Project 1: Predicting Catalog Demand

# Step 1: Business and Data Understanding

***Problem: -*** You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.
You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

## Key Decisions:

1. What decisions needs to be made?

*Yes, this is the main decision to be made. Tip: Add the point from above about meeting profit minimum,*

ANS. Should I send the catalog to these new 250 customers or not?

2. What data is needed to inform those decisions?

ANS. To inform the decision whether to send out these catalogs to new 250 customers or not we need to predict the expected profit (After sending out the catalog to new customers.). And to come up with the expected profit, we need to build the linear regression model. Since, we have to predict the profit, so this business problem is to predict outcome. We are provided with the data of old customers, we have past data on the variable (Avg_Sales_amount of each customer) we're trying to predict, so we're data rich. Our target outcome that we're trying to predict is a number, therefore we should use a numeric model.

*Yes, these are the data dependencies of the project and analysis.*

# Step 2: Analysis, Modeling, and Validation

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.

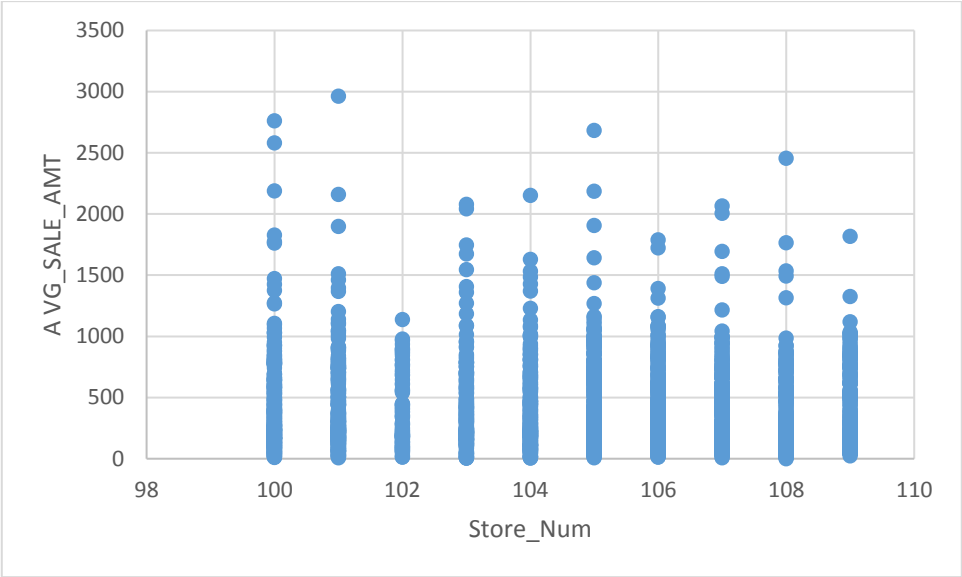
.

ANS:-

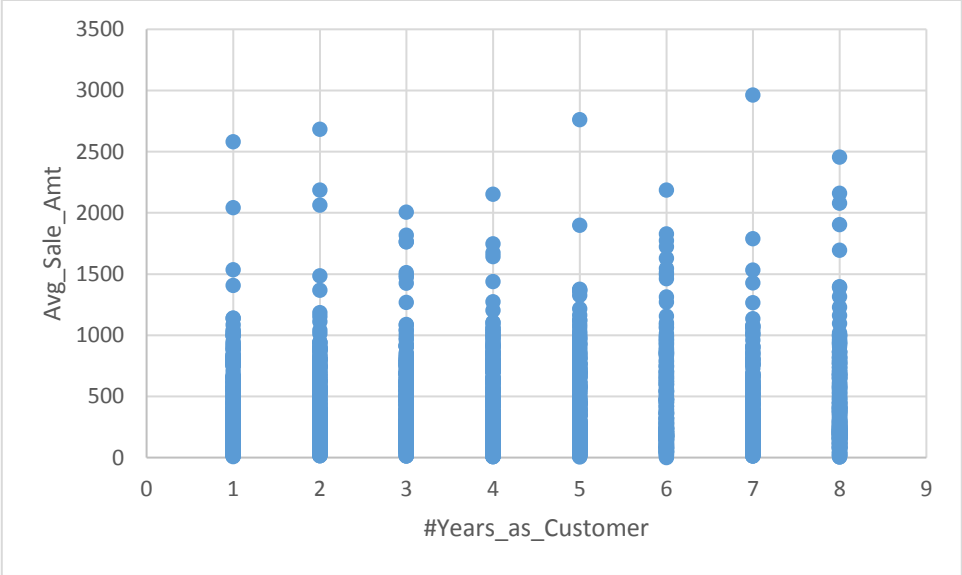
- Our target variable is avg_sales_amt that we will calculate for each customer to predict the total expected profit.
- In the two datasets that are given there are 10 columns that are present in both of them :- Name , customer_segment , customer_id , state, zip ,address ,city , avg_num_products_purchased, store_no, #_Years_As_Customers. Out of these

variables Name, customer_id, state, zip, address, city cannot be predictor variables as they have no relationship with the target variable.

