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 task is to:
   - Predict the profit that the company can expect from sending a catalog to its new customers
   - Whether to send the catalog to its new customer depending on the profit (expected profit should be greater than $10,000)

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

   The data required to make decision are:
   - Average number of products purchased
   - Average sale amount
   - Cost of printing and distributing
   - Profit
   - The probability that the customer will respond to the catalog and make a purchase

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:

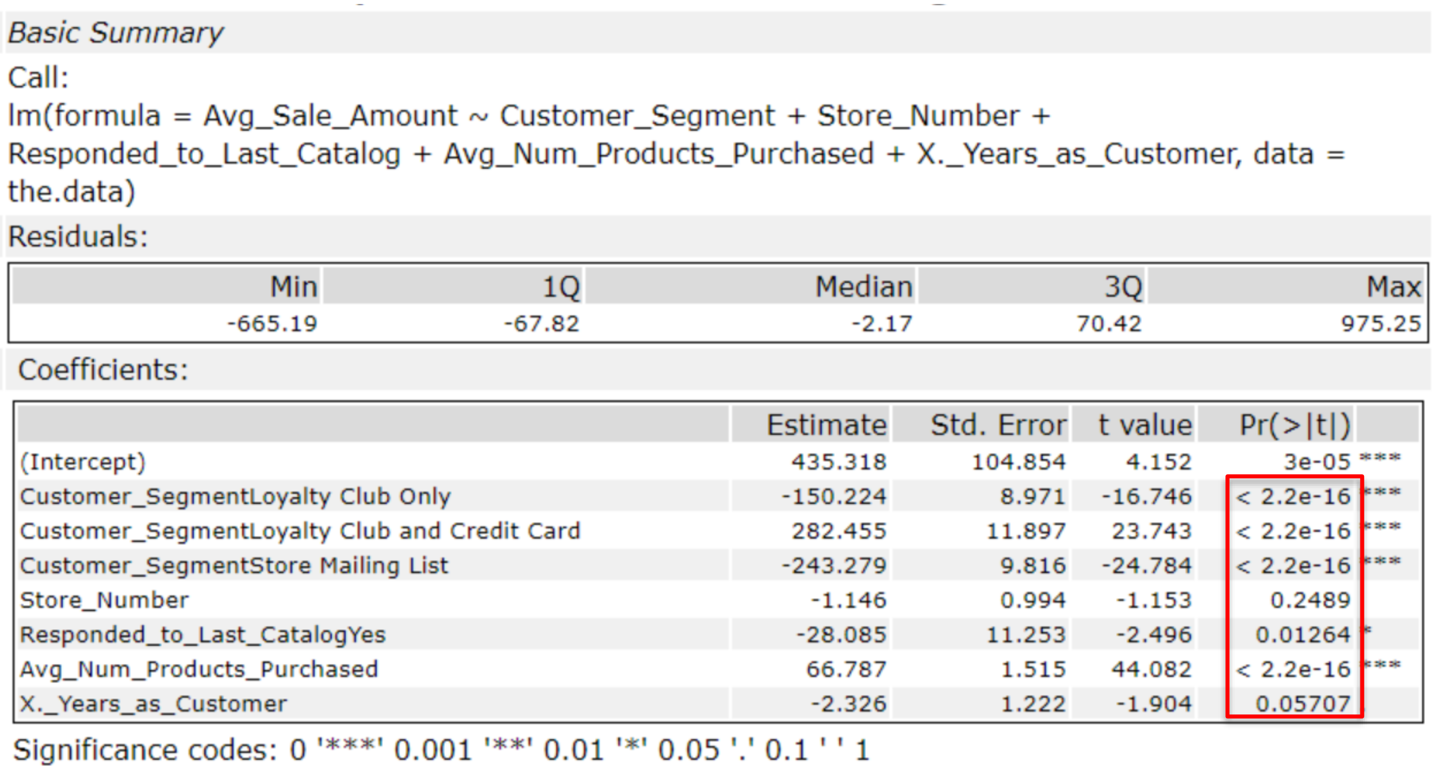
1. **How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and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

   To identify most significant predictor variables:
   - Perform linear regression on all the predictor variable in our dataset against the target variable
   - Keep only those predictor variables having p-value less than 0.05

Below screenshot represent the output of linear regression. Only customer segment and average number of product purchased has p-value less than 0.05

Hence, our final model will be:

Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased



Basic Summary

Call:

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

Residuals:

