Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

I need to predict 250 costumers' sales so I can know the revenue in order to estimate the profit and make sure contribution exceeds $10,000 and decide whether the catalog should be sent or not?

2. What data is needed to inform those decisions?

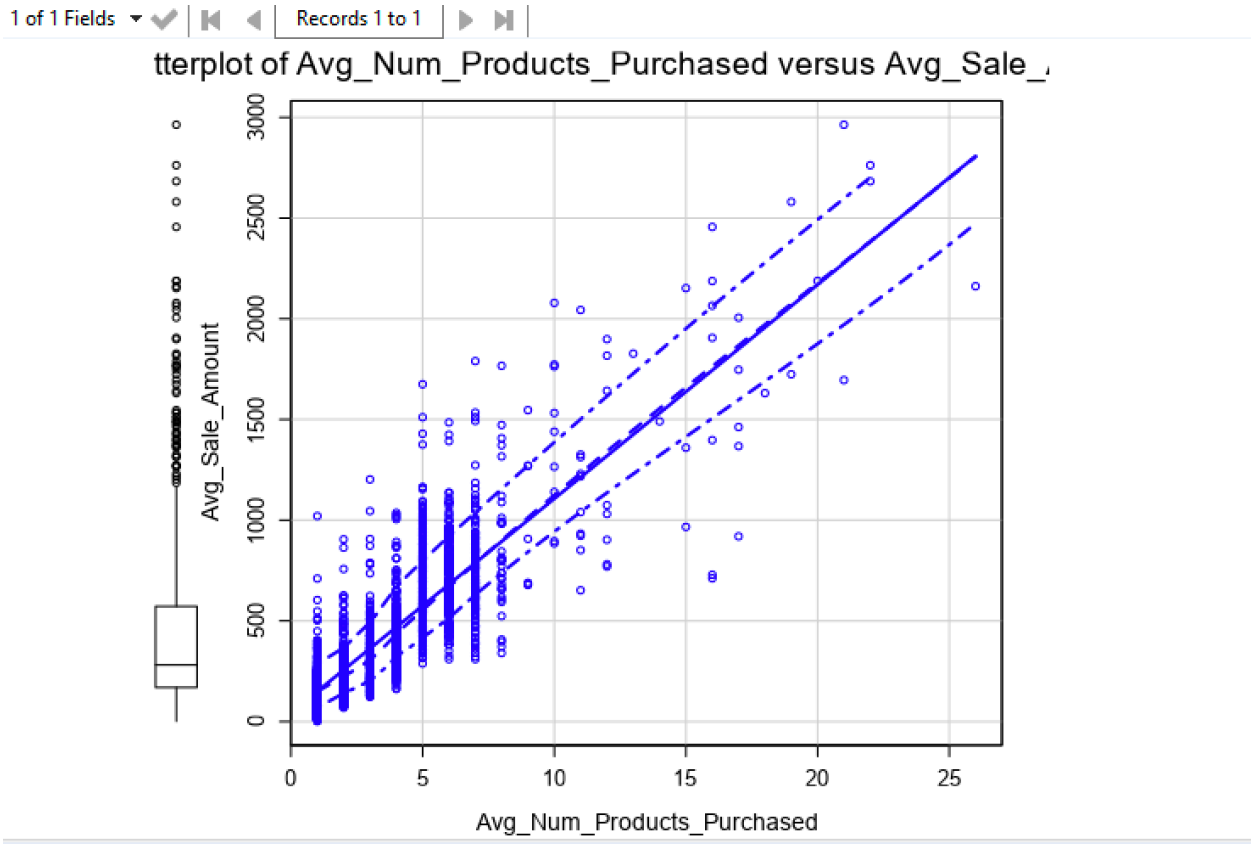
Since there are historical data about sales which includes costs of producing and shipping the catalog, I can predict the sales for current year and multiply it with the probability that a costumer will respond to a catalog and make a purchase (Score_Yes)

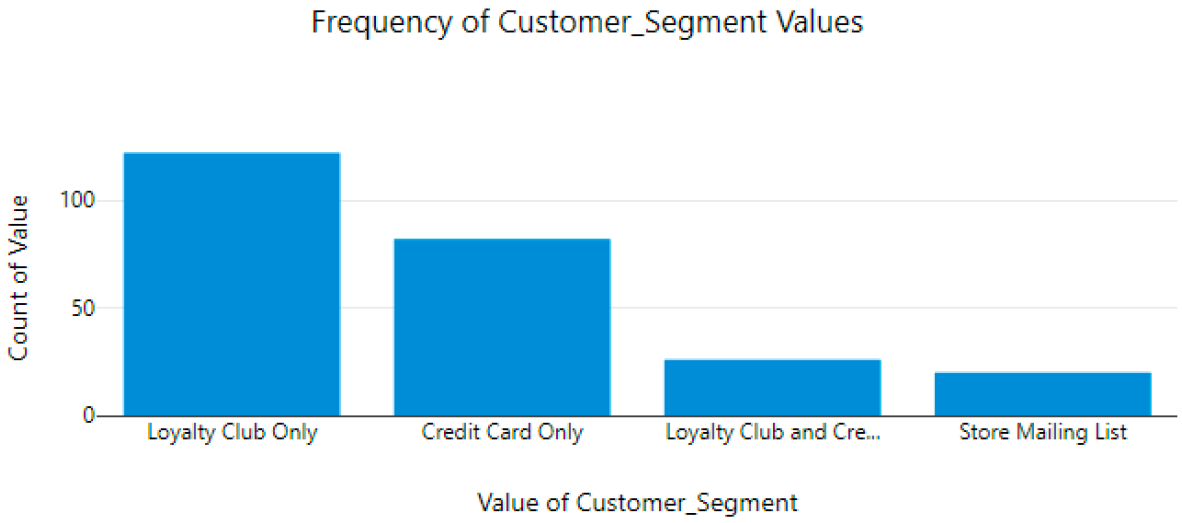
Step 2: Analysis, Modeling, and Validation

