Project2: Predicting Catalog Demand

PROJECT SUBMISSION WRITTEN BY CHIDIEBERE STEPHEN NWOSU

Date: 20th April, 2021

Step1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500word limit)

Key Decisions:

Answer these questions:

1. What decisions needs to be made?

Answer:

The major decision to be made by the company is that, the company want to determine if the expected profit from these 250 new customers will exceed $10,000 and then decide whether to send the catalog to these customers or not to send.

2. What data is needed to inform those decisions

The data that's needed to inform those decisions is as follows:

- Using Alteryx, we can calculate how much the 250 new customers can buy (score) by using the score tool, then multiply score by score_yes to get predicted price. Multiply predicted price by 50%, subtract by $6.5 and get the profit for each 250 new customers, sum it up and compare with the $10,000 profit.
- **Data Rich**: This is when you have data, or a previous data that is similar to the business analysis. From the business problem, since we have data from the P1 customer file, we are data rich.
- **Numeric Model**: Numeric has to do with numbers. From the business point of view, I have been expected to predict profit from the 250 new customers (Numeric) to exceed $10,000. Therefore, a Numeric Model is needed to inform those decisions
- **Continues Model**: Continues model is a continues analysis. It is not interested in forecasting or neither interested in a particular calendar per week. Continues model is needed to inform those decisions because, from the business analysis, we are looking at the outcome of making profit, if we send the catalog to 250 new customers and not interested in the specific calendar week of sending the catalog to the 250 new customers.
- **Linear Regression Model**: A linear regression model shows a linear relationship between the target variable and the predictor variable. Similarly, from the scatter plot in Figure1 below, their exist a very strong linear relationship between the

Avg_Sale_Amount (target variable) and Avg_Num_Product_Purchased (predictor variable). Therefore, the information needed to inform those decisions is also a linear regression model.

- **(Score_Yes)**: Probability of costumer making purchase (Score_Yes) was successful, so is also a data that needed to inform the decision weather or not the company will make profit that exceeds $10,000. Past revenue generated, number of past products purchased, are also needed to inform those decision.

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variable you used and why, and the results of the model. Visualizations are encouraged. (500word limit)

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.

Answer:

I selected Avg_Num_Product_Purchase and Customer segment as the predictor variable because, from the P1-Customers.xlsx file, it is observed that the predictor variable Avg_Number_Product_Purchase and customer segment was chosen because the Avg_Number_Product_Purchase, gives the number of products purchased by the different customers in the Customers segment, which is very important to the business analysis. It is viable and also gives us a clue that, if sending the catalog to the 250 new customers will be feasible. The customer segment gives us the different customers that maid the purchase. These customers are; Store mailing list, loyalty club and credit card, loyalty club only and credit card only. The two predictors variable are very important in the business analysis and also influences the average sale amount.

I chose Avg_Number_Product_Purchase to be the continues predictor variable because, it is the only variable that shows a linear relationship (straight lines) with the target variable Avg_Sale_Amout. This can be seen in Figure1.0, as the X-axis Avg_Number_Product_Purchase (predictor variable) increases, the Y-axis Avg_Sale_Amount (target variable) also increases, indicating the two variables are related.

A non-linear relationship is seen in Figures (2.0, 3.0 and 4.0), Indicating that there are no linear relationships because, as their X-axis increase, their Y-axis does not increase nor move in either

direction showing no relation with each other. No straight line seen. (https://classroom.udacity.com/nanodegrees/nd008t-ent/parts/67b7a57e-f582-44d4-b814-c77b8b4b2ed7/modules/8d88c2b0-39a7-45c5-ae30-d2cbcee2d35b/lessons/9439c39a-6d05-4093-8c55-26c45ebcc9b8/concepts/76a25fba-21bc-498a-9d1c-c2dc172a8f4d).

