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 to be made is whether to send the catalog to these 250 new customers based on expected profit calculated.

2. **What data is needed to inform those decisions?**
   We are given two dataset files (customers.xlxs and mailing.xlsx).
   We need to predict sales and calculate expected profit are Customer Segment, Average Number of Product Purchased, _ScoreYes, Margin and Cost of Catalog.

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?**
   A linear regression study is performed the target variable for Avg_Sale_Amount and the predictor variables are Customer_Segments and Avg_Num_Products_Purchased theses two variables have the p-value less than 0.05 which implies statistical significance. Figure 1 show the report for the linear regression

Report for Linear Model Linear_Regression

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

Figure 1: Linear Regression Report.

The scatterplots below of Customer_Segments Vs Avg_Sale_Amount and Avg_Num_Products_Purchased Vs Avg_Sale_Amount

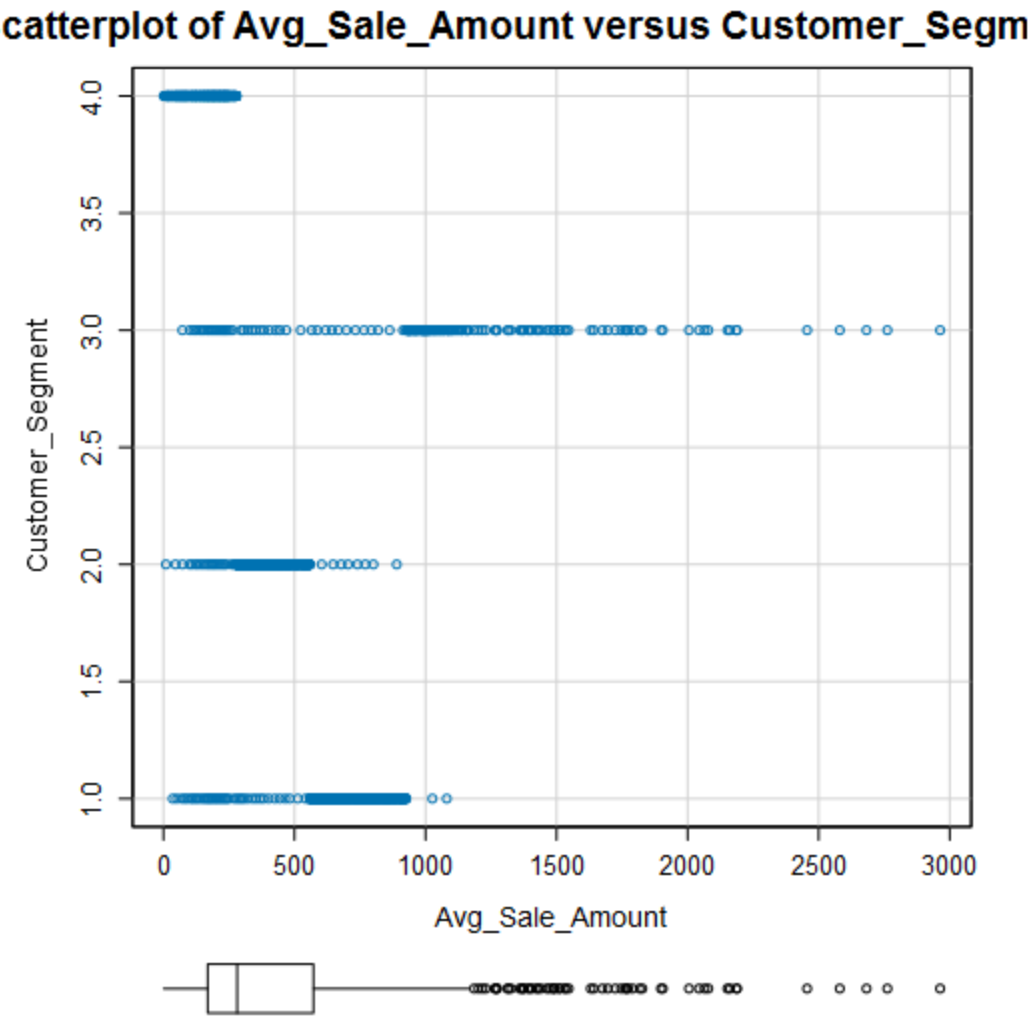

Figure2: Scatterplots of Customer_Segments Vs Avg_Sale_Amount

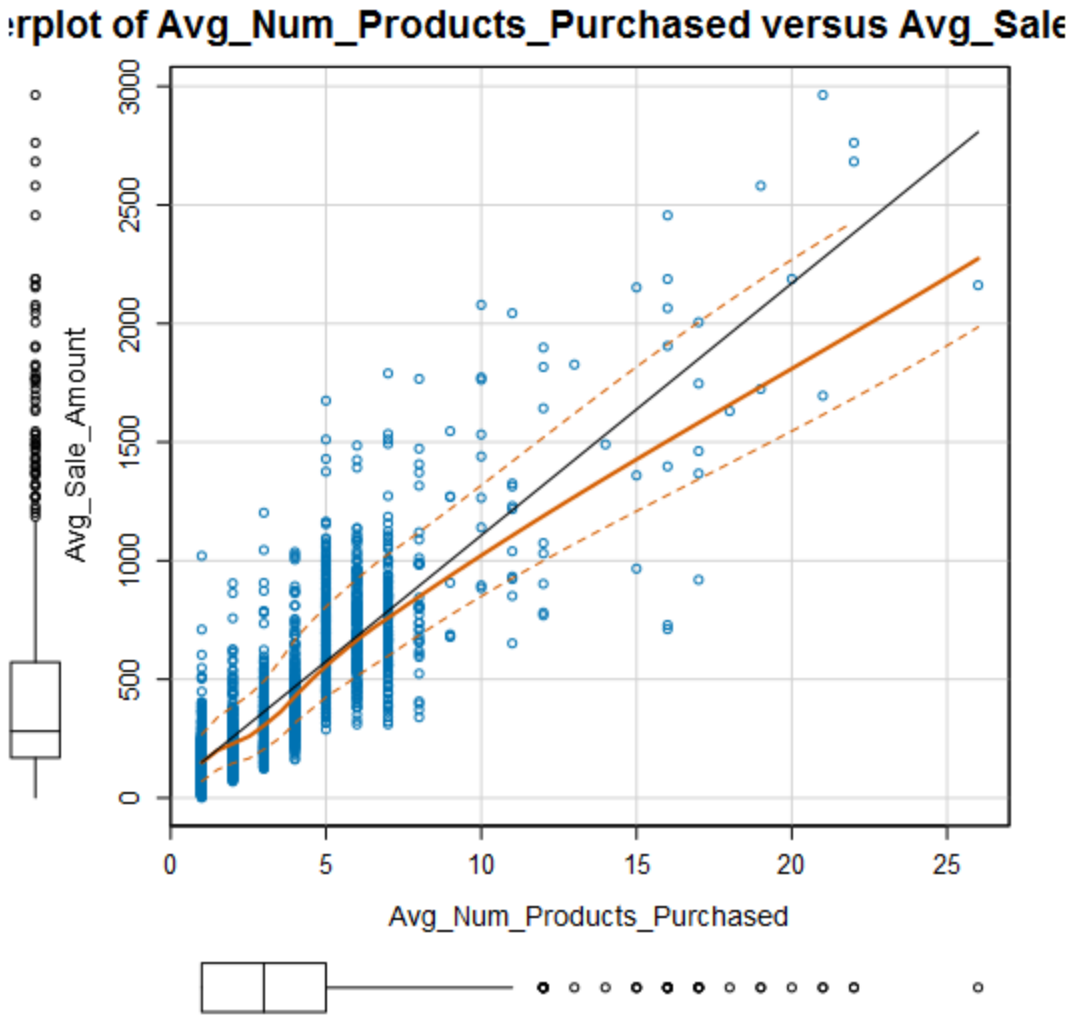

Figure3: Scatterplots of Avg_Num_Products_Purchased Vs Avg_Sale_Amount

2. **Explain why you believe your linear model is a good model.**
   As shown below in the table that have two variables 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 a large value .This implies that our model is a good model because p-value and R-Squared value is statistically significant.

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

Figure4: P-value and R-Squared value.

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 Only and Credit Card)-245.52*(Customer Segment:Store Mailing List)+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,The company should send the catalog to these 250 new customers.

2. **How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)**
   The expected revenue from each customer is determined by multiplying expected sale amount with Score_Yes value. With a gross margin of 50%, is deducted from the sum of expected revenue before the cost of catalog ($6.50) is subtracted to obtain net profit .

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

Expected Profit = (Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)
= (47,225.87 x 0.5) – (6.50 x 250)
= $21,987.44

Alteryx Workflow