Predicting Catalog Demand

Step 1: Business and Data Understanding

Defining Key Decisions:

What decisions needs to be made?

I will analyze how much profit the company can expect from sending a catalog to new customers and if is it worth to send them this catalog. If the expected profit exceeds $10,000 the company must send the catalog to the new customers.

What data is needed to inform those decisions?

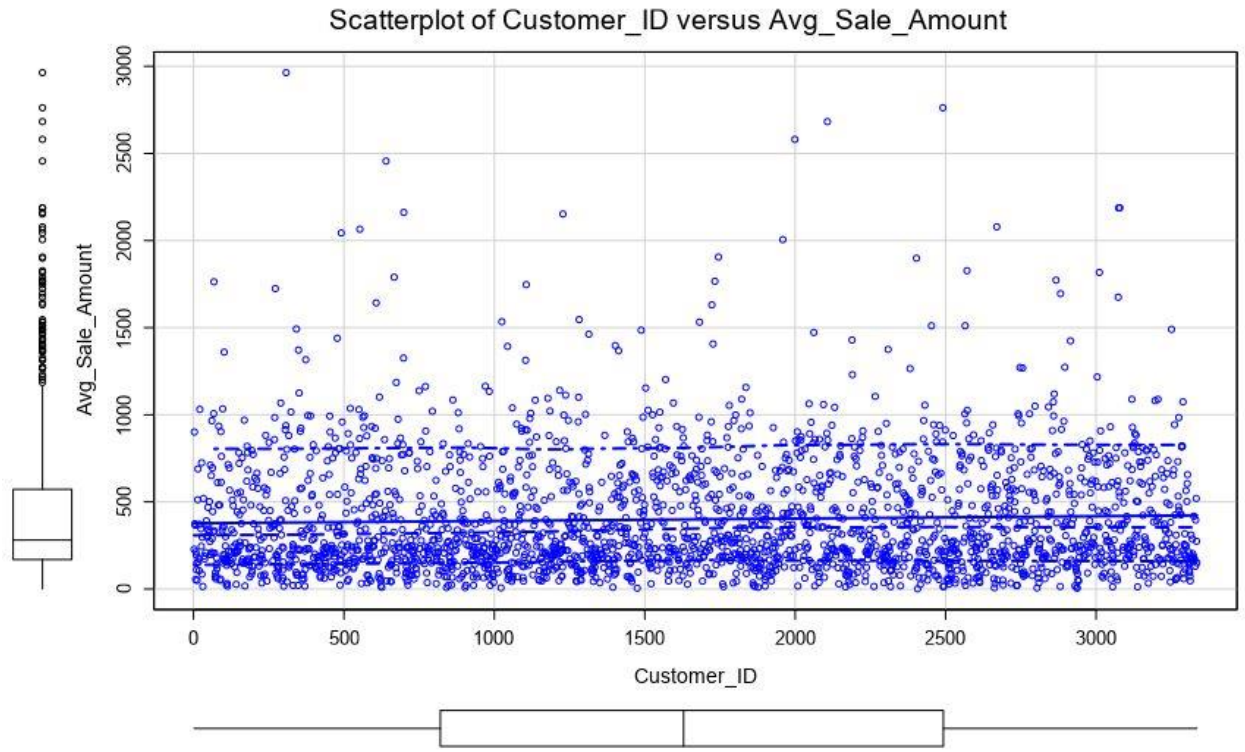
The parameters of the analysis: the minimum amount of profit to send the catalog to the new clients, the cost of printing and distributing per catalog, the average gross margin on all products sold through catalog. Besides that, we need data about current clients and we also need data about new clients, both were provided by the company. With these data we can construct our linear model to predict the average sales amount of the new clients and deduce if we must send the catalog.