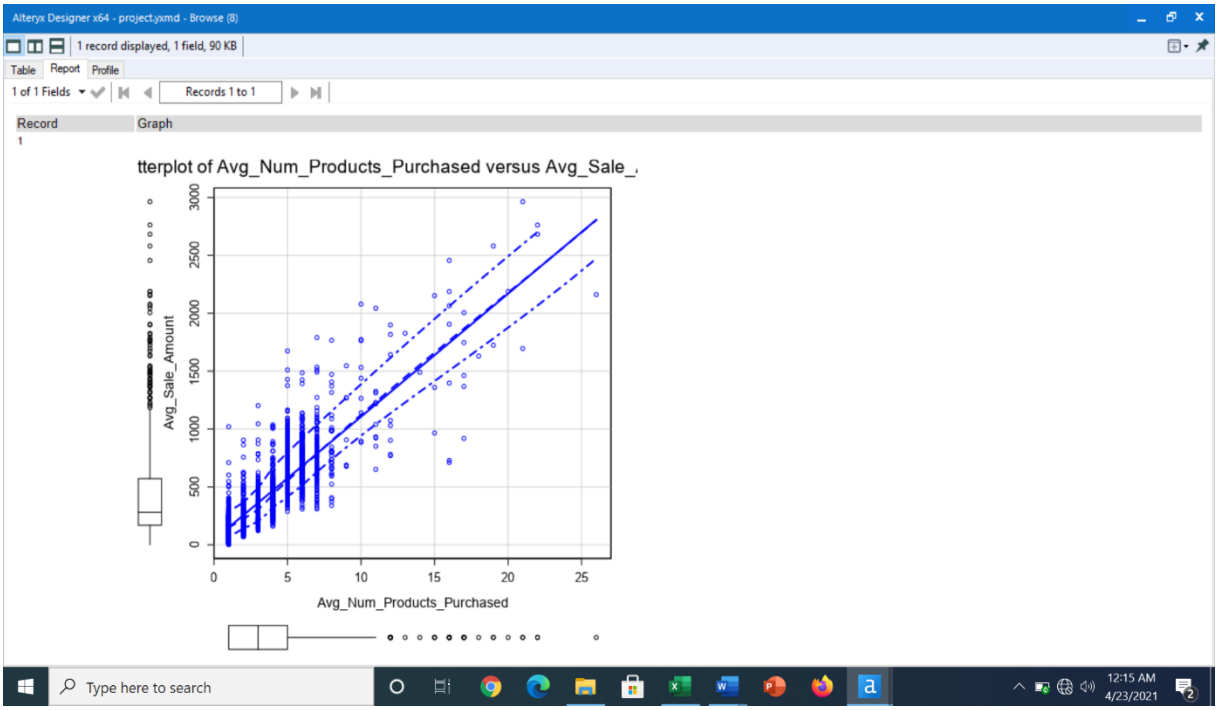

Figure 1.0: Shows a plot of Avg_Num_Product_Purachase (Predictor variable). vs Avg_Sale_Amount (target variable)

Figure 2.0: A Plot of Customer_ID (Predictor variable) vs Avg_Sale_Amout (target variable).

