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?**

   The decision that needs to be made is whether to send the catalog to 250 customers based on the calculated profit or not.

2. **What data is needed to inform those decisions?**

   We are given 2 files of dataset i.e. customers.xlxs and mailing.xlsx. We need Avg_Num_Products_Purchased, Customer Segment, Score_Yes, Mailing, cost of catalogue ($6.50) and gross_margin(50%) to find the 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)

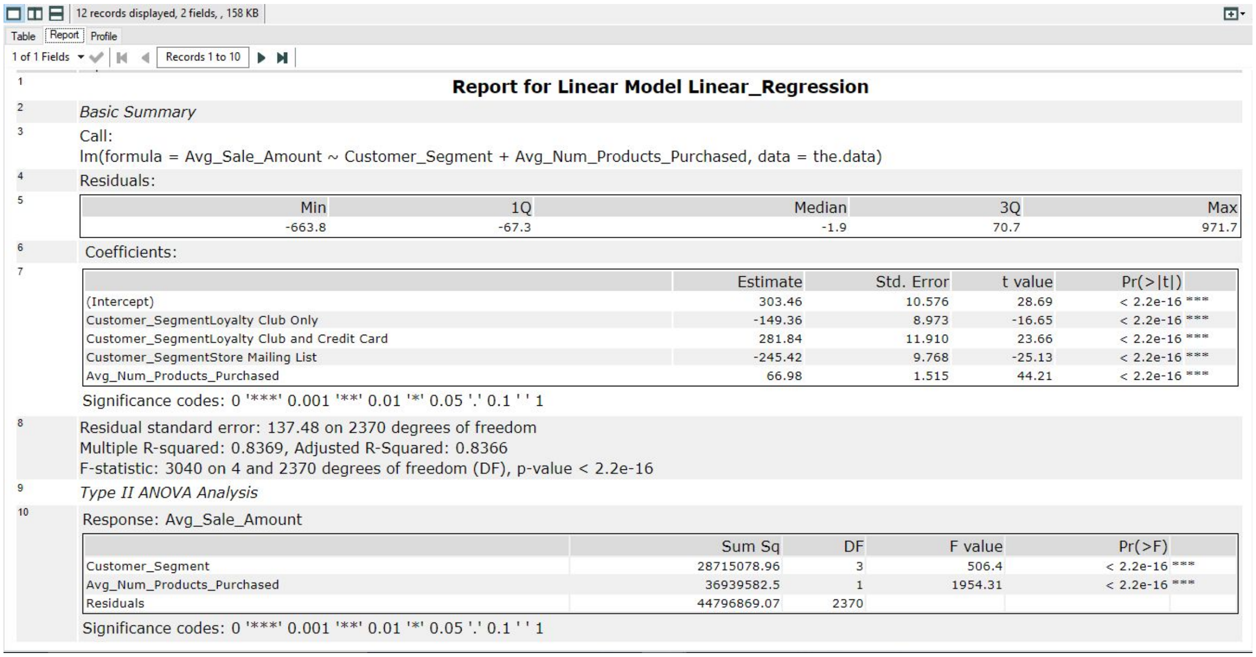
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

1. How and why did you select the predictor variables in your model?

   The target variable for the analysis is Avg_Sale_Amount and the predictor variables selected for the model are Customer_Segments and Avg_Num_Products_Purchased because only these two variables have the p-value less than 0.05 which shows that these two variables are statistically significant.

