PROJECT 1: PREDICTING CATALOG DEMAND

Step 1: Business and Data Understanding

Key Decisions:

1) **What decisions need to be made?**

- Since the business problem is to predict the overall profit by sending catalog to 250 customers. Our decision should be 'Whether to send the catalog or not, based on expected profit exceeding $10,000.

2) **What information do we need to inform the decision?**

- To support the decision first we need to understand whether or not the customer will respond to the catalog and buy products from it. If customer responds and buys the product, we need to predict the worth of products purchased by customers. Now, we need to verify that do we have enough of data to support the above decisions. By going through the excel files which are provided in this project as a data source, we understood that probability of customer response to the catalog is given in *p1-mailinglist.xlsx.*
  Score_No: The probability the customer WILL NOT respond to catalog and not make a purchase.
  Score_Yes: The probability the customer WILL respond to catalog and make a purchase.

  Now, to predict the revenue generated from each customer, we need to predict the worth of products that will be purchased by them, and later multiplying revenue by the average gross margin and subtracting cost of printing and distributing of catalog we can predict our profit. And on the basis of which we can decide whether or not manager should send the catalog to customers.

Step 2: Analysis, Modelling and Validation

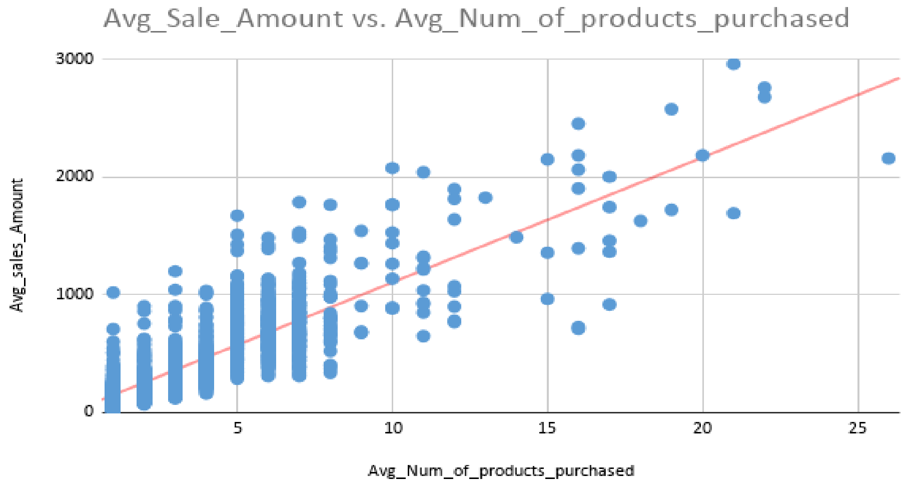
Selection of predictors variables in model from continuous variables:

- In order to select the **predictor variable**, we need to find the coefficient of correlation between each predictor variable and target variable and plot both on scatter plot to observe the relationship between them.
  Since, we need to predict the average sales of product, we'll consider Avg_Sale_Amount as our Target variable which is numeric and we are not suppose to calculate the amount over a time frame so it is Continuous variable.

A problem, where the predicted outcome is Numeric and Continuous, we will use Linear Regression technique to predict the value.

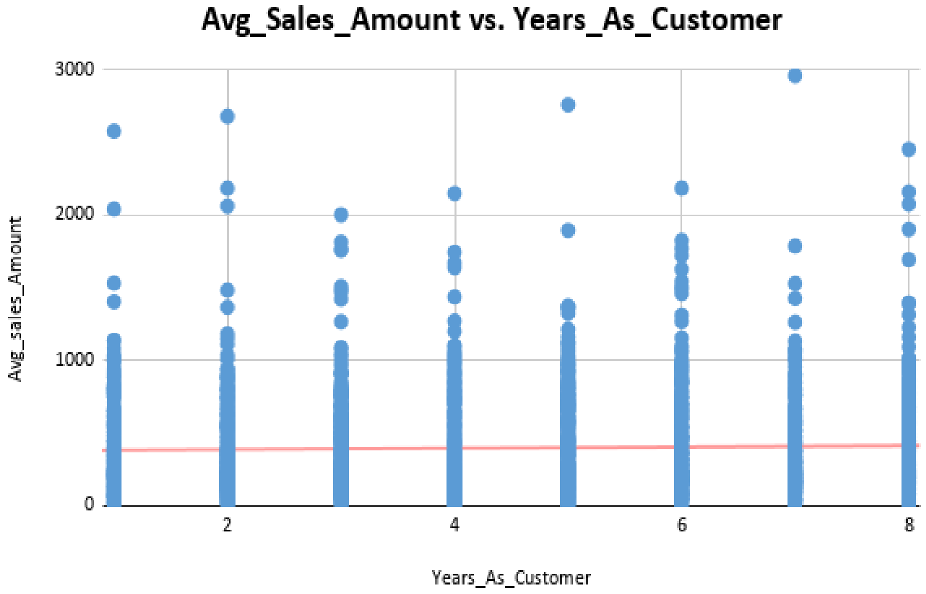
Looking for continuous predictor variables first later we'll check for the categorical value using Alteryx.