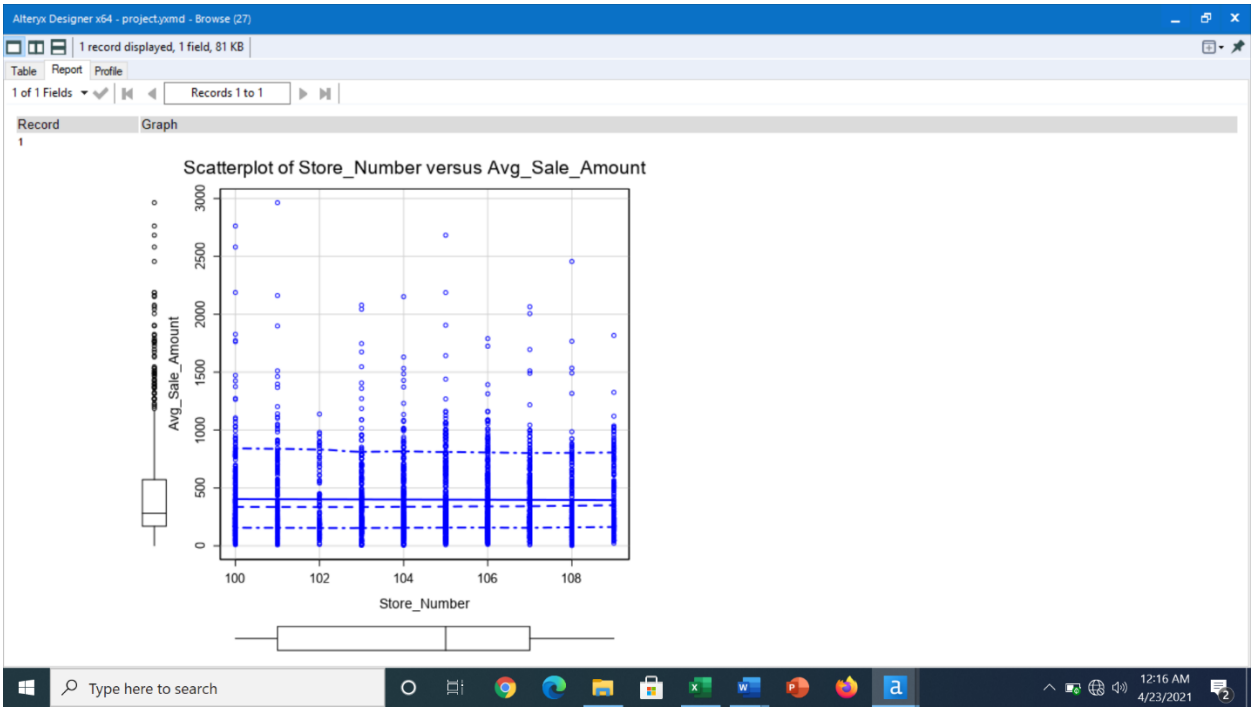

Figure 3.0: A Plot of Store_Number (predictor variable) vs Avg_Sale_Amount (target variable)

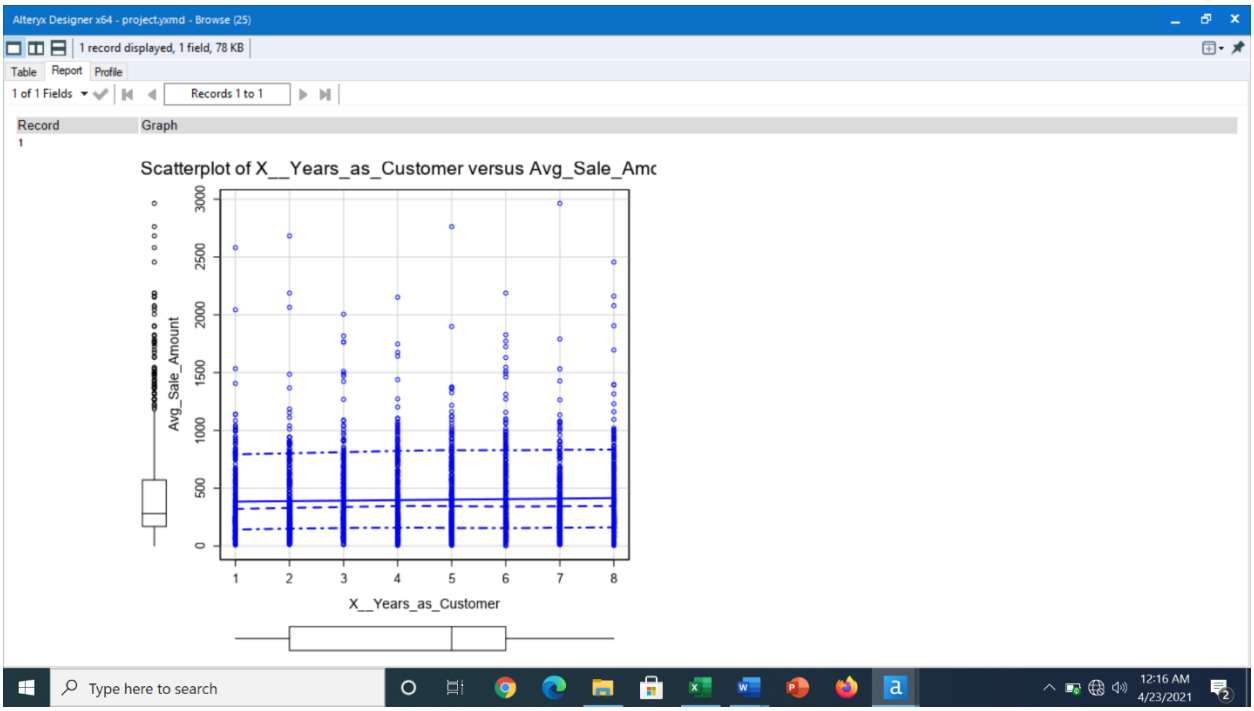

Figure 4.0: A Plot of X_Years_as _Customer (Predictor variable) vs Avg_Sale_Amount (target variable).

