PREDICTING CATALOG DEMAND

Step 1: Business and Data Understanding

The company sells high-end home goods, Last year the company sent out its first print of the catalogue. In the coming months they we will prepare the catalogues for this year send. Management wants to determine if they should send out 250 catalogues to the new customers. If the expected profit exceeds $10,000 then the catalogues will be sent to this new customer group

Key Decisions

Key Decisions: Is the expected profit >10,000, if case is true then the company will send the catalogues. Data required to be able to determine the decision is as follows:

- The cost of printing & distribution per catalogue, $6.50.
- Average gross margin (price-cost) on all products sold through the catalogue is 50%.
- Multiply revenue by gross margin first before subtraction the costs.

Existing & new customer data that includes:

- Average sales amount per customer.
- Customer response to catalogue last year.
- Customer payment method.

We are a data rich company and the existing and new customer details have been provided in two files.

Step 2: Analysis, Modeling, and Validation

We will exclude the following variables:

- Name: sales doesn't depend on a name, and we can get many false results due to duplicated names
- Customer_ID: is assigned to the name and is a unique ID, not having impact on avg_sales
- Address: Also unique to a customer and will not have an impact on avg_sales.

I will explore the data and use the visual exploration method, to check what relationships there are between the target variable: average sale amount.

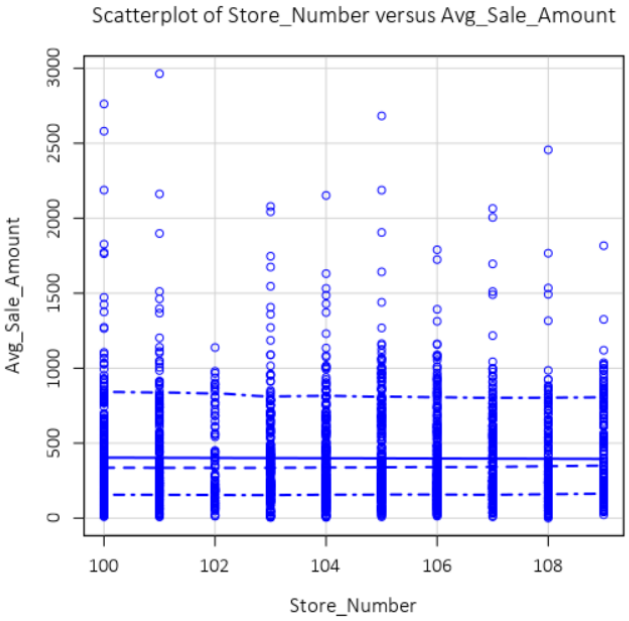
I decided to look at few numeric variables for the predictor, at *post_code* and *store_number and also customer segment* as these are present in both sets of data.

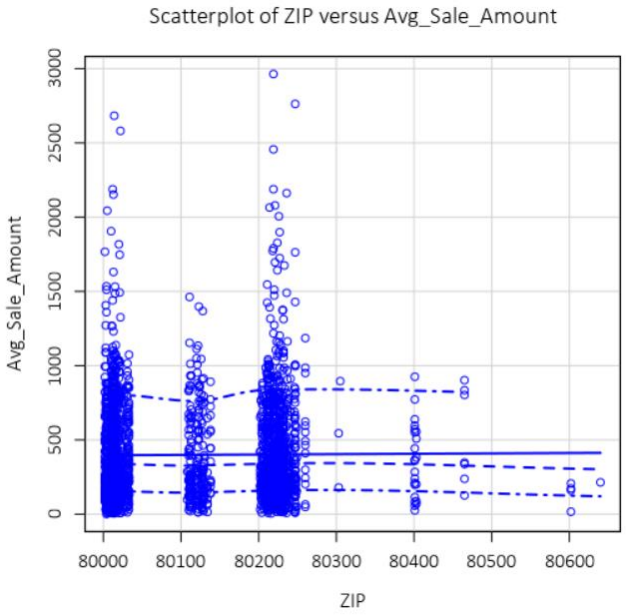
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.396e+03	2.150e+03	-0.6492	0.5163	
Customer_SegmentLoyalty Club Only	-1.498e+02	8.981e+00	-16.6818	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.819e+02	1.191e+01	23.6679	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.458e+02	9.776e+00	-25.1409	< 2.2e-16	***
ZIP	2.249e-02	2.661e-02	0.8453	0.39801	
Store_Number	-9.828e-01	1.007e+00	-0.9761	0.32912	
Avg_Num_Products_Purchased	6.692e+01	1.516e+00	44.1448	< 2.2e-16	***

Looking at the statistical results there is significant coefficients between *Customer_Segment* and A*vg_Num_Products_Purchased* as these are <0.05. We also see a this with *ZIP* and *Store_Number* is not statistically significant as the p_value >0.05.

