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 needs to decide whether to send or not the catalog to these 250 new customers.

2. What data is needed to inform those decisions?

    ✓ To calculate the profit (which must be over $10.000 in order to send catalogues) and build a reliable model, the company needs data included but not limited to:

    - data on current customers (how many items they ordered in the past and how much they spent)
    - data on new customers (chances of ordering a product from catalogue, how much they spent in the past)
    - customer segment
    - catalog printing cost

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.

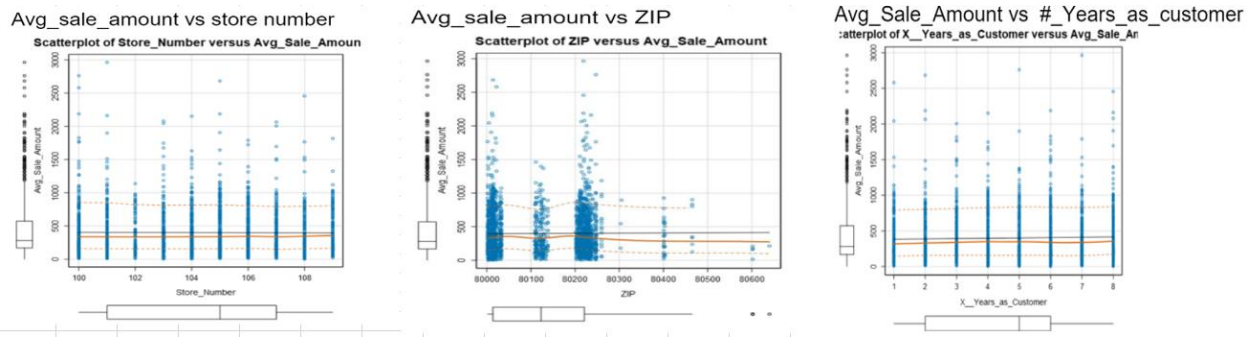
At the minimum, answer these questions:

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.

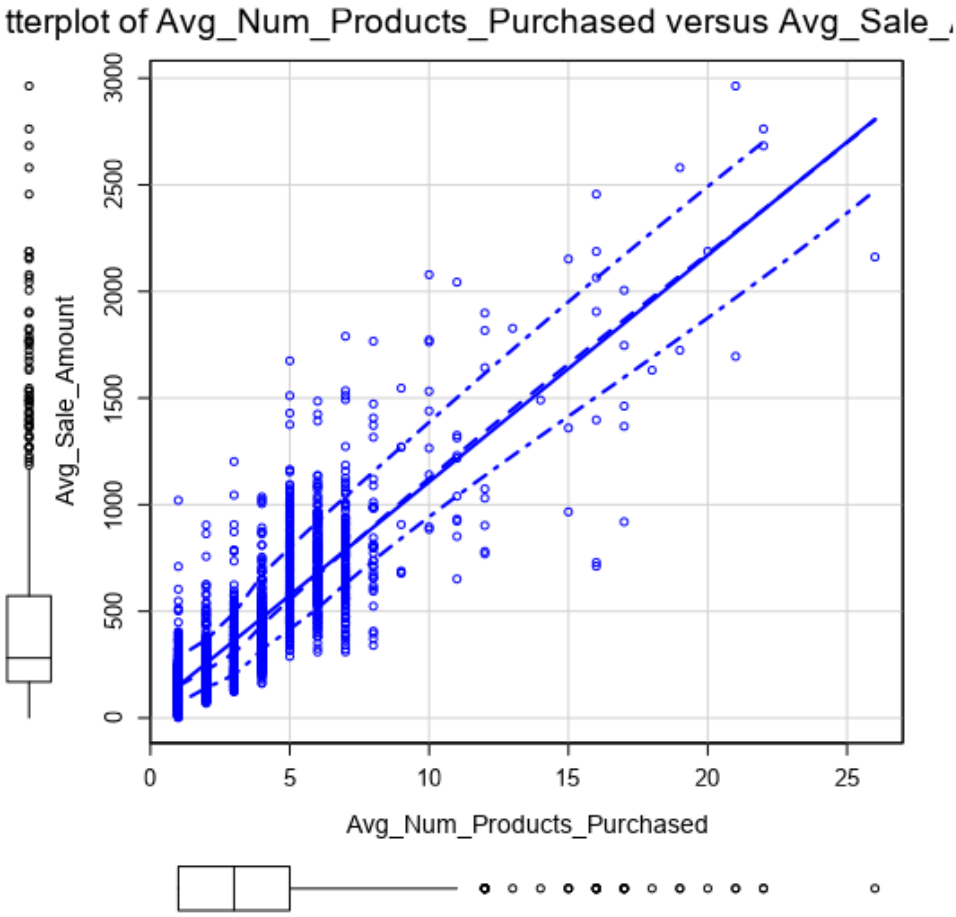
    ✓ For ***Numeric variables*** I used Alteryx scatterplot tool to check the relationship between an individual variable and the target variable and see if any variables

might be a good candidate for a predictor variable. If the relationship is not linear it most likely is not a good candidate.

