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?

   Ans: Should we send our catalog out to our new customers or not.

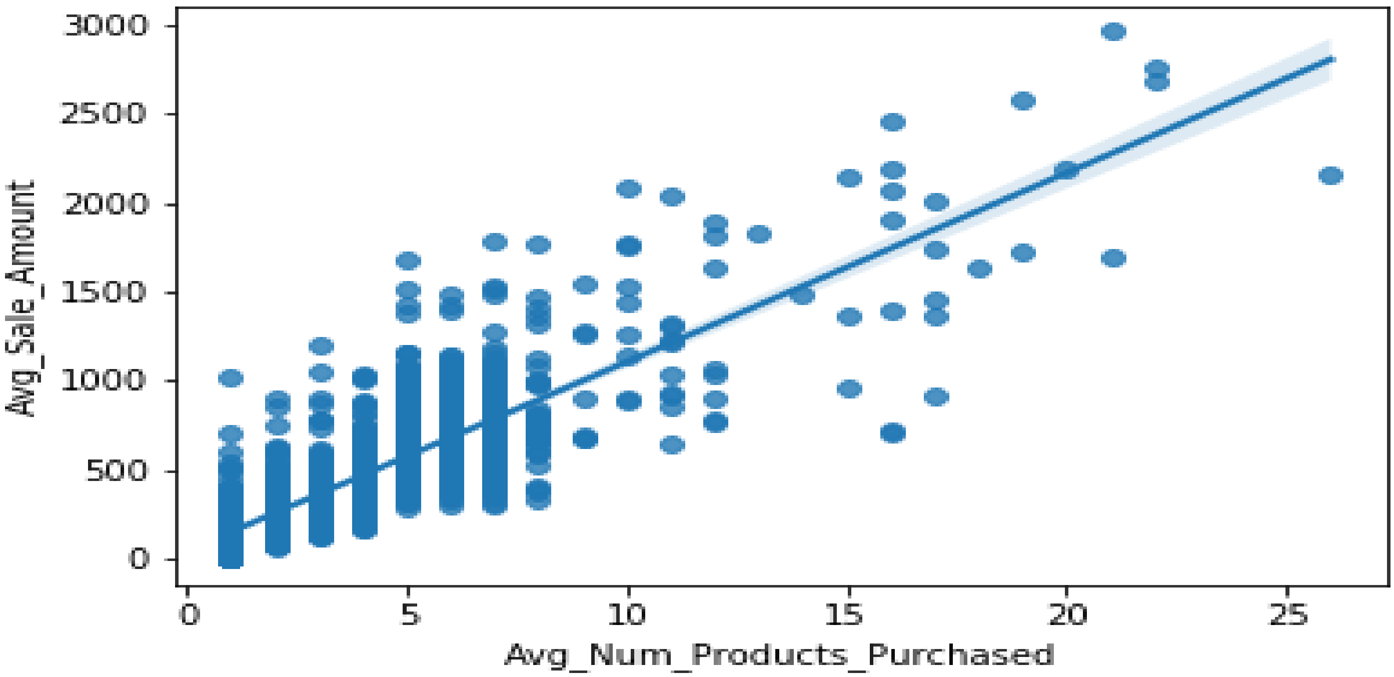
2. What data is needed to inform those decisions?

   Ans: The expected profit from selling our catalog to the new customers.

