Project 1: Predicting Catalog Demand

Key Decisions:

1. What decisions needs to be made?

   The company has to decide whether or not to send the second catalog to a set of 250 new customers to make profits. The company will be profitable only if they make total gross profits over $10,000. So, we need to determine the profit the company is going to make if they decide to send the new catalog to the 250 new customers.

2. What data is needed to inform those decisions?

   In order to determine whether the company is going to make profits by sending the catalog first we need to predict the total profit for the 250 new customers. To calculate profits we need to predict the total sales for the new customers. So first we start by determining the individual average gross sales/revenue per customer in the new dataset.

   To predict the average sales per customer, we need to examine the data of old customers and new customers. In the old customer data, we have several fields like Customer Segment, City OR ZIP, average number of products purchased and number of years as customer and how they responded to the catalog which can be used to understand individual customers. We also have data about the average sales per customer combined with other variables that can be used to predict sales for our new set of customers.

   Now if we examine the new customers' data set, there are similar fields like ZIP OR city, Number of products purchased and number of years as customer which can all be used to build our model if they are also statistically significant.

   We do not have a field for 'Responded to catalog' field in the new customers' data but instead we have a score field for Yes and No values which is probability with the customer will respond. This can be used to calculate the exact sale value for each customer by multiplying the predicted value with the probability the customer responded (score yes) to the catalog which will give us a good estimate of predicted average sales value for each customer.

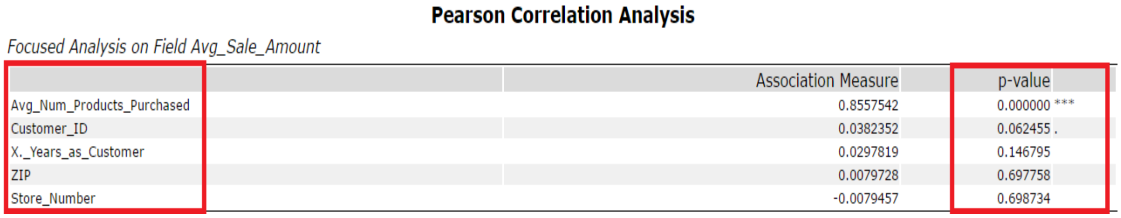
   After examining the existing both datasets, we see that we have previous data for the Average gross sales which is our target variable for new customers.

   Once we have the individual gross sales we can calculate the gross profit value by multiplying the sales with gross margins (50% in this case) and deducting the price of making the catalog ($6.50 here) and he we arrive at final gross profit for individual customers. Next we can sum the gross profits and compare this value with $10,000 which will help us make the decision.

Analysis, Modeling, and Validation

1. How and why did you select the predictor variables (see supplementary text) in your model?

After examining the data in past customer data and understanding the business problem, I selected ‘Avg_Sale_Amount’ as the target variable. I noticed that Customer_ID, ZIP, Store_number, Avg_num_of_products_purchased, #_years_as_customers are numeric fields with ‘double’ datatype. I also did association analysis on Alteryx and examined the p-values.



Pearson Correlation Analysis

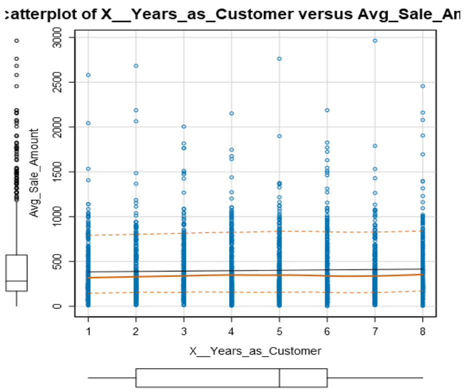
Focused Analysis on Field Avg_Sale_Amount

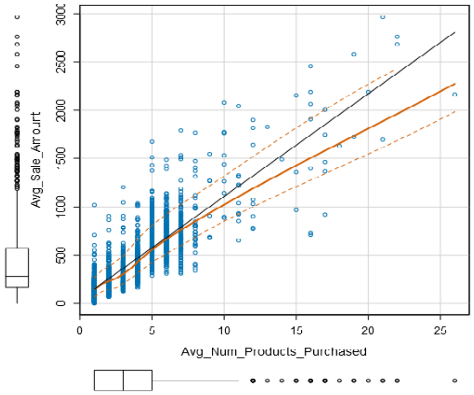
		Association Measure	p-value
Avg_Num_Products_Purchased		0.8557542	0.000000 ***
Customer_ID		0.0382352	0.062455 .
X._Years_as_Customer		0.0297819	0.146795
ZIP		0.0079728	0.697758
Store_Number		-0.0079457	0.698734

Looking at the plots and p-values (which are less than 0.05); we can determine to use only ‘Average_num_of_products_purchased’ as one of our predictor variables.

