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?
   - ➔ Whether the expected profit contribution exceeds $10,000 or not ?
   - ➔ To print and send catalogs to new customers or not ?

2. What data is needed to inform those decisions?
   - ➔ The cost of sending catalogue to each customer.
   - ➔ The gross margin on the products sold.
   - ➔ Probability that a customer will buy goods.
   - ➔ Average number of products bought by each customer
   - ➔ Average sale amount spent by each customer.
   - ➔ Categorical variables which in this case only includes , from which segment the customer is.

# 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.

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.

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……*

**For example:** Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

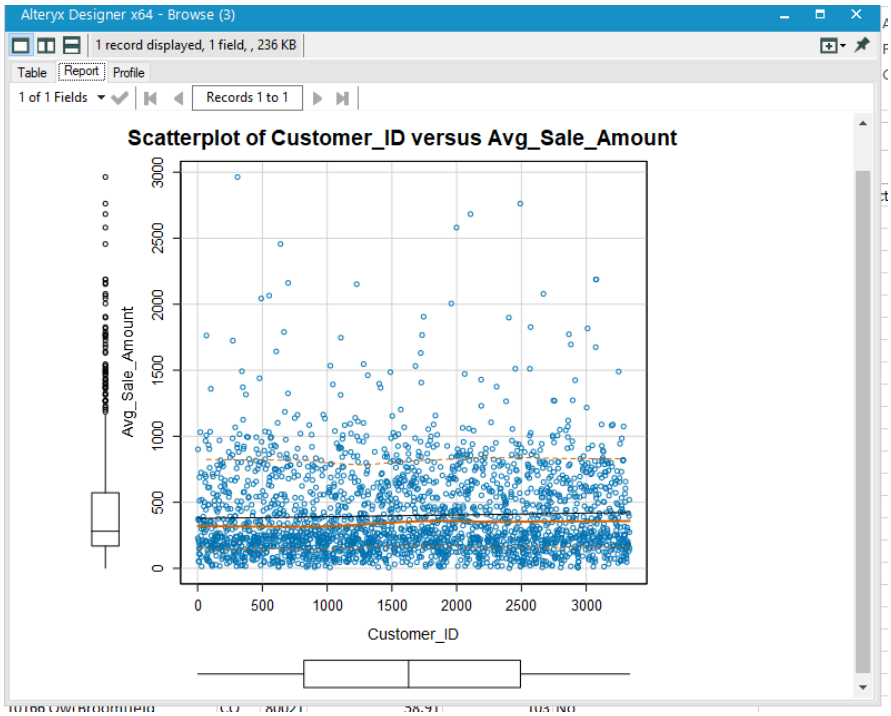
Note that we **must** include the 0 coefficient for the type Cash.

