Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The business problem:
"You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While knowledgeable about data analysis, your manager is not very familiar with predictive models.
You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000."

The main outcome to the project is to predict whether a new catalogue release will generate enough income to justify launching the catalogue.

In order to do this, I would need to be able to predict the amount of sales the catalogue would generate when it is sent out to the company's current mailing list. On top of predicting average sales, the company needs to make a 50% margin and to also generate sales greater than the business' own requirement of $10,000.

Key Decisions:

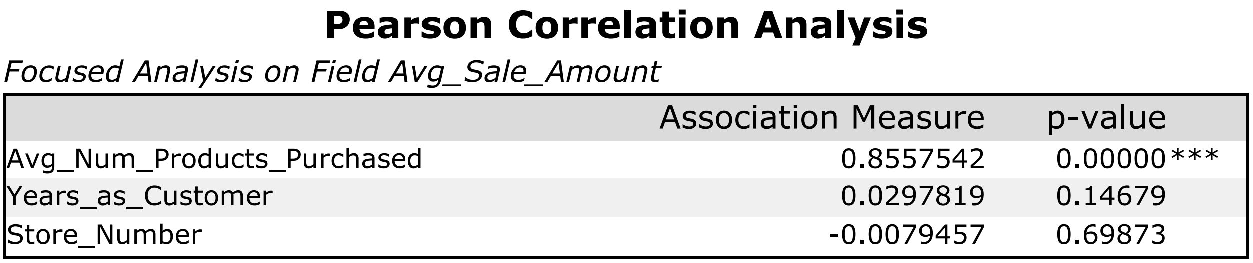
1. **What decisions needs to be made?**
   The business decision here is to decide whether to go with a new catalogue launch for their high-end home goods product range. There is a cost to producing these catalogues and an expected profit margin that needs to be obtained. The sum of the profits after costs and margin needs to also be greater than $10,000 for this catalogue campaign to be successful.

2. **What data is needed to inform those decisions?**
   For me to make my final recommendation, the data needed will be:
   - Avg Sale Amount
   - Avg Num Products Purchased
   - Years as Customer
   - Store Number
   - Customer Segment

   Further analysis of the above data points will be needed to construct an accurate linear model.

Step 2: Analysis, Modeling, and Validation

Firstly, let's look at the Pearson's Correlation between all the continuous predictor variables against the target variable (Average Sales Amount).



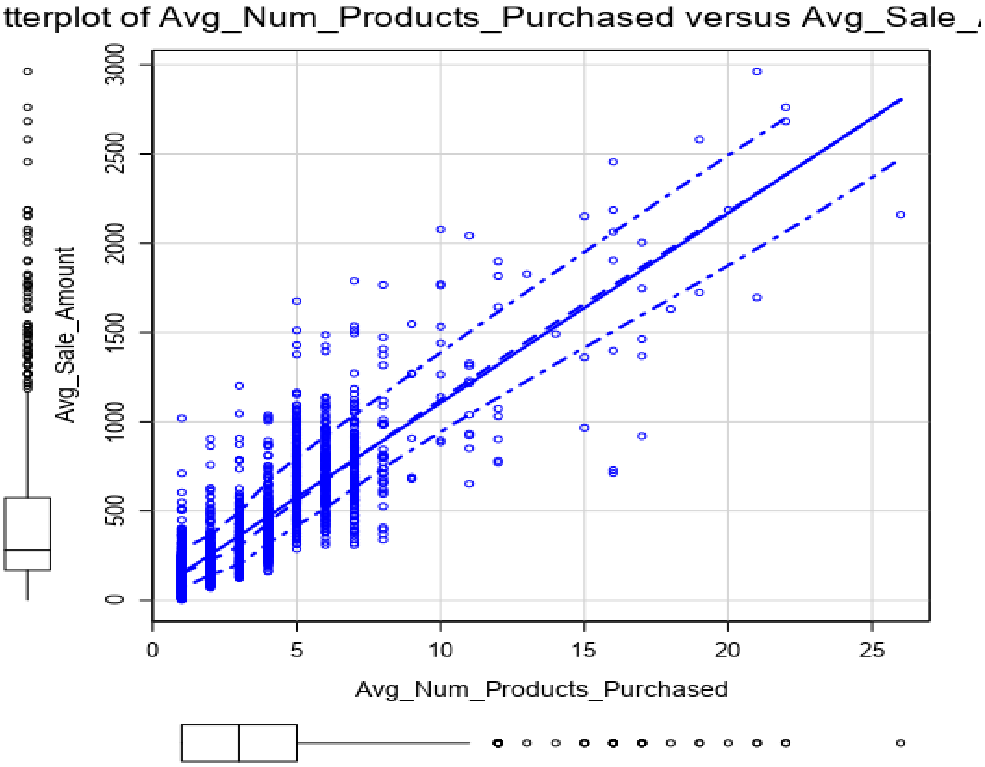
Pearson Correlation Analysis

Focused Analysis on Field Avg_Sale_Amount

	Association Measure	p-value
Avg_Num_Products_Purchased	0.8557542	0.00000 ***
Years_as_Customer	0.0297819	0.14679
Store_Number	-0.0079457	0.69873

From the Pearson Correlation Analysis, I can see that the best variable to use here is "Avg_Num_Products_Purchased". It has a high positive correlation of 0.8557542 with a near 0 p-value and an Alteryx analysis of three-star recommendation.

