Dhuha Baqarish

PAND

Project 1: Predicting Catalog Demand

## **Step 1: Business and Data Understanding**

### **Key Decisions:**

*Answer these questions*

**1. What decisions needs to be made?**

The decision that needs to be made is to send the catalog to 250 clients or not, based on the profit that will be calculated .

**2. What data is needed to inform those decisions?**

We are given two files of dataset ( customers.xlxs and mailing.xlsx. ), From this two files we need :

Avg\_Num\_Products\_Purchased, Customer Segment, Score\_Yes.

In addition to :

- cost of catalogue and that equals($6.50)

- gross\_margin (50%) to find the profit.

## **Step 2: Analysis, Modeling, and Validation**

***Important:******Use the p1-customers.xlsx to train your linear model.***

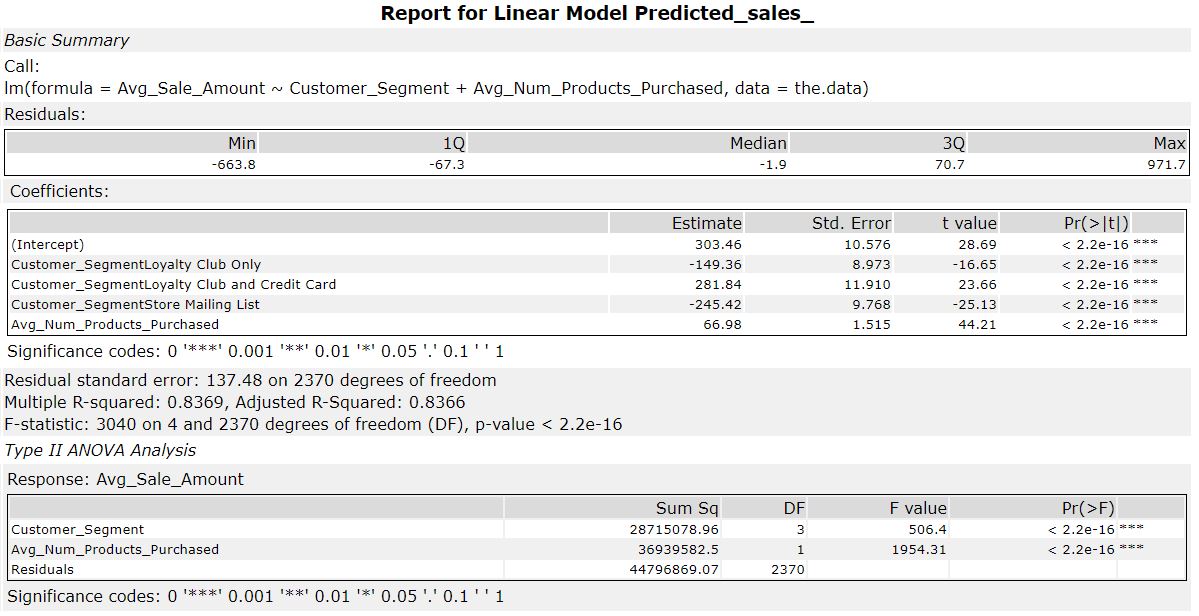
*At the minimum, answer these questions:*

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

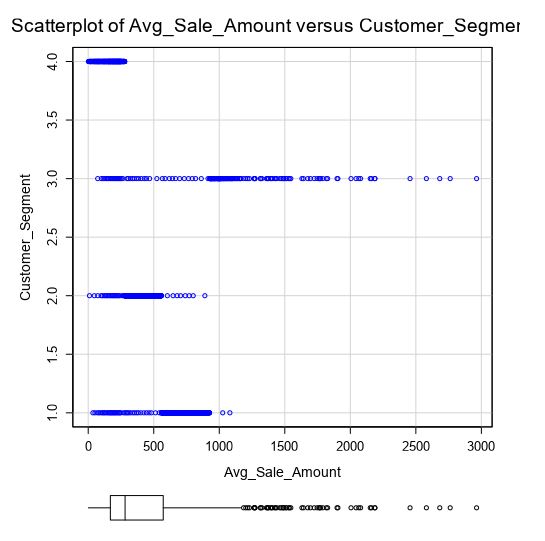
The target variable for the analysis is Avg\_Sale\_Amount .

And the predictor variables selected for the model are Customer\_Segments and Avg\_Num\_Products\_Purchased .

