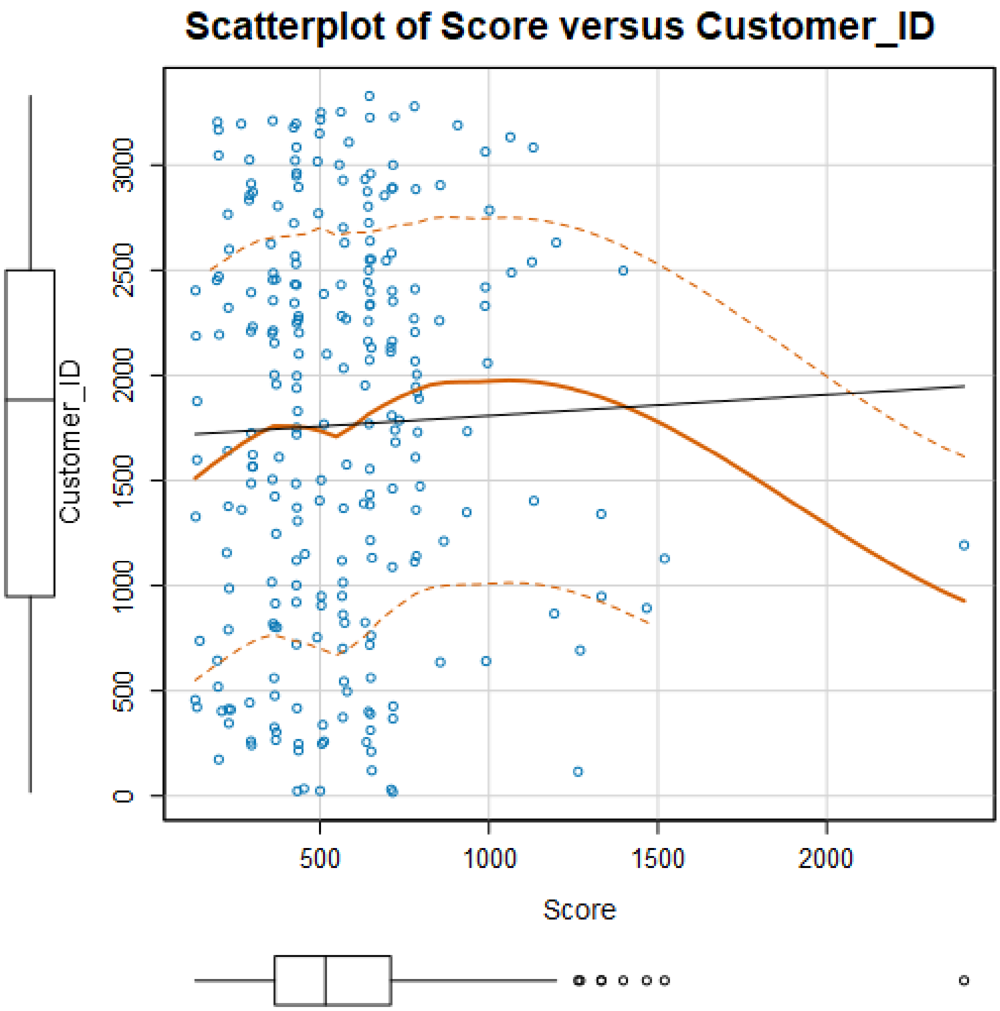
Residual standard error: 137.43 on 2366 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8367
F-statistic: 1522 on 8 and 2366 degrees of freedom (DF), p-value < 2.2e-16

2.

