<u>Project 1: Predicting Catalog Demand</u>

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?

Ans 1: The company needs to predict the expected profit from its 250 new customers, whom it is planning to send out the catalog to. Based on the historical data, the company wants to predict the profit they can make by sending out the catalog to its new customers. The decision to send out the catalog to these 250 new customers is only favorable if the expected profit contribution is greater than $10,000.

2. What data is needed to inform those decisions?

Ans 2: The following data is required to make the decisions mentioned in answer number 1

- Sum of expected profit from 250 new customers
- Historical data to determine predictor variables that has strong relation with Avg_Sale_Amount
- Predicted_Avg_Sale_Amout for each of the 250 new customers
- A formula to calculate the profitability.

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.

Ans 1: Out of 11 columns that could have been the possible predictor variables, I excluded the following (with the reasons attached) to test them in my linear regression and check their P-Values

a. Name

Reason: Because names, despite being possibly repeated, could not fairly determine the average sales amount

b. Customer_ID

Reason: Customer_ID is supposed to be unique to every customer, and hence cannot determine the average sales amount.

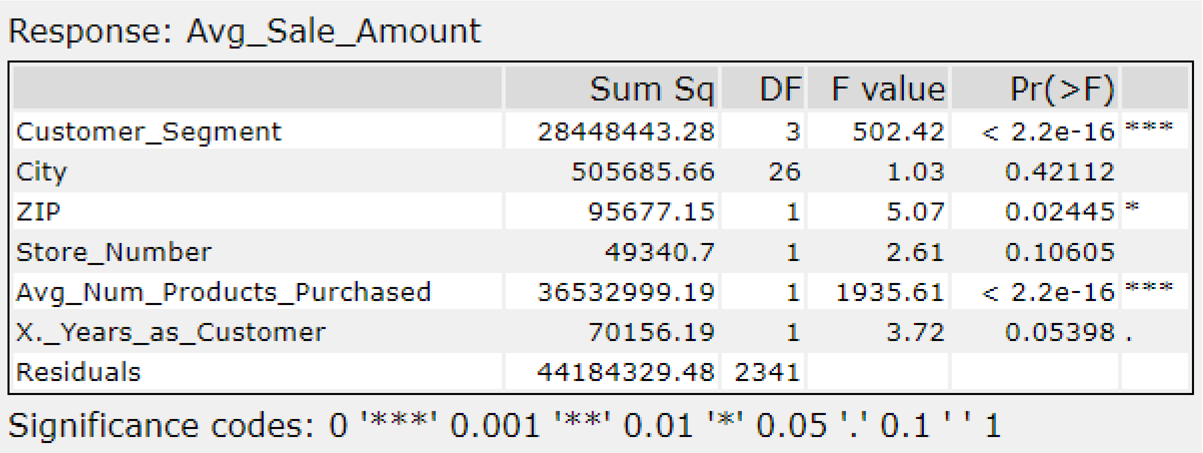
c. Address

Reason: While initially exploring the data, I found the Address column's values quite unique that it could not determine the average sales amount.

d. Responded_to_Last_Catalog

Reason: Since the test data set does not have this variable, we'll simply exclude it from training as well.

I then ran linear regression on remaining variables and the results were as following:



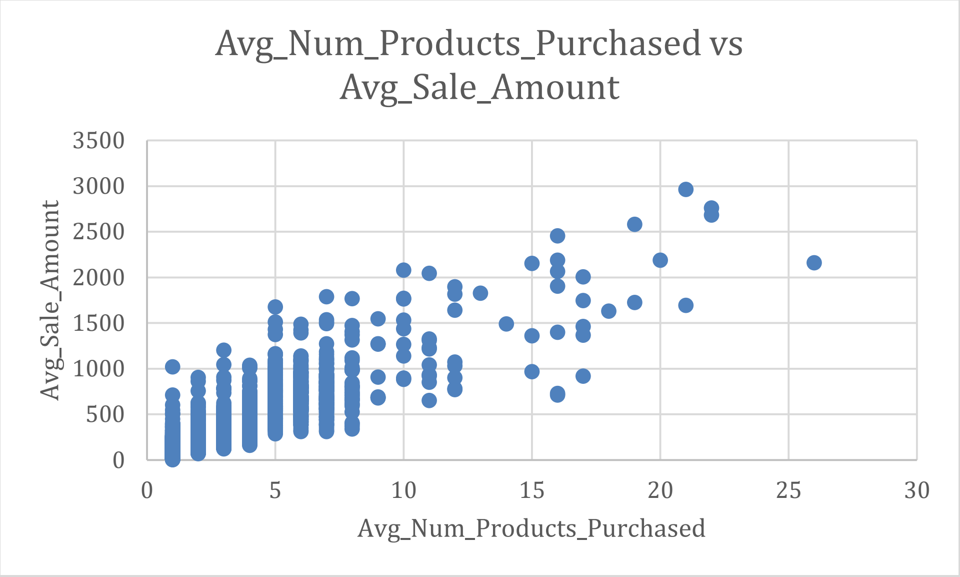
Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28448443.28	3	502.42	< 2.2e-16	***
City	505685.66	26	1.03	0.42112	
ZIP	95677.15	1	5.07	0.02445	*
Store_Number	49340.7	1	2.61	0.10605	
Avg_Num_Products_Purchased	36532999.19	1	1935.61	< 2.2e-16	***
X._Years_as_Customer	70156.19	1	3.72	0.05398	.
Residuals	44184329.48	2341			