**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.

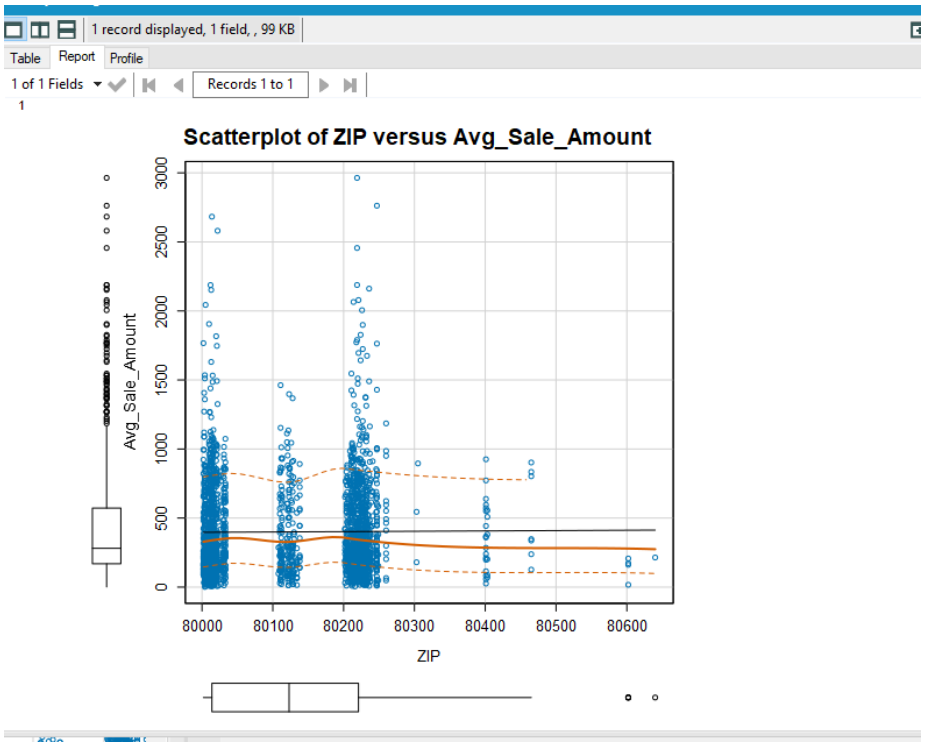
**Solution to this section ->**

- ➔ I have drawn scatter plots for each numeric variable as shown below.
- ➔ Then I have chosen the variable which showed the highest relation with Avg_Sales_Amount.