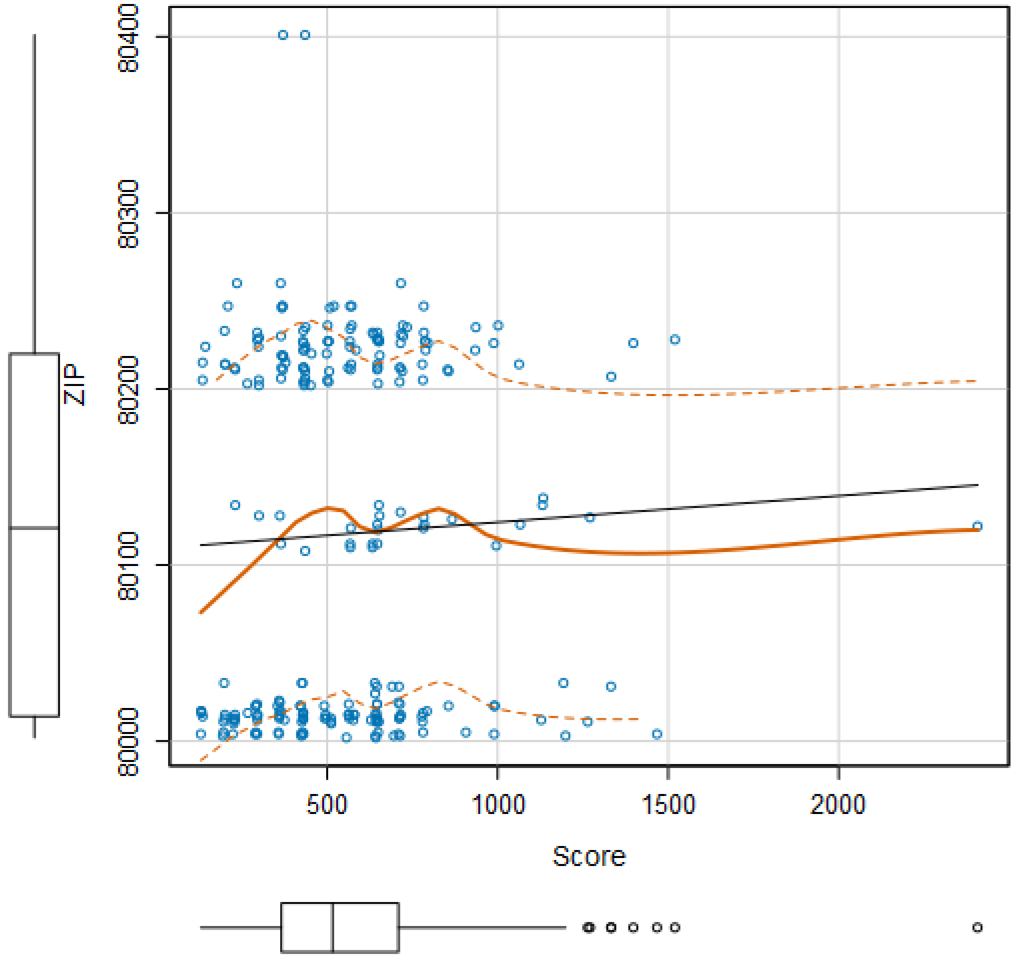
3. Need to only select values with Low P-values (less than .05) and a high R-squared
4. Selected Customer_segment (all 3 categories) & Avg_Num_Products_Purchased (P-value of .00000000000000022) all have Low P-values and a high R-squared
5. Customer_ID has a P value of .64063 which is higher than .05
6. ZIP has a P value of .39814 which is higher than .05
7. Store_Number has a P value of .31539 which is higher than .05
8. X._Years_as_Customer has a P value of .0.554 which is slightly higher than .05

9.