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

I checked the relationship between each variable and sales price wherever there is weak relationship I exclude the variable and when there is a strong linear relationship, I selected it as a predictor Variable.

only Avg_Number_Of_Products_Sold seemed linearly related with target variable And I selected Costumer_Segment and City too in my linear model





After running the linear model, I found out that probability of city's coefficient is going to be 0 is high so I removed it. The other two predictors are significant in deciding the target variable, so I kept them.

Record	Report
1	**Report for Linear Model Linear_Regression_21**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	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

8 Residual standard error: 137.48 on 2370 degrees of freedom

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

The model looks strong because the R-value is strongly high and predicted values are very close to actual values. P-values are a strong evidence of the capability of this model based on the table above.

Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366

2. What is the best linear regression equation based on the available data?

Sales = 303.46 -149.36*LoyaltyClubOnly* + *281.84* LoyaltyClubAndCreditCard - 245.42*StoreMailingList* + *0* Credit_Card_Only + 66.98*Avg_Number_Products_Purchased

Step 3: Presentation/Visualization

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

The profit exceeds $10000 so yes, they should send the catalog.

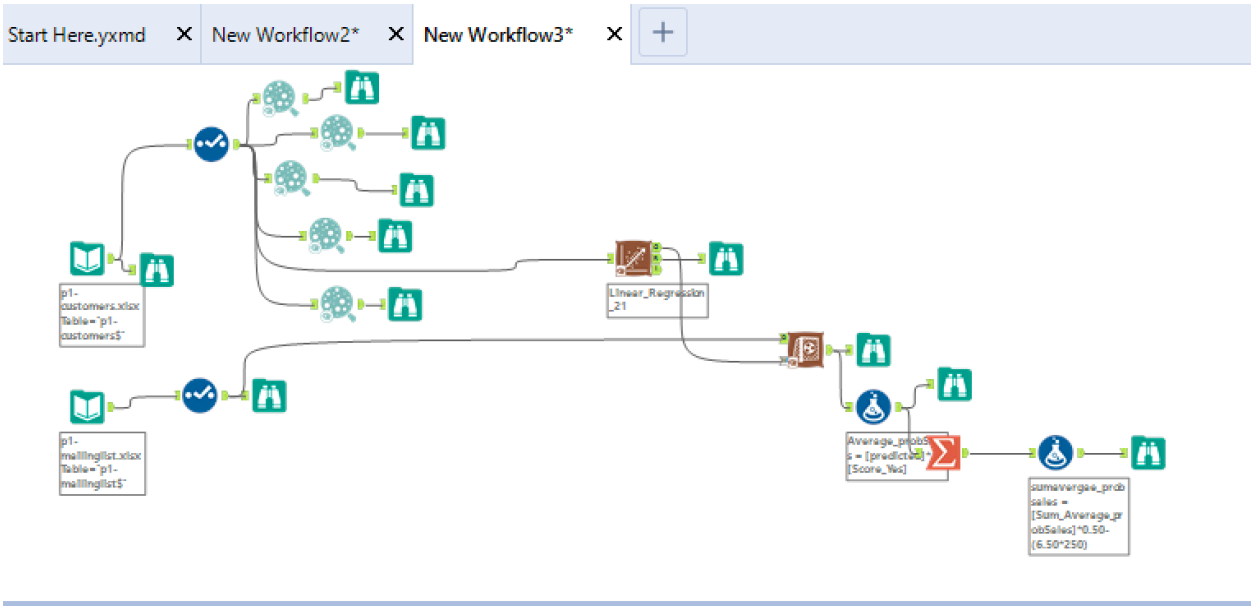
2. How did you come up with your recommendation?

added new column (Avg_ProbSales = PredictedAvgSales * Score_Yes) Given profit margin 50%, cost for each catalog is $6.50, hence calculated for all 250 customers

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

Profit = AvgProbableSales0.5-(6.50*250)**

sumavergae_probsales
21987.435687


Start Here.yxmd
New Workflow2*
New Workflow3*
p1-customers.xlsx Table='p1-customers$'
Linear_Regression_21
p1-mailinglist.xlsx Table='p1-mailinglist$'
Average_probSales = [predicted]*[Score_Yes]
sumaverage_probsales = [Sum_Average_probSales]*0.50-(6.50*250)