   The scatterplots of Customer_Segments Vs Avg_Sale_Amount and Avg_Num_Products_Purchased Vs Avg_Sale_Amount is shown below:



1 **Report for Linear Model Linear_Regression**

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

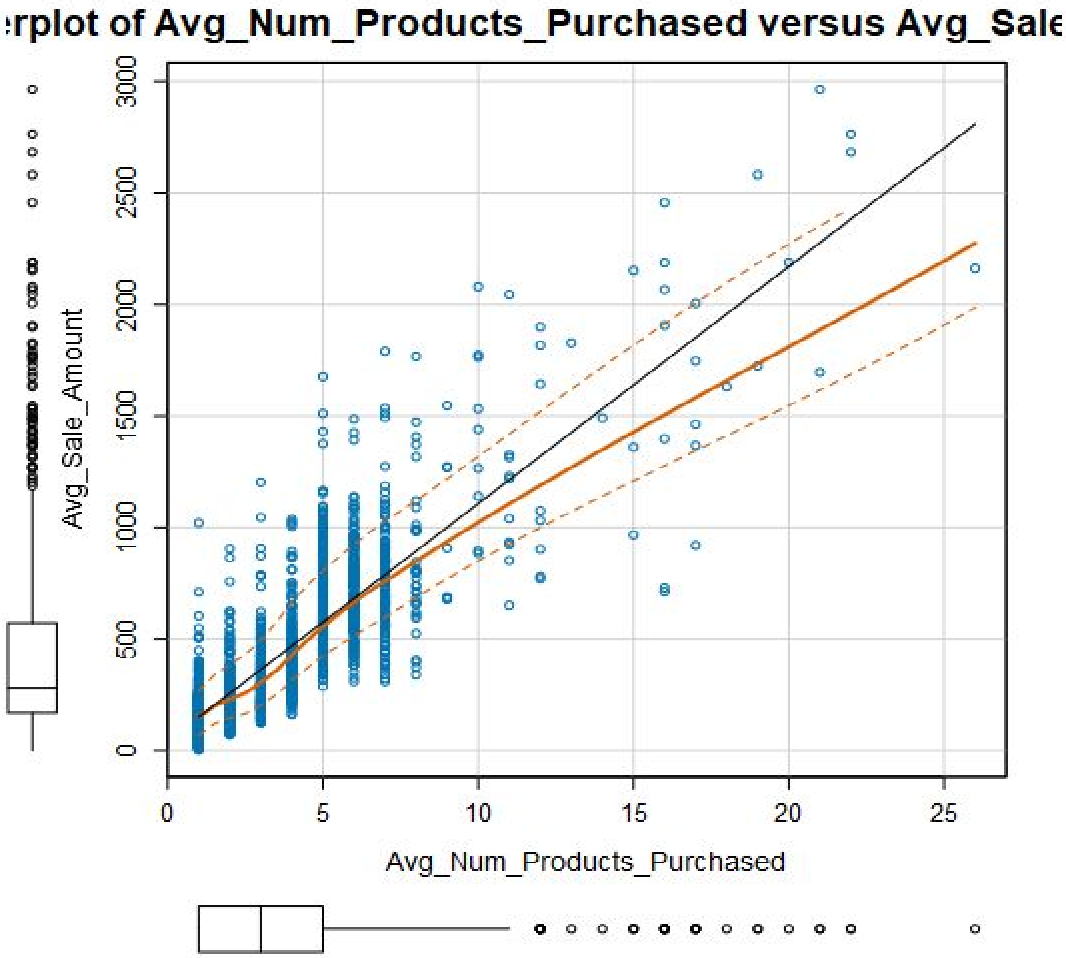
8 Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