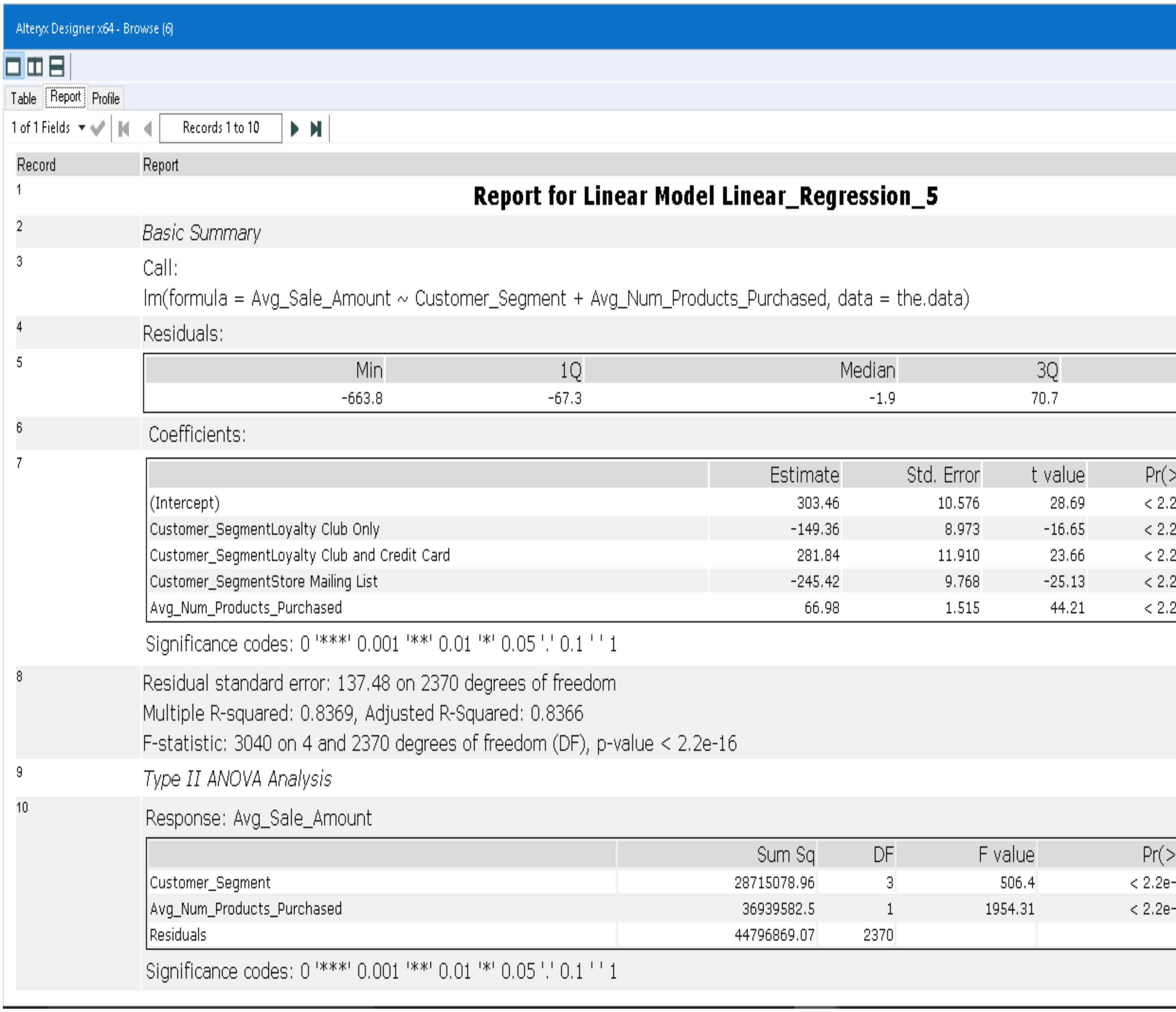
Step 2: Analysis, Modeling, and Validation

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)

Ans: First I imported the input data (“Customers.xlsx) and explored the data to ensure proper data formatting and check for missing values. Next I used scatter plots to check for relationships between the target variable (Avg_Sale_Amount) and the numeric predictor variables.



After selecting the relevant predictor variables using the p-value and the scatter plot diagrams, I used different categorical variables for modeling and settled for the one(s) with p-value lower than 0.05 (Customer_Segment).



Record	Report
1	**Report for Linear Model Linear_Regression_5**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	Residuals:
5	
6	Coefficients:
7	
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

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

	Estimate	Std. Error	t value	Pr(>
(Intercept)	303.46	10.576	28.69	< 2.2
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 2.2

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

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

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

The resulting model has Adjusted R-squared value of 0.8366 hence this is a good model as it shows that over 80% of variance in the target variable is explained by the selected predictor variables.
The linear regression equation of the model is 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) + 0*(Customer_SegmentCredit Card Only) + 66.98*(Avg_Num_Products_Purchased)

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.

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.

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.

Step 3: Presentation/Visualization

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

With the results from my analysis, I recommend that the company should go ahead and send the catalog to the new customers as expected profit is projected to be over $20,000 ($21987.44) which is above the $10000 benchmark.
This figure was gotten by using the validated model to make predictions on the mailing list dataset to obtain the Avg_Sale_Amounts. Next I multiplied the score obtained with the Score_Yes variable to obtain the expected Revenue. Then I obtained the profit by multiplying the sum of revenue ($47224.87) by the average gross margin (0.5) and subtracting the total cost of all the catalogs ($6.50 *250) and I obtained a profit of $21987.44.

At the minimum, answer these questions:

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

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

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

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.