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.

Answer:

My linear model is a good model because, From Figure 5.1, it is observed that the predictor variable Avg_Num_Products_Purchased and the categorical variables – Customer_SegmentLoyalty Club only, Customer_SegmentLoyalty Club and Credit Card and Customer_SegmentStore Mailing List, all have P-value of $2.2e^{-16}$ that are less than 0.05, it shows that it is a good fit for the model. The R-squared value for the model is 0.8369, this figure is an evidence that there is a strong relationship between the target variable and the predictor variables. It also means that all most all the variance is explained by the model since the percent variance is about 84%. In addition, the three stars (***), also prove that the model is statistically significance. In general, a model with R-squared above 0.7 is considered a good model. Customer_Segment, due to the nature of the categorical variable, we cannot use a scatterplot or any other graph to see whether a linear relationship exists or not, so we use p-values to justify its selection. In Figure 5.0, Response: Avg_Sale_Amount, it is observed that both the predictor variables, Customer_Segment and Avg_Num_Product_Purchased are selected and others were discarded because, they both have P-values of $2.2e^{-16}$ justifying the fact that it is lesser than 0.05.

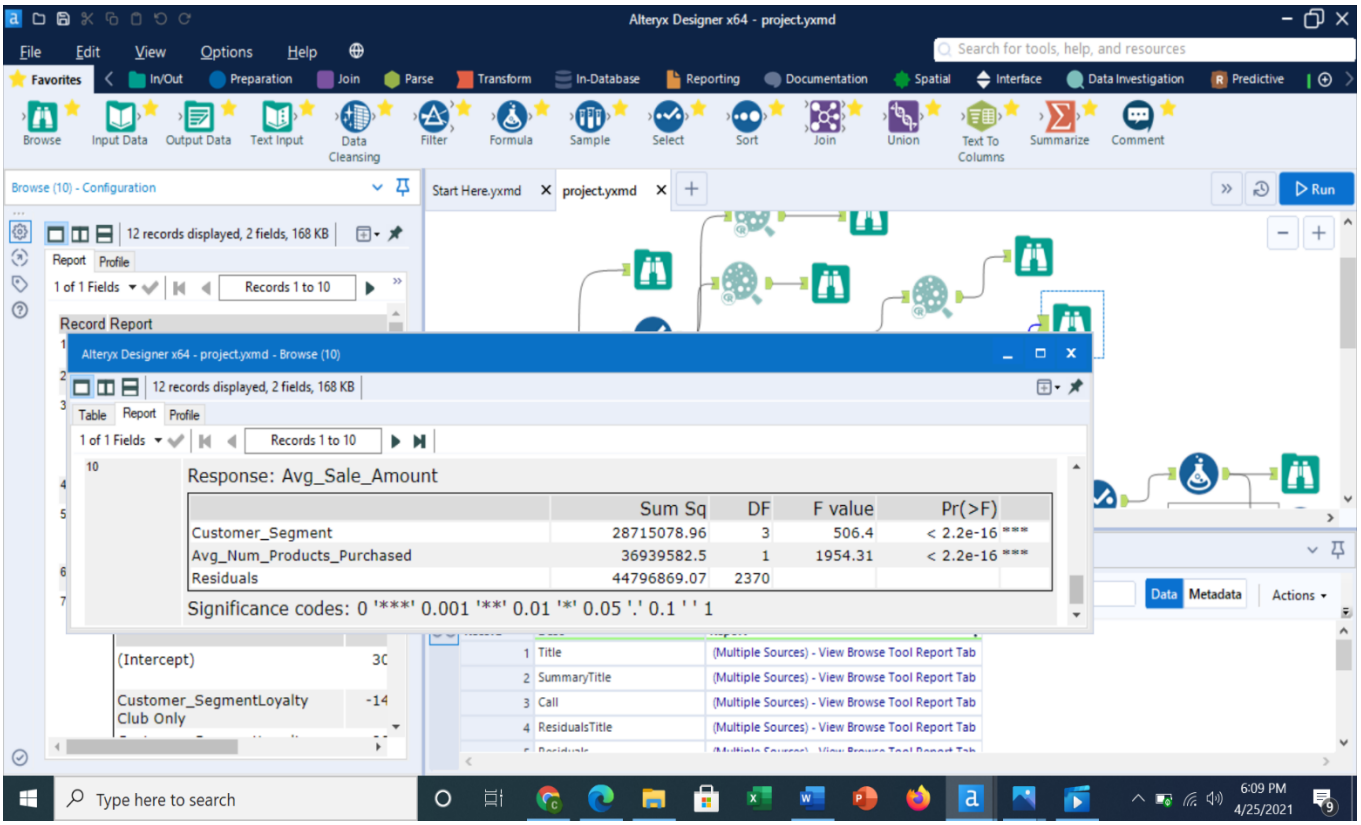

Figure 5.0 Response: Avg_Sale_Amount

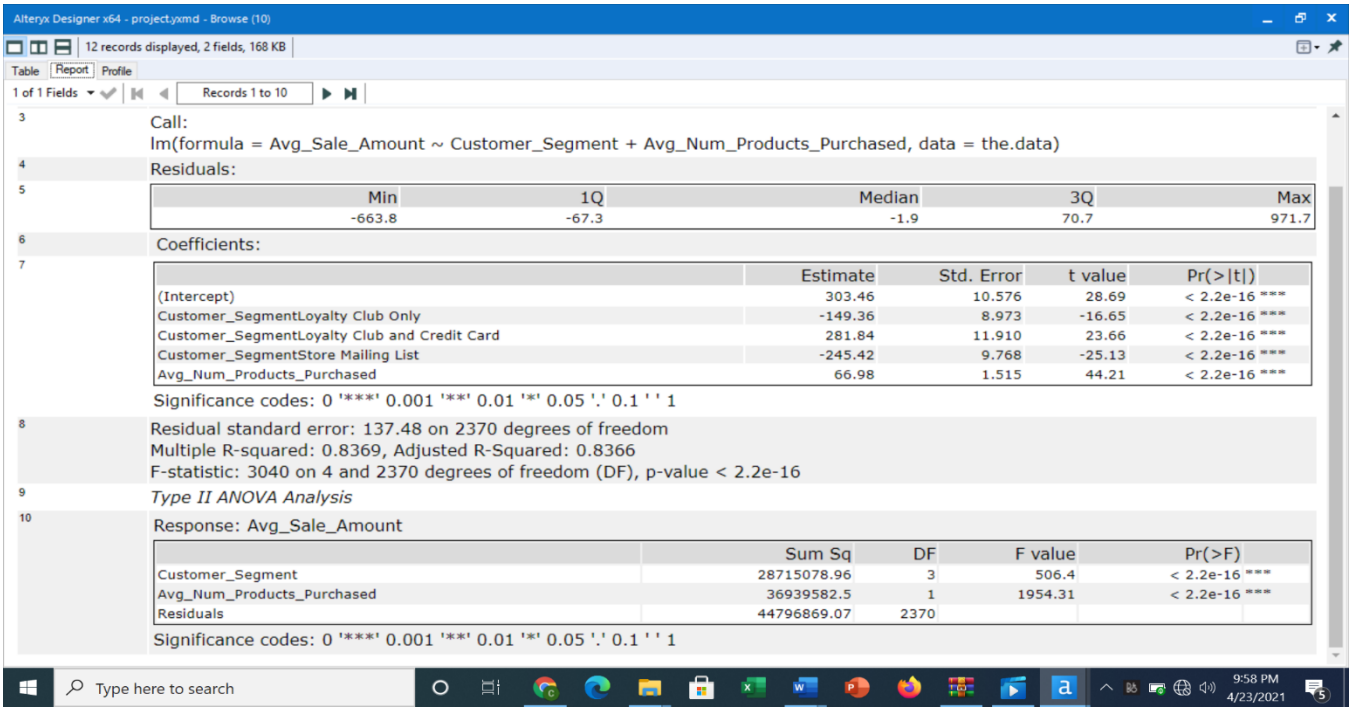

Figure 5.1: Report for Linear Model Linear_Regression

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)