- **Scatter plot – Avg_Sale_Amount on y-axis, and Avg_Num_of_products_purchased on x-axis.**



Coefficient of correlation= 0.855754217

- **Scatter plot- Avg_Sale_Amount on y-axis and years_As_customer on x-axis.**



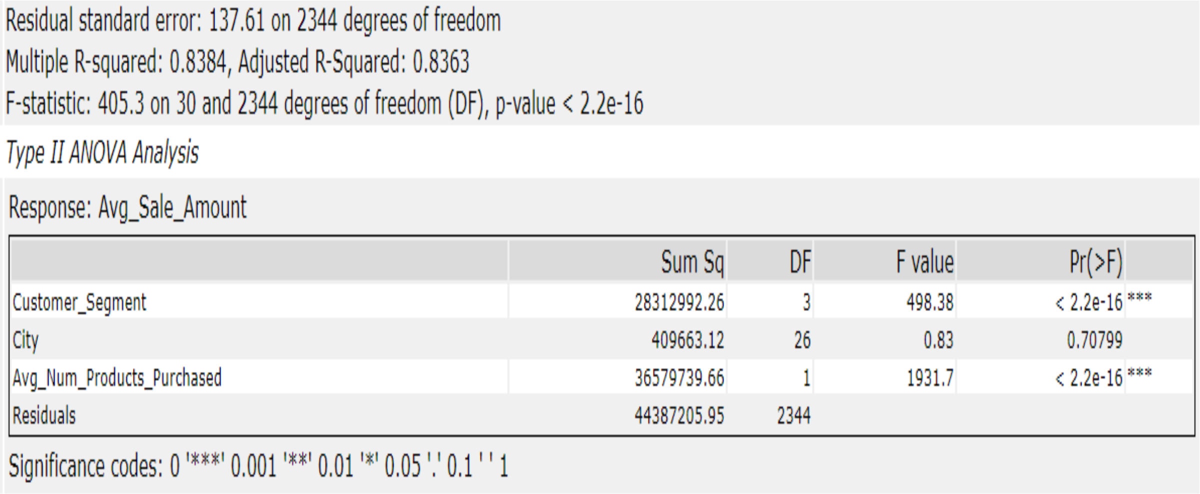
Coefficient of Correlation= 0.029781864

We will take Avg_Num_of_products_purchased out of all continuous variable as a predictor variable because of the strong correlation with the target variable.

Selection of predictors variables in model from categorical variables:

For this , I will use Alteryx to generate dummy variable from the categorical values, and input these values in Linear Regression Model to find their dependency on Target value and How well they linked to Target variable by understanding their r-squared and p-value.