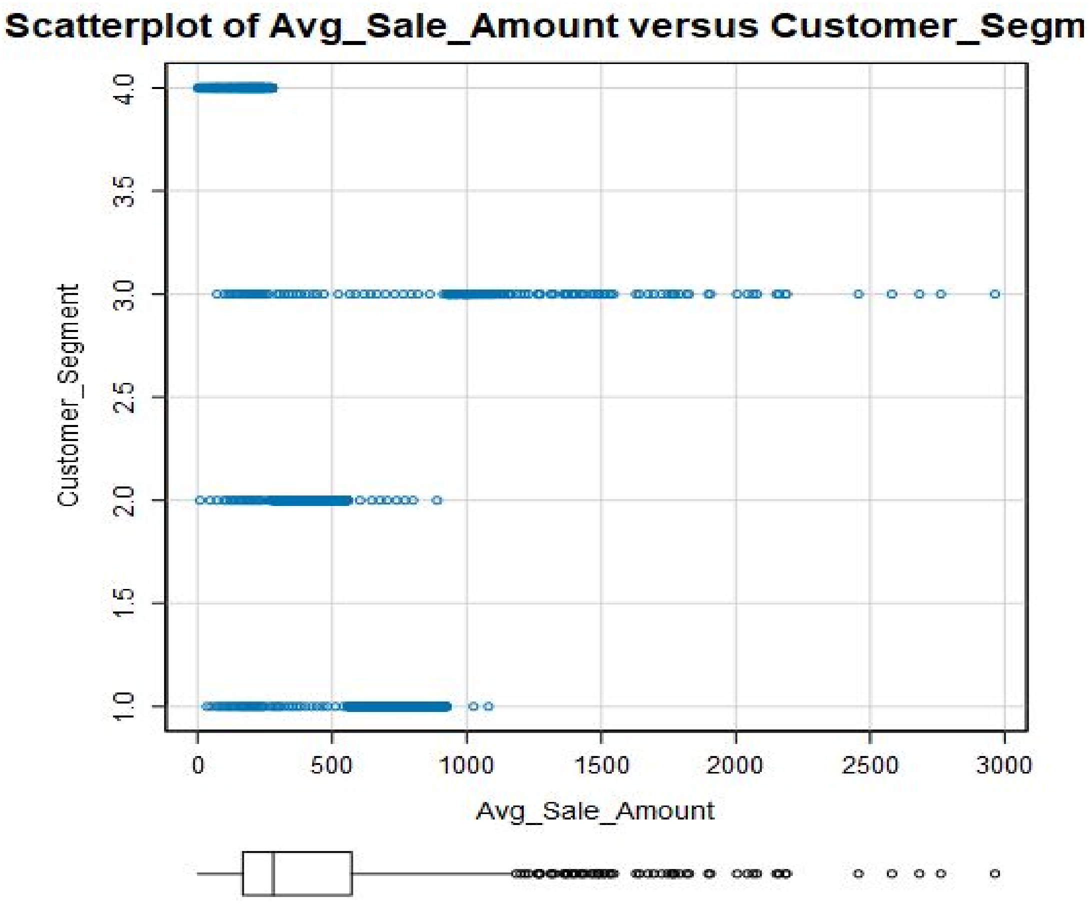
9 *Type II ANOVA Analysis*

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		

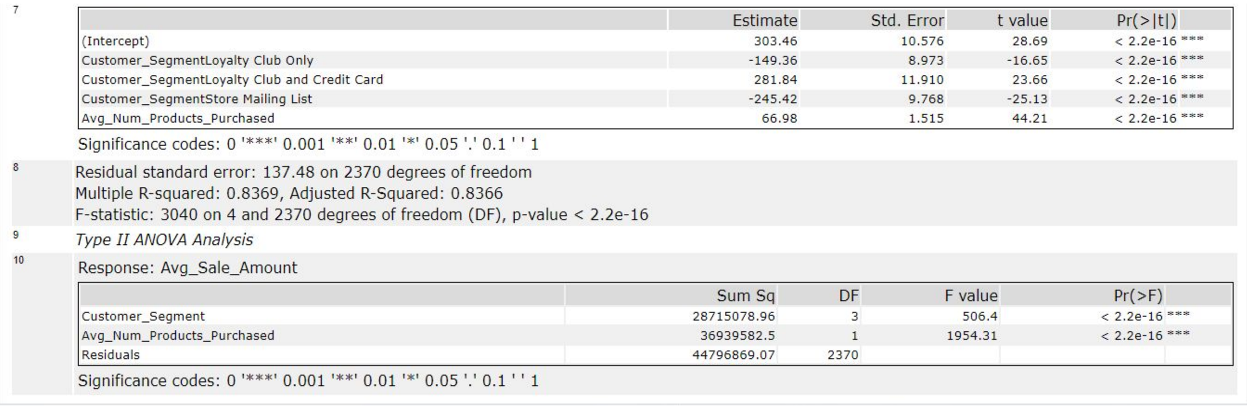
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1





