

MAKE A COPY

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 need to be made?

Calculate the profit of 250 new customers based on new catalog of high-end home goods.

2. What data is needed to inform those decisions?

Based on historical data from previous customers, the data needed are:

1. Cost of printing and distributing the new catalog.
2. The responses of customers to buy from the catalog and the average number of sales.
3. Customer segment.

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.

First, I explored the data we have and tried several plots to see if there is statistical significance supported by lower P-value and high adjusted R-squared which are

(Avg_Num_Products_Purchased) and (Customer_Segment)

After that I removed non-related variables like address and customer name...etc

Report for Linear Model Linear_Regression_15

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X_Years_as_Customer, data = the.data)

Residuals:

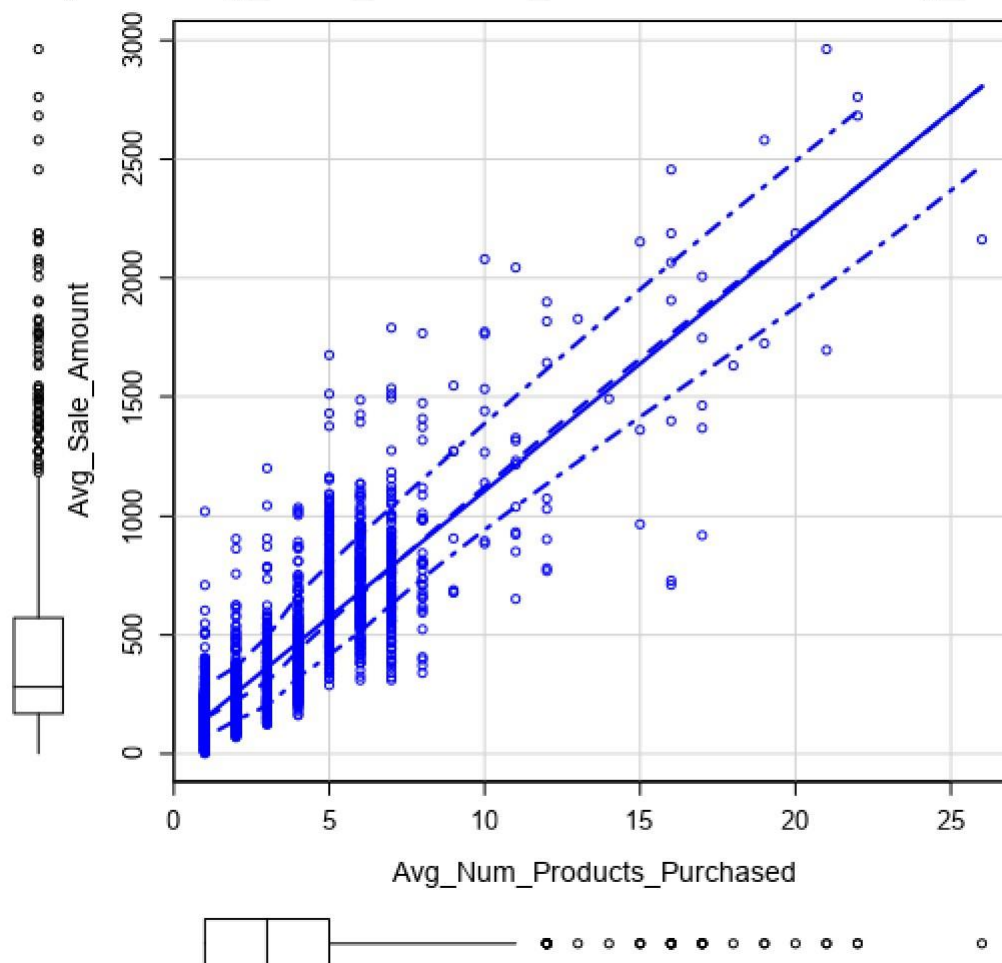
| | Min | 1Q | Median | 3Q | Max |
|--|---------|--------|--------|-------|--------|
| | -665.19 | -67.82 | -2.17 | 70.42 | 975.25 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--|----------|------------|---------|---------------|
| (Intercept) | 435.318 | 104.854 | 4.152 | 3e-05 *** |
| Customer_SegmentLoyalty Club Only | -150.224 | 8.971 | -16.746 | < 2.2e-16 *** |
| Customer_SegmentLoyalty Club and Credit Card | 282.455 | 11.897 | 23.743 | < 2.2e-16 *** |
| Customer_SegmentStore Mailing List | -243.279 | 9.816 | -24.784 | < 2.2e-16 *** |
| Store_Number | -1.146 | 0.994 | -1.153 | 0.2489 |
| Responded_to_Last_CatalogYes | -28.085 | 11.253 | -2.496 | 0.01264 * |
| Avg_Num_Products_Purchased | 66.787 | 1.515 | 44.082 | < 2.2e-16 *** |
| X_Years_as_Customer | -2.326 | 1.222 | -1.904 | 0.05707 . |

