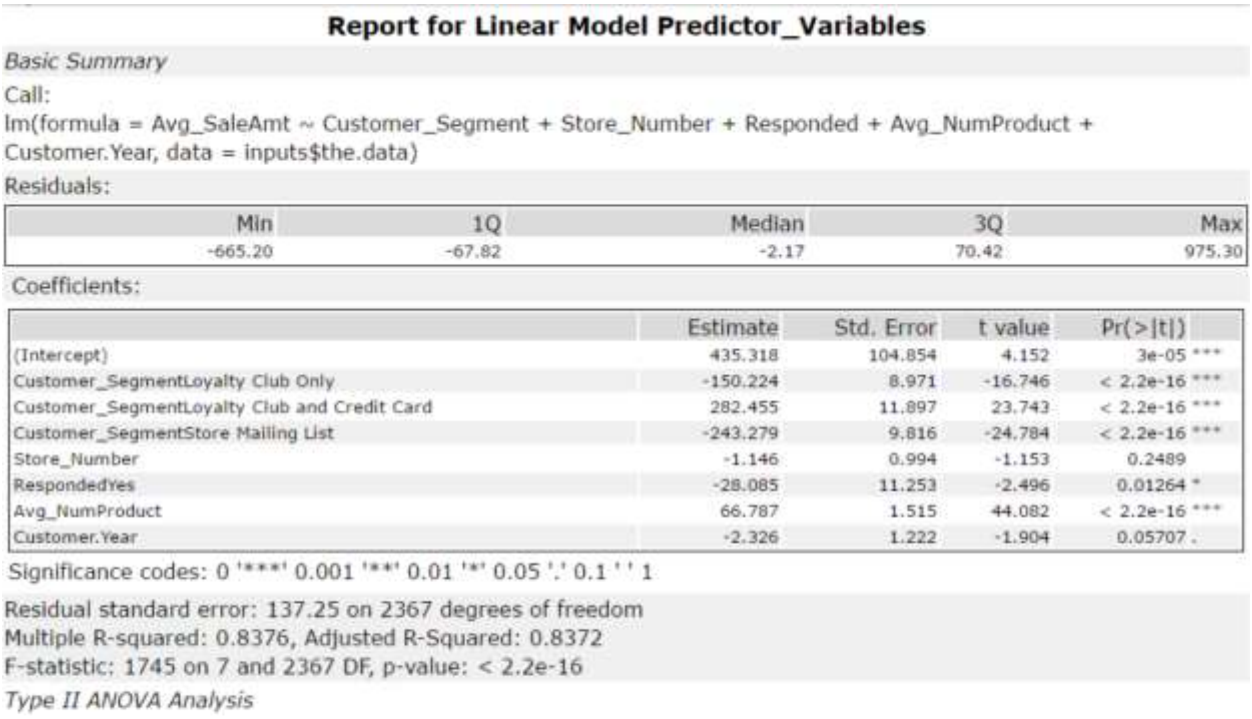
Predicting Catalog Demand

Business and Data Understanding
1. What decisions needs to be made?
The decision is whether to send the catalog to 250 new customers based on expected profit calculated.

2. What data is needed to inform those decisions?
Data needed to predict sales and calculate expected profit are *Customer Segment*, *Average Number of Product Purchased*, _Score*Yes*, *Margin* and *Cost of Catalog*.

Analysis, Modeling, and Validation
1. How and why did you select the predictor variables (see supplementary text) 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.
A linear regression study is performed on all variables against Average Sale Amount. As shown below, only Average Number of Product and Customer Segment have a p-value of less 0.05 which implies statistical significance. Scatterplots of Average Number of Product and Customer Segment versus Average Sale Amount are also plotted to study the linearity.



Report for Linear Model Predictor_Variables

Basic Summary

Call:

```
lm(formula = Avg_SaleAmt ~ Customer_Segment + Store_Number + Responded + Avg_NumProduct +
Customer.Year, data = inputs$the.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-665.20	-67.82	-2.17	70.42	975.30

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	435.318	104.854	4.152	3e-05 ***
Customer_SegmentLoyalty Club Only	-150.224	8.971	-16.746	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.455	11.897	23.743	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-243.279	9.816	-24.784	< 2.2e-16 ***
Store_Number	-1.146	0.994	-1.153	0.2489
RespondedYes	-28.085	11.253	-2.496	0.01264 *
Avg_NumProduct	66.787	1.515	44.082	< 2.2e-16 ***
Customer.Year	-2.326	1.222	-1.904	0.05707 .