Below is the data generated using Alteryx:



```
Residual standard error: 137.61 on 2344 degrees of freedom
Multiple R-squared: 0.8384, Adjusted R-Squared: 0.8363
F-statistic: 405.3 on 30 and 2344 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	28312992.26	3	498.38	< 2.2e-16 ***
City	409663.12	26	0.83	0.70799
Avg_Num_Products_Purchased	36579739.66	1	1931.7	< 2.2e-16 ***
Residuals	44387205.95	2344		

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

So, this shows that only Customer_Segment can be considered as predictor variable from all other categorical values because of its less p value.

So, unmark City from the range of predictor variables, and select only Customer_Segment and Avg_Num_Products_Purchased as are 2 predictor variables for our model.

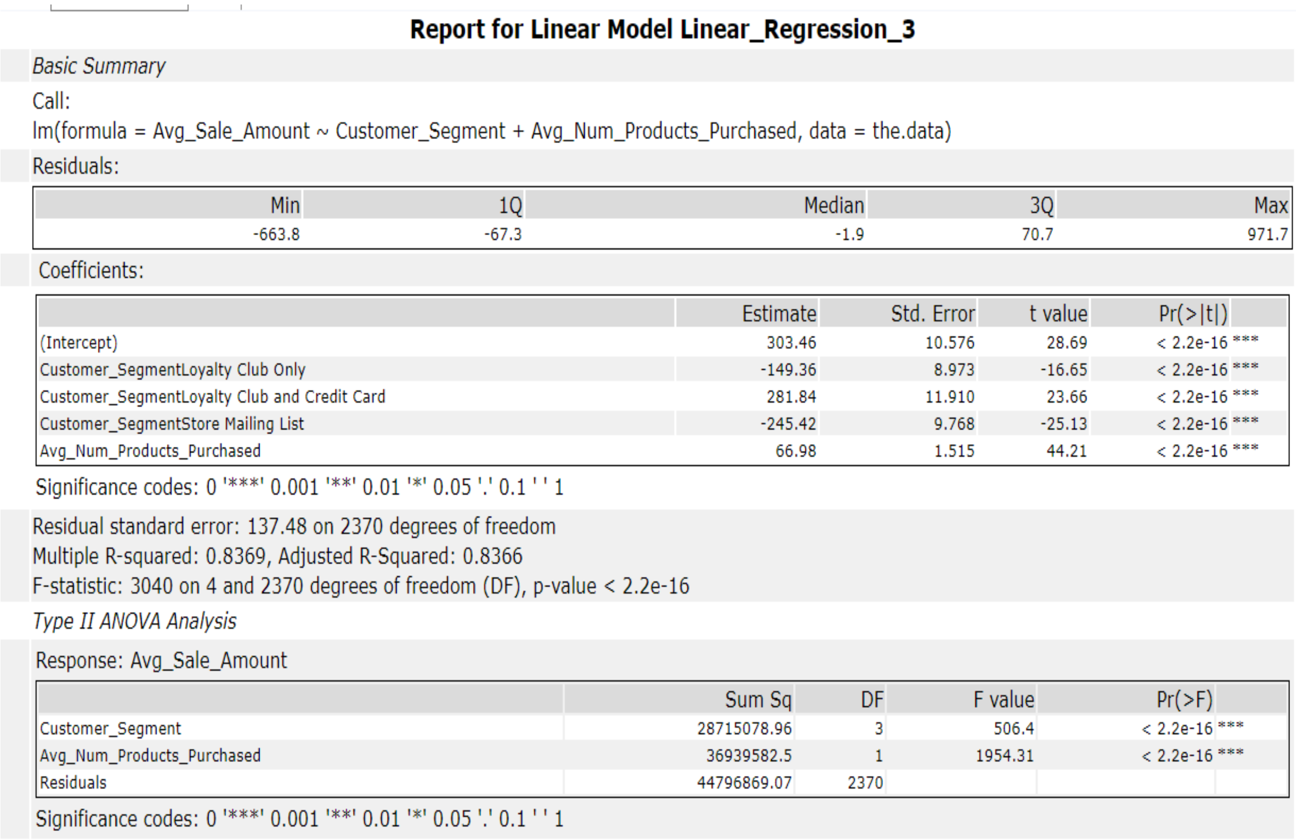
How good is the linear Model?

By finalizing our predictor and target variable, below is our report for Linear Regression Model.

Adjusted R-Squared value = 0.8366 and,

p-value < 2.2e-16

shows the significant and strong relationship between target variable and predictor variables. Also '***' to the right of p-value shows that these are the most significant variables to run across the linear regression 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

The best linear regression equation is:

Y = 303.46 – 149.36*X1 + 281.84*X2 – 245.42*X3 + 66.98*X4