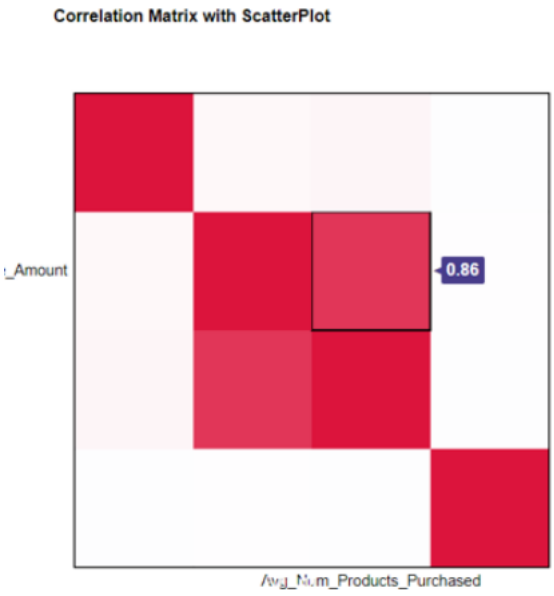
Plot of *ZIP* and *Store_number* to visualize the data.

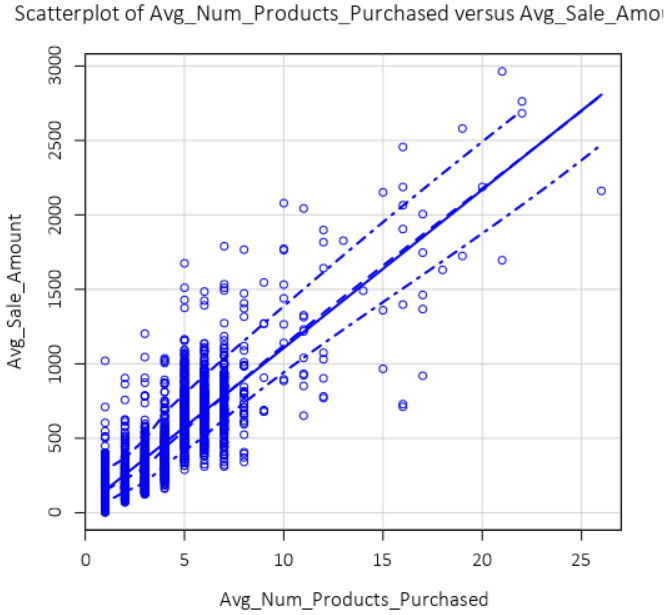




Looking at the plots we see there is no linear relationship as the slope is close to zero when comparing the variable for *Avg_Sale_Amount*.

Plot Avg_Number_Products_Purchased





Using the average number of products and average sales we see that there is a strong linear trend The correlation matrix shows we can further see that the Avg_Num_Products_Purchased the Avg_Sales_Sale is strong with an R value of 0.86. The Scatterplot shows that the is a linear relationship between the two. This is reflected in the Statisical summary showing the coefficients.

We can build our model using the Customer_Segment, Avg_Number_Products_Purchased.

Step 3: Presentation/Visualization

The statistical results of the model.

Report for Linear Model Demand_Catelog

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

The statistical result above show that the regression model is good as each predictor variable has a P value > 0.05. The second reasoning that this is a good model, is that the Adjusted R-Squared value is close to 1, with a value of 0.8366. This suggest that the predictor variables and the target variable in the model are statistically significant.

The best linear regression equation is as follows.

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

Applying Business Rules

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

My Recommendation is that the company should sent the catalogs to the 250 new Customers, as the predicted profit is greater than the stated $10,000.

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

After applying the model to mailing list, to get the predicted_sales_amount we applied the following as per requirement.

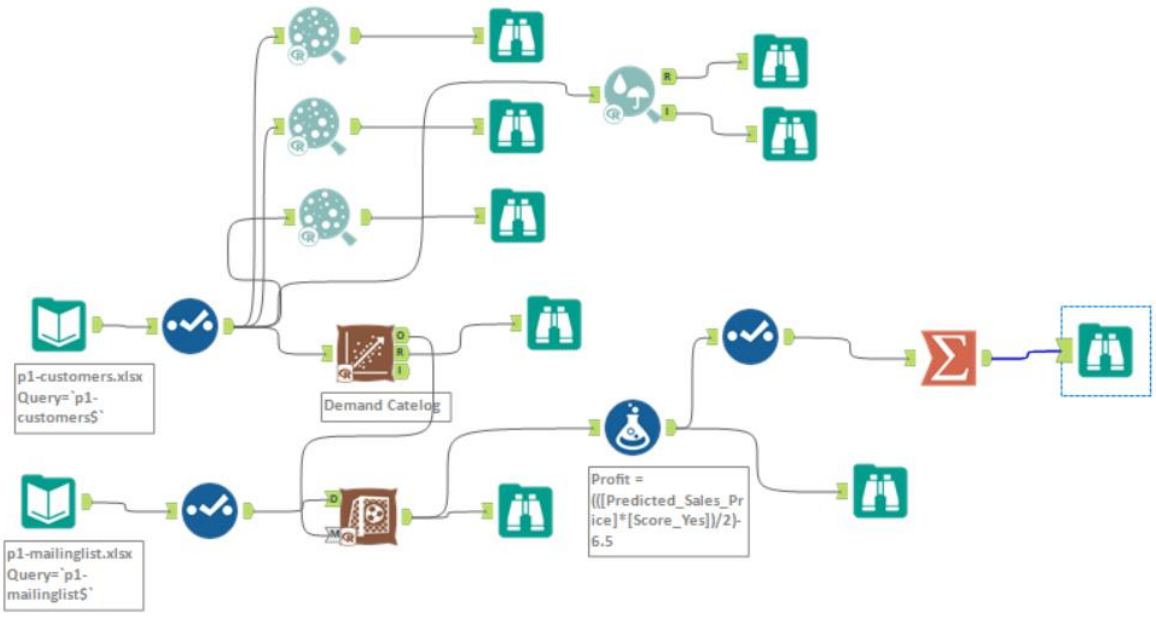
Multiply the predicted_sale_amount by score_yes (Score_yes is the likely hood the customer will buy) for each customer. Sum the amount above and multiply by the 50% gross margin. Then subtract $6.50 per customer.

