MAKE A COPY

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The objective of the project is to predict the profit we can expect by sending out a catalog to the new 250 customers.
The decision to make is whether we should send out a catalog to the new customers? The management wants to send out these catalogs to new customers only if the expected profit contribution exceeds $10,000.
We need data about how much it costs to produce and distribute the catalog. We also need historical data about previous sales. We need details about how many products on an average the customers bought what was their revenue.

Step 2: Analysis, Modeling, and Validation

Target Variable: The average sale amount is going to be our target variable, as we are trying to predict the profit.
Predictor Variables:
To begin, we are first going to take a look at all the variables available and their data type.

Variable	DataType
Name	Categorical
Customer_Segment	Categorical(Nominal)
Customer_ID	Numerical(Discreet)
Address	Categorical
City	Categorical
State	Categorical
ZIP	Numerical(Discreet)
Avg_Sale_Amount	Numerical
Store_Number	Categorical (nominal)
Responded_to_Last_Catalog	Bool
Avg_Num_Products_Purchased	Numerical
#_Years_as_Customer	Numerical

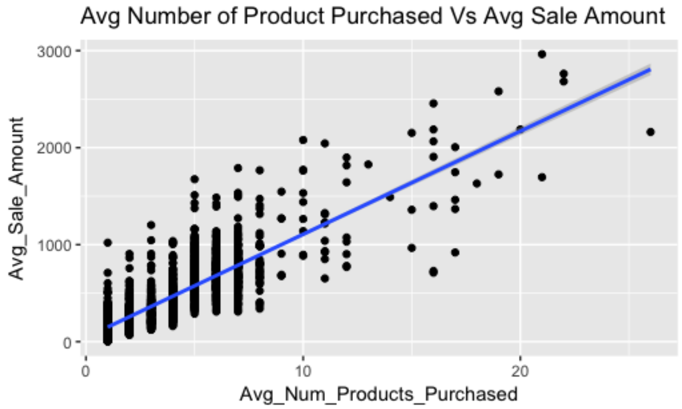
Straight away we are going to drop a few variables from contention, we are dropping these variables based on sound understanding of business as these variables will not affect the sales amount.

1. *Name*
2. *Customer Id*
3. *Address*
4. *City*
5. *State*

Let us take a look at numerical variables first.

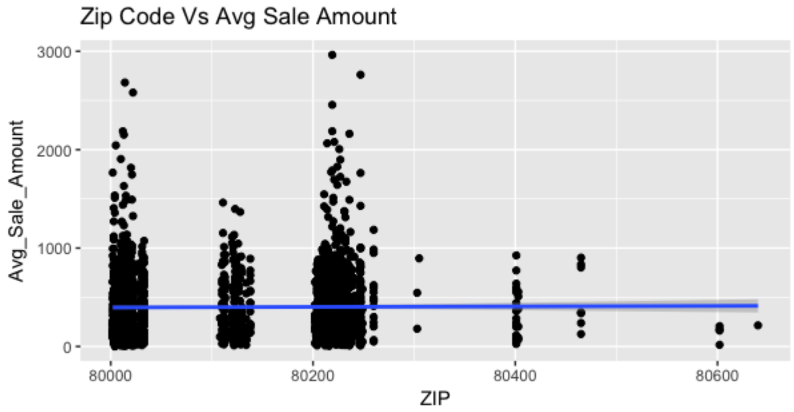
1. *Avg_Num_Products_Purchased.*

The correlation coefficient between the average number of products purchased and the average sale amount is 0.86. Also, you can see the same strong positive linear relationship in the graph below.



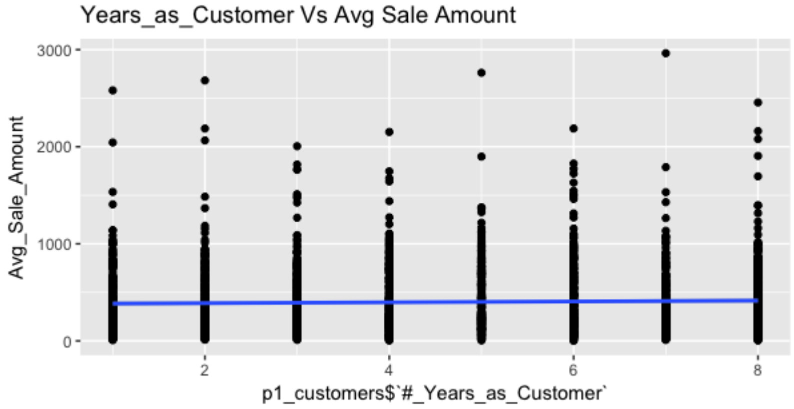
2. *Zip Code*

We could have dropped this variable straight away, as the zip code will not affect the sales amount. But we will check the relationship between zip code and average sale amount by drawing up a scatter plot. The correlation coefficient is 0.0079, which suggests there is no correlation between the two variables and this is further evident in the scatter plot. Thus, we are not going to be including zip code in the model.



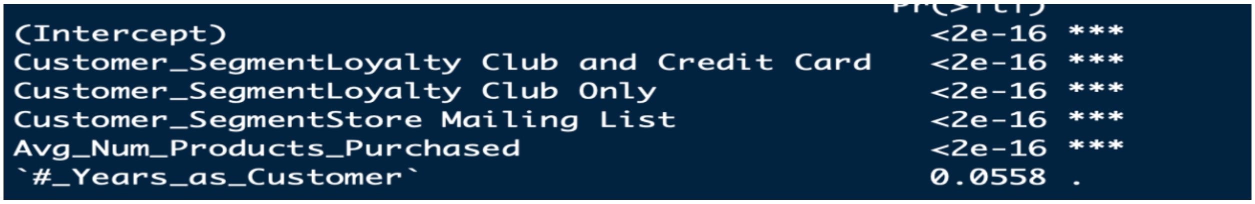
3. #_Years_as_Customer

The correlation coefficient is 0.03, which suggests there is no correlation between the two variables and this is further evident in the scatter plot. But we are going to include the variable in the first iteration of the model and then check if it is statistically significant.



