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?** For this project, I will analyze a business problem in the mail-order catalog business. The company has 250 new customers from their mailing list. They would like to send a catalog to these new customers. I will need to make the following decisions for this project:
    - ✓ Predict how much profit the company can earn from these 250 new customers
    - ✓ Recommend to management whether the catalog should be sent to these new customers or not. The catalog should only be sent if the expected profit exceeds $10,000.
2. **What data is needed to inform those decisions?** A linear regression model will be used to predict sales for the 250 customers. Alteryx will be used to build the best linear model with the data provided. The data included in the following two datasets is needed to make the above decisions:
    - ***p1-customers.xlsx*** includes data on 2375 customers and will be used the regression model.
    - ***p1-mailinglist.xlsx*** contains data on 250 customers that is needed to predict sales. This is the list of customers that the company would send a catalog to. **This dataset will be used to estimate how much revenue the company can expect if they send out the catalog**. It includes all fields from P1_Customers.xlsx except for **Responded_to_Last_Catalog** **so this variable cannot be used in the linear regression model since it could not be applied to the mailing list data set**. It also includes two additional variables.
        - ✓ **Score_No**: The probability that the customer WILL NOT respond to the catalog and not make a purchase.
        - ✓ **Score_Yes**: The probability that the customer WILL respond to the catalog and make a purchase.

Other Details

- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Multiply revenue by the gross margin first before you subtract out the $6.50 cost when calculating your profit.
- Calculate the expected revenue from these 250 people to get expected profit. This means we need to multiply the probability that a person will buy our catalog as well.

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.

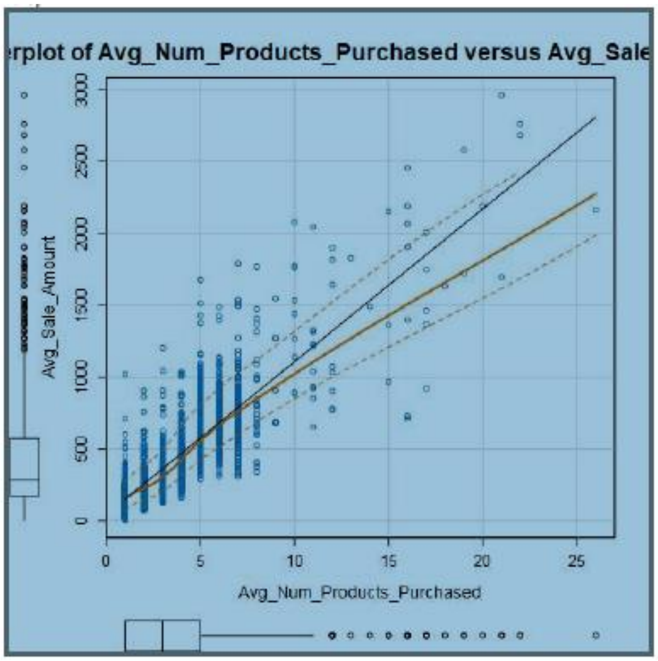
At the minimum, answer these questions:

1. **How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer**.

Selection of Predictor Variables

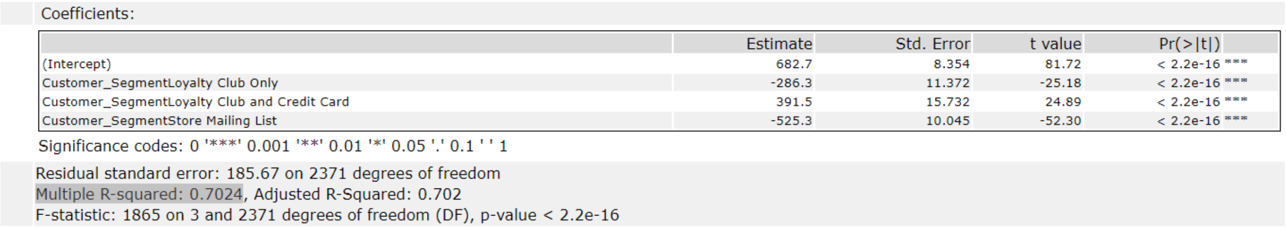
For my linear regression model, I selected the following predictor variables: **Avg_Num_Products_Purchased** and **Customer_Segment**. These predictor variables show a significant linear relationship with the target variable, Avg_Sale_Amount.