Also the fields, ZIP, ‘Customer_ID’ and ‘Store_number’ are categorical variables. These cannot be considered as good predictors because they can have multiple arbitrary values.

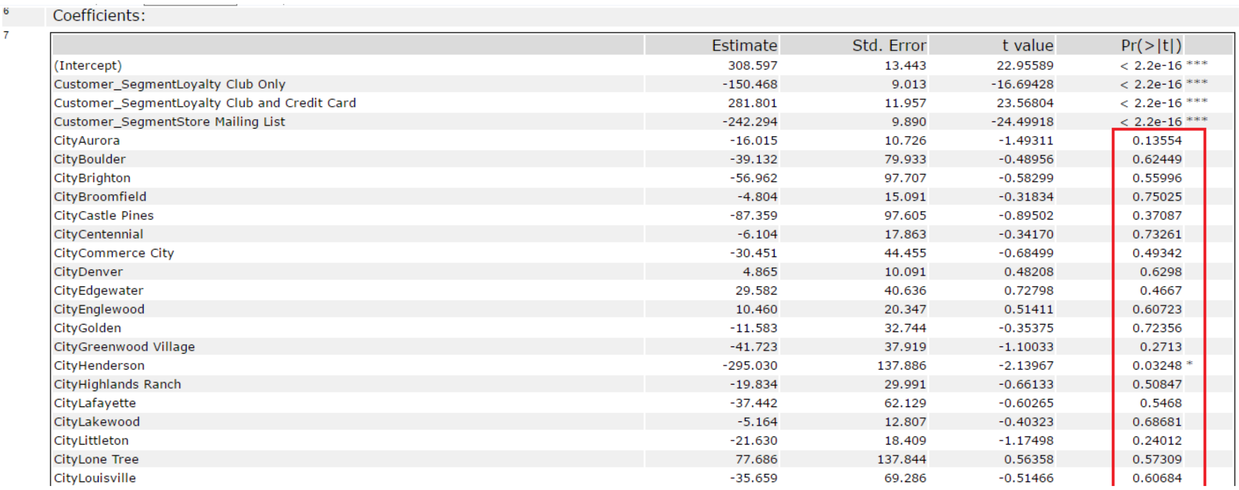
Now I also created a scatter plot with different numeric fields on X-axis and ‘Avg_sale_amount’ on Y-axis to make sure we have a linear relationship.





After examining the plots we see that ‘Average_num_products_purchased’ only has a linear relationship with our target variable hence only this can be our predictor variable.

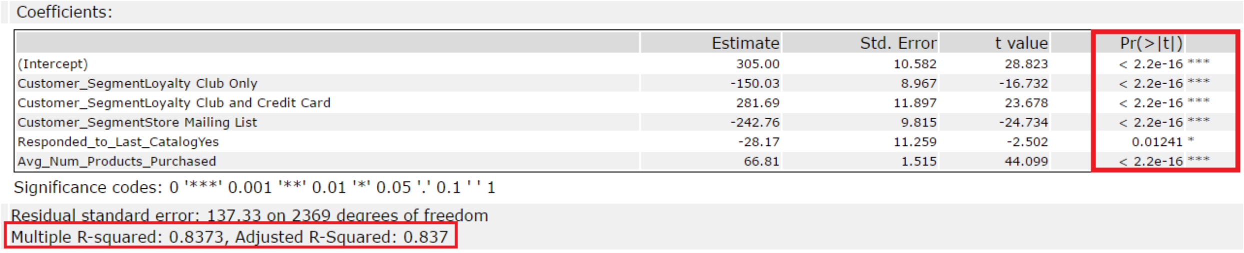
Now looking at the other fields which are ‘Vstring’ datatype, Name and Address fields cannot be considered because there are multiple values and can have any arbitrary values associated with them and there is no particular baseline category we can compare the values to. We cannot consider ‘State’ because it contains only one value.
Now I used Linear Regression tool and built a model using ‘Average_num_of_products_purchased’ and other fields like ‘City’, ‘Customer_Segment’ and ‘Responded_to_last_catalog’.



