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?

The key decision to be made is whether to send the company's catalog for this year out to the new customers (250 new ones from the mailing list). Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

2. What data is needed to inform those decisions?

The data that is needed to inform these decisions is whether the expected profit contribution for these 250 customers exceeds $10,000. This depends on the predicted average sales amount for each customer in the mailing list, which can be obtained by a Multiple Linear Regression model. For this, the factors that may impact the predicted sales amount per customer may be the Customer Segment, Average number of products purchased, Number of Years as a customer, Store number.

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)

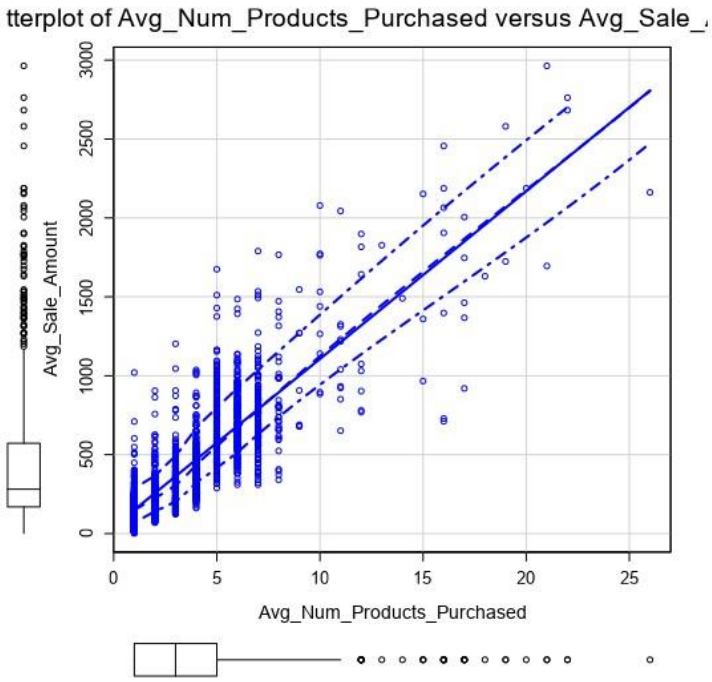
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

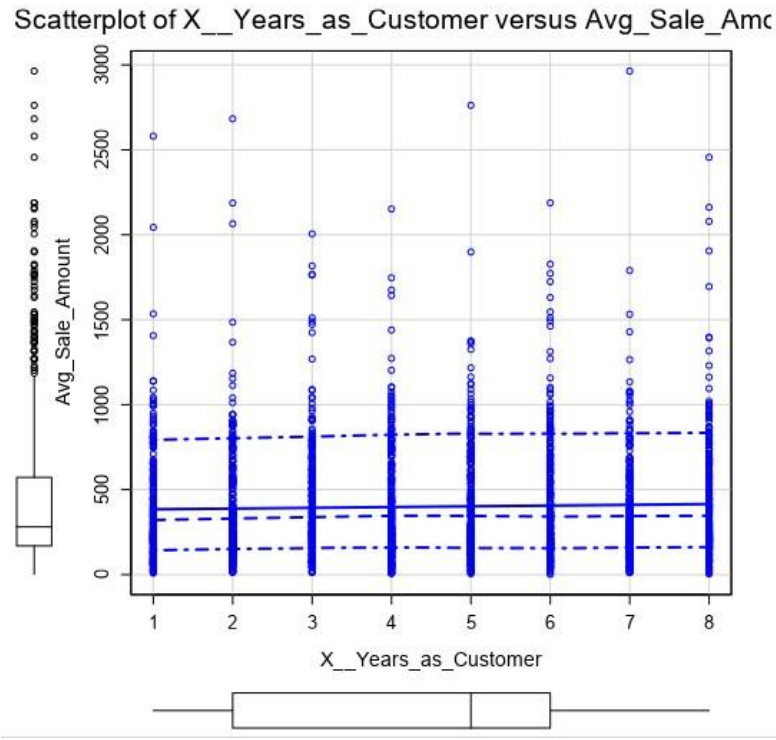
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

Among the categorical variables, Customer Segment was chosen since this would determine loyalty of the customers to the company and be a good determinant of how much they would buy from the company. The geographical variables were left out of the model since much a customer would buy would not depend on where they are from, assuming they are uniformly

