Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

1. What decisions needs to be made?

To decide whether sending this year catalog to the 250 new customers will be profitable or not.

2. What data is needed to inform those decisions?

For this project two datasets have been given (*p1-customers.xlsx* & p1-mailinglist.xlsx) besides some other details. Data about sales that took place earlier with other customers, customers' segments and the probability that the new customer will buy, all these will be captured from these datasets as well as using the given average gross margin 50% & the catalog cost $6.50 to calculate the profit and make our decision.

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables in your model?

For this model equation, the target variable will be Avg_Sale_Amount and the predictor variables are Customer_Segment &Avg_Num_Products_Purchased. The reason behind choosing these variables is explained below:

- Categorical Variables

By trying to run a linear regression model multiple times and check the p value, which must be (p-value <= 0.05) in order to consider that variable to be statistically significant. However, it's worth mentioning that I've ignored some variables and didn't test their significance since they don't have any impact on the sales -logically- such as customer name & ID.

Report

Report for Linear Model Linear_Regression_8

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

Figure 1

- Numeric Variables

By generating scatterplots to illustrate the relationship between the target variable and other variables, as there should be a slope to the line since this indicates the strong linear correlation with the target variable. The graph below shows these relationships between the Avg_Sale_Amount "target" & other variables.

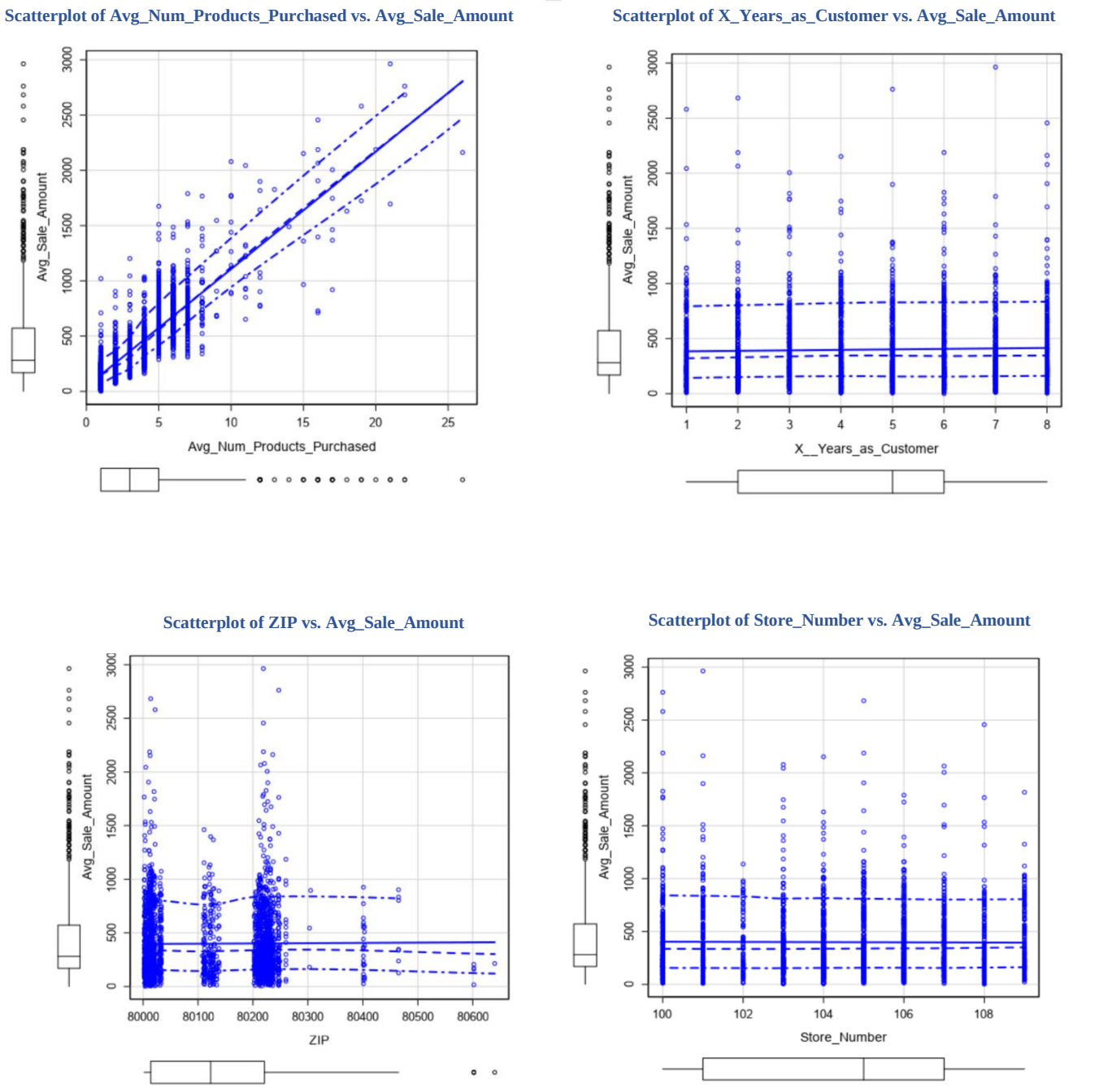

Figure 2

2. Explain why you believe your linear model is a good model.

Referring to Figure 1, the predictor variables (Customer_Segment &Avg_Num_Products_Purchased) have p-value that's less than .05 and the adjusted r-squared equals .8366 which is considered a high value -close to 1-, these two indicators can reveal how reliable this model can be in predicting the target value.