Step 2: Analysis, Modeling, and Validation

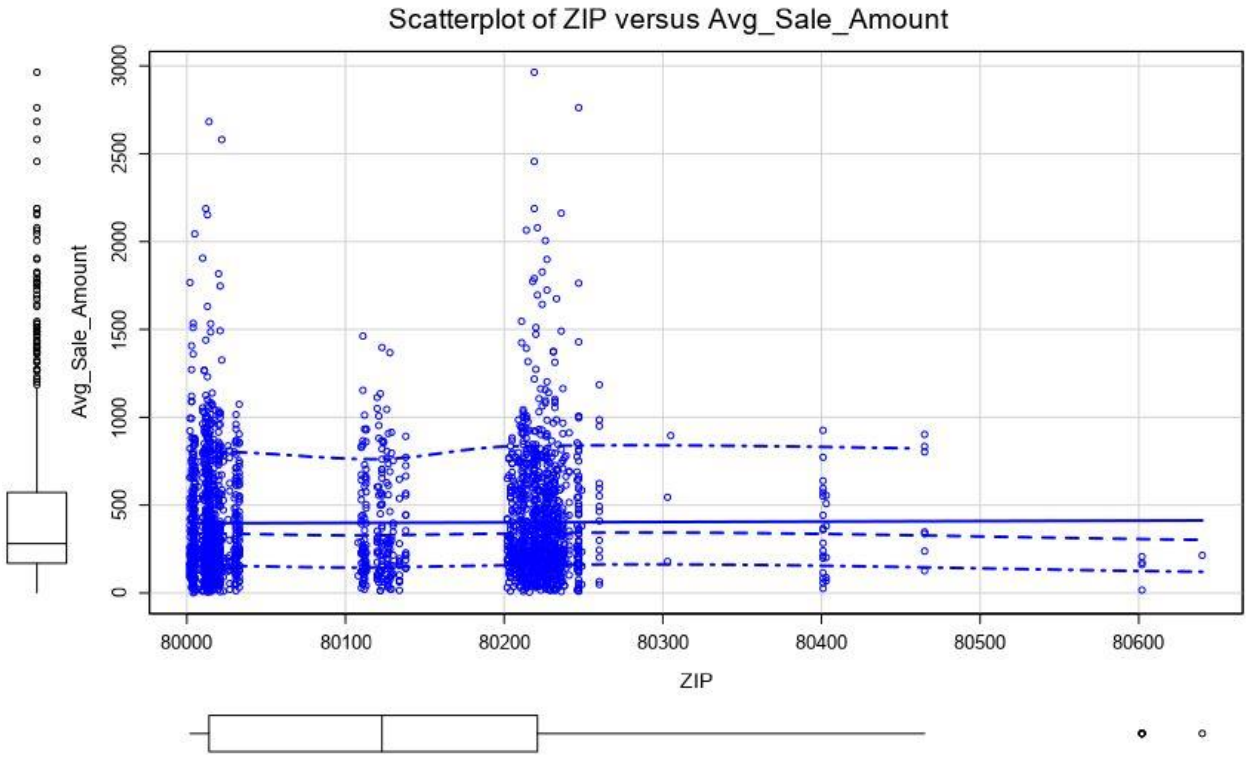
The company provided raw data of the current customers (p1-customers.xlsx at https://github.com/DataGF/business-analytics/blob/main/p1-customers.xlsx) and new customers mailing list (p1-mailinglist.xlsx at https://github.com/DataGF/business-analytics/blob/main/p1-mailinglist.xlsx).

First was applied data cleaning tool to remove nulls and after that was applied select tool to guarantee what fields and data type would be filtered to our model.

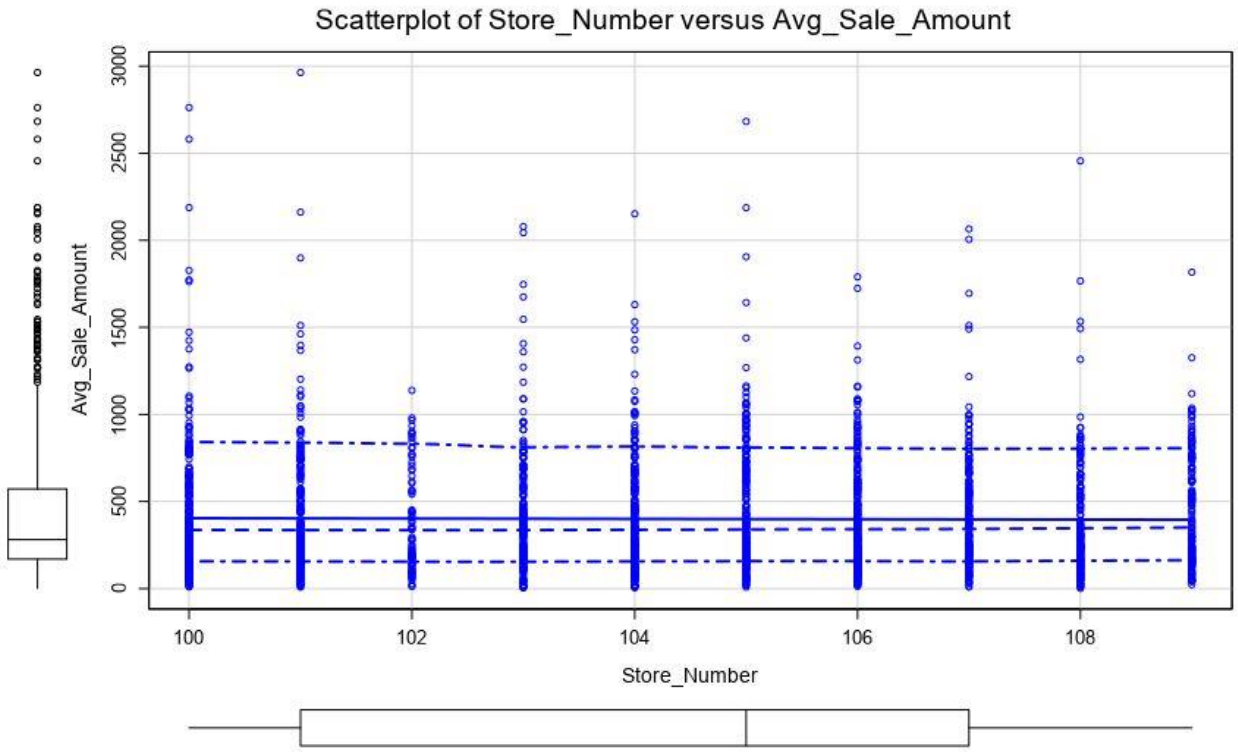
Scatterplots for numeric data types were generated to see if there was linear correlation between the variable analyzed and the average sales amount as follow in the next pages.



