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?

We need to decide either to print the catalogs for 250 new customers or not print the catalog based on the prediction of the profit for the new customers being more than$10,000

1. What data is needed to inform those decisions?

The catalog is printed only if the expected profit is greater than $10,000. In order to find the expected profit, we need to know probability of ordering the catalog, the cost of making the catalog and the revenues. The profit is calculated as the revenue multiplied by the probability of ordering minus the cost of making and sending out the catalog. We are given the cost of printing and the probability of ordering. To calculate revenues, we perform multiple linear regression of customer segment and number of products ordered.

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

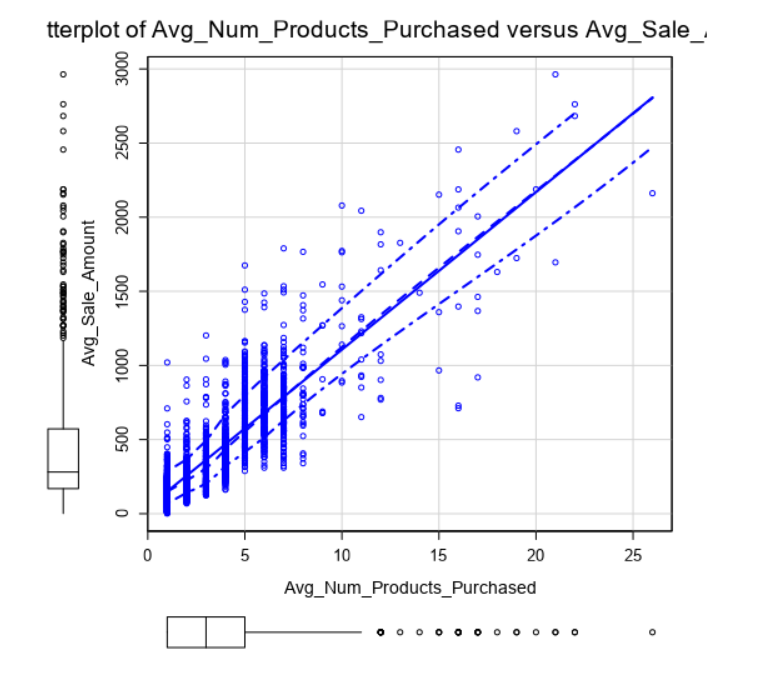
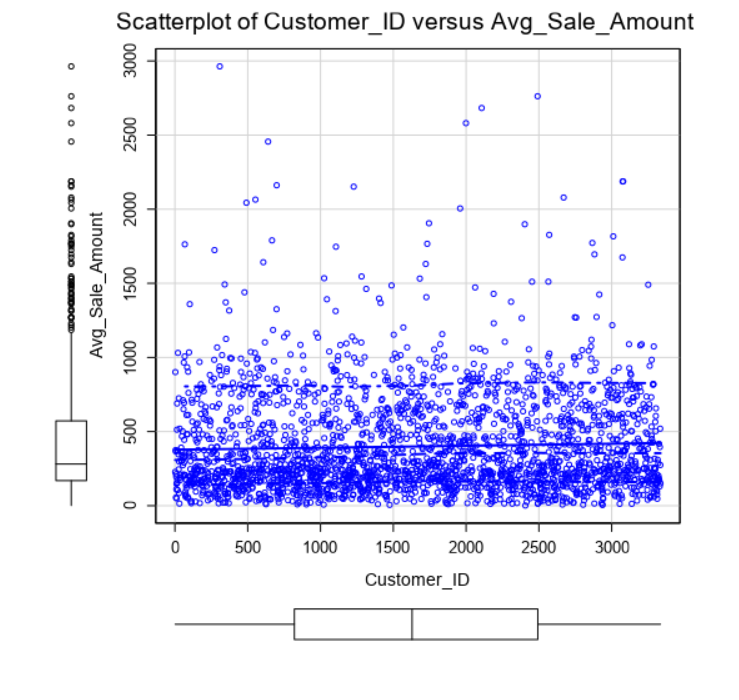
*Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)*