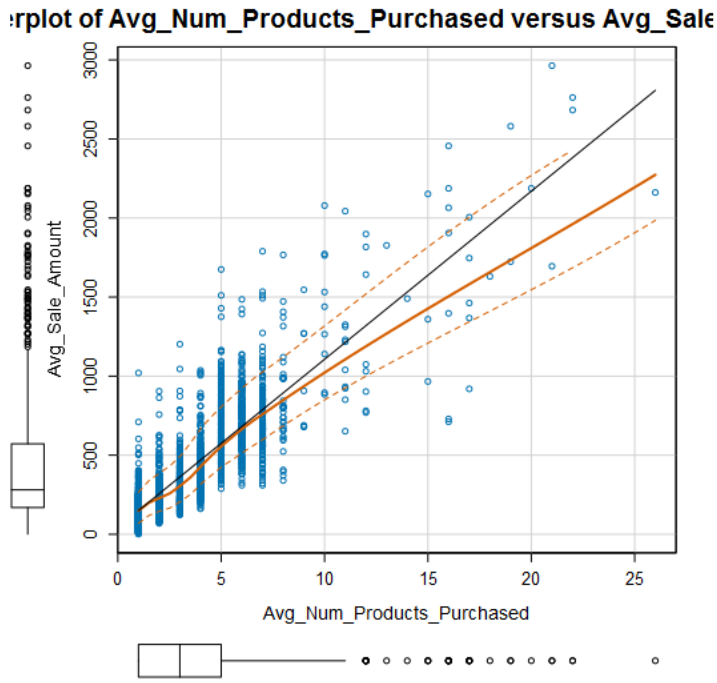
- Store_no and #_Years_As_Customers also be cannot predictor variables as there is no linear relationship between avg_sales_amt and store_no




There is no relationship between avg_sales_amt and #_Years_As_Customers




- Avg_num_products_purchased can be a predictor variable as there exists a linear relationship b/w target variable and avg_num_products_purchased. As avg_num_of_products_purschased increases avg_sales_amt also increases.




- Customer Segment can also be a predictor variable (as it a categorical value so scatter plot cannot be used) as we can see that p-value is less than 0.05 and statistical significance is also high. Therefore , customer_segment is one of our predictor variable

Coefficients:

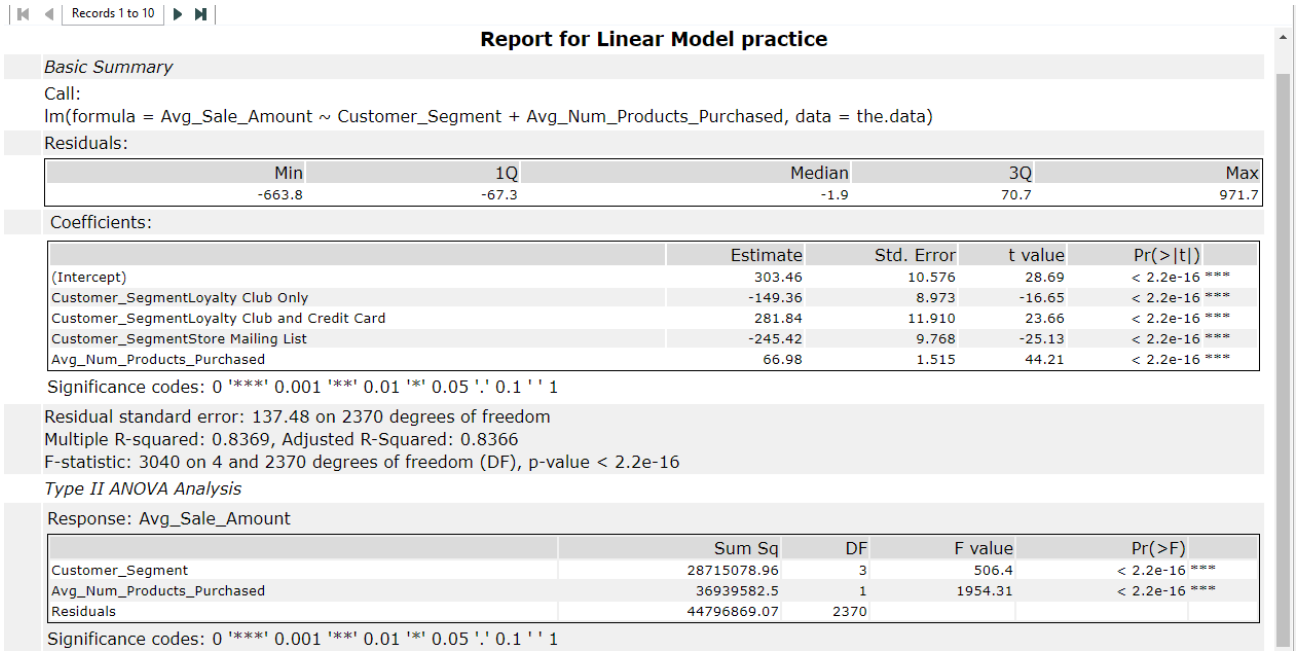
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 303.46 | 10.576 | 28.69 | < 2.2e-16 | *** |
| Customer_SegmentLoyalty Club Only | -149.36 | 8.973 | -16.65 | < 2.2e-16 | *** |
| Customer_SegmentLoyalty Club and Credit Card | 281.84 | 11.910 | 23.66 | < 2.2e-16 | *** |
| Customer_SegmentStore Mailing List | -245.42 | 9.768 | -25.13 | < 2.2e-16 | *** |