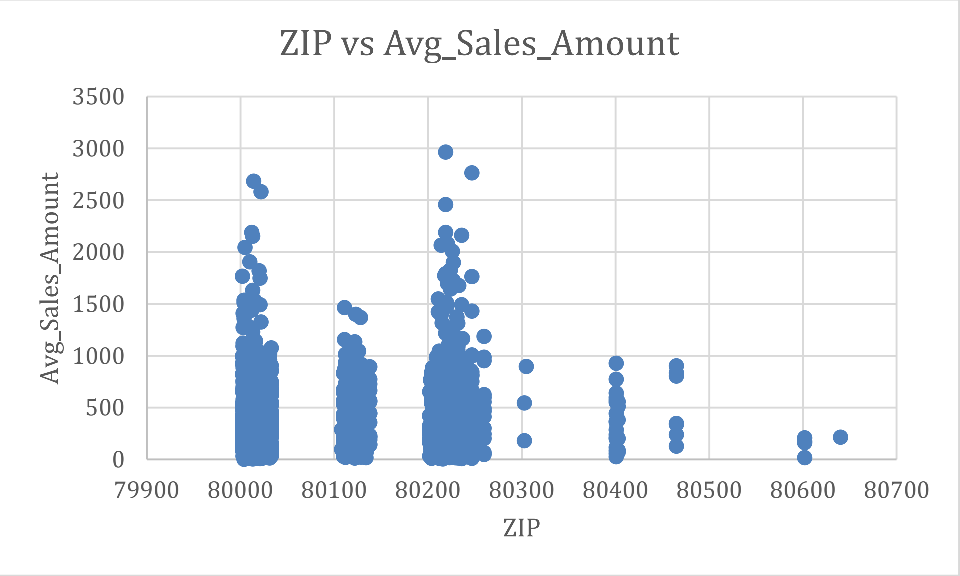
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

These results clearly showed that **Customer_Segment** and **Avg_Num_Products_Purchased** have highest significance and are the best possible predictor variables to predict the **Avg_Sales_Amount** as their P values were very low (less than 2.2e-16).

Also, when we plot Avg_Num_Products_Purchased against Avg_Sale_Amount, we see a linear relationship which tells us that it is a good predictor variable.



Additionally, it is important to mention that ZIP values also showed some significant relationship with the target variable, Avg_Sales_Amount. So we plotted its graphs to study its linear relationship and found the following:

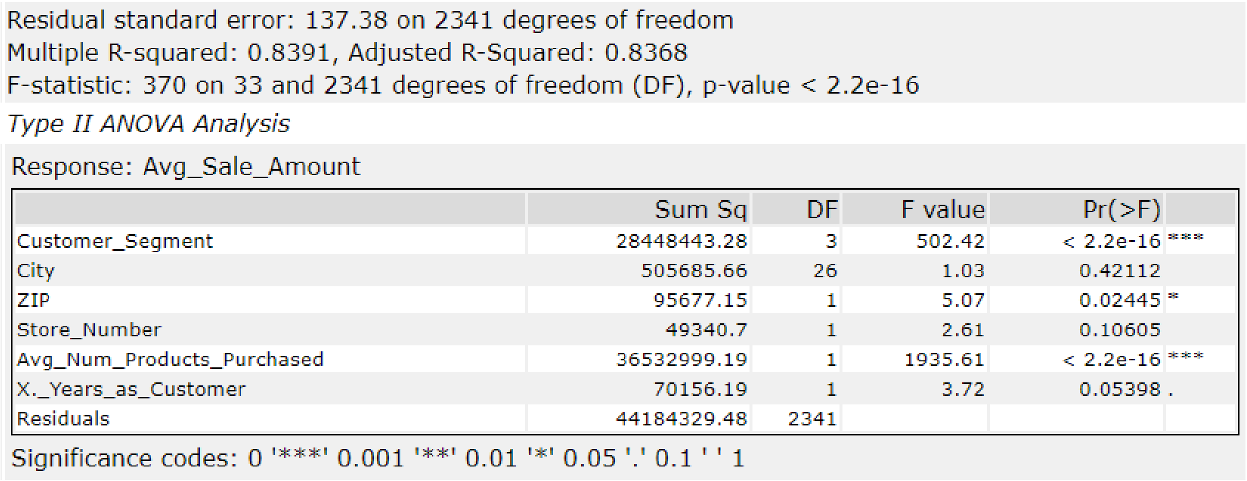


There was no linear relationship between the target (Avg_Sales_Amount) and the predictor (ZIP) variable, hence despite having some significant relationship and P Value less than 0.05, we did not shortlist this variable as a predictor variable.

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.

Ans 2: There are 3 metrics that helps us determine that there is a strong relationship between the target and the predictor variables and that they are statistically significant to make a good model. These 3 metrics are **P Values**, **R Squared Value** and **Adjusted R Squared Value.** The linear regression model built in Alteryx shows the following results:



Residual standard error: 137.38 on 2341 degrees of freedom
Multiple R-squared: 0.8391, Adjusted R-Squared: 0.8368
F-statistic: 370 on 33 and 2341 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

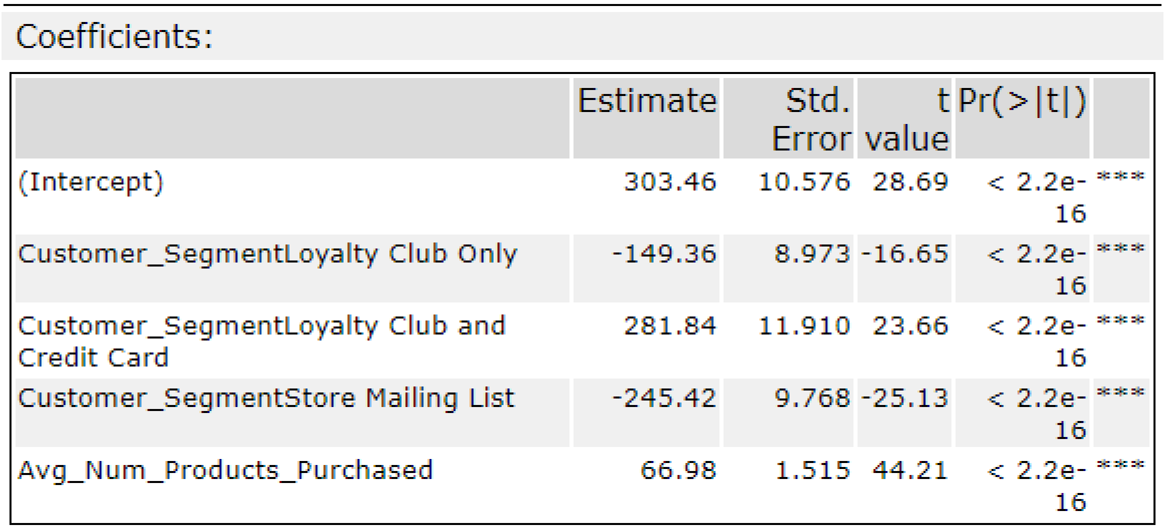
	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28448443.28	3	502.42	< 2.2e-16	***
City	505685.66	26	1.03	0.42112	
ZIP	95677.15	1	5.07	0.02445	*
Store_Number	49340.7	1	2.61	0.10605	
Avg_Num_Products_Purchased	36532999.19	1	1935.61	< 2.2e-16	***
X._Years_as_Customer	70156.19	1	3.72	0.05398	.
Residuals	44184329.48	2341			

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

The table shows that both the predictor variables chosen, Customer_Segment (Categorical) and Avg_Num_Products_Purchased (Continuous) have P Value less than **2.2e-16** which proves that they are statistically very significant. Also, the R-Square value which determines the correlation between the target and the predictor variable is **0.8391** which is near to 1 and signifies positive correlation. Lastly, the adjusted R-Square Value is **0.8368** which also proves that this is a good 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)

Ans 3:



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	***

We built our linear regression model using the chosen predictor variables and the coefficients are calculated and shown in the table above. Using these coefficients, the linear regression equation would be

Avg_Sale_Amount = 303.46 + 66.98(Avg_Num_Products_Purchased) – 149.36(Customer_SegmentLoyalty Club Only) + 281.84(Customer_SegmentLoyalty Club and Credit Card) – 245.42(Customer_SegmentStore Mailing List) + 0(Customer_Segment Credit Card Only)

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.