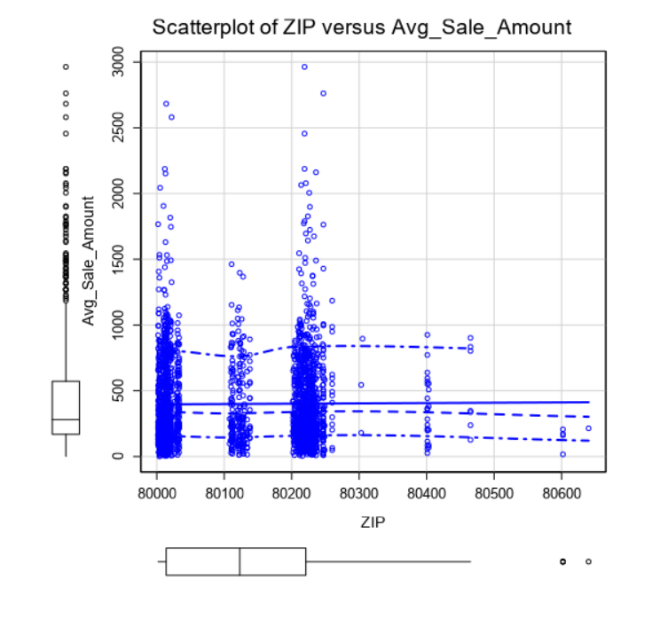
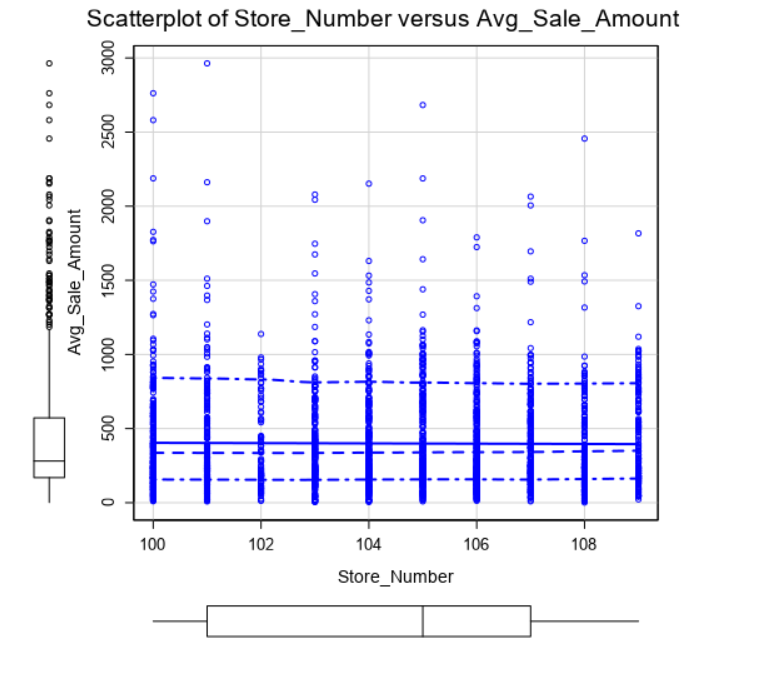
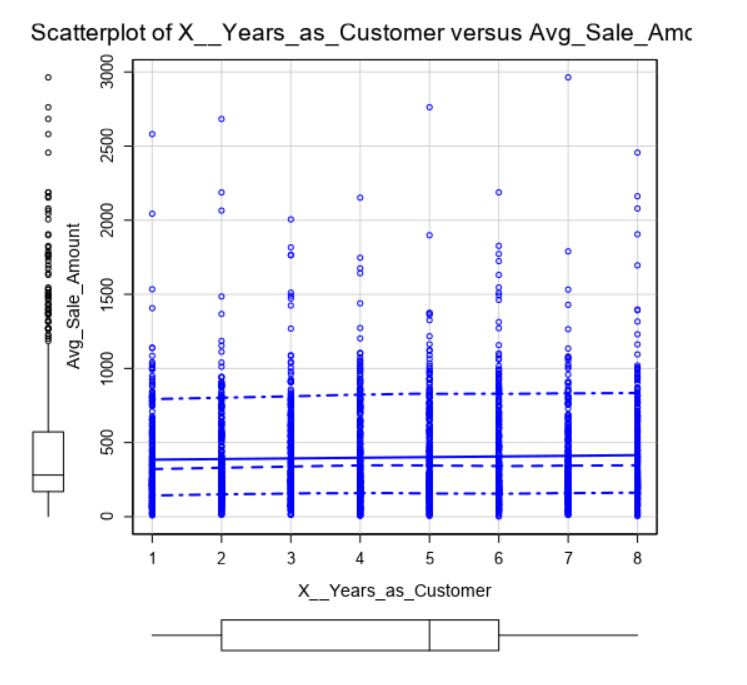
***Important:******Use the p1-customers.xlsx to train your linear model.***

*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.

**ANSWER:** To find out the linear relationship between the predictor variable and the target variable, I produced the scatter plots of numeric predictor variables vs Avg\_dales\_amount and observed it cafefully. Only **‘avg\_num\_products’** tend to have linear relationship. The scatter plots are given below. I used linear regression including non-numeric variables and the selected numeric variable. Then I selected only those where the p value is <=0.05. The selected predictor variables were ‘Customer-Segment’ with p value 0 and avg\_num\_products with p value 0, indicating to be the most significant.

