Project 1: Predicting Catalog Demand

By Matthew Burt

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

### **Key Decisions:**

1. What decisions needs to be made?

The business decision that needs to be made is whether or not the business will send the catalogue out to their new customers.

1. What data is needed to inform those decisions?

At this stage in the analysis, the data that will be needed is the same as that available in the customers data set, excepting the average sales and responded to last catalogue fields. The business will also need to know the costs associated with creating and mailing the catalogue to the new customers.

## **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.***

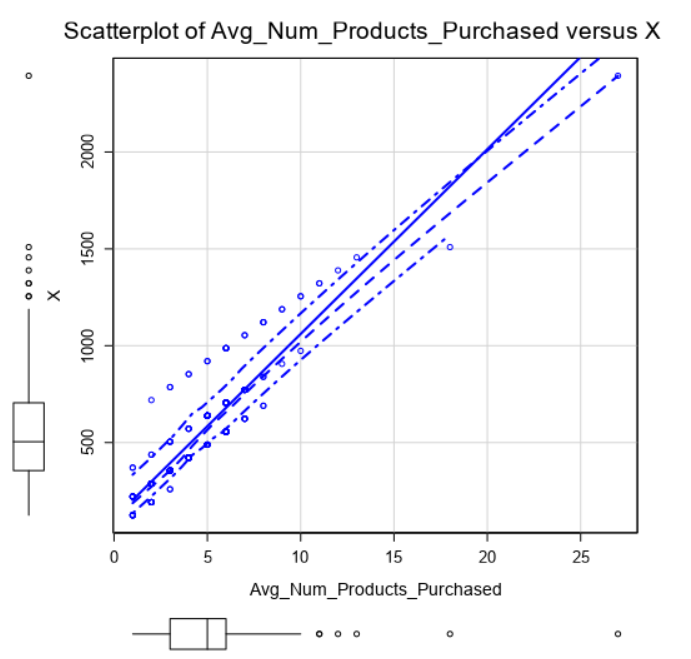
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer back to the “Multiple Linear Regression with Excel” lesson to help you explore your data and use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

I selected the following predictor variables: Customer\_Segment and Avg\_Num\_Products\_Purchased. This was based on the following analysis. I have created scatter plot graphs of each variable to visually determine whether there is a linear relationship present. I have included two of these in this section.

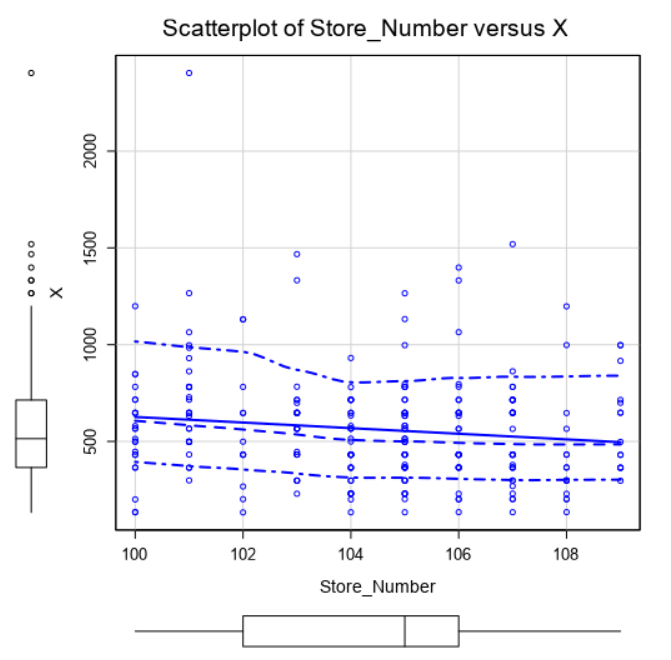
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | -1384.1983 | 2.149e+03 | -0.6441 | 0.51958 |  |
| Customer\_SegmentLoyalty Club Only | -149.5782 | 8.977e+00 | -16.6625 | < 2.2e-16 | \*\*\* |
| Customer\_SegmentLoyalty Club and Credit Card | 282.6768 | 1.191e+01 | 23.7335 | < 2.2e-16 | \*\*\* |
| Customer\_SegmentStore Mailing List | -245.8485 | 9.770e+00 | -25.1625 | < 2.2e-16 | \*\*\* |
| ZIP | 0.0225 | 2.659e-02 | 0.8460 | 0.39761 |  |
| Store\_Number | -1.0002 | 1.006e+00 | -0.9939 | 0.32037 |  |
| Avg\_Num\_Products\_Purchased | 66.9646 | 1.515e+00 | 44.1928 | < 2.2e-16 | \*\*\* |
| X.\_Years\_as\_Customer | -2.3528 | 1.223e+00 | -1.9239 | 0.05449 | . |

As seen above, ZIP, Store\_Number and X.\_Years\_as\_Customer are not statistically significant as the p values are greater than 0.05. I have chosen all variables with a p value less than 0.05.