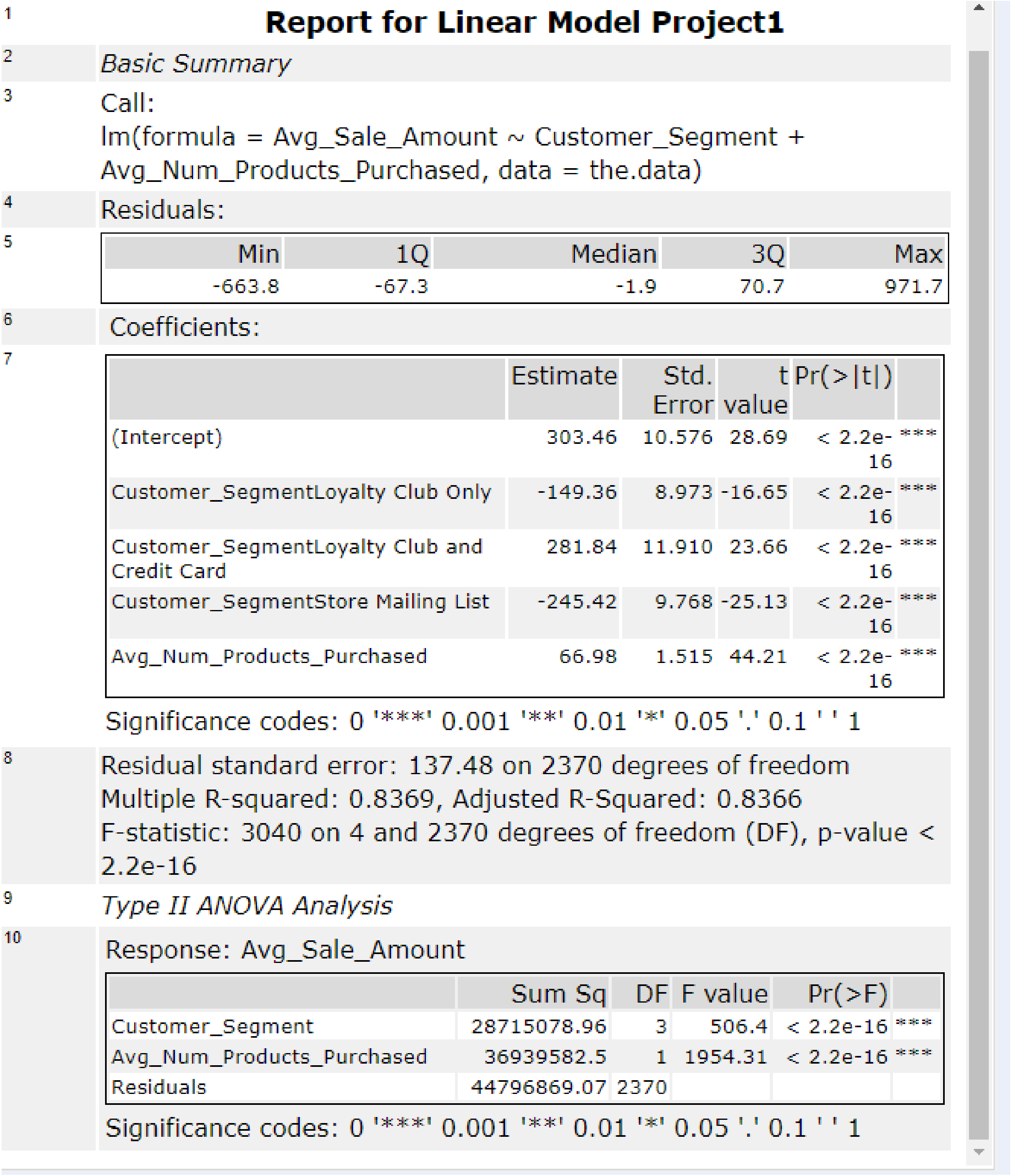

LearningGuide.yxmd
Project1Workflow.yxmd*
Run
p1-customers.xlsx Table=`p1-customers$`
Project1
p1-mailinglist.xlsx Table=`p1-mailinglist$`
Total_Score = [Score]*[Score_Yes]
Margin = [Sum_Total_Score]*
Printing = [Count]*6.5
Total Profit = [Margin]-[Printing]


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

Scatterplot of Score versus ZIP



10. 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.



Report for Linear Model Project1

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

The p-value is the probability that the coefficient is zero. The lower the p-value, the higher the probability that a relationship exists between the predictor and target variable. If the p-value is high, we should not rely on the coefficient estimate. When a predictor variable has a p-value below 0.05, the relationship between it and the target variable is considered to be statistically significant.

R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. The higher the r-squared, the higher the explanatory power of the model.

This Linear model is good because it has low P-values and a high R-squared values.

Customer_segment (all 3 categories) & Avg_Num_Products_Purchased all have a P-value of .00000000000000022 or 2.2^-16
Multiple R-squared=.8369 and Adjusted R-squared=.8366

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……*

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

Note that we **must** include the 0 coefficient for the type Cash.

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.

Avg_Sale_Amount=303.46 - 149.36 * Customer_SegmentLoyalty Club Only + 281.84 * Customer_SegmentLoyalty Club and Credit Card – 245.42 * Customer_SegmentStore Mailing List + Credit Card Only * 0 + 66.98 * 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?
   Yes, sending the catalogs to the 250 new customers will result in $21,987.44 Profit

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   a. Used Linear Regression to figure out formula above
   b. Applied formula to list to mailinglist.xls, using score (138,292.13)
   c. Multiplied Avg Sale Amt for potential cust by Score_Yes (47,224.87)
   d. Summed up and multiplied by .5 (23,612.44)
   e. Subtracted (250*6.5=1625) (21,987.44)
   f. Total expected profit= $21,987.44

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   a. Total expected profit= $21,987.44

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.