*Great job! Predictor selection is well justified using statistical values and plots*

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

- We selected two variables as predictor variables: - customer segment and avg_num_products_purchased. After building the linear model using these predictor variables, the statistical results show that both of our predictor variables have p-value less than 0.05 and their statistical significance is also very high. R-squared value is also greater than 0.7 i.e. 0.837 which means that nearly all the variance in the target variable is explained by the model.

*Excellent- you have explained the model well using the p-values and R-squared values.*



Records 1 to 10

**Report for Linear Model practice**

*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

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 * Variable_3……*

Y = 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).

Base case is set to Credit Card Only.

*Equation is well formulated and complete. Nice work!*

# Step 3: Presentation/Visualization

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

   - After doing all the necessary calculations, the expected profit is $21987.436 which greater than $10,000 . So, the company should send the catalog to these new 250 customers.

   *Exactly - your recommendation is specific and valid.*

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

- After calculating sales_amount for each customer using linear regression model, we multiply the each amount calculated for each customer with the probability that the customer will buy from the product from catalog. After this we again multiply the each result (obtained after first multiplication) with the gross margin 0.5 (50%). After second multiplication we subtract from each result the cost of printing and sending the catalog i.e. $6.50 from each result for each customer. And at last we add all the results and obtained the expected profit $21987.436.

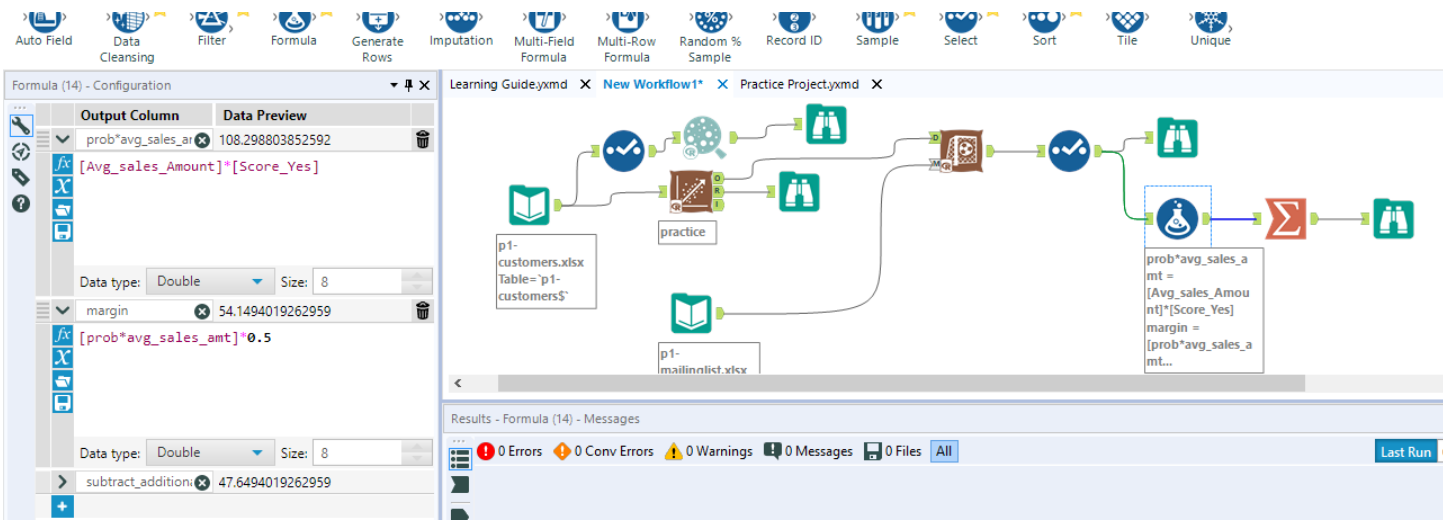
*Steps to calculate profit are well outlined.*

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

- The expected profit is $21987.436.

*Profit value is correct.*

MY ALTERYX WORKFLOW:-




*Good job including the workflow.*