To confirm that Avg\_Num\_Products\_Purchased is a good predictor, the scatter plot below shows a strong positive linear relationship with the Avg\_Sale\_Amount target variable.

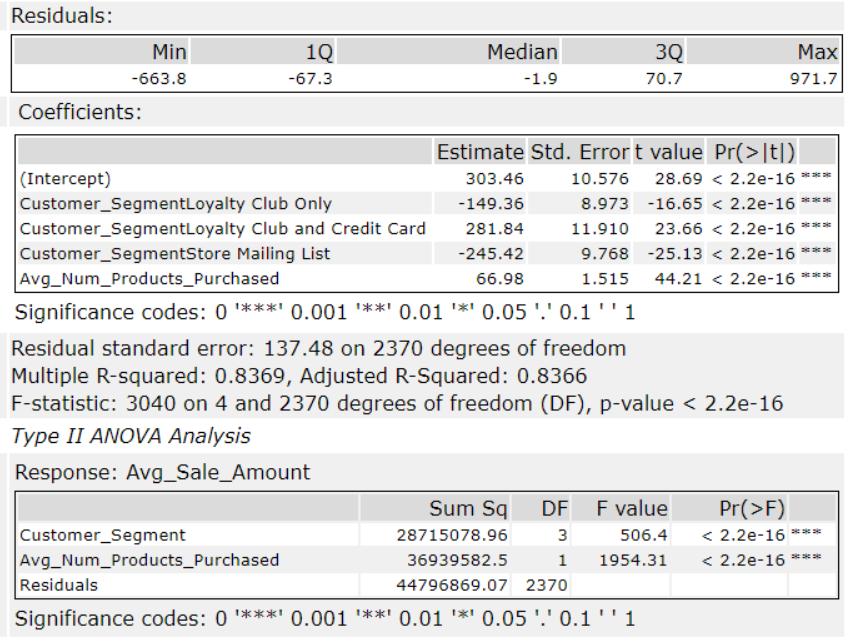


Alternatively if we plot store number against the Avg\_Sale\_Amount target variable, there is no obvious linear relationship present, so I did not choose this variable. I have done the same for the rest of the variables.



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.

The following is a report on the linear model with the variables I have chosen.



The linear model output above shows that my model is a good model. The p values are all much lower than 0.05, which shows that the variables I have selected are statistically significant. The adjusted R-Squared value is also 0.8366. This is strong evidence that the model explains a large amount of the variance in the predictor variable.

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 best linear regression equation is:

X = 303.36 + 0 (if: Credit Card Only) - 149.36 (if: Customer\_SegmentLoyalty Club Only) + 281.84 (if: Customer\_SegmentLoyalty Club and Credit Card) - 245.42 (if: Customer\_SegmentStore Mailing List) + 66.98 \* Avg\_Num\_Products\_Purchased.

## 

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

*Use your model results to provide a recommendation. (500 word limit)*

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

The company should send the catalogue to the 250 new customers as the expected profit is $21,987.44 which is greater than the $10,000 threshold.

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

I used the linear regression model to estimate the expected profit of sending the catalogue to the new customers. I had to account for the cost of revenue of 50%, then I then used the catalogue costs to arrive at a cost of sending the catalogue.

