Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   The Management of the company that manufactures and sells high-end home goods wants to determine whether they should send the printed catalogs to the 250 new customers from their mailing list based on the criteria that only if the expected contribution from these new customers exceeds $ 10,000, the company would send catalogs or else not.

2. What data is needed to inform those decisions?

   a) Predicted revenue from the 250 new customers
   b) Cost of goods sold to the 250 new customers
   c) Expected profit contribution from the 250 new customers
   d) Score value from the 250 new customers
   e) Justified predictor and target variables from the historical customers data set

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

   To start with, I explored the historical customers data and ruled out Name, Customer ID, Address, City, State, Responded to Last Catalog since they don't affect the revenue. Name, Address and Customer ID are unique fields for each customer and it's uncommon that two customers have the same identities. State is selfsame across the data set. I ran a frequency table over the responded to last catalog and ascertained that a large portion of customers did not respond to the catalog in the previous instances and cannot be appended to the new customers mailing list data set.

   I then ran the association analysis to determine the pearson correlation by taking the avg sale amount as target field and zip, store number, avg sale amount, avg num of products purchased and years as customers as associated fields.

Pearson Correlation Analysis

Focused Analysis on Field Avg_Sale_Amount

	Association Measure	p-value
Avg_Num_Products_Purchased	0.8557542	0.00000 ***
X._Years_as_Customer	0.0297819	0.14679
ZIP	0.0079728	0.69776
Store_Number	-0.0079457	0.69873

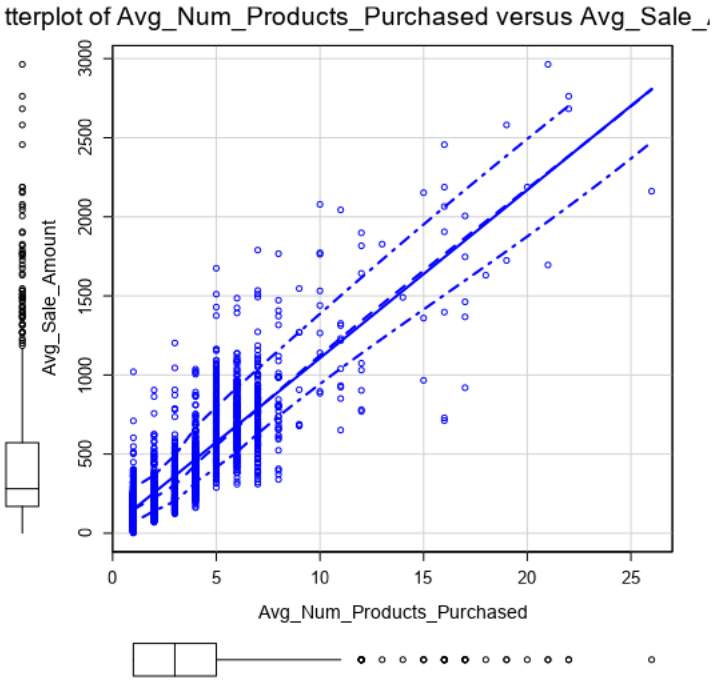
Full Correlation Matrix

	Avg_Sale_Amount	ZIP	Store_Number	Avg_Num_Products_Purchased	X._Years_as_Customer
Avg_Sale_Amount	1.0000000	0.0079728	-0.0079457	0.8557542	0.0297819
ZIP	0.0079728	1.0000000	-0.1489063	0.0017896	0.0016432
Store_Number	-0.0079457	-0.1489063	1.0000000	-0.0115250	-0.0095729
Avg_Num_Products_Purchased	0.8557542	0.0017896	-0.0115250	1.0000000	0.0433464
X._Years_as_Customer	0.0297819	0.0016432	-0.0095729	0.0433464	1.0000000

Matrix of Corresponding p-values

	Avg_Sale_Amount	ZIP	Store_Number	Avg_Num_Products_Purchased	X._Years_as_Customer
Avg_Sale_Amount		6.9776e-01	6.9873e-01	0.0000e+00	1.4679e-01
ZIP	6.9776e-01		3.0154e-13	9.3054e-01	9.3621e-01
Store_Number	6.9873e-01	3.0154e-13		5.7454e-01	6.4101e-01
Avg_Num_Products_Purchased	0.0000e+00	9.3054e-01	5.7454e-01		3.4659e-02
X._Years_as_Customer	1.4679e-01	9.3621e-01	6.4101e-01	3.4659e-02	

Based on the above, avg number of products purchased shows a significant coefficient with avg sale amount. I plotted the variable on the scatterplot and apparently there is a linear relationship between avg number of products purchased and avg sale amount.



I then ran a linear regression tool to determine my categorical predictive variable taking customer segment as the sole predictor variable as other variables are inappropriate as categorical variables.

Report for Linear Model Linear_Regression_28

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-1001.85	-71.66	3.08	73.02	1889.33

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	682.7	8.354	81.72	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-286.3	11.372	-25.18	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	391.5	15.732	24.89	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-525.3	10.045	-52.30	< 2.2e-16 ***

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

Based on the above, customer segment depicts a high statistical significance with avg sale amount. In conclusion, I considered the target variable as avg sales amount, avg number of products purchased along with customer segment as quantitative and categorical predictor variables sequentially.