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

Residual standard error: 137.25 on 2367 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8372
F-statistic: 1745 on 7 and 2367 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Figure 1: Report for Linear Model Predictor_Variables

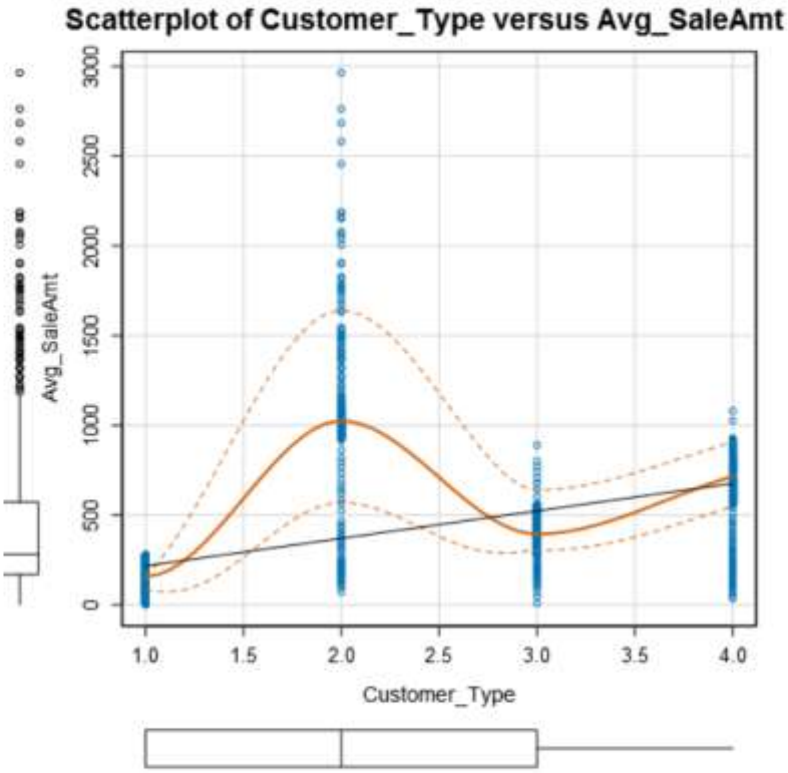


Figure 2: Scatterplots of Avg Number of Products Purchased

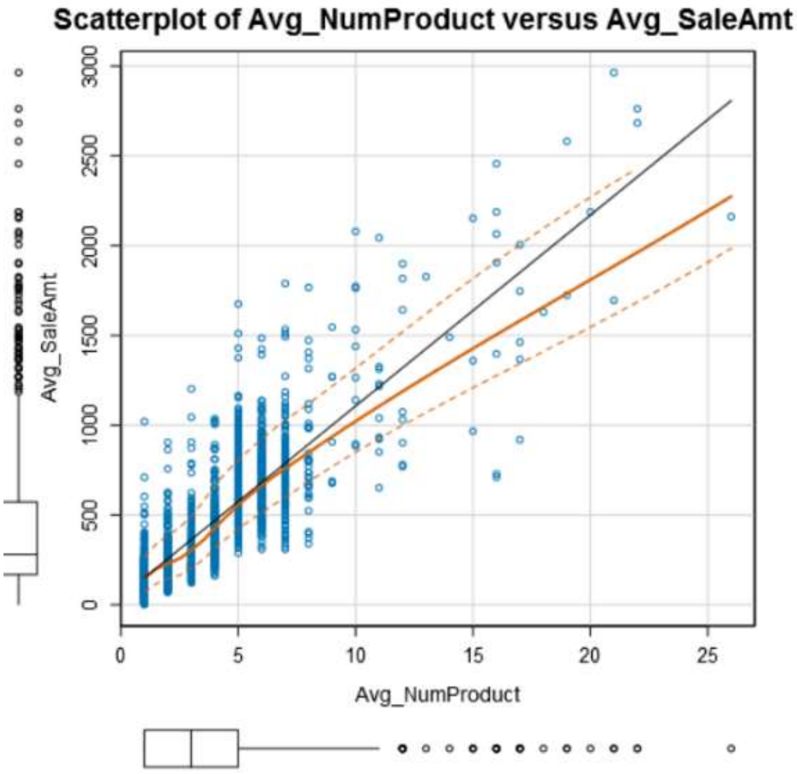
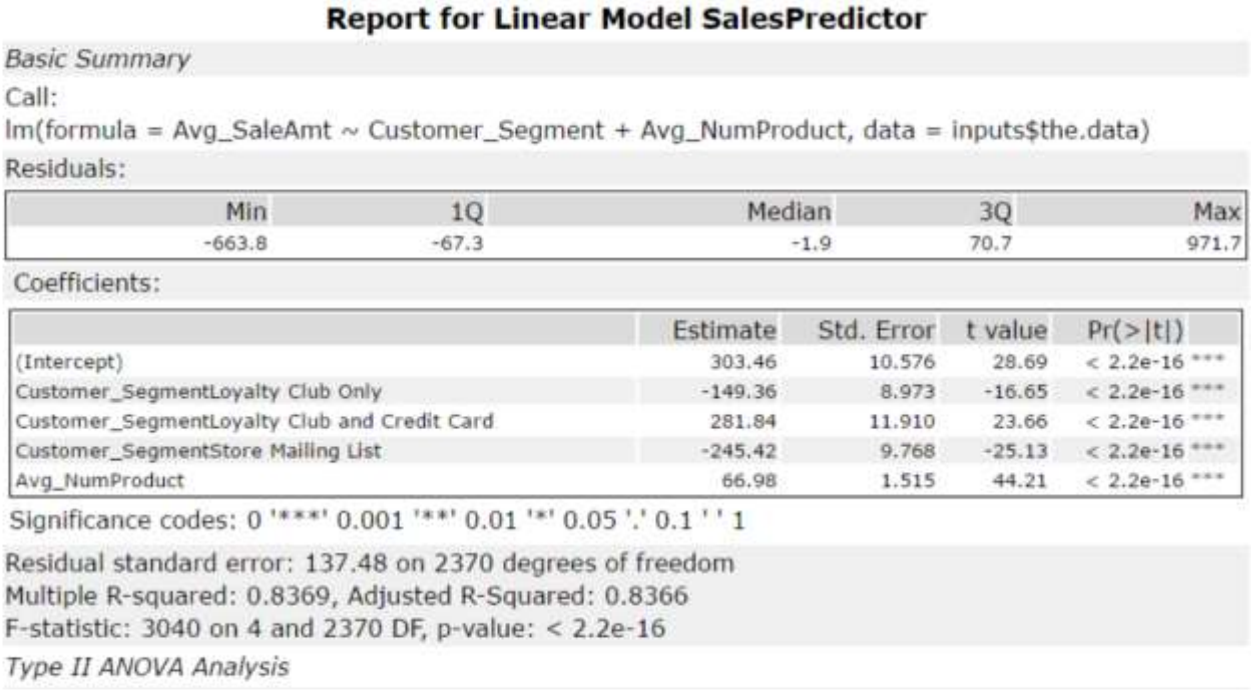


Figure 3: Scatterplots of Customer Segment vs Avg Sale Amount

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.

The Alteryx linear regression function is used to determine the strength of the linear and the statistical result shows an adjusted R-squared value of 0.8366 which is a high value. Customer Segment and Average Number of Products also have a p-value lower than 0.05, implying their statistical significance. Thus, the model is considered a good one.



Report for Linear Model SalesPredictor

Basic Summary

Call:
lm(formula = Avg_SaleAmt ~ Customer_Segment + Avg_NumProduct, data = inputs$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_NumProduct	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

Type II ANOVA Analysis

Figure 4: Report for Statistical Result

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 – 149.36 x (If Type: Loyalty Club Only) + 281.84 x (If Type: Loyalty Club and Credit Card) – 245.42 x (If Type: Store Mailing List) + 0 x (If Type: Credit Card Only) + 66.98 x (Avg_Num_Products_Purchased)

Presentation/Visualization

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

The company should send the catalog to these 250 new customers.

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

Using linear regression model, the expected revenue from each customer is determined by multiplying expected sale amount with Score_Yes value.

With a gross margin of 50%, 50% is deducted from the sum of expected revenue before the cost of catalog ($6.50) is subtracted to obtain net profit.

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

Expected Profit = *(Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)*
= (47,225.87 x 0.5) – (6.50 x 250)
= 23,612.44 – 1,625
= $21,987.44

Variables Distribution

Variables such as *Address*, *Name*, *State*, *Customer ID*, *Store Number* and *ZIP* are not important predictor variables as they are either unique to each value or irrelevant in predicting the sale using common sense.

City, *Responded to Last Catalog* and *# Years as Customer* might seem to be a good predictor as they are not unique ID but linear regression model showed that they are statistically insignificant.

More data from category of items purchased, items turnover duration will be helpful to understand customer's buying behavior where we can exploit it to segment our customers and customize the catalog.