Expected Profit = Expected Revenue - Cost of Revenue - Cost of Catalogue

Expected Revenue = $47,224.87, Cost of Catalogue = $6.5 \* 250 = $1,625

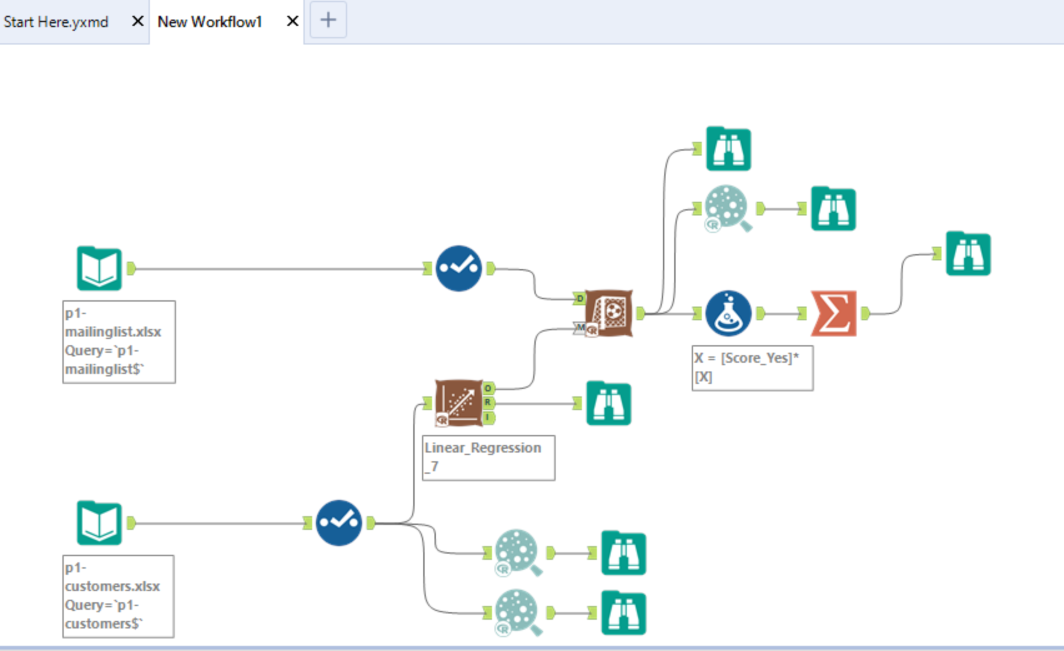
Expected Profit = $47,224.87 - 0.5\*$47,224.87 - $1,625 = $21,987.44

Expected Profit > $10,000

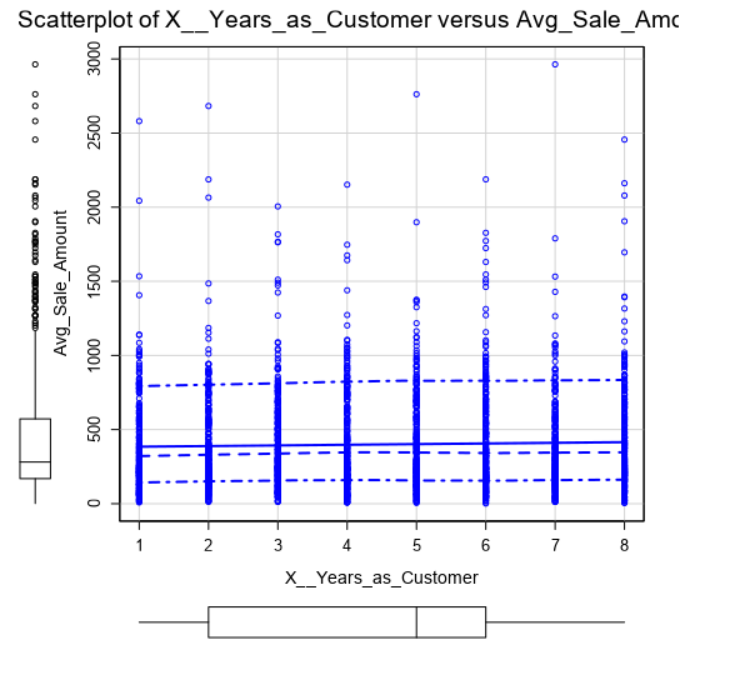
1. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