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.

Report for Linear Model Linear_Regression_18

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 above statistical results, the P-value for each variable is less than < 2.2e-16 or 0.05 indicating that the relationship between the predictor variables and target variable is statistically significant. Also, the R-squared value is 0.8369 and Adjusted R-squared value is 0.8366 which are above 0.7 indicating a strong relationship. All things considered; I conclude this linear regression model as 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)

   Avg_Sale_Amount = 303.46 + 66.98*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_Segement: Store Mailing List) + 0 (If Customer_Segment: Credit Card Only)

Step 3: Presentation/Visualization

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

   My Recommendation would be that the company should send its first print catalogs to these 250 new customers since the expected profit is $ 21,987.41 which is notably higher than $ 10,000.

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

   To begin with, I derived a formula using the linear regression model.

   Avg_Sale_Amount = 303.46 + 66.98*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_Segement: Store Mailing List) + 0 (If Customer_Segment: Credit Card Only)

   I then ran the score tool to apply this model to the mailing list data set and to arrive at the score value. Following that, I used the formula tool to multiply the score value by Score_Yes (Probability to buy) for each customer to determine the predicted revenue.

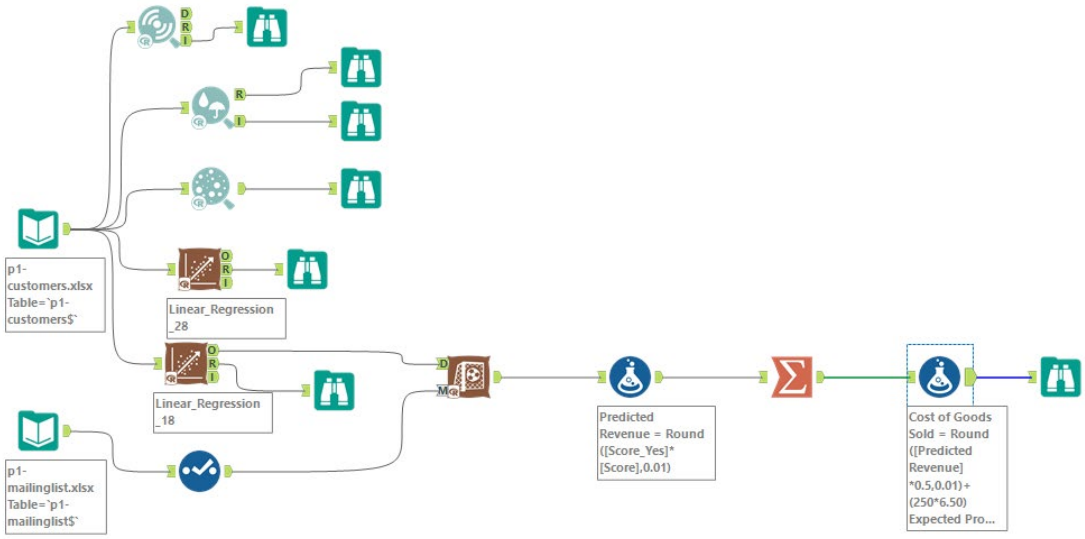
   Later, I ran the summarize tool to sum the predicted revenue. Ensuing this, I ran the formula tool again to calculate the cost of goods sold and to ultimately arrive at an expected profit.

   Cost of Goods Sold = Round([Predicted Revenue]*0.5,0.01)+(250*6.50)

   PS: Where in 50% is the average gross margin and costs of printing and distributing per catalog is $ 6.50

   Expected Profit = [Predicted Revenue] - [Cost of Goods Sold]

See the Alteryx Workflow below:



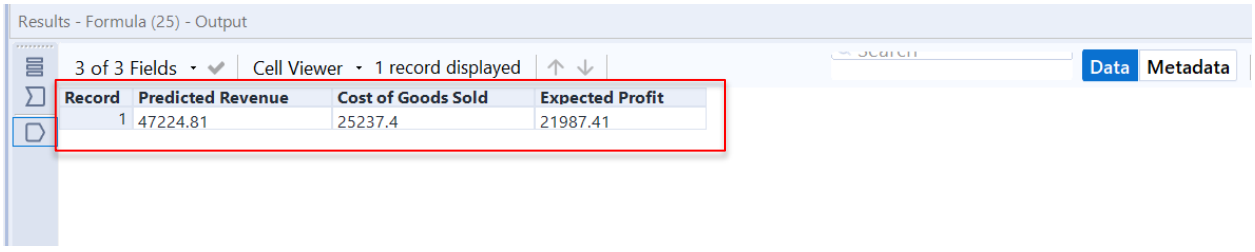
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

   The expected profit from the new catalog is $ 21,987.41.

   Predicted Revenue = Sum (Score*Score_Yes) = $ 47,224.81

   Cost of Goods Sold = ($47,224.81*0.5) + (250*$6.5) = $ 25,237.40

   Expected Profit = $ 47,224.81 - $ 25,237.40 = $ 21,987.41



Record	Predicted Revenue	Cost of Goods Sold	Expected Profit
1	47224.81	25237.4	21987.41