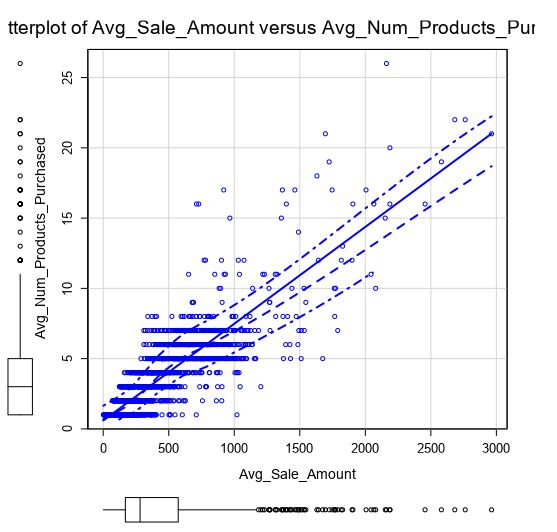
The reason for we selected this two variables as predictor variables is because their p-value less than 0.05 which that mean these two variables are statistically significant.



The relationship between Avg\_Sale\_Amount and Customer\_Segments represent by scatterplot :



The relationship between Avg\_Sale\_Amount and Avg\_Num\_Products\_Purchased represent by scatterplot :



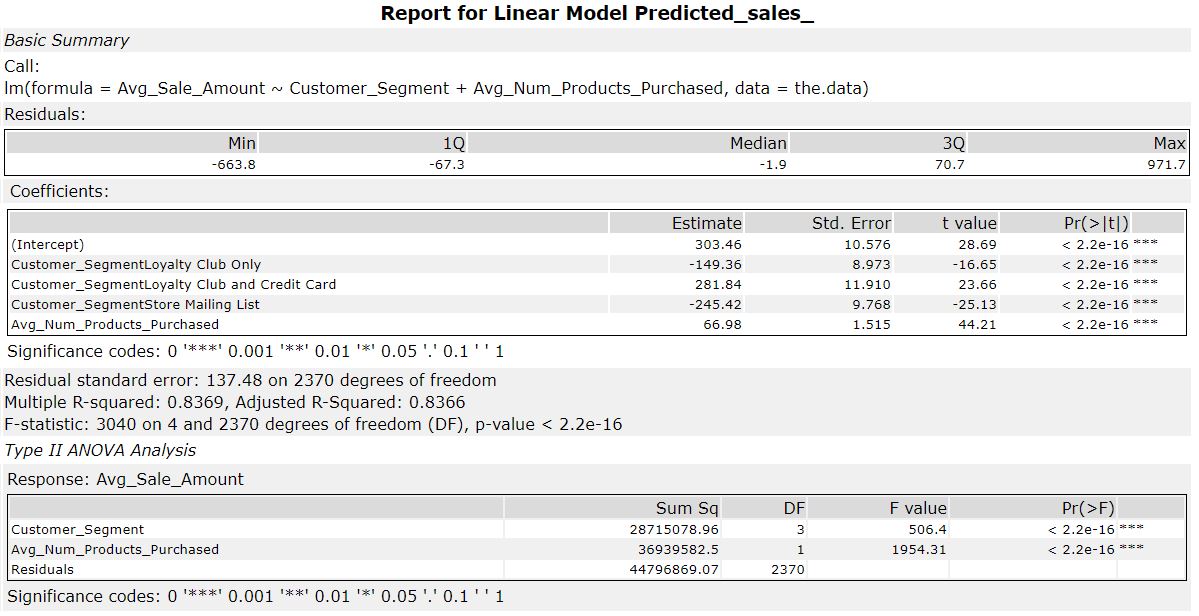
**2. Explain why you believe your linear model is a good model.**

As shown below :

- The Customer\_Segment and Avg\_Num\_Products\_Purchased have p-values less than 0.05.

- The Adjusted R Squared value isequal 0.8366 which is quite a large value.

This mean that our model is a good model because p-values and R-Squared value is statistically significan



**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)**

**The regression equation form:**

*Y = Intercept + b1 \* Variable\_1 + b2 \* Variable\_2 + b3 \* Variable\_3……*

Avg\_Sales\_Amount = 303.46 + ( -149.36 \* Customer\_Segment :Loyalty Club Only ) + ( 281.84 \* Customer\_Segment :Loyalty Club and Credit Card ) + ( -245.42 \* Customer\_Segment :Store Mailing List ) + ( 66.98 \* Avg\_Num\_Products\_Purchased )

## **Step 3: Presentation/Visualization**

*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 these catalogues to these 250 customers.

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

I will explain the process in steps ::

1. calculated predicted\_sales\_amount using the linear regression and score tools (linear regression model) .
2. After that, I created a new column Predicted\_Average\_Sales = predicted\_sales\_amount \* Score\_Yes , by using formula tool .
3. Then the profit is calculated with the given margin to be 50% and cost of each catalogue as $6.50, for all the 250 customers , sush as :

Profit = ([Profit\_avg]\*0.50)-(250\*6.50)

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

Profit = ([Profit\_avg]\*0.50)-(250\*6.50) = 21987.4356865455 $

**Alteryx Workflow :**