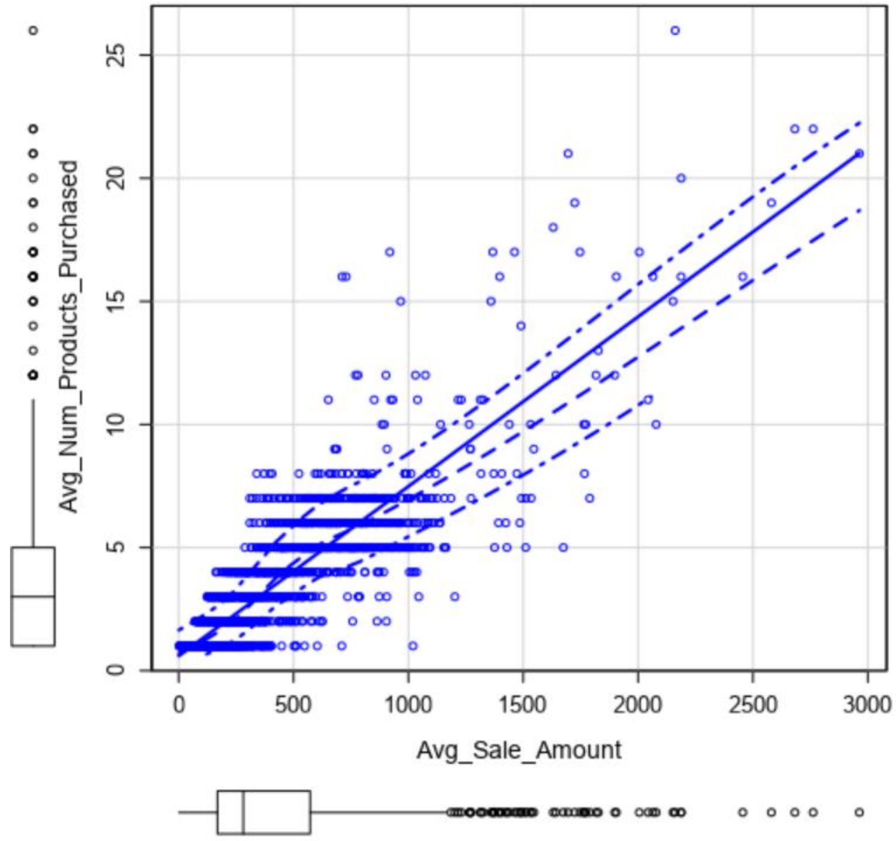
Min	1Q	Median	3Q	Max
-665.19	-67.82	-2.17	70.42	975.25

Coefficients:

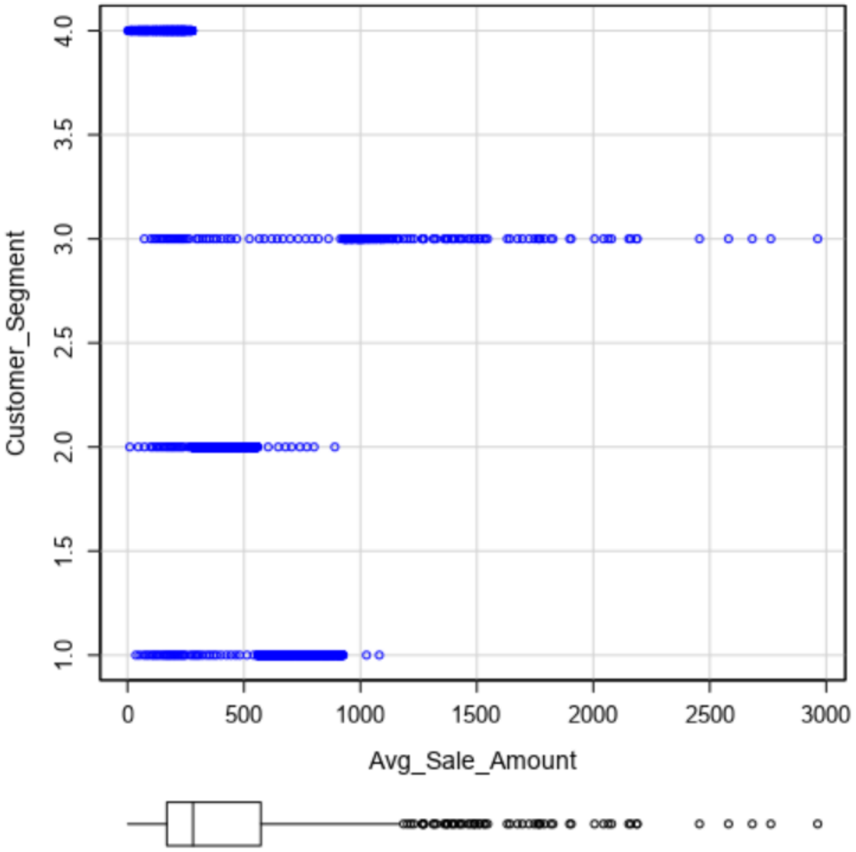
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	435.318	104.854	4.152	3e-05	***
Customer_SegmentLoyalty Club Only	-150.224	8.971	-16.746	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.455	11.897	23.743	< 2.2e-16	***
Customer_SegmentStore Mailing List	-243.279	9.816	-24.784	< 2.2e-16	***
Store_Number	-1.146	0.994	-1.153	0.2489	
Responded_to_Last_CatalogYes	-28.085	11.253	-2.496	0.01264	*
Avg_Num_Products_Purchased	66.787	1.515	44.082	< 2.2e-16	***
X._Years_as_Customer	-2.326	1.222	-1.904	0.05707	

