**AYAME GOD’SWILL CLAUDE**

**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?

2. What data is needed to inform those decisions?

**Answers:** Ultimately, the company needs to decide whether or not it should send a catalog to the 250 new customers.

To make this decision, the company needs to determine how much money it can expect to earn from sending out a catalog to the new customers. It needs to predict the potential profit of sending out the catalog. The expected profit contribution must exceed $10,000, otherwise the catalog will not be sent out.

To make this prediction, we need a historical dataset (p1-customers.xlsx) to build a linear regression model, which will then be applied to the new dataset (p1-mailinglist.xlsx) and used to predict the potential profit of sending out the catalog for the 250 new customers.

From the p1-customers.xlsx data set, we can initially assume the followings:

Target variable: Average\_Sale\_Amount

Predictor variables: Name, Customer\_Segment, Customer\_ID, Address, State, City, Zip, Store\_Number, Avg\_Num\_Products\_Purchased, #\_Years\_as\_Customer.

**NB:** Because the field **Responded\_to\_Last\_Catalog**is only contained in the p1-customers.xlsxbut not in the p1-mailinglist.xlsx, it was not used **in the linear regression model since it could not be applied to the mailing list data set.**

## 

## **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 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.

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……*

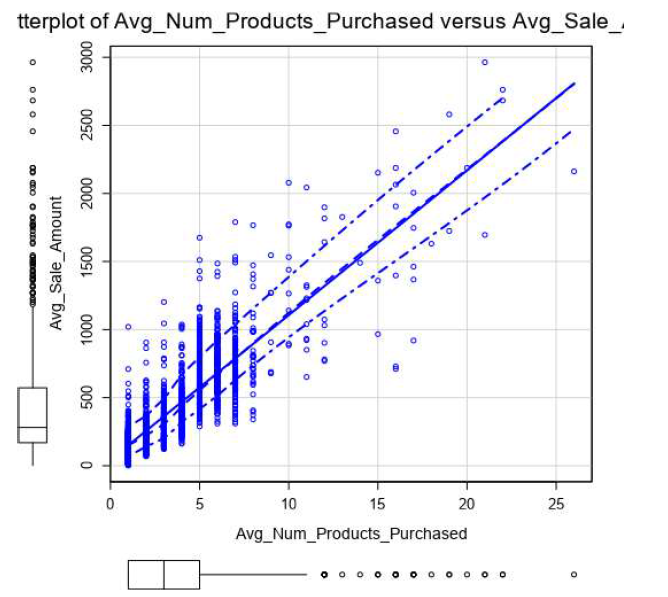
**For example:** Y = 482.24 + 28.83 \* Loan\_Status – 159 \* Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

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

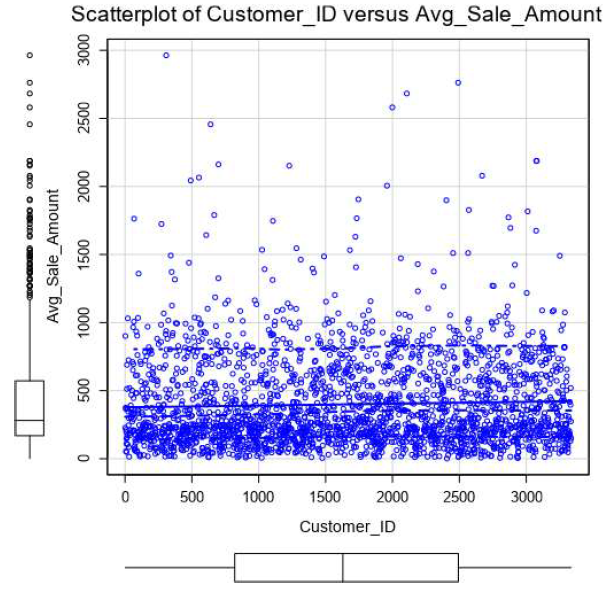
**Note**: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