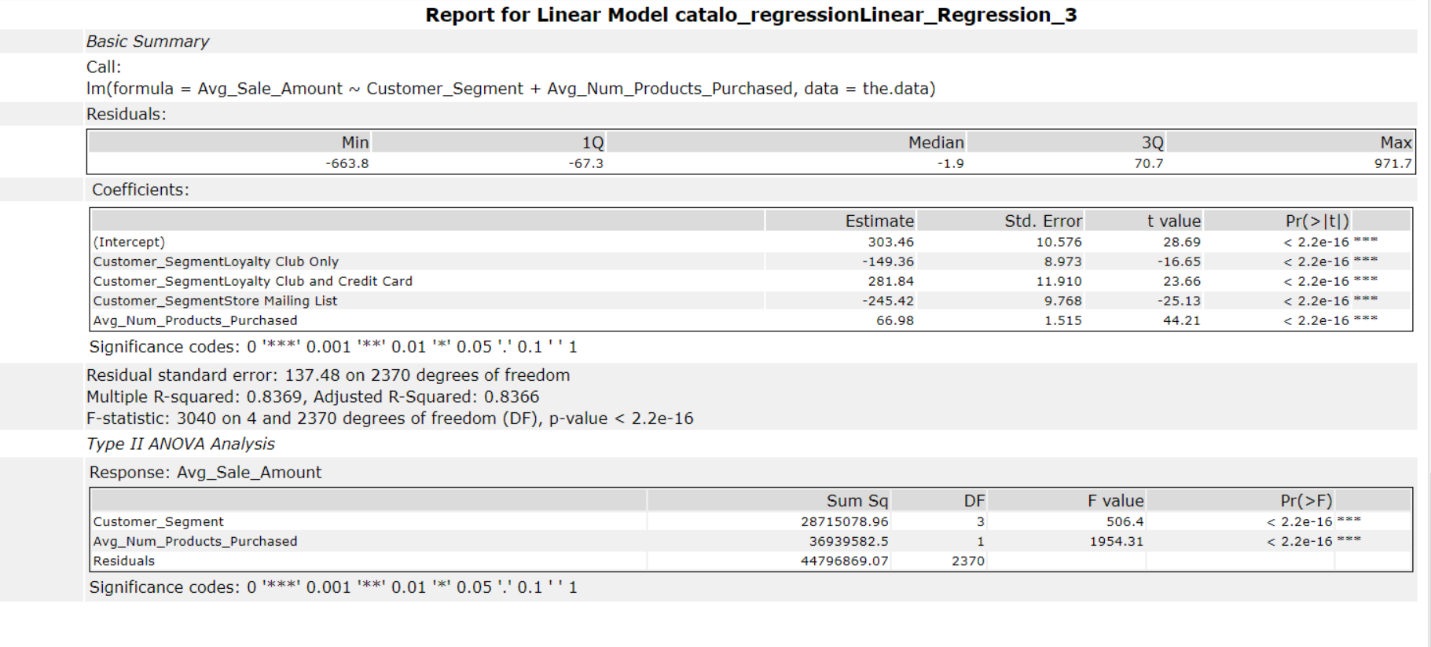




Only the scatter plot of avg\_num\_products vs Avg\_sales\_Amounts tends to have linear re

1. 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.

**ANSWER:** The linear model is a good fit as only the predictor variables with p <=0.05 were chosen and the high adjusted r-squared value is 0.837.



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)

*Avg\_Sales\_Amount = 303.46 + 0.00 (If Type: Customer\_SegmentCredit Card Only)*

*- 149.36 (If Type: Customer\_SegmentLoyalty Club Only)*

*+ 281.84 (If Type: Customer\_SegmentLoyalty Club and Credit Card)*

*- 245.42 (If Type: Customer\_SegmentStore Mailing List)*

*+ 66.98 \* Avg\_Num\_Products\_Purchased*

**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.

## **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?

ANSWER: Yes, the company should send catalog to the 250 customers as the predicted profit is much higher than $10,000.

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

ANSWER: In order to recommend whether to print catalogs or not, I followed the directive given to print catalog for 250 new customers if the predicted profit exceeds $10,000.

To get the predicted profit, the avg\_sale\_amount was calculated by running a linear regression with the predictive variables customer\_segment and the avg\_num\_products which were chosen as mentioned above. Also, we had the probability of the customer ordering and not ordering. Using the probabitly of ordering with the avg\_sale\_amount(revenue) minus the cost of printing , I calculated the profit per customer.i.e. Profit = New\_Customer\_revenue \* Probability of ordering -$6.50.Summing up the profts of 250 new cusomer gave the result $21,987.44.

Since, this value is greater than $10,000, the recommendation to make the catalog was made.

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

ANSWER: The expected profit is $21,987.43