3. What is the best linear regression equation based on the available data?

Referring to Figure 1 as well, the equation will be as follow:

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

Step 3: Presentation/Visualization

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

Yes, they should send the catalog to these 250 new customers as the expected profit will exceed $10,000

2. How did you come up with your recommendation?

In order to reach that conclusion, the steps mentioned below have been followed:

1. Calculating Avg_Sale_Amount by running the linear regression model on Alteryx, then using score function to apply than on the p1-mailinglist.xlsx dataset to get ExpectedSales.
2. Creating and calculating a new column "Ave_ExpectedSales = ExpectedSales * Score_Yes, then getting the sum of that new value.
3. Calculating the expected profit by using the given inputs:

Expected Profit= ∑ Ave_ExpectedSales * gross margin – (cost of catalog * number of new customers)
Expected Profit= ∑ Ave_ExpectedSales * .5 – (6.50 * 250)

A screenshot of my Alteryx workflow is provided on the next page.

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

By using the equation mentioned in the previous question,
Expected Profit= ∑ Ave_ExpectedSales * .5 – (6.50 * 250)
Expected Profit= (47,224.871373 * .5) – (6.50 * 250) = **$21,987.4356865** > $10,000

- Alteryx Workflow

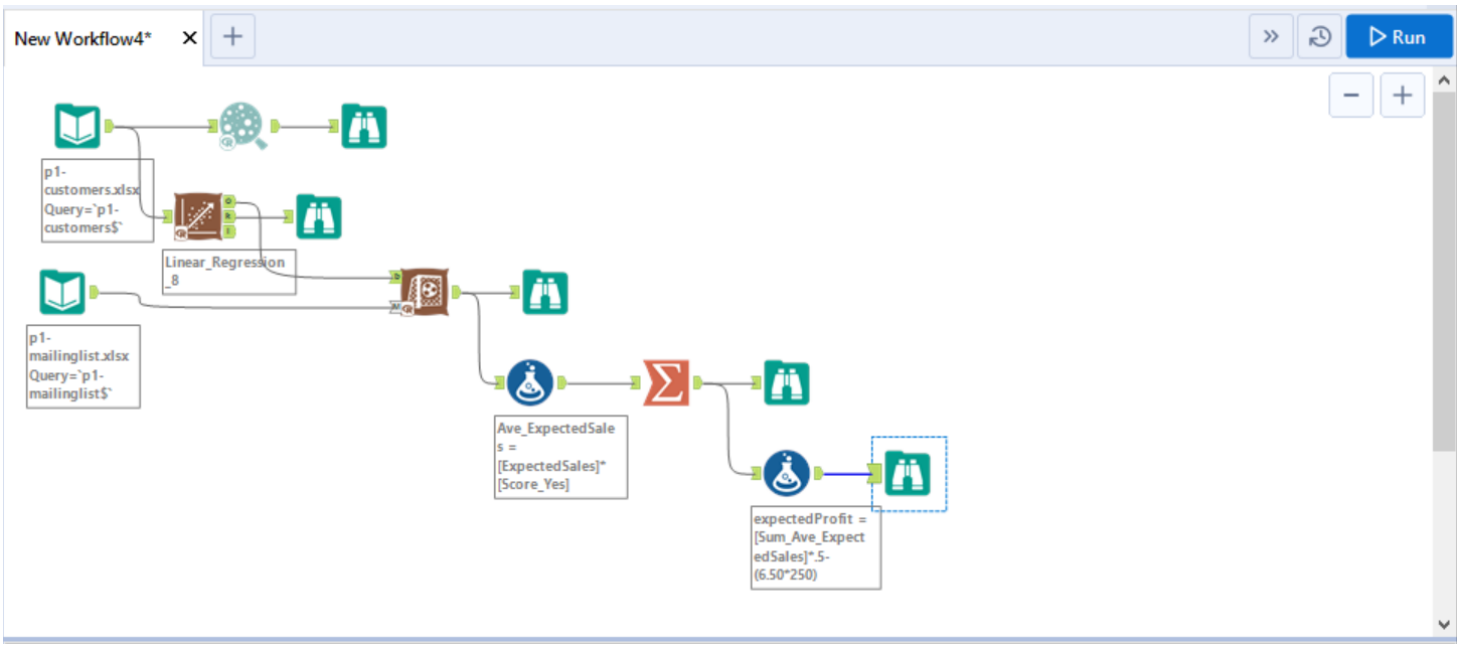

Figure 3