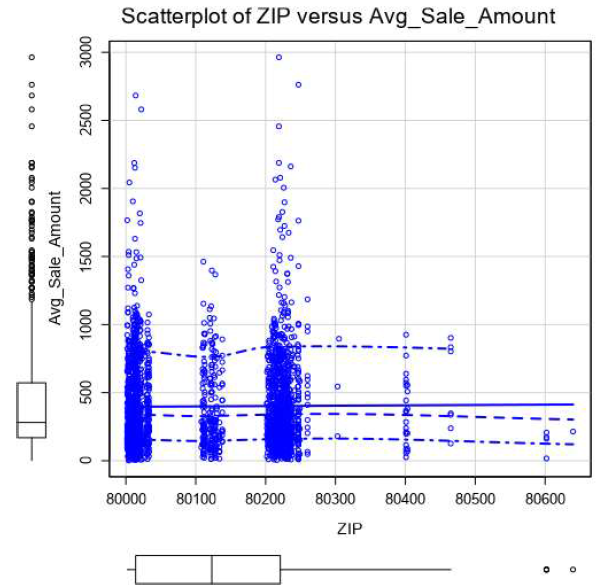
**Answers:** Customer\_ID, Zip, Store\_Number, Avg\_Num\_Products\_Purchased, #\_Years\_as\_Customer are all numeric variables. Hence, scatterplot was used to check their usefulness in predicting the target variable. The followings were discovered:

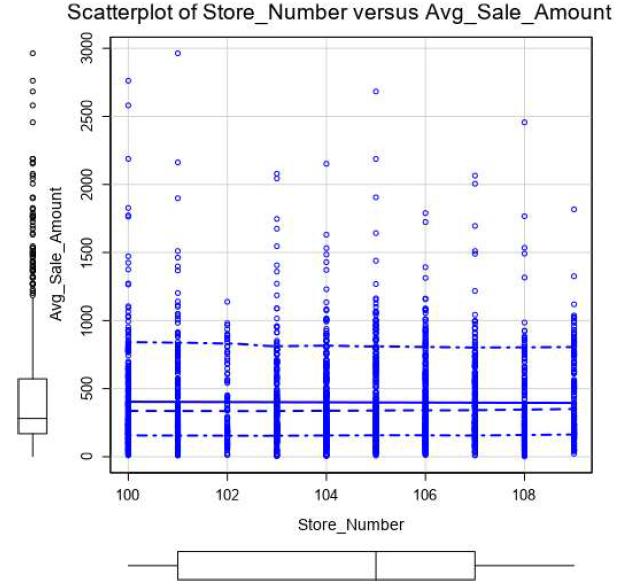
* The scatterplot of the predictor variable Avg\_Num\_Products\_Purchased versus the target variable Average\_Sale\_Amount shows a sloped line. More so, the linear regression trial shows a p-value of 2.2e-16 (which is far less than 0.05) and with three stars to the right Hence, a strong linear relationship exists between the predictor variable Avg\_Num\_Products\_Purchased and the target variable and will be used in the model.

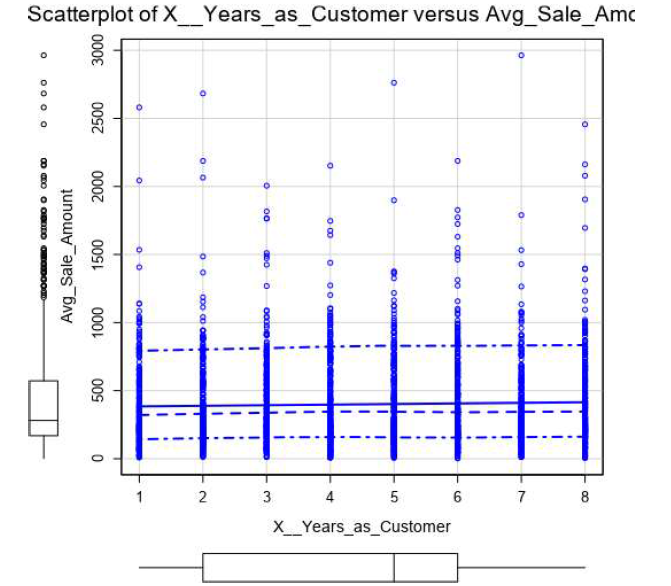
****

* The scatterplot for the rest of the numeric variables (Customer\_ID, Zip, Store\_Number, #\_Years\_as\_Customer) versus the target variable Average\_Sale\_Amount show no slope. Hence, no linear relationship exists between these variables and the target variable. More so, investigations carried out by plugging these variables into the linear regression tool gave p-values greater than 0.05 with no stars to the right. As such, they are not good predictor variables for the target variable and will not be included in the model.