1) Customer_Id vs Avg_Sale_Amount



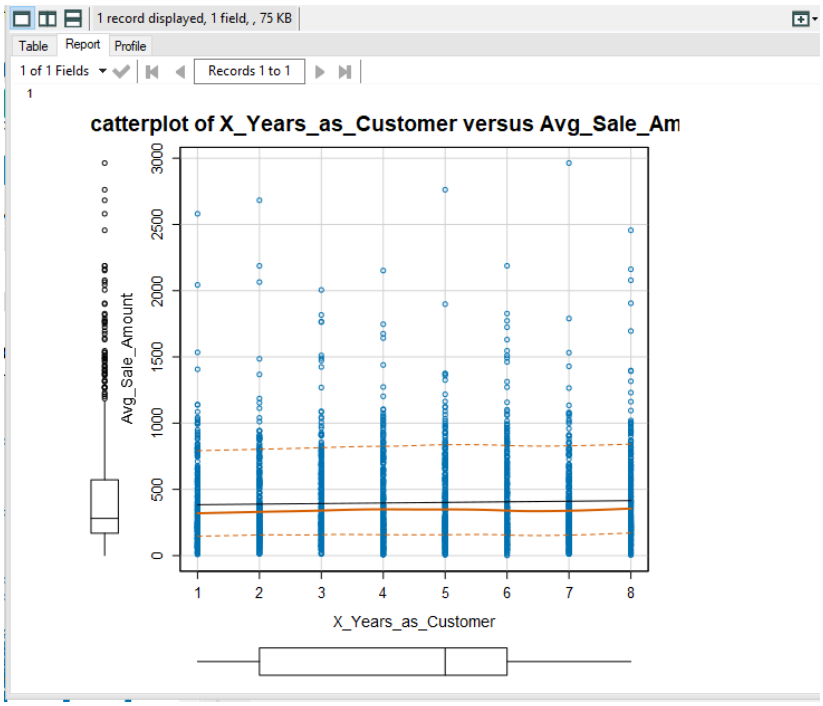
2) ZIP vs Avg_Sale_Amount



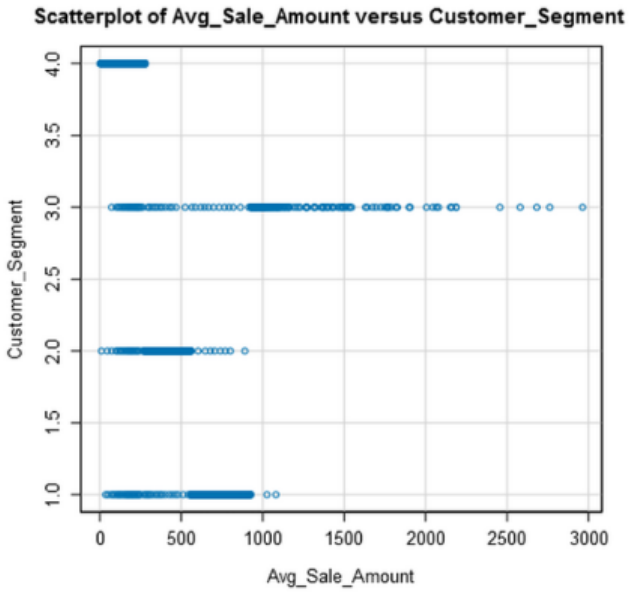

3) Store number VS Avg_Sale_Amount

4) Avg_Num_Products VS Avg_Sale_Amount


5) No of years as customer vs Avg_Sales_Amount

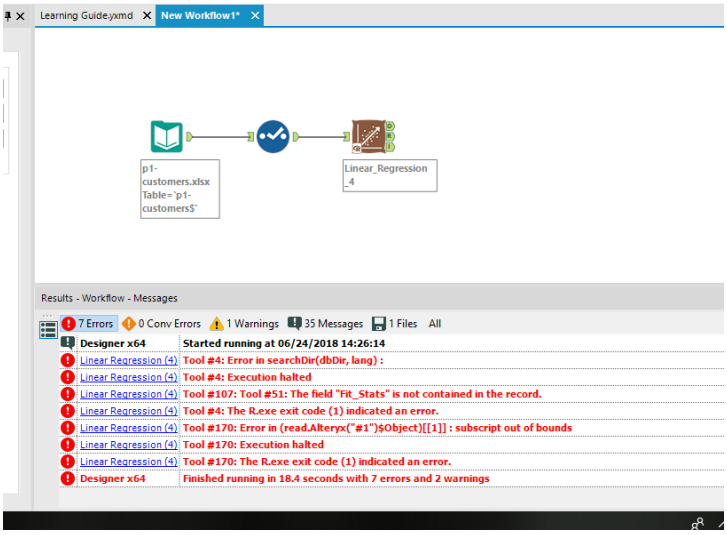


6) Also I have converted the customer segments into numerical values and plotted scatter-plot .