Y = Intercept – b1 * Variable_1 + b2 * Variable_2 – b3 * Variabl_3 + b4 * Variable_4

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) + 66.98 * (Avg_Num_Products_Purchased)

Step 3: Presentation/Visualization

Use your model result to provide a recommendation. (500word limits)

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

Answer:

I recommend that the result obtained from scoring the model is viable. Yes, the company should send the catalog to the 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)

Answer:

The result obtained from scoring the regression model developed to the 250 new customers tested datasets is, the expected revenue generated was calculated by multiplying the probability that the customers will be attracted to the catalog and make purchase, (Score_yes) by predicted amount of each of the 250 new datasets (score) to give the Prdicted_sales. The profit was calculated by multiplying the predicted revenue by gross margin, and then subtracting the cost of printing the catalog form it. The predicted total profit was calculated by summing up the profit across the 250 new customer datasets. The predicted total profit calculated is $21,987.435687, exceeds the expected profit of $10,000 expected by the company. Figure6.0 shows the analysis of the predicted sale.

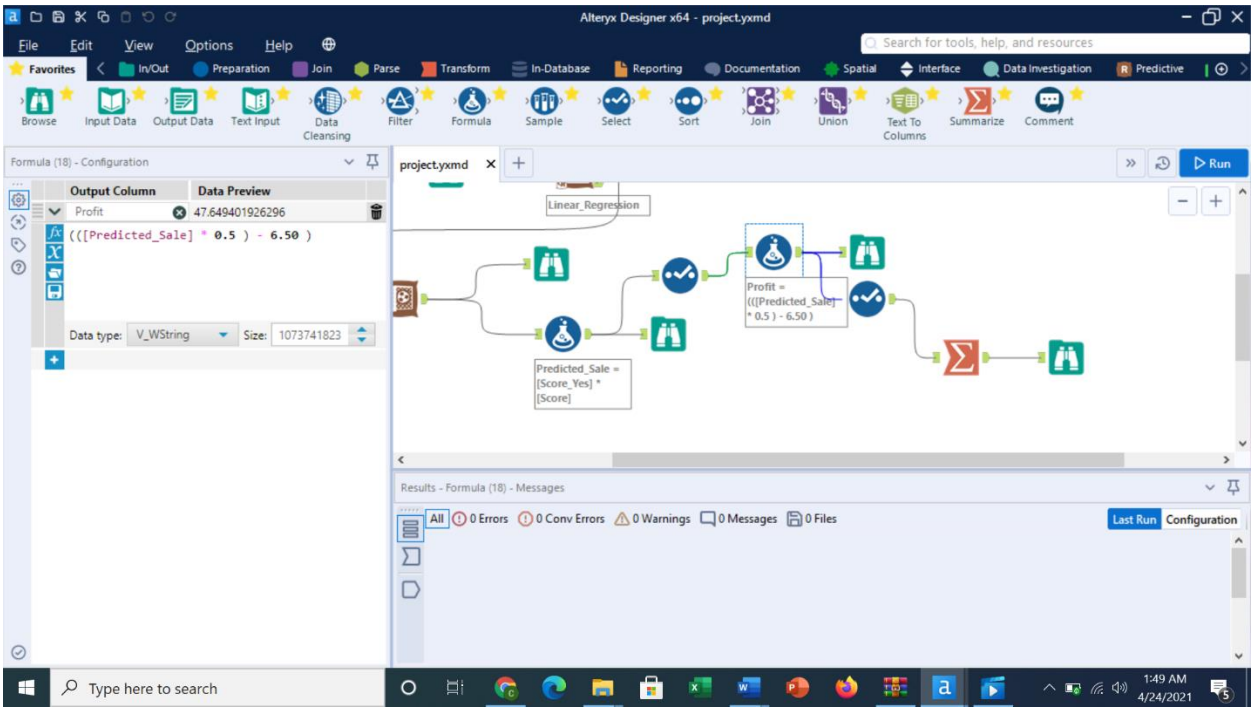

Figure 6.0: Predicted_Sale

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

   Answer:
   The expected profit from the new catalog is $21,987.435687. This is seen in figure7.0 below. The more data we get the more revenue generated for the company.