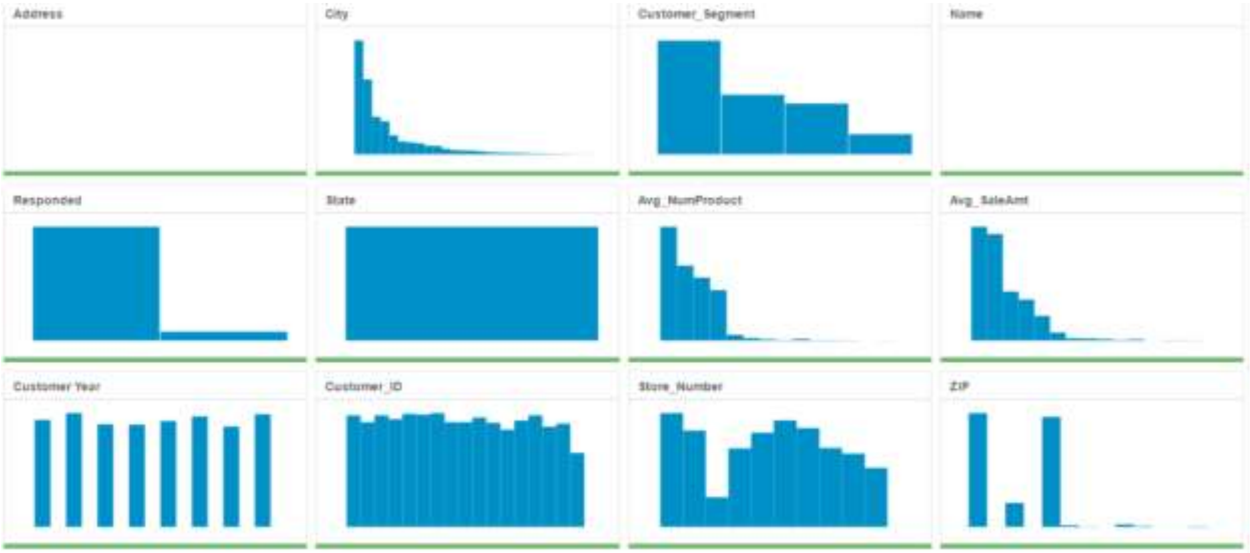
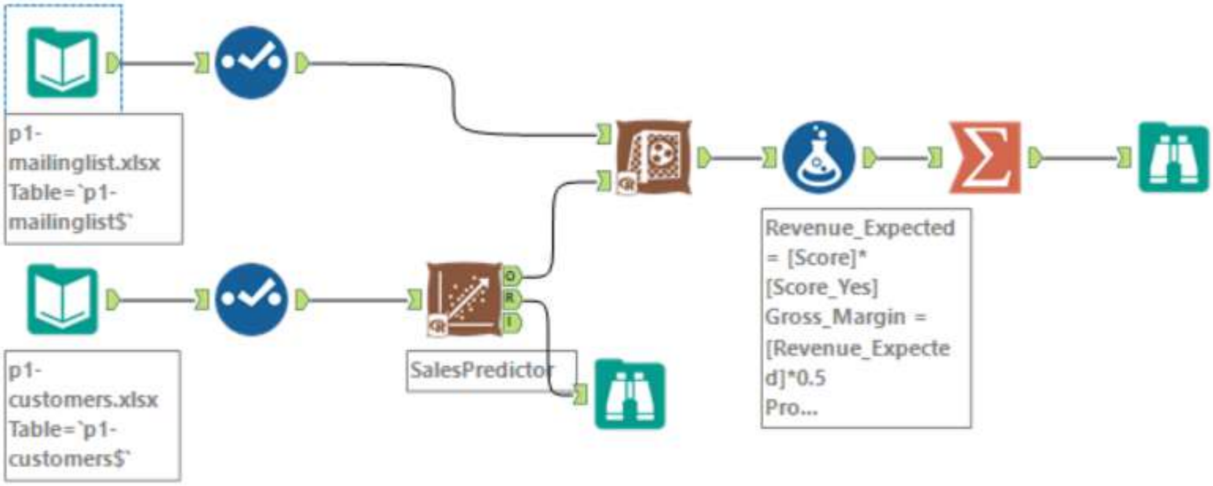


Figure 5: Distribution of each variable in the Customer List dataset

Alteryx Workflow


p1-mailinglist.xlsx
Table=`p1-mailinglist$`
p1-customers.xlsx
Table=`p1-customers$`
O
R
I
SalesPredictor
Revenue_Expected = [Score]*[Score_Yes]
Gross_Margin = [Revenue_Expected]*0.5
Pro...