The non linear results are shown in the following picture:



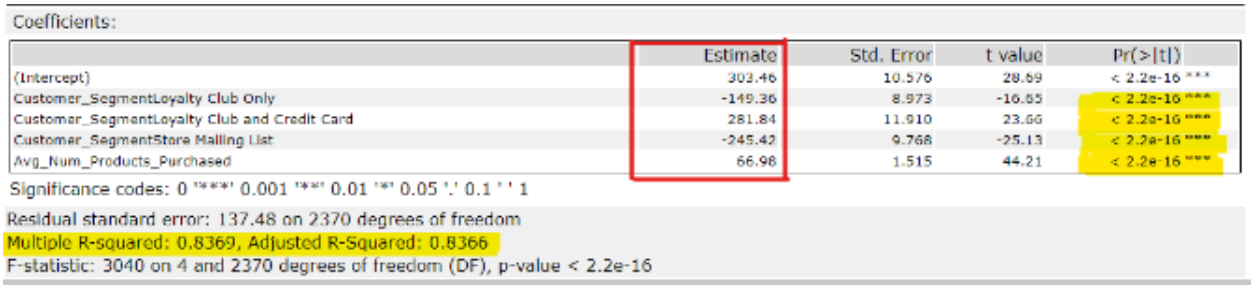
- ✓ The only linear result found (shown in the following picture) was between **Avg_Num_Products_Purchased** and decided to use it as a predictor variable (the p value <.05 later on will show it is statistically significant)



For ***categorical variables*** I will have to use trial and error to see if they are statistically significant. (On the later session I will demonstrate that only **Customer_segment** is statistically significant with a p value < .05)

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.

- ✓ The results of the model in the following screenshot:



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

- ✓ Based on the results, I would say this is a good model, since each of the predictor variables are statistically significant (p-value<0.05) and *** Adjusted R^2** is about **0.84** This means the model explains 84% of the total variation of the data. Also the adjusted R^2 is so close to the R^2 meaning that the additional predictor has contributed positively in our prediction.

* *Using Adjusted R^2 for unbiased estimation and also to demonstrate the explanatory power of regression model since it contains more than one predictor.*

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……*

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.

- ✓ The linear regression equation as it follows:

Avg_Sale_Amount = **303.46** + **66.98** Avg_Num_Products_Purchased **– 149.36** (If type: Loyalty Club Only) + **281.84** (If type: Loyalty Club and Credit Card) **– 245.42** (If type: Store Mailing List) **+ 0** (If type: Credit Card Only)

reference group, used as baseline

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?

- ✓ I'd recommend the company to send the catalogue to these 250 customers

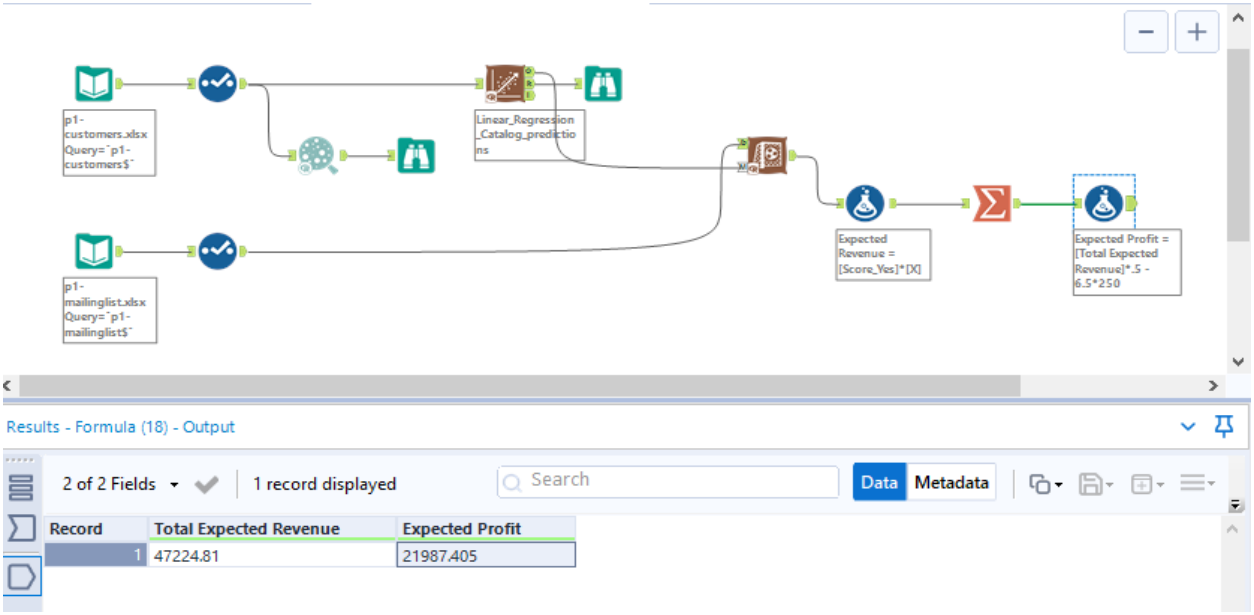
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

- ✓ I first constructed the linear regression model with the data of current customers (customers.csv file) using Customer_Segment and Avg_Num_Products_Purchased as predictor variables with Avg_Sale_Amount being the target variable.

- ✓ Then by using the Alteryx Score Tool I made the prediction on the Avg_Sale_Amount for the 250 customers (mailinglist.csv file)

- ✓ Then by using Formula and Summarize tool, I calculated the profit of **$ 21987.4 which is above the $10.000 threshold and that's why I'd recommend the company to send the catalogue to these 250 customers**

- ✓ Below you can find the screenshot of my Alteryx Workflow



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

- ✓ The expected profit is **$ 21987.4**

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.