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 company has a list of 250 new customers and would like to understand if sending out catalog (with sending cost of $ 6.50 per customer) to them will get expected profit contribution exceeding $10,000. If the profit does not exceed the amount, the company will drop this attempt.

2. What data is needed to inform those decisions?

We need 1) the probabilities of a customer will purchase the item after receiving a catalog; 2) the costs of printing and distributing, which is $6.50 per catalog; 3) the average gross margin (price - cost) on all products sold through the catalog is 50%. If we can estimate the successful purchase revenue through sending catalog, we can then deduct the costing of the catalog sending to all customers on the list, as well as deducting the 50% of the product price to get the overall profit of sending out catalog to the new list of customers.

Other variables important for this analysis include customer data such as Customer_Segment and Avg_Num_Products_Purchased of a customer.

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.

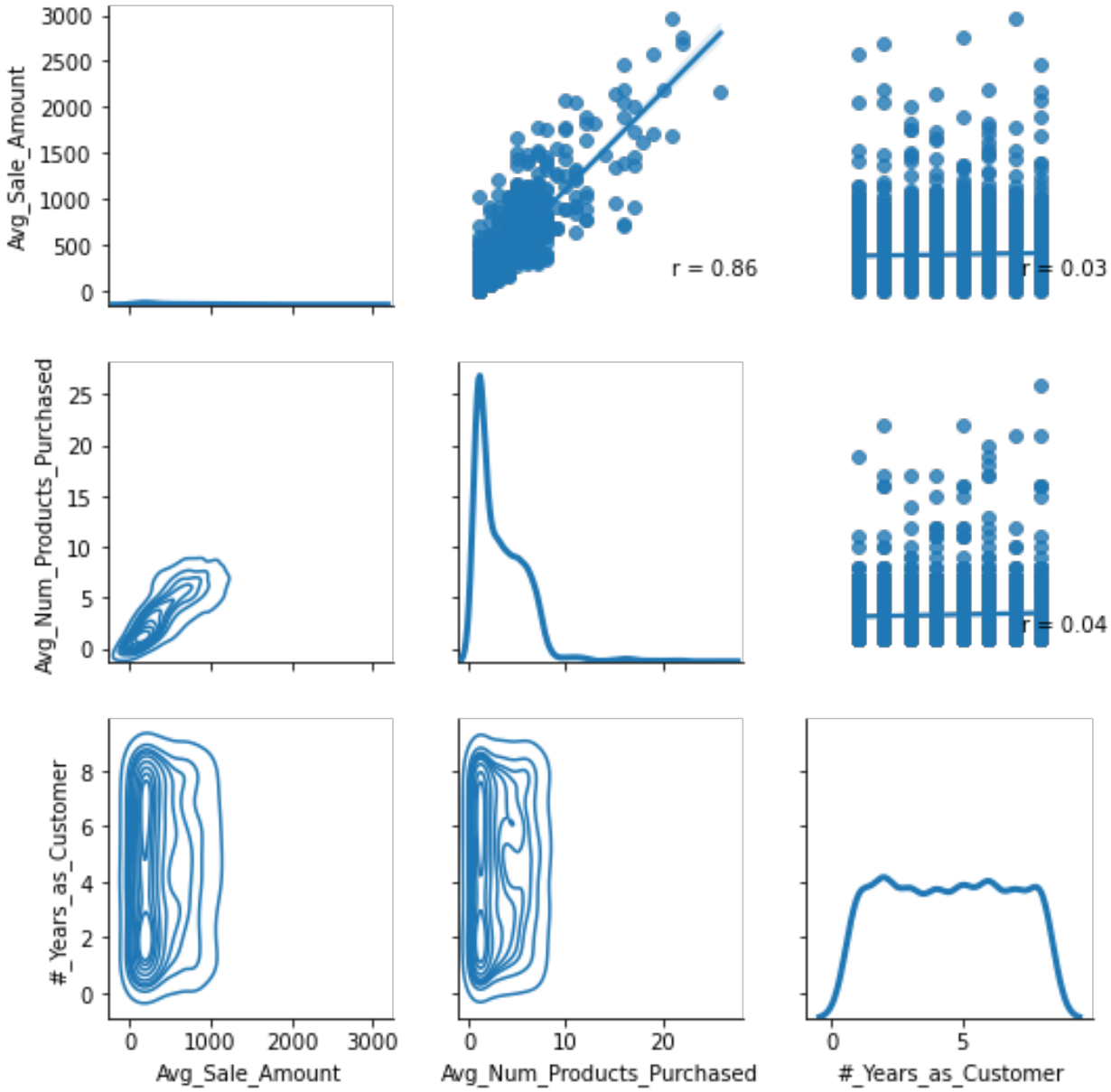

Figure 1. Pairplots included scatterplots with linear regression and R (Upper), histograms (diagonal), Kernel Density Estimates (Lower) for all continuous independent and dependent variables.

