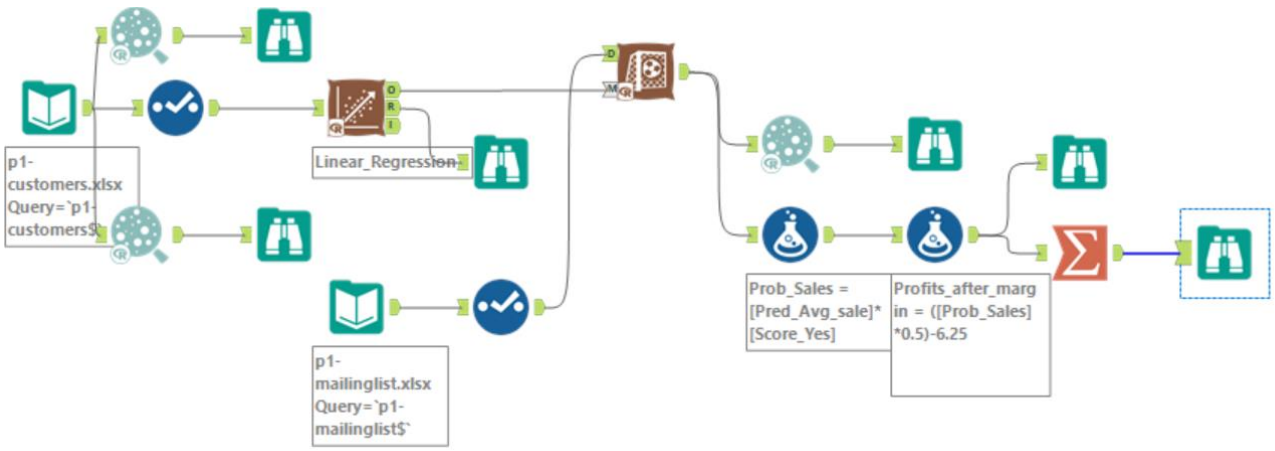
Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
    - Whether to send the catalogs to the new customers or not?
2. What data is needed to inform those decisions?
    - The expected profit by sending the catalogs to the new customers in mailing list
    - We Compute profit by using sales amount data
    - The sales amount data for the new customers is to be predicted using regression modelling using the previous customer data as training set

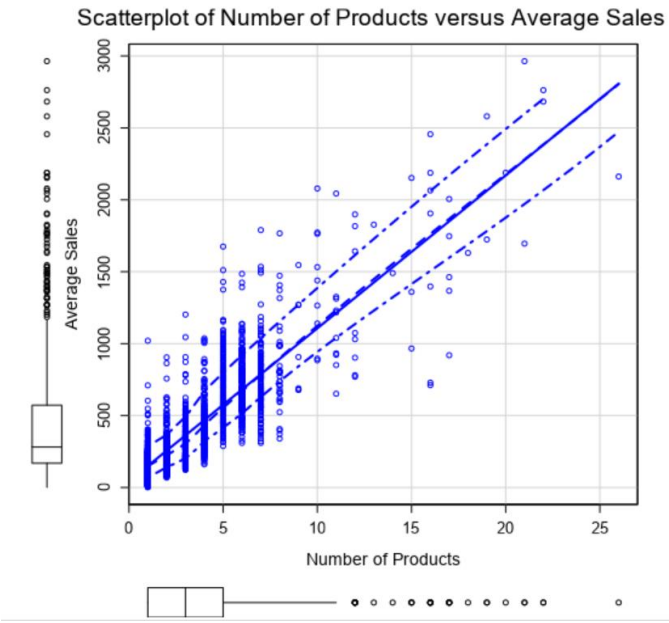
Step 2: Analysis, Modeling, and Validation



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

For the target variable Average Sales amount, I selected 2 of the variables in the training set as predictors for the model

1. Average number of products purchased: From a Scatter Plot between this variable and the target variable, I could infer that they have a strong linear relationship (The same was not observed with 'Number of years as customer' variable)



2. Customer Segment: After running a test run by using this variable as predictor variable. The Multiple R- squared value is greater than 0.7 and the high Co-efficient value indicated its strong linear relation with target variable. (Refer the image in next question)

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.

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	***

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

P Values of all the variables are significantly lesser than 0.05, hence they have significant relationship with the target variable, the high value of R-squared indicates the high degree to which the data is explained by the 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)

The regression equation is:

Average Sales Amount = 303.46 -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)
*+ 66.98 * Average number of Products purchased*

Step 3: Presentation/Visualization

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

    Yes, the company should send the catalog to these 250 customers

2. How did you come up with your recommendation?

    From the results of the predictive analysis, it was found that the profit the company can make was greater than the minimum threshold of $ 10,000 the management had decided

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

    The expected profit after considering the margin is **$ 22,049.93**