****

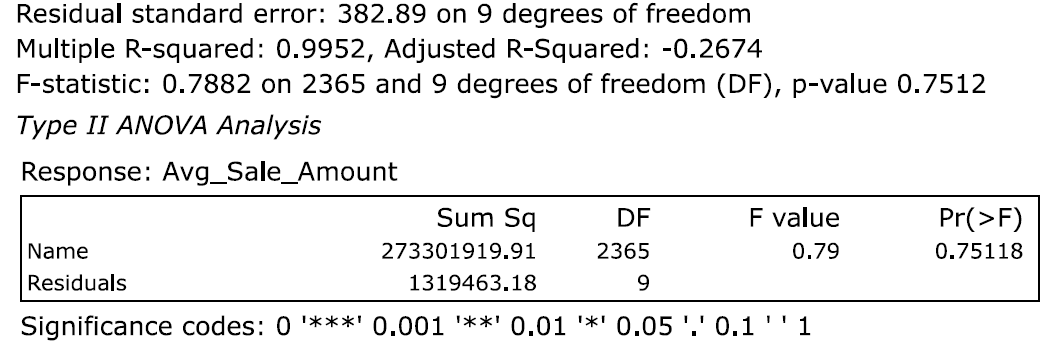




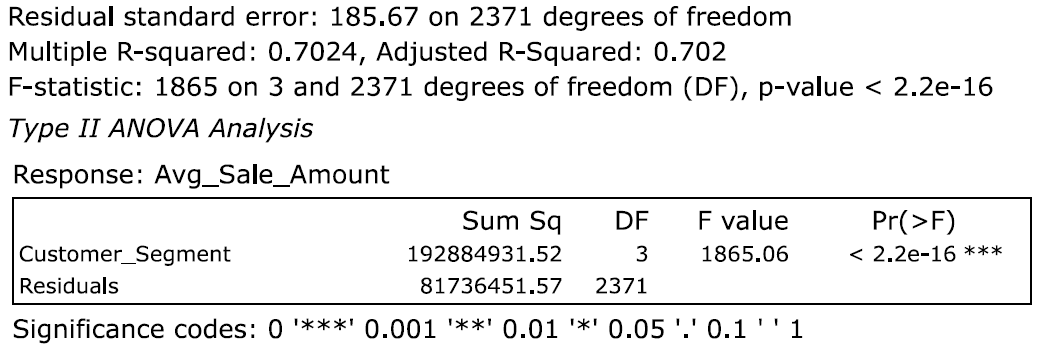


As for the categorical variables Name, Customer\_Segment, Address, State and City, the linear regression tool in Alteryx was used as a trial-and-error measure to determine which is/are statistically significant, one after the other, and the following results were obtained:

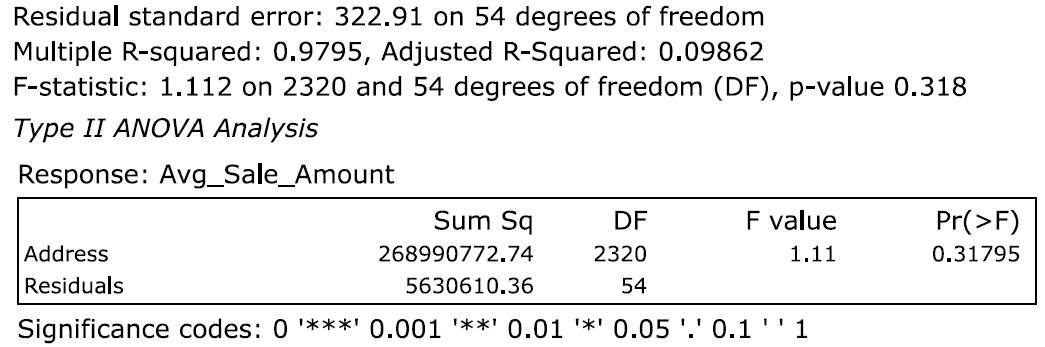
* For Name, the p-value is 0.7512 (which is greater than 0.05) and with no stars. Hence it is not statistically significant.



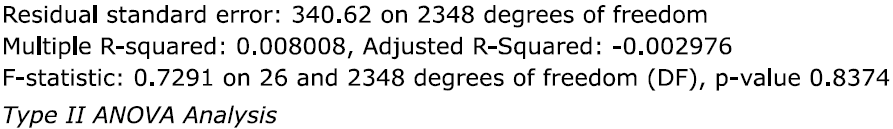
* For Customer\_Segment, the p-value is 2.2e-16 (which is far less than 0.05) and with three stars to the right. Hence it is very much statistically significant.

