Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

- Management from the company that manufactures and sells high-end home goods want to know if they should send a printed catalog to their 250 customers or not.
  They won't it out to these new customers unless the expected profit contribution exceeds $10,000.

2. What data is needed to inform those decisions?

- Predict the expected revenue from these 250 new customers using linear regression.
- Make sure to multiply Avg_Sale_Amount by Score_Yes to get predicted revenue.
- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin on all products sold through the catalog is 50%.
- The global expected profit = SUM(Predicted Avg_Sale_Amount * Score_Yes) * 50% – $6.50 * 250
- If the global expected profit >= $10,000, the company will send the catalog. If the expected profit < $10,000, the company won't send the catalog.

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables in your model?

- First, just by exploring the data, I did not select Name, Customer_ID, Address, State, Responded_to_Last_Catalog. Name and Customer_ID do not affect the sales amount. Address is so unique to the customers that it's rare that two customers have the same address. State is always CO in the dataset. Responded_to_Last_Catalog cannot be used in the linear regression model since it could not be applied to the mailing list data set.
- I run the linear regression model in Alteryx by selecting the target variable as Avg_Sale_Amount, and the predictor variables as Customer_Segment, City, Zip, Store_Number, Avg_Num_Products_Purchased and #_Years_as_Customer.

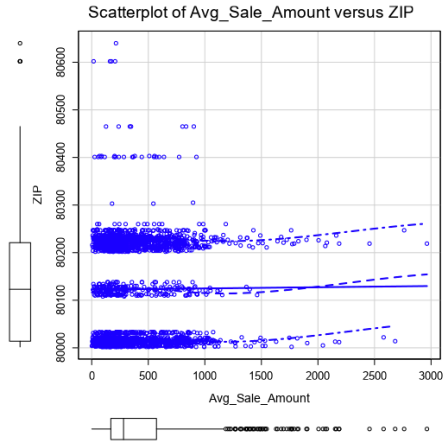
Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28108396.61	3	497.67	< 2.2e-16	***
City	513643.12	26	1.05	0.3956	
ZIP	97379.62	1	5.17	0.02304	*
Store_Number	49929.74	1	2.65	0.10355	
Responded_to_Last_Catalog	130123.59	1	6.91	0.00862	**
Avg_Num_Products_Purchased	36240138.84	1	1924.95	< 2.2e-16	***
X._Years_as_Customer	69147.95	1	3.67	0.05543	.
Residuals	44054205.89	2340			

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

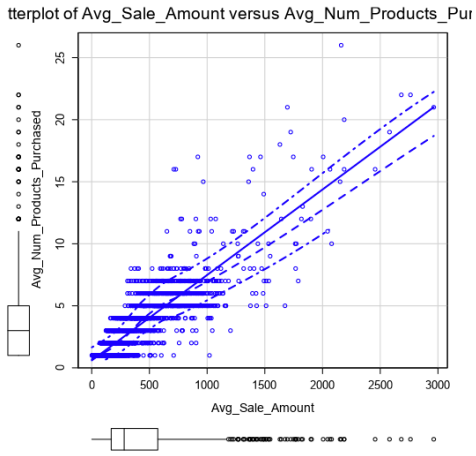
Based on the result above, seems that Customer_Segment, Zip and Avg_Num_Products_Purchased have a significant coefficient with Avg_Sale_Amount.

- **Plot Zip:**



- Apparently, there is no linear relationship with Avg_Sale_Amount.

- **Plot Avg_Num_Products_Purchased:**



- As expected, there is a linear relationship between Avg_Num_Products_Purchased and Avg_Sale_Amount.

- Customer_Segment is a categorical variable. The P-value is less than 0.05, so the relationship between Customer_Segment and Avg_Sale_Amount is considered to be statistically significant.
- In conclusion, the target variable is Avg_Sale_Amount; the predictive variables are Customer_Segment and Avg_Num_Products_Purchased. Customer_Segment is a categorical variable and Avg_Num_Product_Purchased is a continuous variable.

2. Explain why you believe your linear model is a good model.

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

Type II ANOVA Analysis

Response: Avg_Sale_Amount

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

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

- Based on the statistical results above, we can see, the P-value for each variable is less than 2.2e-16. Since the predictor variables have a p-value below 0.05, the relationship between it and the target variable Avg_Sale_Amount is considered to be statistically significant. Also, the adjusted R-squared value is 0.8366, which is good. Considering P-value and R-squared, the linear regression model is a good model.

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)

- Here are the coefficients from our linear equation from the above report:
  - Intercept: 303.46
  - Avg_Num_Products_Purchased: 66.98
  - Customer_Segment(Loyalty Club Only): 149.36
  - Customer_Segment(Loyalty Club and Credit Card):281.84
  - Customer_Segment(Store Mailing List):245.42
  - Customer_Segment(Credit Card Only):0

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)

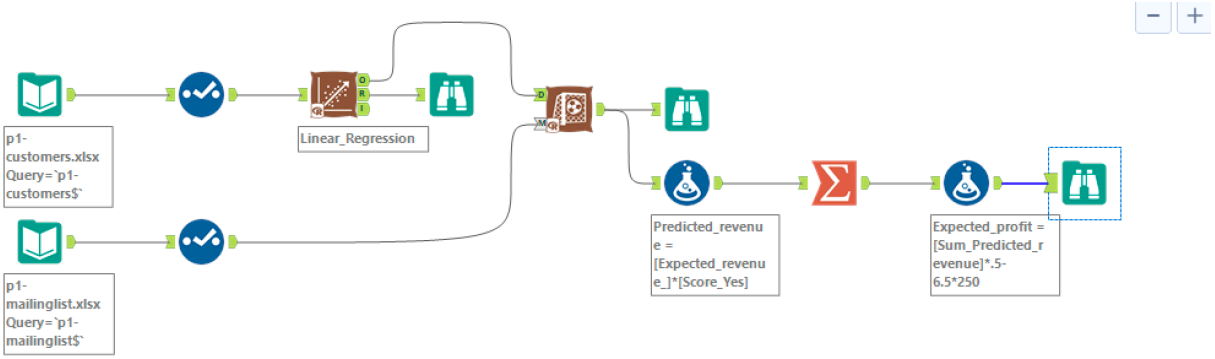
Step 3: Presentation/Visualization

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

- My recommendation is that the company should send the catalog to these 250 customers. The expected profit is $21,987.44, which is higher than $10,000, so the company should send the catalog to these 250 new customers.

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

Below is an an Alteryx workflow for our problem:



- After configuring the linear regression, I applied the model to the mailing list data set to get the expected revenue. Then multiply expected revenue by Score_Yes (which is the probability to buy) for each customer to get predicted revenue. Then I multiplied this value by 50% which is the gross margin and subtracted the catalog cost ($6.5 *250). So, the expected profit = SUM(Predicted Avg_Sale_Amount * Score_Yes) * 50% – $6.50 * 250.

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

- Using the formula and the process that I explained in the previous question, the expected profit from the new catalog is $21,987.44.

Here is how I come up with the result (it is done in Alteryx):

- The expected profit = SUM(Predicted Avg_Sale_Amount * Score_Yes) * 50% – $6.50 * 250
- The predicted Revenue = SUM(Predicted Avg_Sale_Amount * Score_Yes) = $47,224.87
- The expected profit = $47,224.87 * 50% – $6.5 *250 = $21,987.44