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)

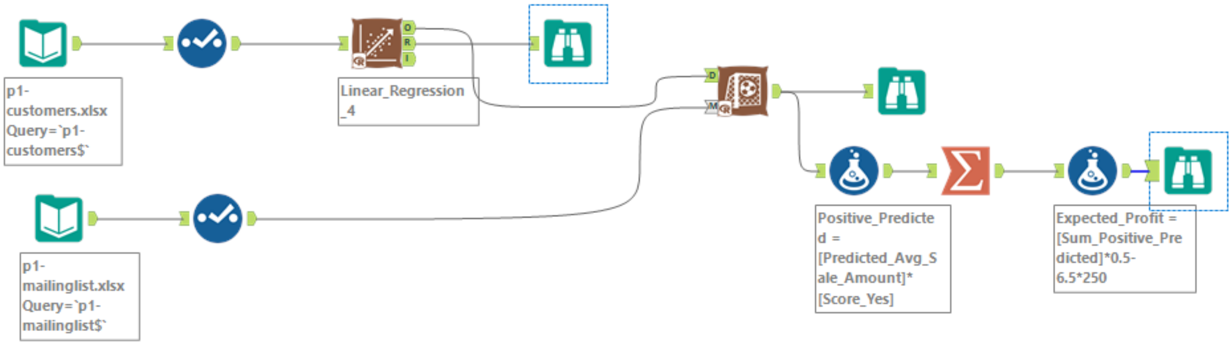
At the minimum, answer these questions:

1. What is your recommendation? Should the company send the catalog to these 250 customers?

Ans 1: We trained our model based on customer list and then applied new customers' data set on the trained model. Based on the model's predictions, the total expected profit would be **$21,987.44** which is greater than the benchmark of $10,000. Hence, my recommendation would be that the company should definitely go ahead with sending the catalog to 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)

Ans 2: Following workflow was made on Alteryx:


p1-customers.xlsx Query=`p1-customers$`
Linear_Regression_4
O
R
I
D
M
p1-mailinglist.xlsx Query=`p1-mailinglist$`
Positive_Predicted = [Predicted_Avg_Sale_Amount]*[Score_Yes]
Expected_Profit = [Sum_Positive_Predicted]*0.5-6.5*250

Step 1 > Input the customers list and select the columns; Avg_Sale_Amount, Customer_Segment and Avg_Number_Products_Purchased

Step 2 > Build a linear regression model using Avg_Sale_Amount as target variable and Customer_Segment and Avg_Number_Products_Purchased as predictor variable

Step 3 > Input mailinglist data and select the columns; Customer_Segment, Score_Yes and Avg_Number_Products_Purchased

Step 4 > Run the trained model to predict the values for the column 'Predicted_Avg_Sale_Amount'

Step 5 > Multiple Predicted_Avg_Sale_Amount with 'Score_Yes' and sum all the values

Step 6 > Find the expected profit by applying the formula SUM(Predicted_Avg_Sale_Amount * Score_Yes) * 0.5 - $6.50 * 250

Note: All these steps were followed after selecting the predictor variables that were statistically significant.

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

Ans 3: The expected profit is $21,987.44. This value is calculated using the following formula:

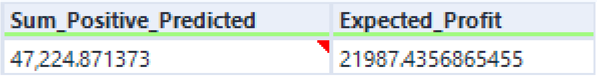
Expected_Profit = SUM(Predicted_Avg_Sale_Amount * Score_Yes) * 0.5 - $6.50 * 250

Where the value of SUM(Predicted_Avg_Sale_Amount * Score_Yes) is **$47,224.87**

Now, when we plug in all the values in the Expected_Profit formula, we get:

=$47,224.87 * 0.5 - $6.50 * 250

=$21,987.44



Sum_Positive_Predicted	Expected_Profit
47,224.871373	21987.4356865455