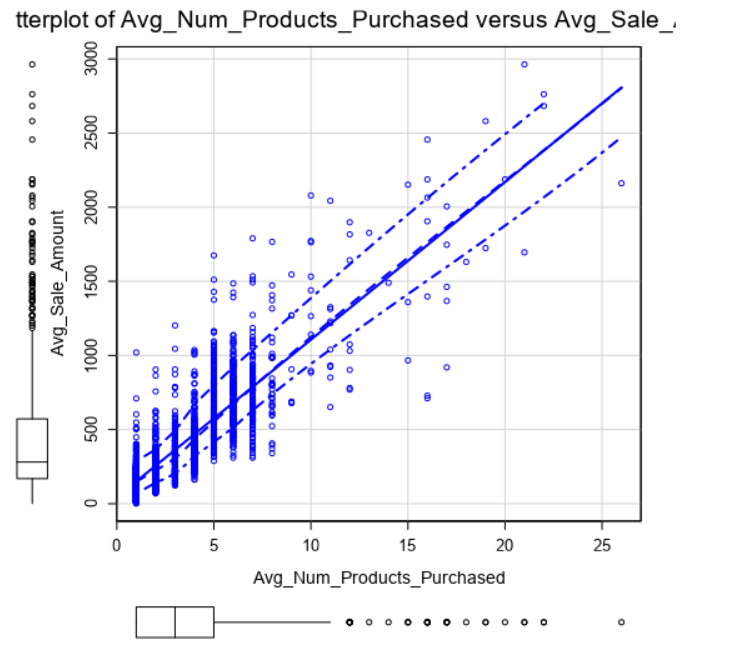
Expected Profit is $21,987.44

Appendix - Alteryx Output

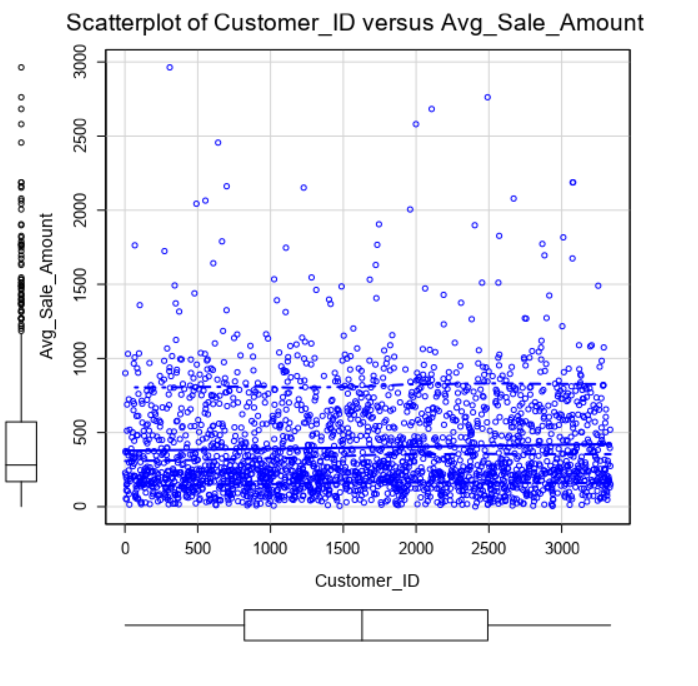
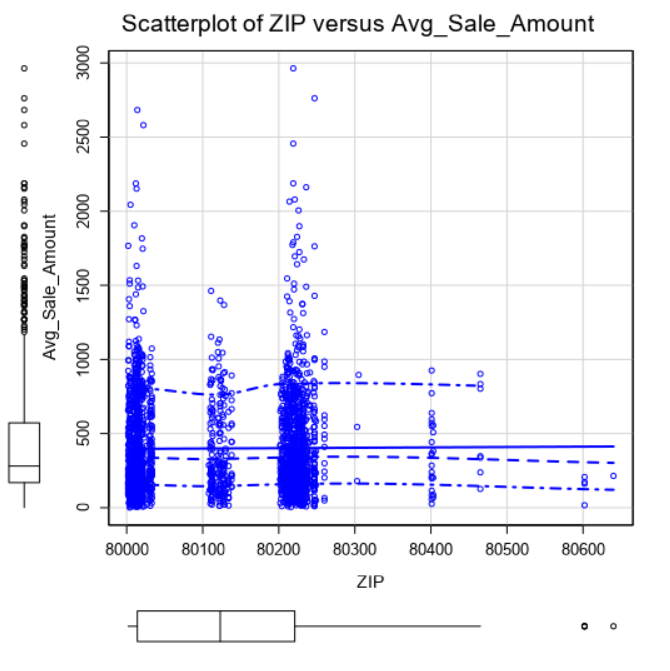


Other Scatterplots:

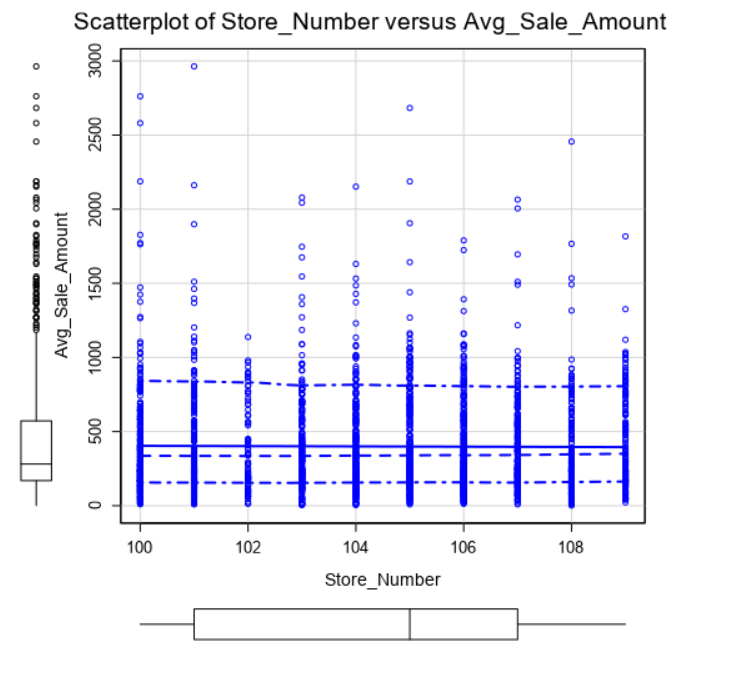




Strong Positive Relationship No relationship



No relationship No relationship



No relationship