Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

A mail order catalog business that sells high-end home goods has tasked us with analyzing the viability of an upcoming campaign. The recommendation to continue with the campaign will be based on calculating the potential profit to be generated from sales of the products in the upcoming product catalogue.

We are required to have at least two sets of data. The first being the test dataset, providing the observations in question to make the prediction, and the second, is the training dataset, containing historic data to train our model.

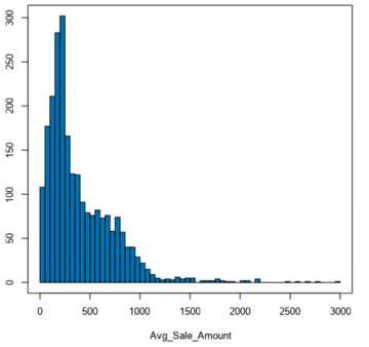
- The test data, with the new observations, contains the same variables as the training dataset. The test dataset also includes a field for the probability or likelihood of purchase which will further enhance the accuracy of the profit calculation before making a recommendation.
- A much larger dataset with historic customer sales data is used to train the predictor model. This is done by determining significant correlations between average sales and the different customer related variables. The historic dataset is data rich with over twenty thousand numeric and categorical observations, making it possible to formulate a linear relationship between the predictor variables and target variable

Step 2: Analysis, Modeling, and Validation

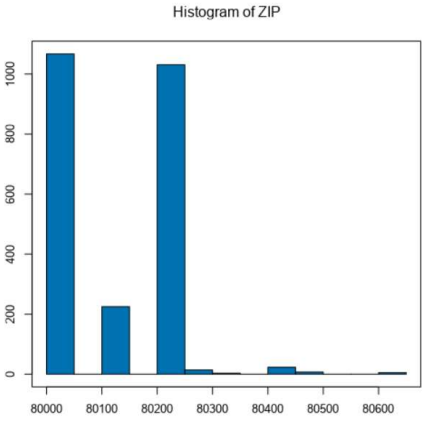
Alteryx was the only tool used throughout the complete Analysis process. Alteryx was used for gathering as well as exploring the data. Alteryx was further used to build the regression model, which was ultimately used to predict profits based on new customer data.

After pulling in the training dataset with historic customer sales data, we used the Alteryx *Select* tool to observe and in turn further format the dataset. A minor format to improve manual readability was done by moving the target variable as the final column. Unique ‘Customer ID’ and ‘Address’ variables were dropped from the data set before proceeding the analysis

First step to building the model is exploring the different numerical variables and identify key continuous variables to be used in the model. The variable ‘Avg_Sale_Amount’ is a continuous variable as seen in the slightly skewed histogram below. This is close to the bell-shaped curve we would expect from a normal distribution to be used in a linear regression model.

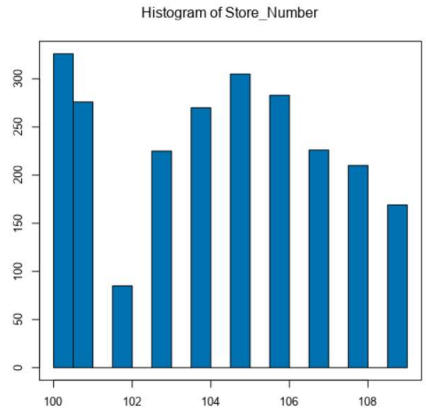


The distributions for most of the other numerical variables do not show variations in the distribution and are discrete. The *histogram* plots and the aggregated values shown below identify the discreteness of various numerical variables [Zip; Store Number; Years as customer]



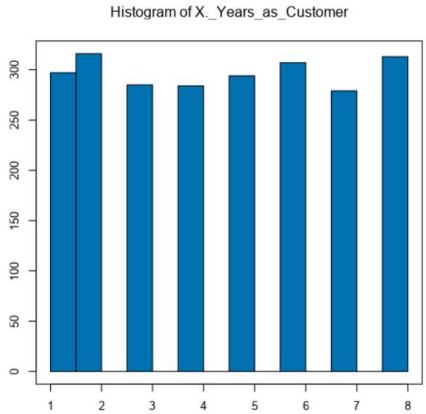
ds ▾ 86 records displayed, 1865 bytes

ZIP	Count
80013	127
80219	93
80015	90
80012	88
80020	87
80004	80
80005	71
80014	71
80226	71
80247	57