2. Explain why you believe your linear model is a good model.

As shown above in the table that 2 variables namely Customer_Segment and Avg_Num_Products_Purchased have p-values less than ***0.05*** *and the Adjusted R Squared value is* ***0.8366*** *which is quite a large value. This implies that our model is a good model because p-values and R-Squared value is statistically significant.*



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)

> Avg_Sale_Amount = 303.46 - 149.36 * (Customer_Segment : Loyalty Club Only) + 281.84 * (Customer Segment : Loyalty Club and Credit Card) - 245.52 * (Customer Segment : Store Mailing List) + 66.984 * (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, the company should send these catalogues to these 250 customers.*

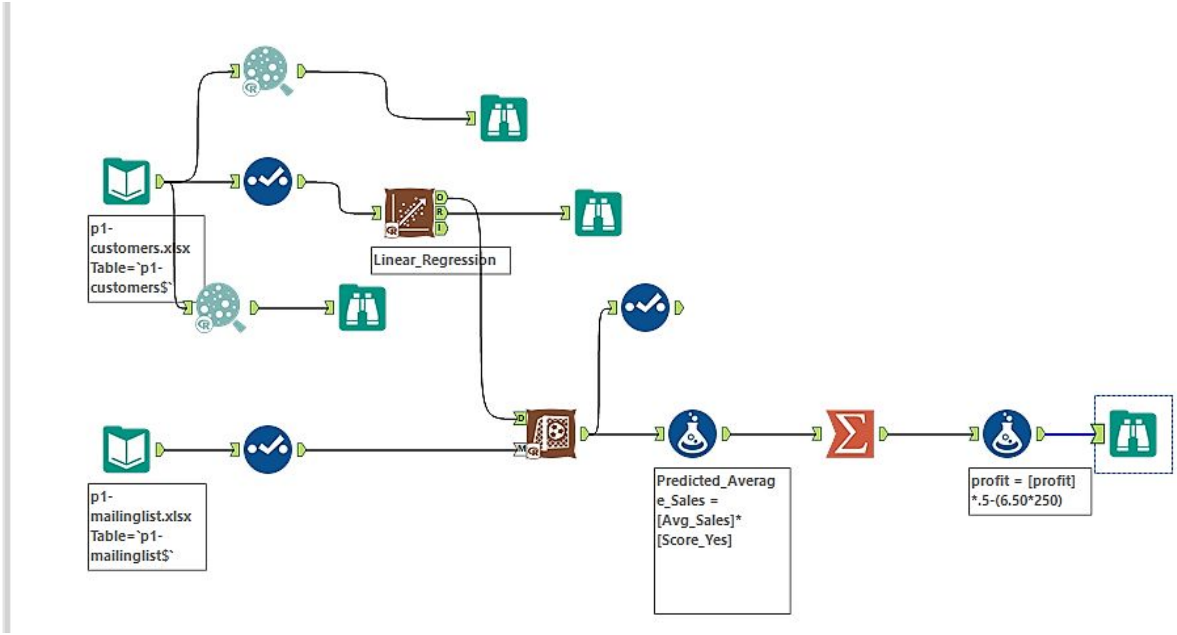
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

> *Firstly I calculated Avg_Sales using the linear regression model. Then I created a new column Predicted_Average_Sales = Avg_Sales * Score_Yes. Then the profit is calculated with the given margin to be 50% and cost of each catalogue as $6.50, for all the 250 customers.*

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

> *Profit = (profit * 0.5) - (Cost of catalog * 250)*
>
> *= $ 21, 987.43587*

Alteryx Workflow:



p1-
customers.xlsx
Table=`p1-
customers$`

Linear_Regression

p1-
mailinglist.xlsx
Table=`p1-
mailinglist$`

Predicted_Average_Sales =
[Avg_Sales]*
[Score_Yes]

profit = [profit]
*.5-(6.50*250)