Next let's plot the Average Sales vs Average Number of Products Purchased.



The scatter plot would indicate as predicted from the Pearson correlation a positive relation between our two continuous predictor variables.

Customer Segment is a categorical variable found in our data and a variable that needs to be tested for to see if there is any association with the target variable. Using Alteryx and running a linear regression using Average Number of Products Purchased and Customer Segment, we get the below summary.

Report for Linear Model Linear_Regression_3

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 results above show that the coefficients generated from the Customer Segments indicate an association between the three categorical variables and the target variable and all three variables also have a very low p-value (<2.2e-16).

The Adjusted R-Squared Value for the model is 0.8366 indicating that 83.66% of the variance can be explained by the model itself.

My analysis indicates that I should use “Average Number of Products Purchased” and “Customer Segments” as my predictor variables.

The actual linear model we should use is:

Y (or Predicted Sales) = 303.46 + 66.98(Avg_Num_Products_Purchased) - 149.36(Customer_Segment Loyalty Club Only) + 281.84(Customer_Segment Loyalty Club and Credit Card) – 245.42(Customer_Segment Store Mailing List) + 0(Customer_Segment Credit Card Only)

Step 3: Presentation/Visualization

Results

In order to calculate the actual profit, we need to calculate the predicted profit from the catalogue, multiple by 50% due to the margin required and then subtract the cost of printing one catalogue per person.

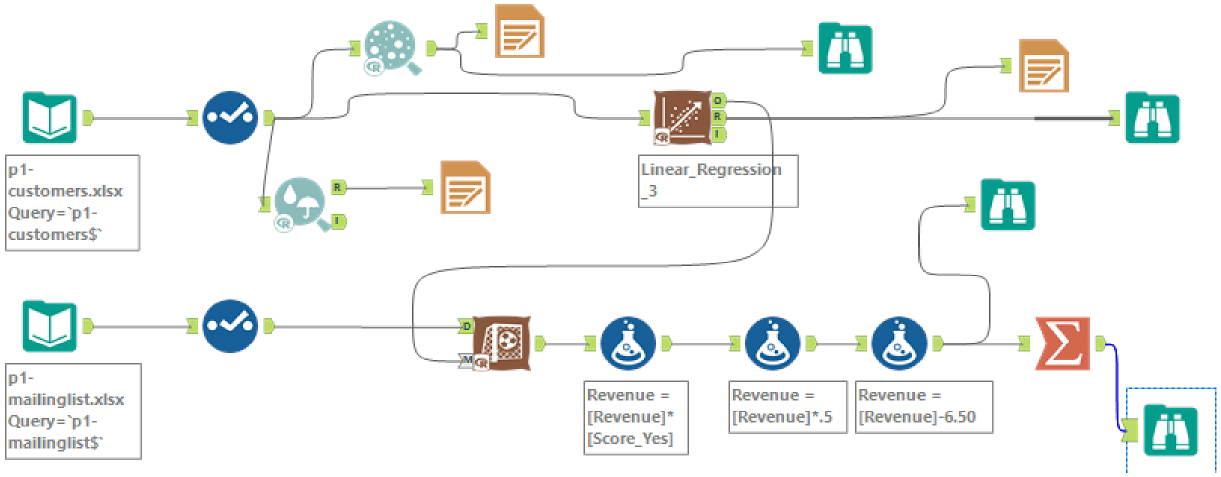
The total profit generated from the catalogue is $21,987.44.

Conclusions

In conclusion, I would recommend sending out the catalogue as it is predicted to return a profit of $21,987.44 after all costs, which is above the $10,000 required.

The reason for my recommendation is that the predicted sales generated from my linear regression model using “Average Number of Products Purchased” and “Customer Segments” as my predictor variables are predicted to generate enough profit after both margin and costs are taken away

Appendix.
Below is the Alteryx workflow.


p1-customers.xlsx Query=`p1-customers$`
Linear_Regression_3
O
R
I
R
I
D
M
p1-mailinglist.xlsx Query=`p1-mailinglist$`
Revenue = [Revenue]*[Score_Yes]
Revenue = [Revenue]*.5
Revenue = [Revenue]-6.50