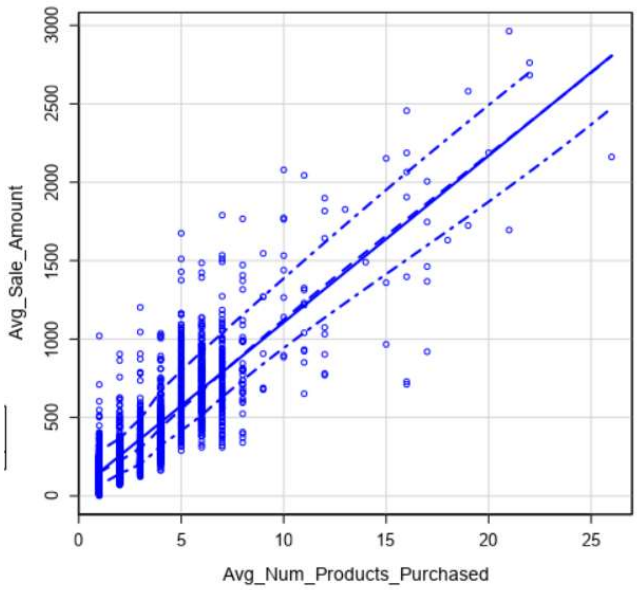


lds ▾ 10 records displayed,

	Store_Number	Count
1	100	326
2	105	305
3	106	283
4	101	276
5	104	270
6	107	226
7	103	225
8	108	210
9	109	169
0	102	85



The numerical value for Average number of products purchased, while seeming discrete in its distribution, does indicate a trend line when plotted as a scatter plot (which none of the other numerical correlations indicated)



Next step in identifying the correlation explores the non-numerical categorical variables to be used as predictor variables in our model. ‘State’ has only one value and cannot be used to indicate correlation across categories. Whereas ‘Customer Segment’ includes significant distributions across the different customer segments types, justifying using it the linear regression model (as seen below).

Customer_Segment	Count
Credit Card Only	494
Loyalty Club Only	579
Loyalty Club and Credit Card	194
Store Mailing List	1108

Next, a model is built using the Linear Regression Model in Alteryx setting the predictor variables: *‘Customer_Segment’* and *‘Avg_Num_Products_Purchased’*, along with the target variable: *‘Avg_Sale_Amount’*. Alteryx automatically denotes the ‘dummy variable’ for the ‘Credit Card Only’ customer segment and establishes a linear relationship for the other three customer segments when formulating the linear equation using categorical variables.

The results from running the model are shown below. The p-values indicate a statistically significant relationship between the target variable and the predictor variables, as its much lower than 0.05 (or 5%). The R-squared and Adjusted R-Squared are both above 0.7 and very close to 1, indicating a strong correlation between the predictor variables and the target variable.

	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	***

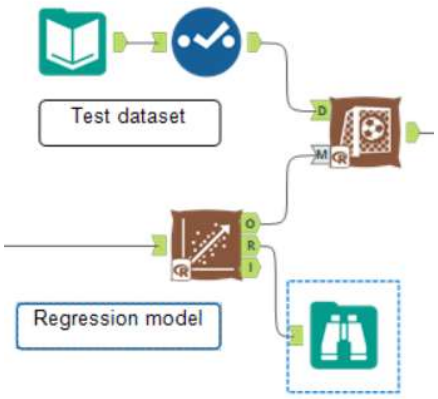
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

The Linear Regression equation, can then be written as follows:

Avg_Num_Products_Purchased = 303.46 – 149.36 x [Customer_SegmentLoyalty Club Only]
+ 281.84 x [Customer_SegmentLoyalty Club and Credit Card]
– 245.42 x [Customer_SegmentStoreMailing List]
+ 66.98 x [Avg_Num_Products_Purchased]

With the regression model complete, it is then used to predict the average sale amount. This is done seamlessly in Alteryx using the *Score* Tool. The score tool (shown below with the goal post), takes the test dataset as one input and the regression model as the other input, and outputs the predictions for the average sales based on the formulated linear regression model



Step 3: Presentation/Visualization

The output from the score tool generated the predicted 'Average Sale Amount' for each observation based on the Linear Regression model. Profit is then calculated following these steps:

- Multiply the average sale amount by the probability of the customer actually buying the product

```
[predicted_amount_purchased]*[Score_Yes]
```

- Then, calculate the average gross margin by taking 50% of predicted average sale amount and then subtract $6.50 for each catalog sent to each customer, to account for the cost or printing and distributing each catalogue

```
[total_predicted_amount_purchased]*0.5 - 6.5*250
```

We recommend that the catalogues be sent to the 250 customers, since our model **has predicted the profit amount at $21,987.**