Below is a scatterplot which shows a linear relationship between Avg_Num_Products_Purchased and Avg_Sale_Amount:



As the Avg_Num_Products_Purchased increase, the Avg_Sale_Amount increase. In addition, Avg_Num_Products_Purchased has a significant P-Value in the linear regression model. The P-Value is < 2.2e-16. This value .00000000000000022 is much smaller than .05 P-Value is much smaller than the conventional value of .05 that is often used as a criterion for statistical significance. Since the P-Value is the probability that observed results occurred by chance, and that there is no actual relationship between the predictor and target variable, the probability that the coefficient is zero. Since the P-Value is very low, there is a high probability that a relationship exists between Avg_Num_Products_Purchased and Avg_Sale_Amount.

Since the Customer_Segment predictor variable is categorical, I did not use a scatterplot to see whether a linear relationship exists for this categorical predictor variable. Instead, I ran Customer_Segment separately through the regression model and see if the coefficients had high multiple-R-squared and low P values. Please find the results below:



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	682.7	8.354	81.72	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-286.3	11.372	-25.18	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	391.5	15.732	24.89	< 2.2e-16	***
Customer_SegmentStore Mailing List	-525.3	10.045	-52.30	< 2.2e-16	***

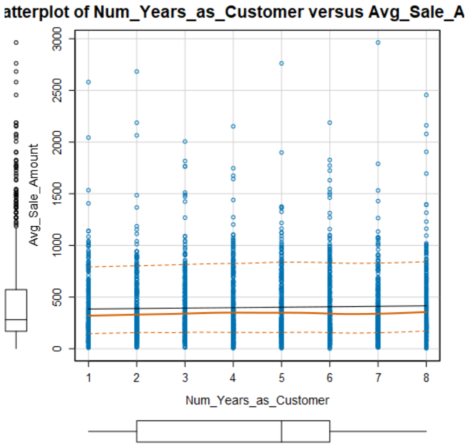
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 185.67 on 2371 degrees of freedom
Multiple R-squared: 0.7024, Adjusted R-Squared: 0.702
F-statistic: 1865 on 3 and 2371 degrees of freedom (DF), p-value < 2.2e-16

The multiple r-squared value is 0.70, which is relatively high and thus shows a linear relationship with the target variable, Avg_Sale_Amount. In addition, the P-Value for each the Customer_Segment coefficients is very low as well, 2.2e-16 (.0000000000000022), which also shows a high probability that a relationship exists between Customer_Segment and Avg_Sale_Amount.

I did not find that a linear relationship existed (used Scatterplots) with the Customer_Id, Zip, Store_Number, and Num_Years_As_Customer variables in the ***p1-customers.xlxs*** dataset and decided not to use the variables as predictor variables for this linear regression model.