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

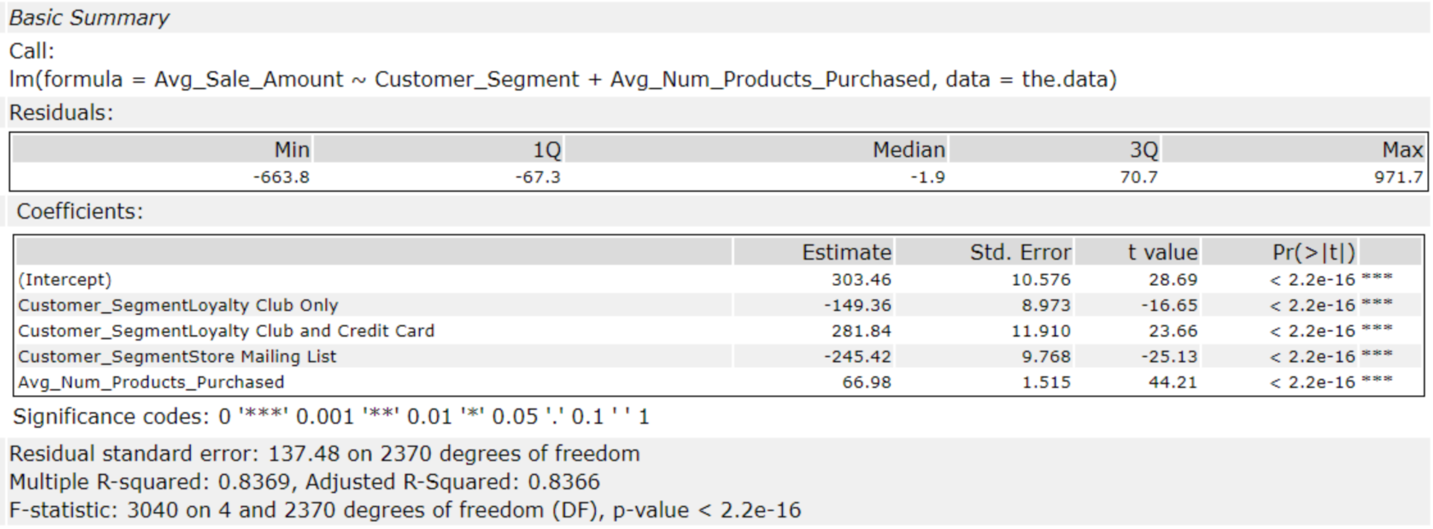
Scatterplot of Avg_Sale_Amount versus Avg_num_Products_Purchased



Scatterplot of Avg_Sale_Amount versus Customer_Segment



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.**



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

I believe that my linear model is a good model because:

- The R-Square value is 0.8369 (close to 1)
- The p-value for all the variable is less than 0.05, indicating significance

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)**

   The linear equation is:
   Avg_Sale_Amount = 303.46 - 149.36 * (If Type: Loyalty Club Only) + 281.84 * (If Type: Loyalty Club and Credit Card) – 245.42 x (If Type: Store Mailing List) + 0 * (If Type: Credit Card Only) + 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 customers since the profit exceeds $10000

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

- The linear regression model will give us three columns: Score_No, Score_Yes, and Average_Predicted_Sales
- Calculate Expected_Revenue = Avg_Predicted_Sales * Score_Yes
- Profit margin = 50%, and cost for each catalog = $6.50, hence for all 250 customers
  Profit = Price – Cost
  = (Sum of Expected_Revenue * 0.5) - (6.50 * 250)

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 * Gross Margin) – (Cost of Catalog * 250)
   = (47,224.87 * 0.5) – (6.50 x 250)
   = $21,987.43