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	308.597	13.443	22.95589	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-150.468	9.013	-16.69428	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.801	11.957	23.56804	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-242.294	9.890	-24.49918	< 2.2e-16 ***
CityAurora	-16.015	10.726	-1.49311	0.13554
CityBoulder	-39.132	79.933	-0.48956	0.62449
CityBrighton	-56.962	97.707	-0.58299	0.55996
CityBroomfield	-4.804	15.091	-0.31834	0.75025
CityCastle Pines	-87.359	97.605	-0.89502	0.37087
CityCentennial	-6.104	17.863	-0.34170	0.73261
CityCommerce City	-30.451	44.455	-0.68499	0.49342
CityDenver	4.865	10.091	0.48208	0.6298
CityEdgewater	29.582	40.636	0.72798	0.4667
CityEnglewood	10.460	20.347	0.51411	0.60723
CityGolden	-11.583	32.744	-0.35375	0.72356
CityGreenwood Village	-41.723	37.919	-1.10033	0.2713
CityHenderson	-295.030	137.886	-2.13967	0.03248 *
CityHighlands Ranch	-19.834	29.991	-0.66133	0.50847
CityLafayette	-37.442	62.129	-0.60265	0.5468
CityLakewood	-5.164	12.807	-0.40323	0.68681
CityLittleton	-21.630	18.409	-1.17498	0.24012
CityLone Tree	77.686	137.844	0.56358	0.57309
CityLouisville	-35.659	69.286	-0.51466	0.60684

Due to high p-values for ‘City’, we removed this as a predictor variable and obtained the following report with a significant R-squared value.



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	305.00	10.582	28.823	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-150.03	8.967	-16.732	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.69	11.897	23.678	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-242.76	9.815	-24.734	< 2.2e-16 ***
Responded_to_Last_CatalogYes	-28.17	11.259	-2.502	0.01241 *
Avg_Num_Products_Purchased	66.81	1.515	44.099	< 2.2e-16 ***

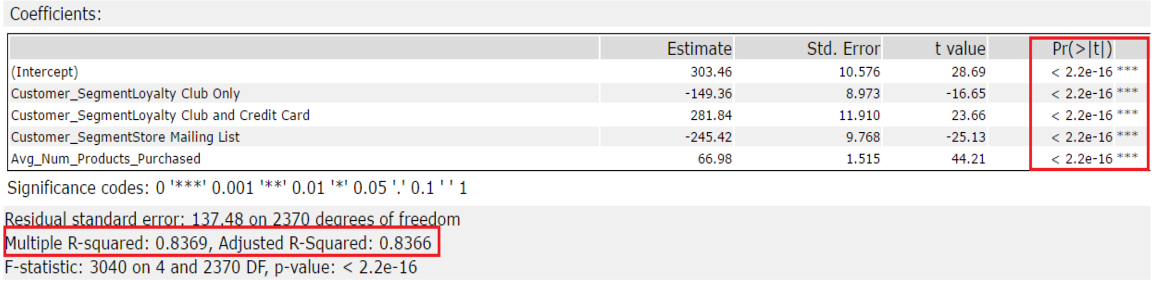
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.33 on 2369 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.837

However after looking at the new dataset to which we need to apply the linear equation, I see that there is no field corresponding to ‘Responded_to_the_catalog’. So we cannot use this variable in our equation.
Hence, we can conclude that only ‘Customer_Segment’ and ‘Average_num_of_products_purchased’ are our two statistically significant predictor variables.

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

After finalizing the predictor variables, I arrive at the following p-values and R-values



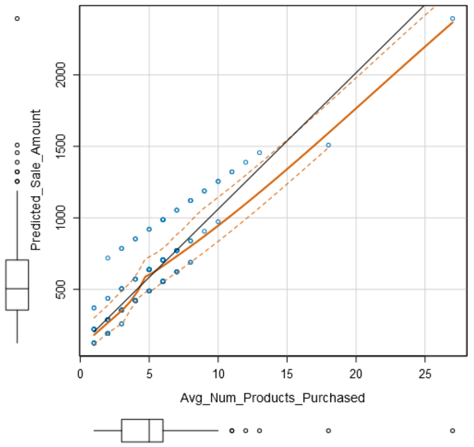
Coefficients:

	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 DF, p-value: < 2.2e-16

Customer_Segment and Avg_num_products_purchased (p-value) is <2.2e-16 and significance ***. Adjusted R-squared value is 0.8366.
The adjusted R-value here is 0.8366, which is close to 1 hence I can say that my model is a good fit. Also all the p-values of the predictor variables are less than 0.05 which suggest that the relationship between the target and predictor variable is statistically significant.
While a high R-squared is not a guarantee that the model is good, in this case I also created a scatter plot between the predicted sales on Y-axis and 'Avg_num_products_purchased' on X-axis as shown below



After examining the scatter plot as well as the R-squared value, we can say that this is a good model for predicting sales.