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ttestplot of Avg_Num_Products_Purchased versus Avg_Sale_Amount



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.

- 1- The model resulted in low P-value. Which is bellow 0.05
- 2- The model resulted high adjusted R-squared.

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

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)

$$\text{Avg sale} = 303.46 - 149.36 * \text{Loyalty Club Only} + 281.84 * \text{Loyalty Club and Credit Card} - 245.42 * \text{Store Mailing List} + 0 * \text{Credit Card only} + 66.98 * \text{Avg_Num_Products_Purchased}$$

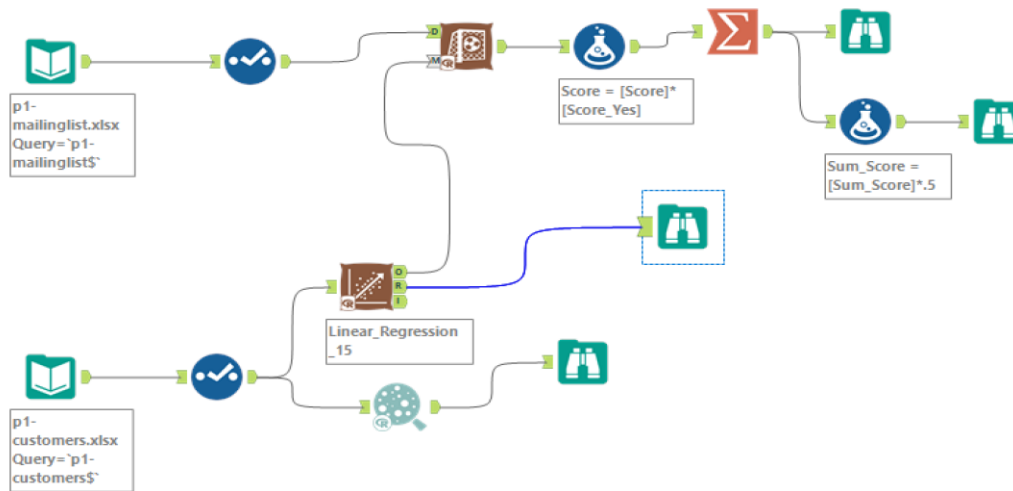
Step 3: Presentation/Visualization

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

They should send the catalog to the 250 customers.

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

- 1- Find the linear regression from predictor variables based in historical data.
- 2- Use linear equation to the new data to predict the needed objective.
- 3- Calculate the total profit with a percentage cross margin and with additional costs (cost of catalog).



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

Expected Profit = (Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)

$$= (47,225.87 \times 0.5) - (6.50 \times 250)$$

$$= 23,612.44 - 1,625$$

$$= \$21,987.44$$

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](#) here. Reviewers will use this rubric to grade your project.