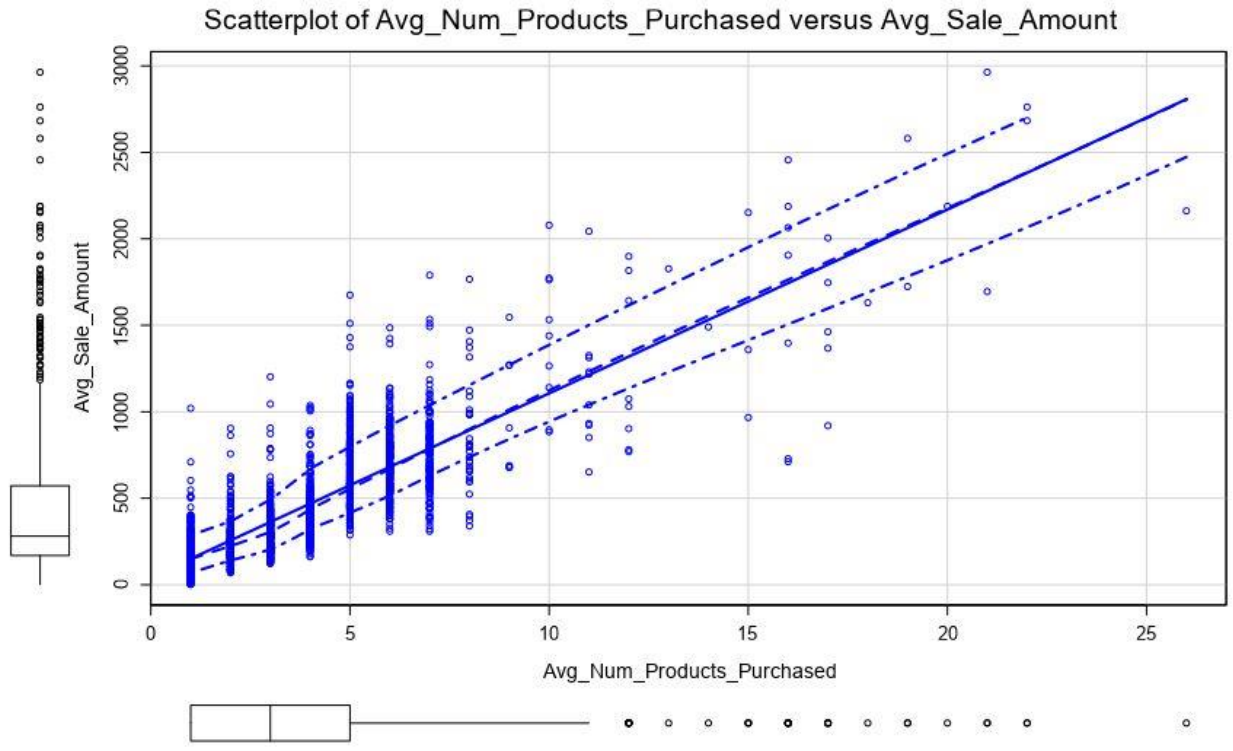
Scatterplot 1 – Customer ID versus Average Sales Amount