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)

   Avg_Sale_Amount = 303.46 +66.98 * (IF Avg_Num_Products_Purchased) -149.36 * (IF Customer_Segment Loyalty Club Only) +281.84 *(IF Customer_Segment Loyalty Club and Credit Card) -245.42 *(IF Customer_Segment Store Mailing List) + 0 * (IF Customer_Segment Credit card only)

<u>Presentation/Visualization</u>

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

   Using the linear regression model, we are able to predict the total average gross profit from the 250 new customers will be approximately $21987.44. Since the expected gross profit is greater than $10,000 it is recommended we send the second catalog to the set of 250 new customers in the list.

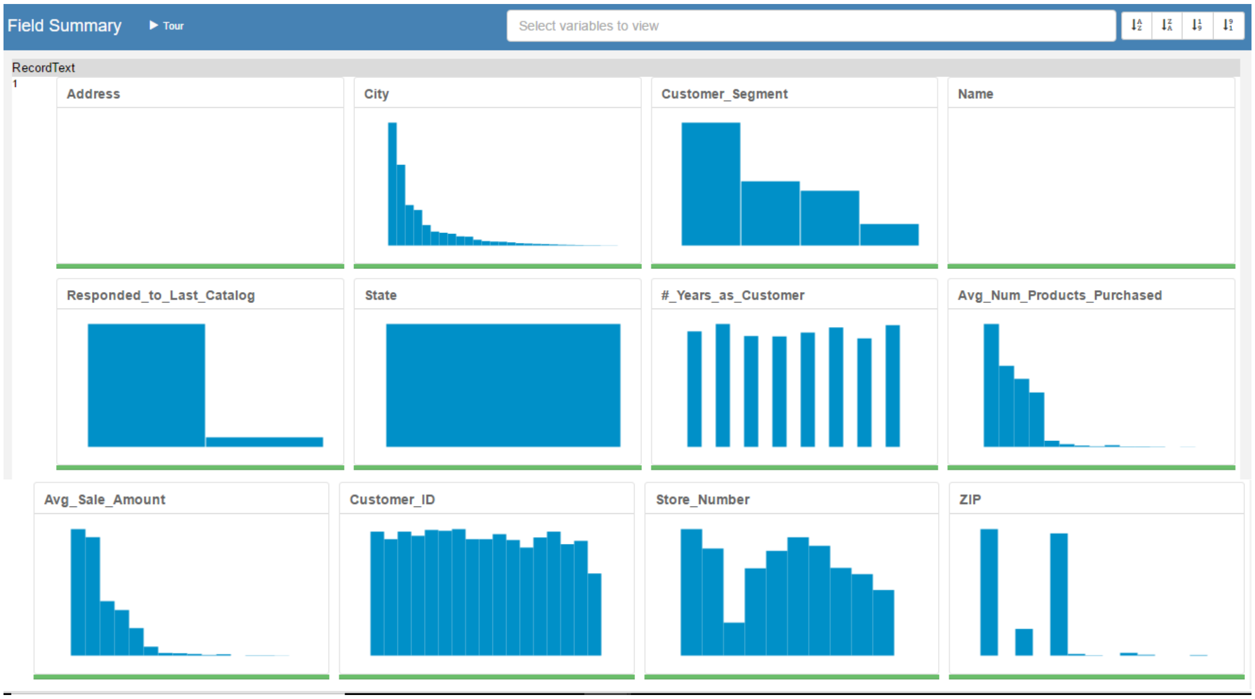
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

Once we have the predicted sales amount for each customer, in order to calculate the exact revenue we need to consider the probability, which means we need to multiply the predicted sales with probability which is Score_yes value for each customer. Now we calculate the average gross profit per customer by multiplying the average sales amount with 0.5 (since 50% is gross margin) and deducting the $6.50 price for catalog. Now that we have average gross profit per customer, we can summarize by adding the total average gross profit which is $21987.44.

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

   Assuming we send out the catalog to these 250 new customers, we can expect an average gross profit of $21987.44.

Rubric Question: Show the distributions for each variable in the Customer List dataset.


Field Summary
▶ Tour
Select variables to view
RecordText
1
Address
City
Customer_Segment
Name
Responded_to_Last_Catalog
State
#_Years_as_Customer
Avg_Num_Products_Purchased
Avg_Sale_Amount
Customer_ID
Store_Number
ZIP