Where,

X1= Customer_SegmentLoyality Club Only

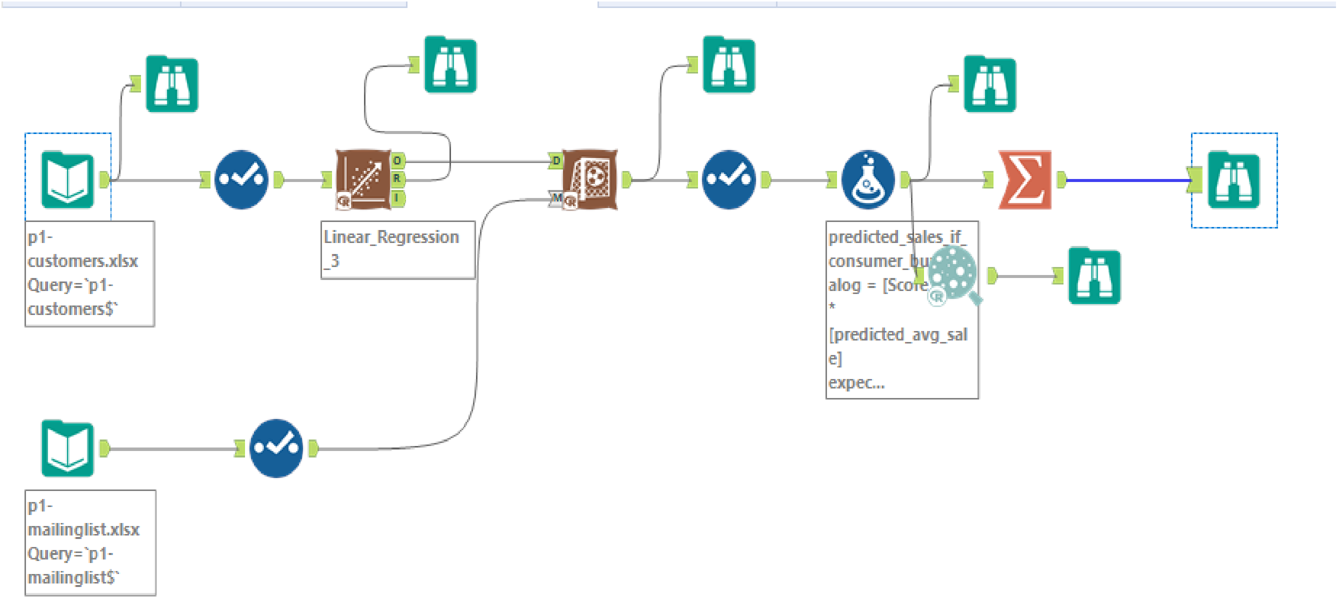
X2= Customer_SegmentLoyality Club and Credit Card

X3= Customer_SegmentStore Mailing list

X4= Avg_Num_of_Products_Purchased

Step 3: Presentation And Visualization

Below is the flow of entire model.


p1-customers.xlsx Query=`p1-customers$`
Linear_Regression_3
O
R
I
D
M
predicted_sales_if_consumer_bu... alog = [Score... * [predicted_avg_sale] expec...
p1-mailinglist.xlsx Query=`p1-mailinglist$`

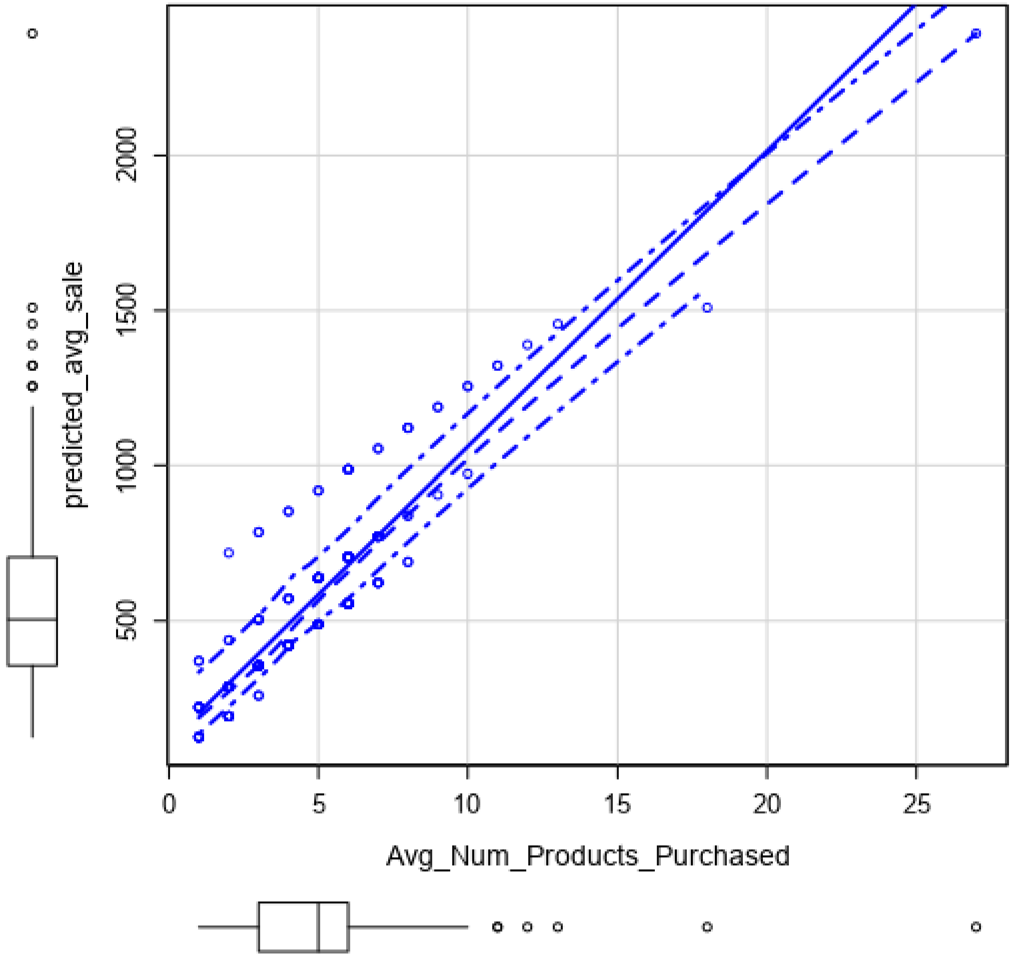
Quick Breif of the Workflow:

1. First, drop the input value on the canvas, and connect it to the p1-customers.xlsx file. Adding select tool to it, to check their datatypes and whether renaming of column is required or not. I have also added browser to it, to keep the track of data at different stages.
2. Adding linear regression tool to the output of 'select tool', and add the target and predictor variables. This will give us the linear regression equation, r-squared value and p-value.
3. Now, since we want to predict the Avg_sales_amount of testing dataset i.e p1-mailinglist.xlsx, we'll add this to another input tool. To calculate the values using equation which we predicted from our training dataset, we'll add score tool and connect its 'O' terminal with the 'O' terminal of Linear Regression tool and connect 'M' terminal of score with another input tool's terminal. This will give us the predicted avg_sales_Amount.
4. Since we need to find the profit, we'll do certain calculations. For this we add formula tool in front of select tool and add 2 columns.
   Predicted_sales_if_consumer_buy_catalog=['Score_Yes']*['predicted_Avg_sales']
   Reason we multiply with ['Score_yes'] because we want to calculate the expected revenue from 250 customers in order to find expected profit. This means we need to multiply the probability that a person will buy our catalog as well.

   Now, the average gross margin on all products sold through catalog is 50%. It means, manager is earning 50% of the predicted sales if consumer buy catalog. But this is not the profit, as there is some printing cost and distributing cost of catlog which we have to subtract from it, which will then give us the total profit from each customer.

Expected_profit= [‘Predicted_sales_if_consumer_buy_catalog’] * 0.5 -6.50

5. Adding summation tool in front of formula tool to find the sum of expected profit from each customer.
6. Finally, our total calculated profit is : **21987.435**



RESULT:

So, after a thorough analysis, I will definitely recommend the manager to send the catalog to these 250 customers as our net profit predicted is $ 21987.435 which is greater than $ 10,000.