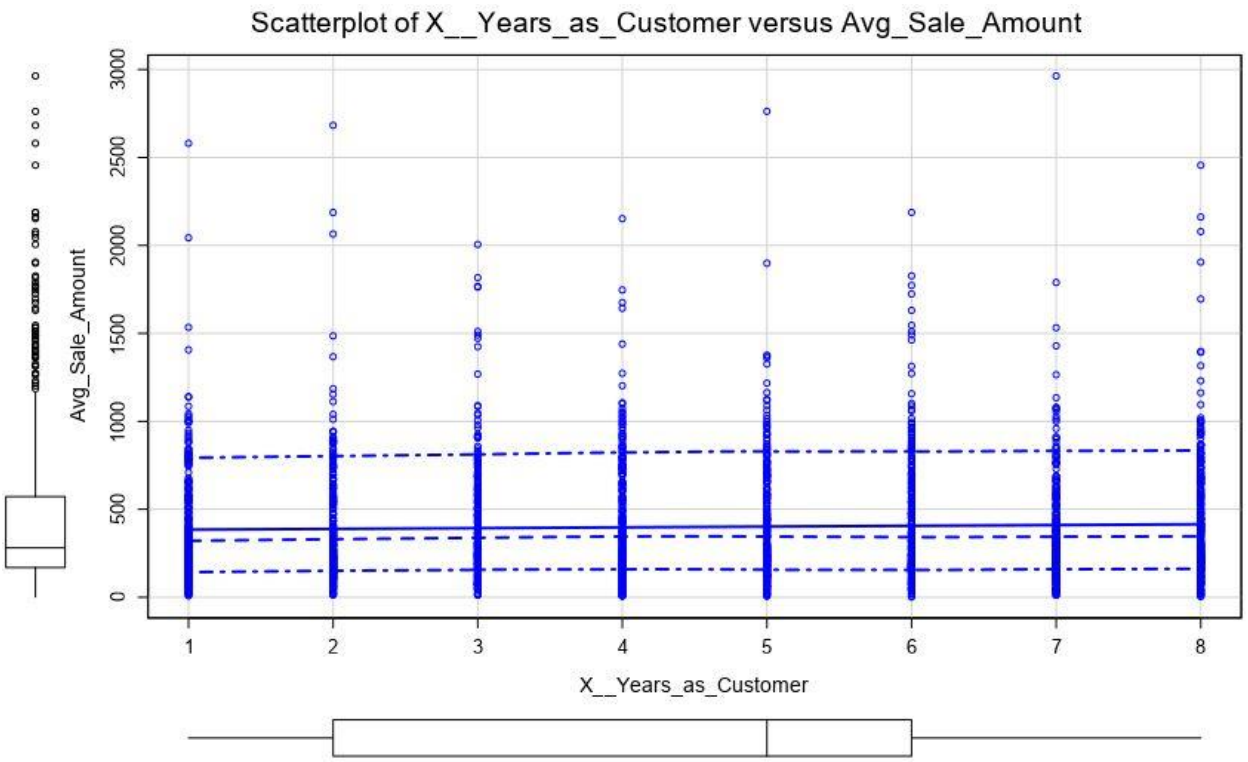
Scatterplot 2 – ZIP versus Average Sales Amount



Scatterplot 3 – Store Number versus Average Sale Amount



Scatterplot 4 – Average Number Products Purchased versus Average Sale Amount



Scatterplot 5 – Years as Customer versus Average Sale Amount

As was observed the only scatterplot that shows a linear correlation with average sale amount is the scatterplot 4. This indicates that the average number of products purchased influences at the average sale amount. The slope of the curve is positive, this tell us that the greater average number of products purchased the bigger is the average sale amount. Said that, the average number of products purchased is a predictor variable for the target variable average sale amount. Customer ID, ZIP, Store Number, Years as Customer are not good predictors to our model, they don't show any trend related to average sale amount.

The process of trial and error was adopted for finding relevant categorical data types. Reports were generated to see if the P-values were statistically significant.

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Name	273301919.91	2365	0.79	0.75118
Residuals	1319463.18	9		

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

Report 1 – Name versus Average Sale Amount

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	192884931.52	3	1865.06	< 2.2e-16	***
Residuals	81736451.57	2371			

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

Report 2 – Customer Segment versus Average Sale Amount

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Address	268990772.74	2320	1.11	0.31795
Residuals	5630610.36	54		

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

Report 3 – Address versus Average Sale Amount

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
City	2199299.15	26	0.73	0.83744
Residuals	272422083.94	2348		

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

Report 4 – City versus Average Sale Amount

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Responded_to_Last_Catalog	10914470.43	1	98.22	< 2.2e-16	***
Residuals	263706912.66	2373			

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

Report 5 – Responded to Last Catalog versus Average Sale Amount

P-values are statistically significant when its values are less or equal to 0.05. The only two significant p-values were encountered at reports 2 and 5. This indicates that customer segment and responded to last catalog are categorical variables important to the model. Unfortunately, it is impossible to know the responded to last catalog data for new clients, it was not sent any catalog for them. Said that, this variable will not be part of the model. Customer segment is the only categorial variable that will compose the model to predict if we should send catalogs to new clients or not.