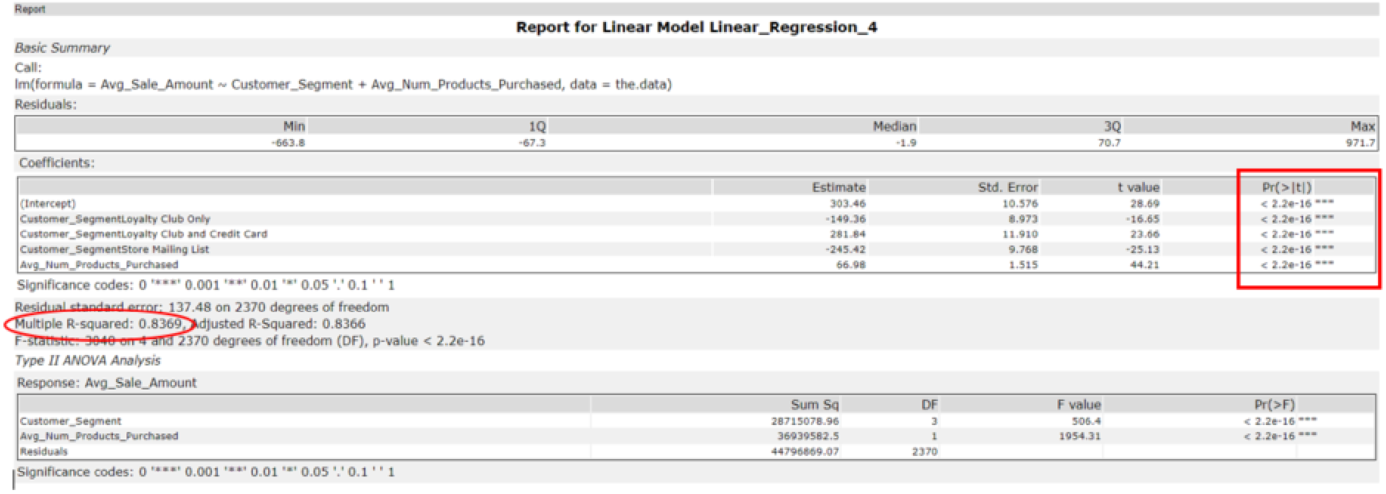
i.e.



In addition, after running the other variables (Address, City, State) in the ***p1-customers.xlxs*** dataset through the regression model, it did not show a significant linear relationship (high p values and low multiple r squared values) between these variables and Avg_Sale_Amount. As a result, I did not use these variables as predictor variables for this linear regression model either. The **Responded_to_Last_Catalog** variable was not used in the regression model since this variable does not exist in the *p1-mailinglist.xlsx* dataset.

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.**

Overall, I believe that my linear model is good. After running the predictor variables that I selected for this model, **Avg_Num_Products_Purchased** and **Customer_Segment,** a significant linear relationship is shown with the target variable, Avg_Sale_Amount. The Model produced the following P-Values and R-Squared values using these two predictor variables:



Report

Report for Linear Model Linear_Regression_4

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

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		

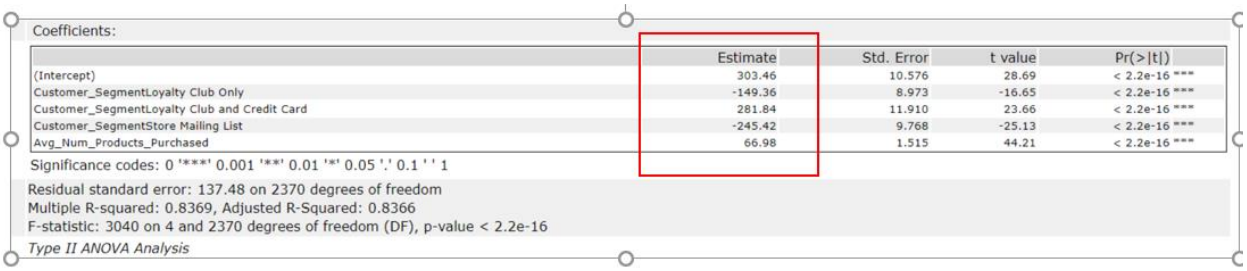
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Customer segment is a categorical variable and has three categories, Loyalty Club Only, Loyalty Club and Credit Card, and Store Mailing List. The P-Value for each of the predicator variables, Customer_SegmentLoyalty Club Only, Customer_SegmentLoyalty Club and Credit Card, Customer_SegmentStore Mailing List, and Avg_Num_Products_Purchased, is < 2.2e-16 (.00000000000000022).   This value is much lower than a P-Value of .05 which shows a high statistical significance. There is a high probability that a relationship exists between these predictor variables and the target variable (Avg_Sale_Amount).

