Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
 - Whether the company should send out catalogs to the new customers.
2. What data is needed to inform those decisions?
 - The total expected profits from sending out catalogs to the new customers.

Customer segment, average number of product purchased, score_yes, margin and cost of catalog are needed to drive the final results.

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)

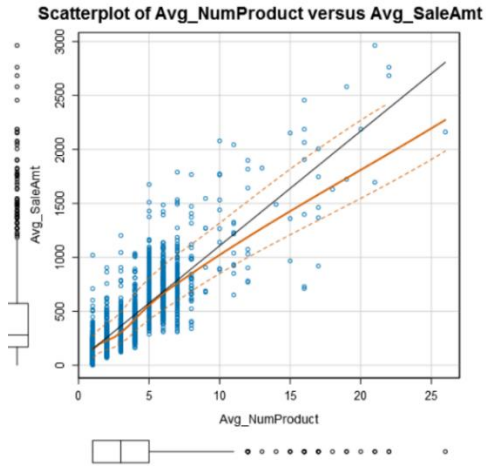
Important: Use the p1-customers.xlsx to train your linear model.

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 align with the final business goal, first I need to build a model of the target variable, which is the continuous average sales amount data.
- Then I imported the customer data from customers.xlsx to Alteryx to gain the intuitive from the scatter plot, several pairs of predictor variables to average sales amount.
- Lastly, I kept the variables with significant p-value which is < 0.05 in the model.

Here is example of scatter plot which shows positive correlation between avg_number_product variable and avg_salesamount variable.



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.

- All the predictor variables are with significant p-values and the R-squared value for the model is 0.8369. So I conclude this is a good model as it shows that over 83% of variance in the target variable is explained by the selected independent variables.

-

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

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……*

Y = 303.46 - 149.36 customer_segmentLoyaltyClueOnly) + 281.84 (Customer_segmentloyaltyclub and credit card) – 245.42 (Customer_segmentstoremailinglist) + 66.98 (Avg_num_products_purchased)

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)

At the minimum, answer these questions:

1. What is your recommendation? Should the company send the catalog to these 250 customers?
 - The company should send the catalog to these 250 customers as the expected return is above the $10000.
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
 - I started with the idea of getting the predicted average amount from these 250 customers.
 -Then I run a linear regression model based on customer data using the relevant variables and generate the coefficient for these variables.
- Then the expected revenue from each customer is determined by multiplying expected sale amount with Score_Yes value.
- With a gross margin of 50%, 50% is deducted from the sum of expected revenue before the cost of catalog ($6.50) is subtracted to obtain net profit.
- The result is $21987.44.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
 - The expected return of profit is $21987.44

Here is my alteryx workflow