Step 1: Business and Data Understanding

"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.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

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. "

Here, in this analysis project, the main aim is to find "How much profit can we expect from the new 250 customers, if we send our printed catalogue to them?" And this expected profit should be more than $10.000 when summing up profits from all 250 customers. Because, as instructed by the manager "Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000"

DATASET NEEDED TO PERFORM ANALYSIS:

- Historical Customer Dataset – which includes customer's behavioral related features, which either directly or indirectly correlated to the profit. A features which contains customer's historical buying patterns, frequency of their purchases, profits they have contributed to the company and a few other features which has a relationship with the profits earned.

  THE FEATURES IN THE PROVIDED HISTORICAL DATASET ARE

  ```
  'Name', 'Customer_Segment', 'Customer_ID', 'Address', 'City',
  'State', 'ZIP', 'Avg_Sale_Amount', 'Store_Number',
  'Responded_to_Last_Catalog', 'Avg_Num_Products_Purchased',
  '#_Years_as_Customer'
  ```

  THE FEATURES WHICH MAY CONTRIBUTE TO THE PROFITS BASED ON THEIR NATURE:
  8 features are selected that may contribute to the profit of the company.
  [] Customer Segment
  [] City
  [] State
  [] ZIP
  [] Average Sale Amount [] Average Number of Products Purchased
  [] Responded to Catalog [] # Years as a customer

AFTER FILTERING FEW FEATURES FROM THE ABOVE BY COMAPRING THE FEATURES PROVIDED IN TEST DATASET:

6 features are selected by filtering out **2** features.
[] Customer Segment
[] City
[] State
[] ZIP
[] Average Number of Products Purchased
[] # Years as a customer

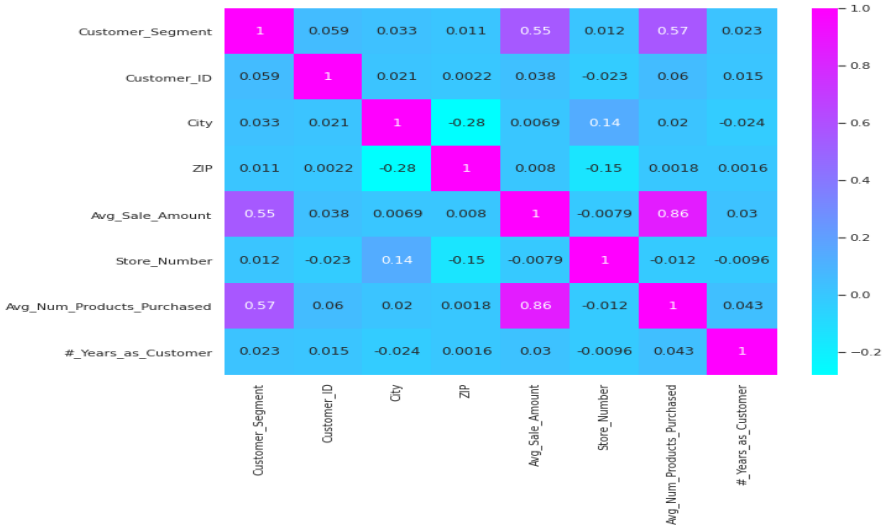
The above 6 features are chosen from the before 9 features by comparing the features of historical dataset and the new customer dataset. The above 6 features **may contribute** to the profit earned from the customers.

Apart from the above 6 features, additional features such as "Score_Yes" & "Average_Sale_Amount" can also be used.

Step 2: Analysis, Modeling, and Validation:

The above six features are then analyzed to see whether they contribute to the profit of the company or not before we use them to build our regression model. Below are the visualization plots which I've built to analyze the linearity between each of the six features with respect to the Avg_Sale_Amount - a dependent variable which we have to predict in order to derive profit from it.