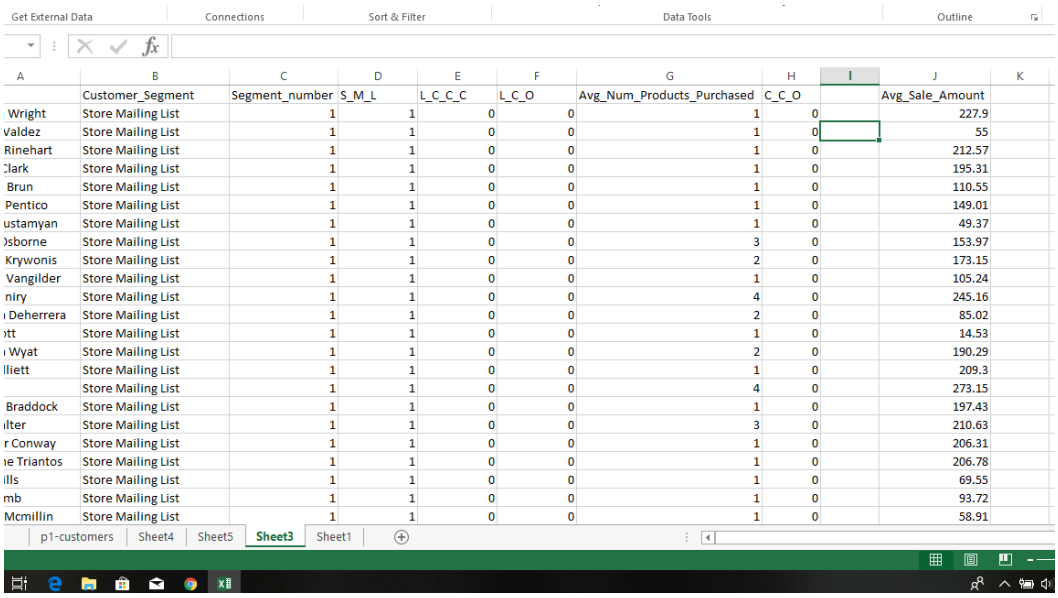



- ➔ Only Avg_Number_Of_Products_Sold seems linearly related with target variable out of the numeric variables.
- ➔ Also the Customer Segment also seems to be related to the target variable.
- ➔ Variable city was not significant according to the regression table.
- ➔ All other categorical variable seems to have no affect on the target variable.
- ➔ On the basis of regression table, “Avg_Number_Of_Products_Sold”, “Customer_Segment”, both are very much significant.

My Alteryx version is not totally working proper , I was facing issues with my regression feature as shown below :

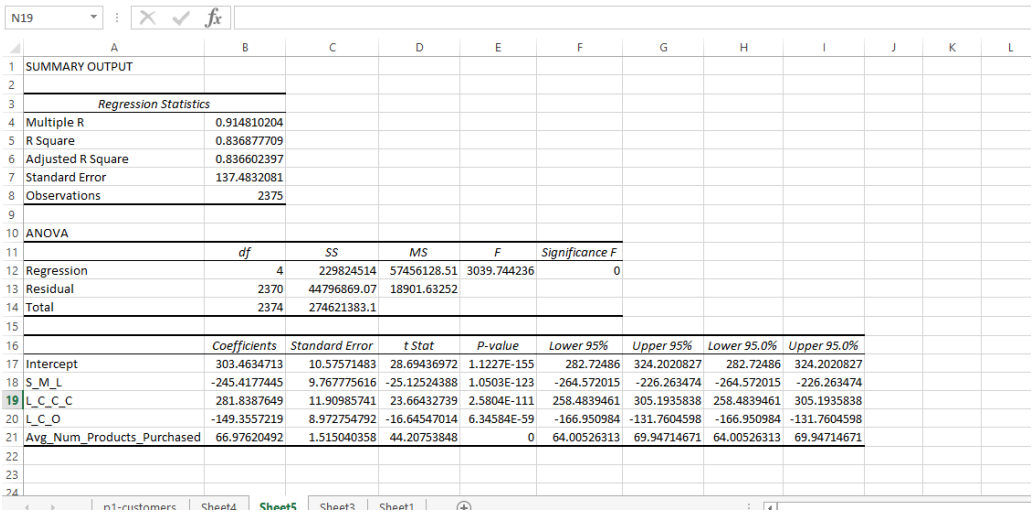


➔ So I have used Excel to calculate the regression table:-



| | Customer_Segment | Segment_number | S_M_L | L_C_C_C | L_C_O | Avg_Num_Products_Purchased | C_C_O | Avg_Sale_Amount |
|---|---|---|---|---|---|---|---|---|
| Wright | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 227.9 |
| Valdez | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 55 |
| Rinehart | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 212.57 |
| Clark | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 195.31 |
| Brun | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 110.55 |
| Pentico | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 149.01 |
| ustamyan | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 49.37 |
| Osborne | Store Mailing List | 1 | 1 | 0 | 0 | 3 | 0 | 153.97 |
| Krywonis | Store Mailing List | 1 | 1 | 0 | 0 | 2 | 0 | 173.15 |
| Vangilder | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 105.24 |
| niry | Store Mailing List | 1 | 1 | 0 | 0 | 4 | 0 | 245.16 |
| Deherrera | Store Mailing List | 1 | 1 | 0 | 0 | 2 | 0 | 85.02 |
| tt | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 14.53 |
| Wyat | Store Mailing List | 1 | 1 | 0 | 0 | 2 | 0 | 190.29 |
| lliett | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 209.3 |
| | Store Mailing List | 1 | 1 | 0 | 0 | 4 | 0 | 273.15 |
| Braddock | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 197.43 |
| lter | Store Mailing List | 1 | 1 | 0 | 0 | 3 | 0 | 210.63 |
| r Conway | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 206.31 |
| e Triantos | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 206.78 |
| lls | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 69.55 |
| mb | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 93.72 |
| Mcmillin | Store Mailing List | 1 | 1 | 0 | 0 | 1 | 0 | 58.91 |