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

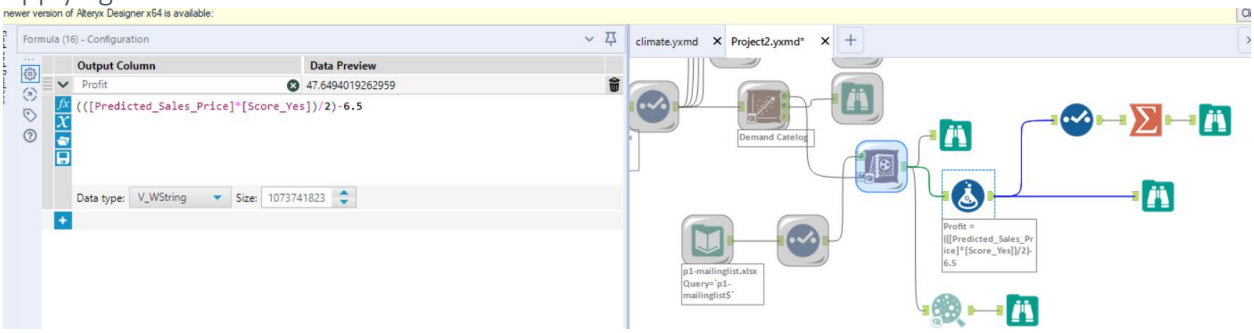
$21,987.44 = (sum([Predicted_Sale_Amout]*[Score_Yes])*.5)-6.5

Alteryx Workflow

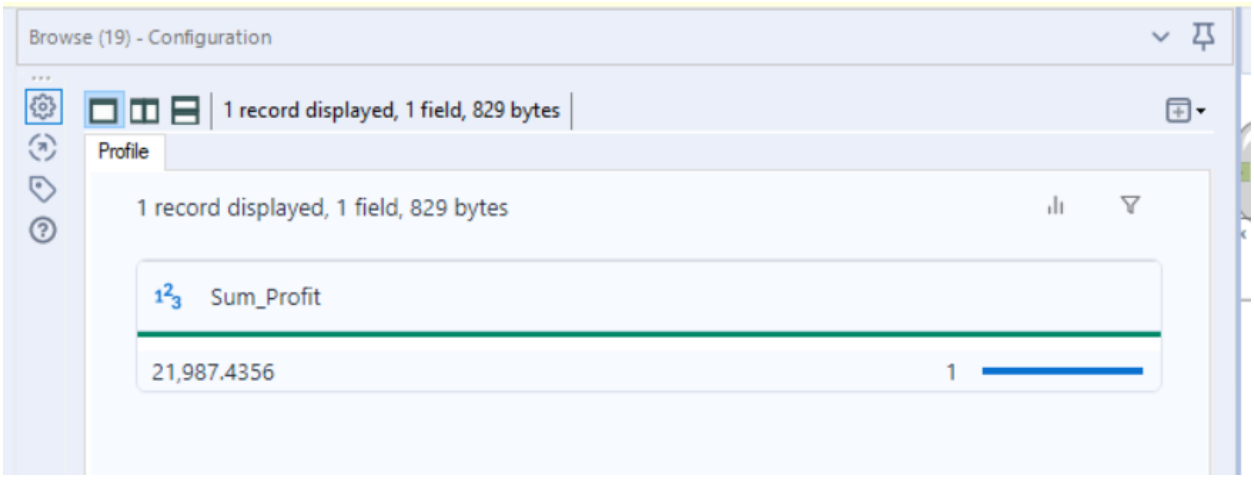
Workflow



Applying Business Rules



Calculated Profit



Resources

https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

https://courses.lumenlearning.com/wm-macroeconomics/chapter/interpreting-slope/

Supporting Model in Excel, to check usage of Alteryx