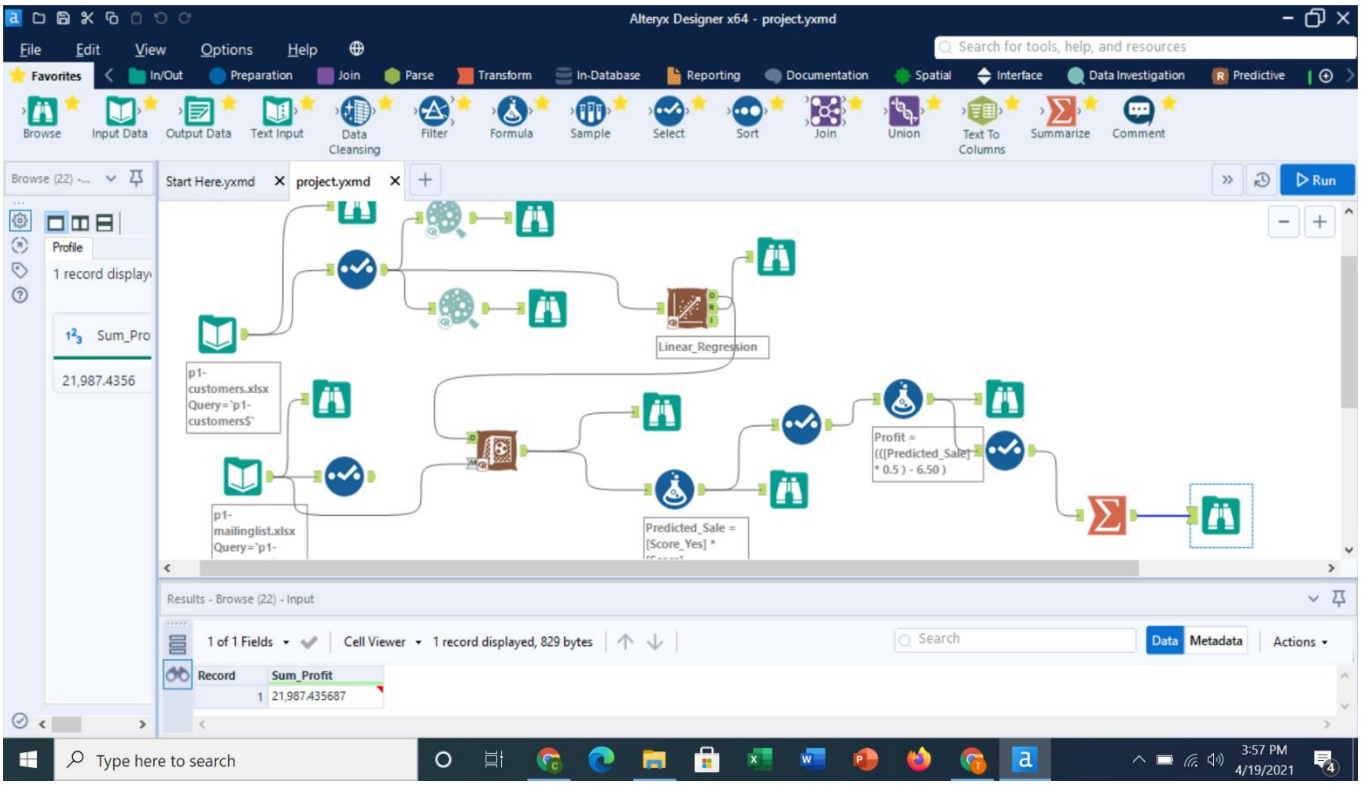

Figure 7.0: Expected Profit.

Conclusion:

In conclusion, it is advisable that the company send the catalog the 250 new costumer because it will yield more profit and increase the company's costumer base.