➔ Regression Table:-



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | | | | | |
| *Regression Statistics* | | | | | | | | | |
| Multiple R | 0.914810204 | | | | | | | | |
| R Square | 0.836877709 | | | | | | | | |
| Adjusted R Square | 0.836602397 | | | | | | | | |
| Standard Error | 137.4832081 | | | | | | | | |
| Observations | 2375 | | | | | | | | |
| ANOVA | | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | | |
| Regression | 4 | 229824514 | 57456128.51 | 3039.744236 | 0 | | | | |
| Residual | 2370 | 44796869.07 | 18901.63252 | | | | | | |
| Total | 2374 | 274621383.1 | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* | |
| Intercept | 303.4634713 | 10.57571483 | 28.69436972 | 1.1227E-155 | 282.72486 | 324.2020827 | 282.72486 | 324.2020827 | |
| S_M_L | -245.4177445 | 9.767775616 | -25.12524388 | 1.0503E-123 | -264.572015 | -226.263474 | -264.572015 | -226.263474 | |
| L_C_C_C | 281.8387649 | 11.90985741 | 23.66432739 | 2.5804E-111 | 258.4839461 | 305.1935838 | 258.4839461 | 305.1935838 | |
| L_C_O | -149.3557219 | 8.972754792 | -16.64547014 | 6.34584E-59 | -166.950984 | -131.7604598 | -166.950984 | -131.7604598 | |
| Avg_Num_Products_Purchased | 66.97620492 | 1.515040358 | 44.20753848 | 0 | 64.00526313 | 69.94714671 | 64.00526313 | 69.94714671 | |

➔ The model is strong since the p-values for the chosen predictor variable is very low .
➔ Also the adjusted r – Square value is 0.8366 which tends that model is very good .
➔ Best Predicted Model:- ( Upto 2 decimal places )

Avg_Sales_Predicted = 303.46 + (66.98* Avg_Number_Products_Purchased)
– (149.36* Loyalty_Club_Only)
+( 281.84* Loyalty_Club_And_Credit_Card)
– (245.42* Store_Mailing_List)
+( 0* 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?
   - ➔ Yes, the company should send catalog to new customers as it will increase the company's profit.
   - ➔ As said, the predicted profit is more than $10,000 so the company should send catalogues to new customers.

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

   - ➔ As mentioned by me earlier, I was not able to use the regression tool in alteryx and it was not solved by the alteryx community.
   - ➔ So,I designed the linear regression formula in Excel.
   - ➔ Rest of the work I have done in Alteryx.
   - ➔ First I designed a new column named "Avg_Sale_Amount" = 303.46 + 66.98 *[Avg_Num_Products_Purchased] -149.36 * [L_C_O] + 281.84 * [L_C_C_C] -245.42 *[S_M_L]
   - ➔ Then , I created the column "prob_sales" which tell the probability of the amount spend by the customer which is calculated by using formula: [Avg_Sale_Amount]*[Score_Yes]
   - ➔ Then column "Final_sale" was created in which the gross margin was deducted and the cost of catalogue for each customer was also deducted. Formula used: Final_sale=0.5 *[prob_sales] -6.50
   - ➔ Using summarize tool, the sum was calculated which comes out to be $21987.9570281758

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

➔ Profit = $21987.95