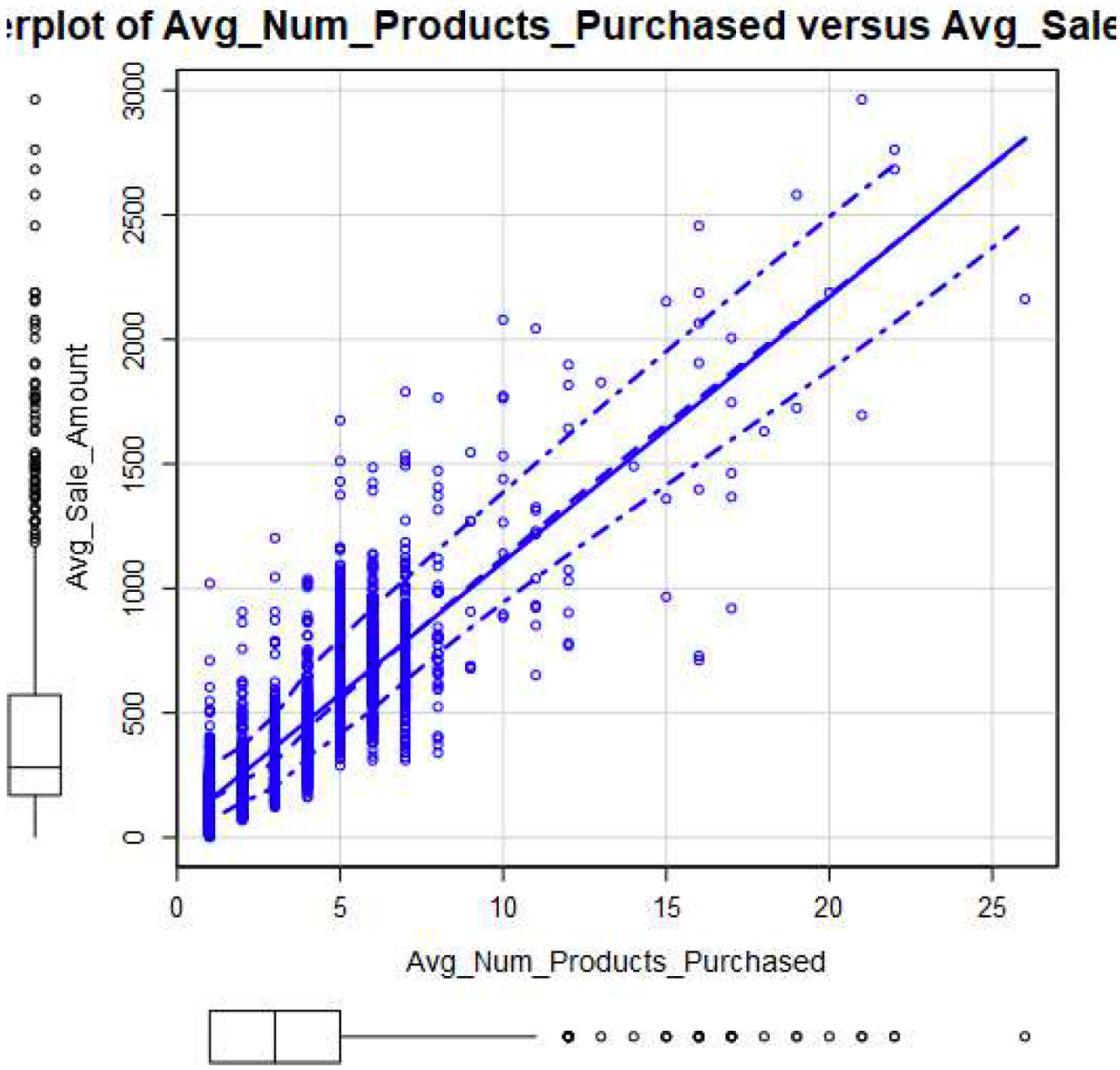

Figure 2. Scatterplot and linear-fitted line for Avg_Sale_Amount VS Avg_Num_Products_Purchased.

There are several methods we can use to select correlated continuous variables in this multiple regression study, including but not limited to visual evaluation as the **Figure 1** illustrated above. Here, we can visually observe that the Avg_Num_Products_Purchased by a customer is positively correlated with the Avg_Sale_Amount of the customer (r=0.86), that is as Avg_Num_Products_Purchased, Avg_Sale_Amount also increases (**Figure 2**).

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 3. Type II ANOVA Analysis

Besides, we can also use p-value (if less than the predefined alpha of 0.05) in Type II ANOVA (**Figure 3**) and/or multiple linear regression models (**Figure 4**) to select for both continuous or categorical variables. Here, we found that both Customer_Segment (Categorical) and Avg_Num_Products_Purchased (Continuous) were significantly associated with Avg_Sale_Amount – Both with p < 2.2e-16.

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.

Figure 3 shows the multiple linear regression model including Avg_Num_Products_Purchase and Customer_Segment (with Credit Card Only as reference/base). The 4 variables were significant association with p-value much lesser than 0.05, with Adjusted R-squared values of model showing 0.8366 (Generally R-squared>0.7 is considered good).

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

Figure 4. Ordinary least square regression model.

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.

Y = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149.36 (If Customer_Segment: Loyalty Club Only) + 281,84 (If Customer_Segment: Loyalty Club and Credit Card) – 245.42 (If Customer_Segment: Store Mailing List) + 0 (If Customer_Segment: Credit Card Only)

Note: For Customer_Segment, the reference/base case to Credit Card Only.

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 the 250 customers.

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

As we want to calculate the expected revenue from the 250, we need to multiply the probability that a person will buy our catalog with the predicted average average sale amount of that person. For example, if a customer were to buy from us, we predict this customer will buy $450 worth of products. At a 30% chance that this person will actually buy from us, we can expect revenue to be $450 x 30% = $135. After obtaining the gross revenue by summing up all the expected revenue from the 250 customers, we will need to multiply the gross revenue by 0.5 to obtain the profit margin. Finally to obtain the overall net profit, we need to subtract the cost of the catalogs sending out to the 250 customers ($6.50 per customer).

By using the following formula, we can obtain:

$$Overall\ Net\ Profit, OP = 0.5 \sum_{i}^{n} P_i S_i - 6.5n$$

where:-
n =. Total customers (250)
i = each customer
P = Probability that the customer will response to the catalog
S = Predicted average of sales amount for the customer

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

The expected profit from the new catalog project sending to the 250 customers is 21987.44.

Appendix:

Alteryx Flow