In addition, the Multiple R-Squared value is 0.8369. R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. Since we have a high R-Squared value (value above .70), it can be determined that the model is strong.

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……*

I used Alteryx to build the regression equation below using the **p1-customers.xlsx** dataset:



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

I have highlighted the Coefficient Estimates in the screenshot above. These values were plugged into the formula below:

***Y =** 303.46 - 149.36 * LoyaltyClubOnly + 281.84 * LoyaltyClubAndCreditCard – 245.42 * StoreMailingList + 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?** My recommendation to management is that the company send the catalog to these 250 new customers. The catalog should only be sent if the expected profit exceeds $10,000.

2. **How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)**
   I arrived at the recommendation by following the process below:
   - ✓ Used Alteryx software to build a linear regression model using the **p1-customers.xlsx** dataset based on Customer_Segment and Avg_Sale_Amount. I determined which predictor variables to use in the equation by analyzing each variable by using either a scatterplot or running the variable through the liner regression model to determine if a liner relationship existed between the predictor variable and the target variable (Avg_Sale_Amount). For the variables run through the linear regression model, those with low P-Values (<.05) and high multiple R-

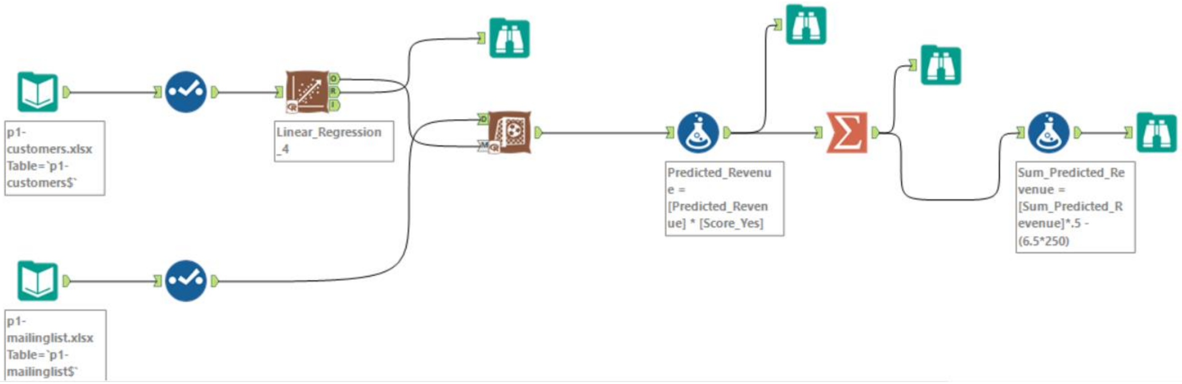
Squared values (> .70) show a significant relationship with the target variable Avg_Sale_Amount and use in the linear regression model.

- ✓ Calculated predicted revenue (sales) by using the score tool in Alteryx using my linear regression equation to predict sales for the individual people in the mailing list.
- ✓ Used the formula tool in Alteryx to calculate expected revenue from these 250 people to get expected profit. Score_Yes, data provided in the mailing list dataset, is the probability that the customer WILL respond to the catalog and make a purchase.

  Expected_Revenue = Predicted_Revenue * Score_Yes
- ✓ Used the sum tool in Alteryx to sum expected revenue on all products
- ✓ Used the formula tool in Alteryx to multiply the expected revenue (profit) sum by the gross margin and then subtract cost of goods sold ($6.50 cost for each catalog) to get the final expected revenue (profit) amount.

  SUM(Predicted_Revenue) *.5 - (6.5*250)

See Alteryx workflow below:



2. **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, assuming that it is sent these 250 customers, is **$21,987.44**. Since this value exceeds $10,000, the catalog should be sent to these new customers.