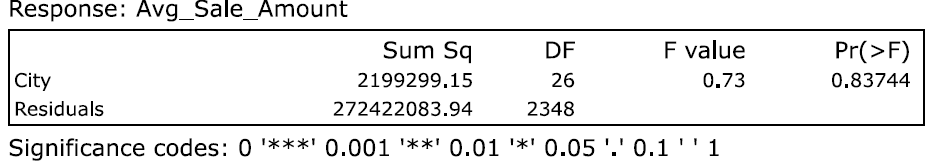


* For Address, the p-value is 0.318 with no stars, which is greater than 0.05. Hence it is not statistically significant.



* For City, the p-value is 0.8374 with no stars, which is greater than 0.05. Hence it is not statistically significant.





* The variable State could not run in Alteryx, and looking at the fact that there is no variation in its values both in the historical data set and the dataset to be predicted, we can safely conclude that it is not statistically significant.

Conclusively, the only two statistically significant predictor variables that were used to build the model are: Avg\_Num\_Products\_Purchased and Customer\_Segment.

Fixing in and running the linear regression tool in Alteryx using these two predictor variables, the following regression formula (or model) was obtained:

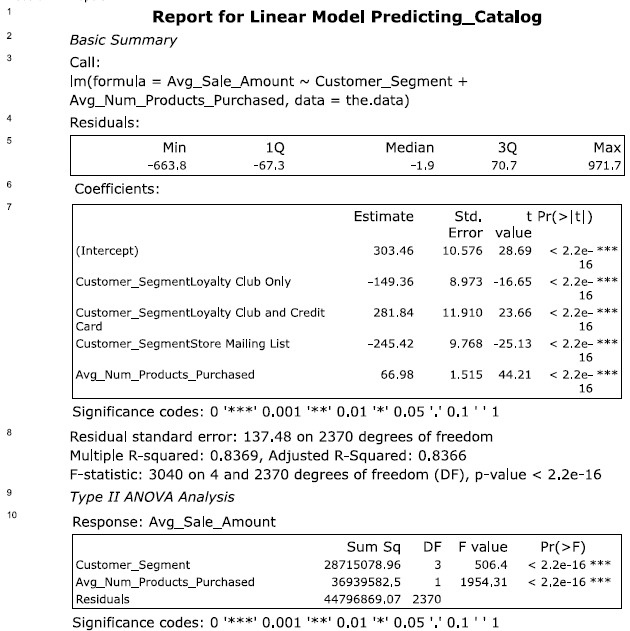
Average\_Sale\_Amount = 303.46 + 66.98(Avg\_Num\_Products\_Purchased)

– 149.36(Customer\_SegmentLoyalty Club Only)

+ 281.84(Customer\_SegmentLoyalty Club and Credit Card)

– 245.42(Customer\_SegmentStore Mailing List)

+ 0 (Customer\_SegmentCredit Card Only).



Looking at the statistical results from the model, it can be seen that the p-values of all the predictor variables used to build the model are all 2.2e-16 (which is far less than 0.05) and all with significance codes \*\*\*. Hence, the predictor variables are all statistically significant, and we can leave them all in. The Multiple R-squared is 0.8369 and the Adjusted R-Squared is 0.8366 (which are all close to 1). All of these indicates a very strong model with which a Data Analyst should be confident of making reliable predictions.

## **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?

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

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

* **Answers:** Using a score tool, I was able to get the predicted sales amount per customer.
* Next, using a scatter plot tool, I observed the correlation between the score (or predicted sales amount) versus the Avg\_Num\_Products\_Purchased (the best numeric variable in the model), and I noticed a very strong positive correlation.
* Then using a formula tool, I computed the expected revenue per customer as Expected\_Revenue = [Score\_Yes] × [Score].
* Next, I used a summarize tool to sum up all the expected revenues for all 250 customers. This gave $47,224.8713.
* Finally, using another formula tool, I calculated the final profit that can be accrued from sending a catalog to the 250 customers as follows:

Profit = (Sum\_Expected\_Revenue × 50% average gross margin) – (6.50 costs of printing and distributing per catalog × 250 customers)

= $ 21987.4356865455

Since the company will only send out the catalog if the expected profit exceeds $10 000, and the trusted model here is predicting a profit more than twice this amount, I would recommend that the company sends out the catalog to the 250 new customers.