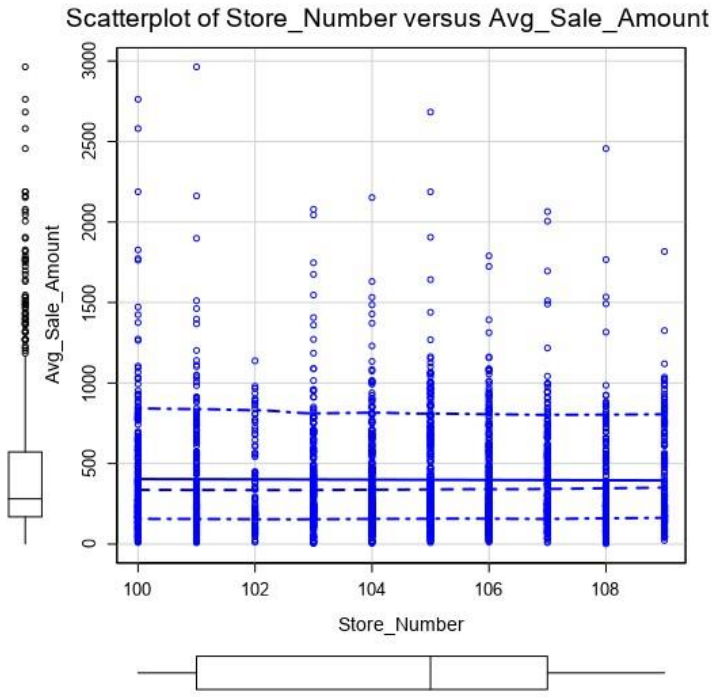
distributed. This assumption can be tested by seeing if the store number correlates with the average sales of a customer.
Among the numeric variables, the Average Number of products purchased, Average number of years as customer and Store Number were tested against the Average Sales by plotting scatter plots for each to test correlation.



Average number of products purchased shows a linear correlation with Average Sale amount.



Number of Years as Customer shows no linear correlation with Average Sale.



Store number, even though a categorical variable, was tested to see if there is any relationship with the Average sale amount.

The correlation shown in the last 2 plots above were confirmed by also running a Regression model and the p-values returned for Number of Years and Store number were higher than the threshold 5%.
Therefore, only Average number of products purchased was chosen in addition to the Customer segment as predictor variables.

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.

The linear model is a good model because of two reasons:

i. The p-value for all the selected variables is returned as 2.2e-16 or 0.00000000000000022. This is very much lower than the 5% threshold alpha rate for Type I errors and means that the model is a very good fit.
ii. The Adjusted R-squared value is 0.8366, which is high and above the 0.7 level for a strong correlation and hence indicating a strong fit for the model.

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_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased - 149.36 * (If Type: Loyalty Club Only) + 281.84 * (If Type: Loyalty Club and Credit Card) - 245.42 * (If Type: Store Mailing List) + 0 * (If Type: Credit Card Only)

Where the base case is Credit Card only.

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?

My recommendation is that the company should send the catalog to these 250 customers. This is based on the expected profit contribution of these 250 customers, which is $ 21987.44. This is above the threshold decision level of $ 10,000 set by the management.

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

First, the business decision was determined, and the required data was understood. Through an initial exploratory analysis, it was determined that the Customer Segment and the Average number of products purchased are correlated with the average sale per customer. Using these two predictor variables, a Multiple Linear Regression model was built. This model was shown to be a good fit, given the low p-values for each of the variables (2.2e-16) and the high Adjusted R-squared value for the model (0.8366).
This model was then used to predict the expected average sales of the customers from the mailing list database (250 customers) using the Score technique in Alteryx. The predicted average sales for each customer was then multiplied with the probability that a person will buy the catalog, given in the 'mailing list' table as 'Score_Yes', to get the expected revenue.
The expected revenue was then converted to profit by multiplying the revenue with the gross margin (50%) and then subtracting out the $6.50 cost per catalog (of printing and distributing). This gave the final expected profit per customer. This was then summed across all 250 customers in the mailing list to get the total expected profit from the new catalog. This value was arrived at as $ 21987.44.
The company's decision point was to send the catalog out to these new customers if the expected profit contribution exceeds $10,000. Since the expected profit contribution is more than double the threshold value, we can safely recommend the company to send out the catalog.

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

The expected profit from the new catalog is $ 21987.44