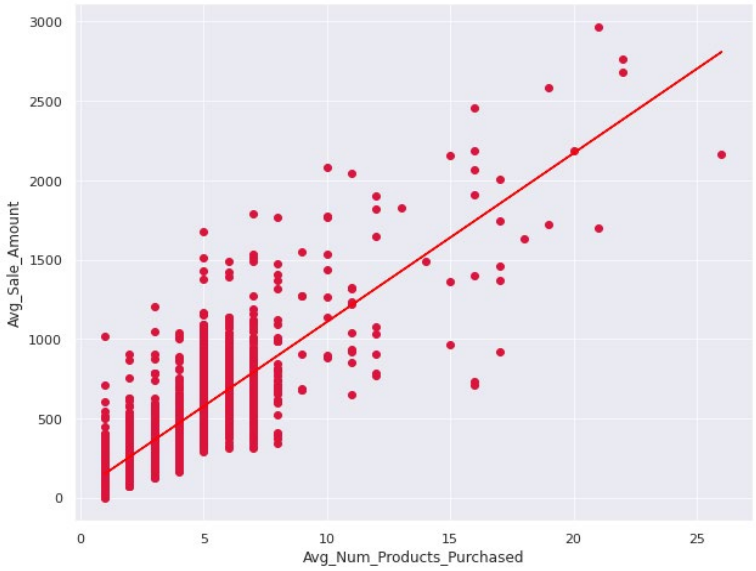
Visualizing Correlation using Heat map



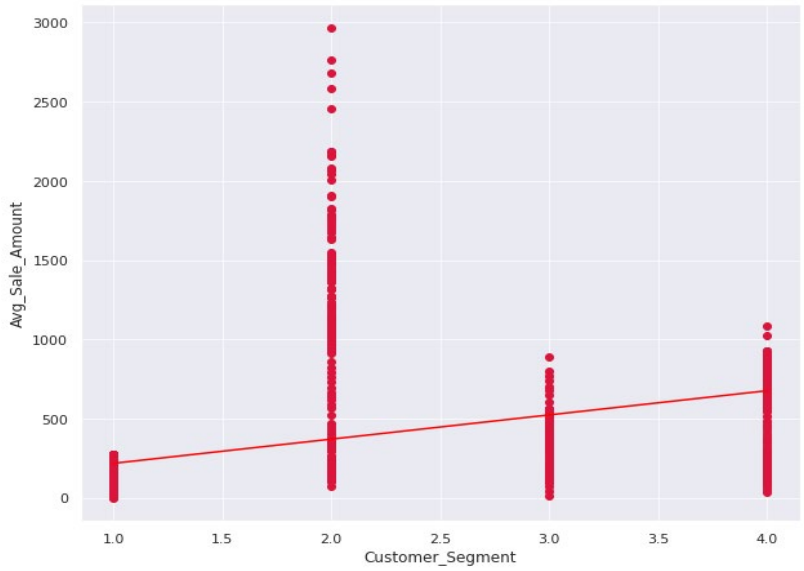
From the above correlation heatmap, it is clear that the dependent variable **"Avg_Sale_Amount"** shares a good relation with features **"Customer_Segment"** and **"Average_Number_of_Products_Purchased".**

Now, let's visualize the linear relation between two correlated variables with respect to **"Avg_Sale_Amount"**

Average number of products purchased w.r to Average sale amount:

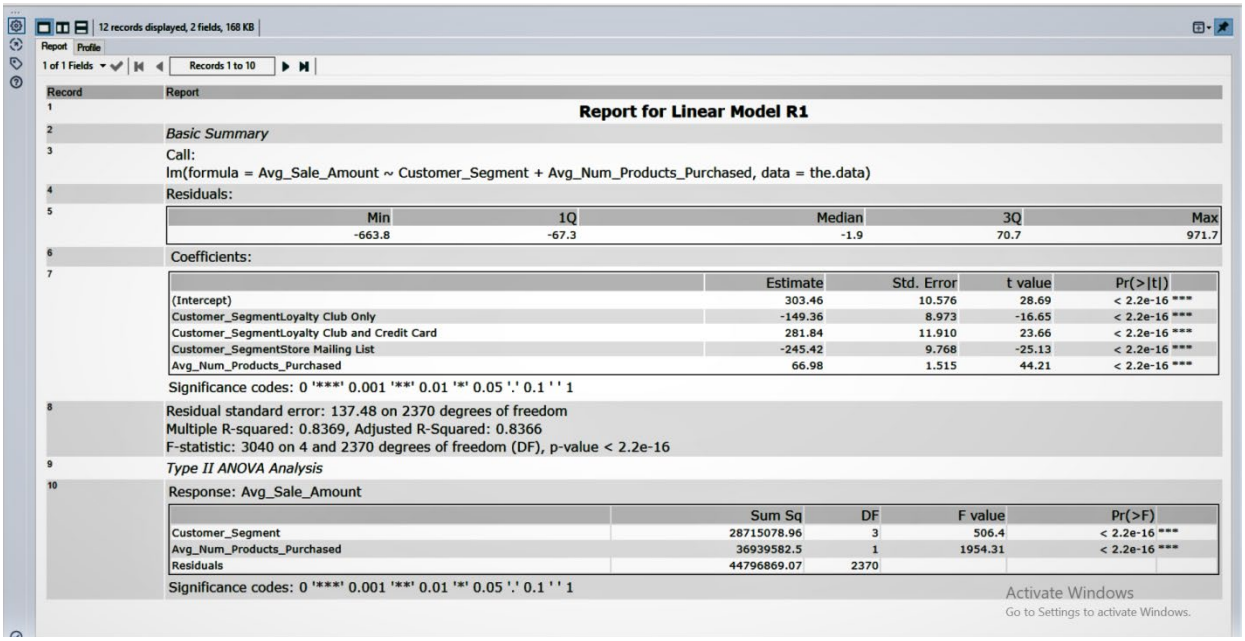


Customer Segment w.r to Average sale amount: (plotted by converting categorical data to numerical data)



From the above graphs it is very much clear that both the “Customer_Segment” and “Avg_Num_Products_Purchased” shares a good linear relationship with “Avg_Sale_Amount”.

Here is the further analysis we have done using Alteryx to better understand their statistical relation.



12 records displayed, 2 fields, 168 KB

Report | Profile

1 of 1 Fields | Records 1 to 10

Report for Linear Model R1

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

Activate Windows
Go to Settings to activate Windows.

INTERPRETATION OF ABOVE REPORT:

- Adjusted R squared – it is a statistical method which signifies how well our fitted regression model is? The ideal r-squared value is 1. The more the r - square value closer to 1, the better our regression model is. The adjusted r-squared value of our fitted model is 0.8366, which is very closer to 1. Hence, our regression model is very highly significant.

- P – value in our model is <2.2e-16, which is less than the significant level of 0.05. Hence we can conclude that there is very good relationship between all the features which we have selected to build our model.

- Linear Regression Equation:

  Prediction (Average_Sales) = 303.46 – 149.36(Customer_SegmentLoyality 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

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

2. **How did you come up with your recommendation?**

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

Yes, the company should send the catalogue to these 250 customers as our linear regression model predicts the expected profit from sending these catalogues to all 250 customers to be **$21987.9570**, which is more than the threshold **$10,000** dollars said by the manager.

Step 1: Average Sales amount for 250 customers are predicted with the help of our regression model.

Step 2: All the predicted average sales amount values for each 250 customers are then multiplied by the probability of customers buying that product ("Score_Yes").

Step 3: Then, from the obtained probability average sales values, I predicted the profit values for each 250 customers by using the below formula

$$\textbf{Profit} = (\text{Prob\_predicted\_sales\_amount} * 0.5) - 6.50$$

Step 4: Then I summed up all the profits obtained from 250 customers and the resulted overall profit of sending catalogue to all 250 customers are **$21987.9570**, which is more than **$10,000** said by the manager.

TOOLS USED: Python & Alteryx