Summarizing, the numeric variable average number of products purchased, and categorical variable customer segment will be the predictors to this model for the reasons explained.

Now I will show why this is a good model based on statistical results of this linear regression. I will justify each variable that I selected based on p-values and R-squared values that was produced by this model.

Report for Linear Model Avg_Sale_Amount_LR_Model

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Avg_Num_Products_Purchased + Customer_Segment, 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	***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 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	***

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

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16	***
Customer_Segment	28715078.96	3	506.4	< 2.2e-16	***
Residuals	44796869.07	2370			

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

Report 6 – Customer Segment and Average Number Products Purchased versus Average Sale Amount

A good model is defined by the significance of its predictors. The first parameter to look at is P-value. In this model the P-value is less or equal to 0.05, this means, for each predictor (average number of products purchased and customer segment) we can have at least 95% of confidence that there is a relationship between the predictors and target variable. The second parameter to a good model confirmation is R-squared. In this case we pay attention to Adjusted R-squared of 0.8366. For this case the adjusted value is preferable because there is more than one predictor. The value for Adjusted R-squared of this model means that 83,66% in the output variable is explained by the input variables.

Said that, he best linear regression equation based on the available data is:

Y = 306.46 – 149.36 * X1 + 281.84 * X2 – 245.52 * X3 + 66.98 * X4, where:

Y = Predicted Average Sale Amount.
X1 = (If Customer Segment: Loyalty Club Only).
X2 = (If Customer Segment: Loyalty Club and Credit Card).
X3 = (If Customer Segment: Store Mailing List).
X4 = (If Average Number of Products Purchased).

Step 3: Presentation/Visualization

My recommendation, based in this predictive business analysis and the parameters established by the company, is that the catalog should be sent to its 250 new clients. The process that I used to come up with this recommendation is explained bellow.

Step 1 – Finding predictors: for numeric predictors I looked for predictors with scatterplots that showed a linear correlation with average sale amount. For categorical ones, I looked for predictors with 95% (or more) of confidence related with average sale amount.

Step 2 – Measuring the statistics significance of the model: after finding the predictors and the linear regression model, I measured the statistical significance of the model. It has more than 95% of confidence that there is a relationship between the predictors and average sale amount and 83,66% of confidence that the average sale amount is explained by the input variables.

Step 3 – Calculating the expected profit: given by the sum of the chance that a client buy motivated by the catalog multiplied by the score of the model and the gross margin of 50% and then subtracted by the cost to send the catalog.

The expected profit from the new catalog is: $ 21,987.44 if the catalog is sent to these 250 new clients.

Expected_Profit	
21,987.4356	1

Figure 1 – Expected profit given by the model

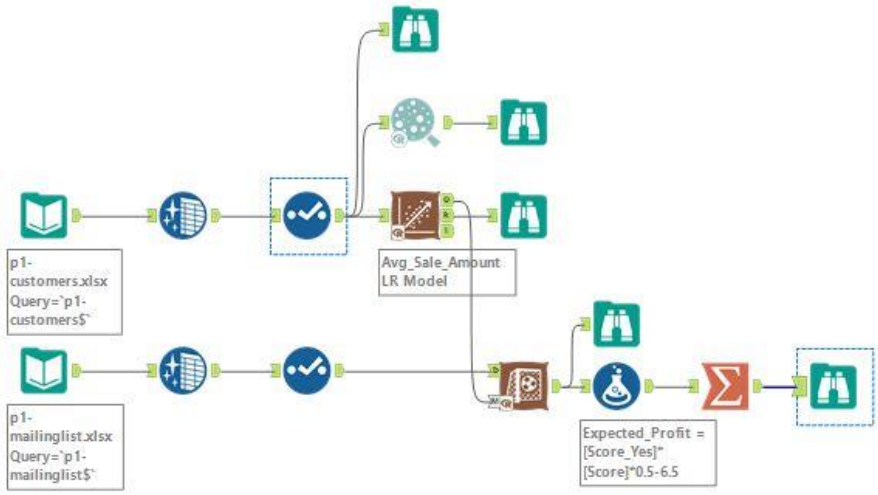


Figure 2 – Linear regression model at alteryx workflow