Modelling

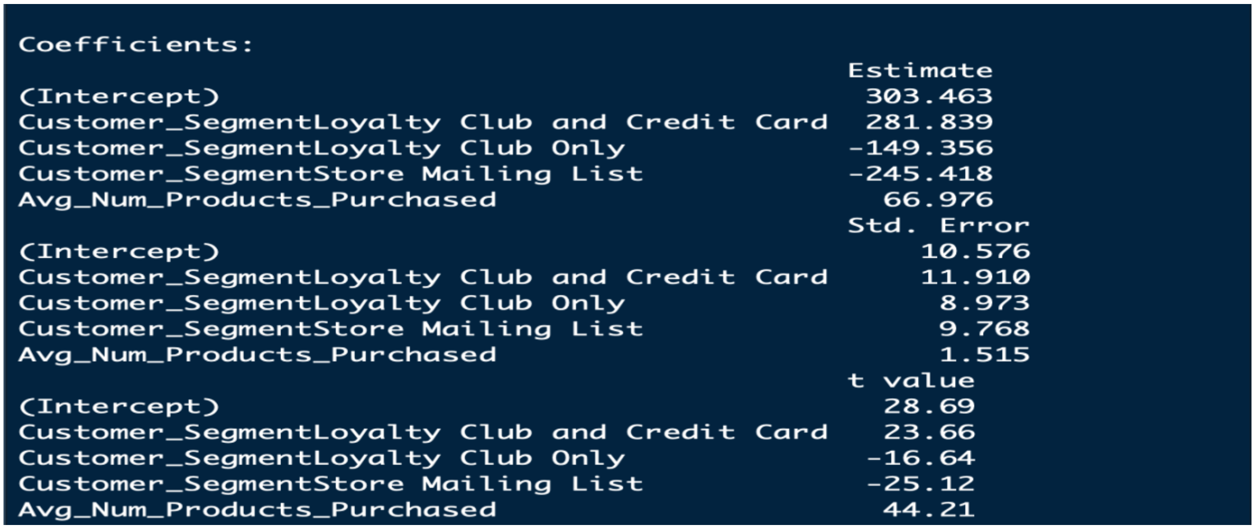
Let look at the first iteration of the model with number of years part of the model. If you look at the p values, number of years as customer is not statistically significant as confidence level of 95%. So we are going to drop the variable from the model. And also, not consider this model.



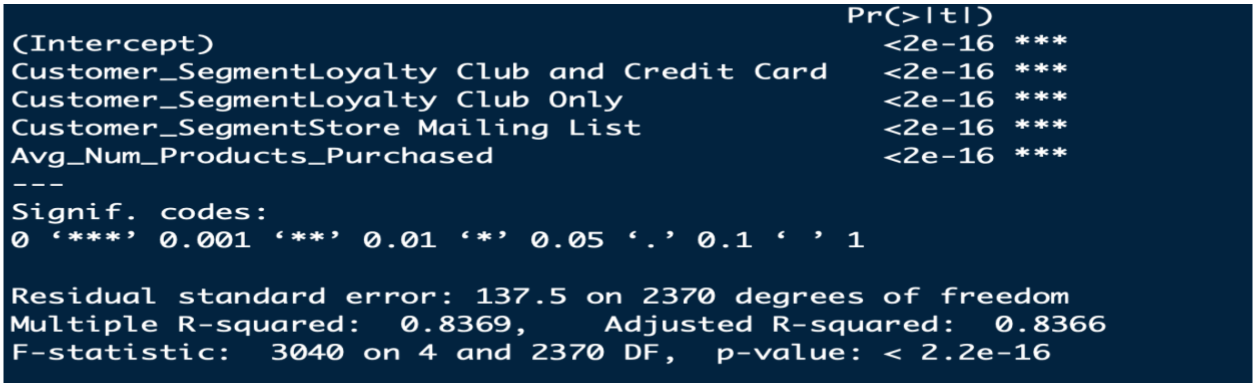
```
                                                  Pr(>|t|)
(Intercept)                                        <2e-16 ***
Customer_SegmentLoyalty Club and Credit Card       <2e-16 ***
Customer_SegmentLoyalty Club Only                  <2e-16 ***
Customer_SegmentStore Mailing List                 <2e-16 ***
Avg_Num_Products_Purchased                         <2e-16 ***
`#_Years_as_Customer`                              0.0558 .
```

Now we run the model customer segmentation, and average number of products purchased as predictor variables. And after looking at the summary statistic all the predictor variables have p-value less than 0.05 thus statistically significant and the R_2 and Adjusted R_2 value is 0.8369 and 0.8366 respectively, which is fairly high and thus explain more than 83% of the variability.

Equation for the model: **Avg_Sale_Amount = 303.46 + (281.84 * If Customer_Segmentation is Club and Credit Card) – (149.36 * If Customer_Segmentation is Loyalty Club Only) – (245.42 * If Customer_Segmentation is Store Mailing List) + (66.98 * Avg_Num_Products_Purchased) + (0* If Customer_Segmentation is Credit Card Only)**

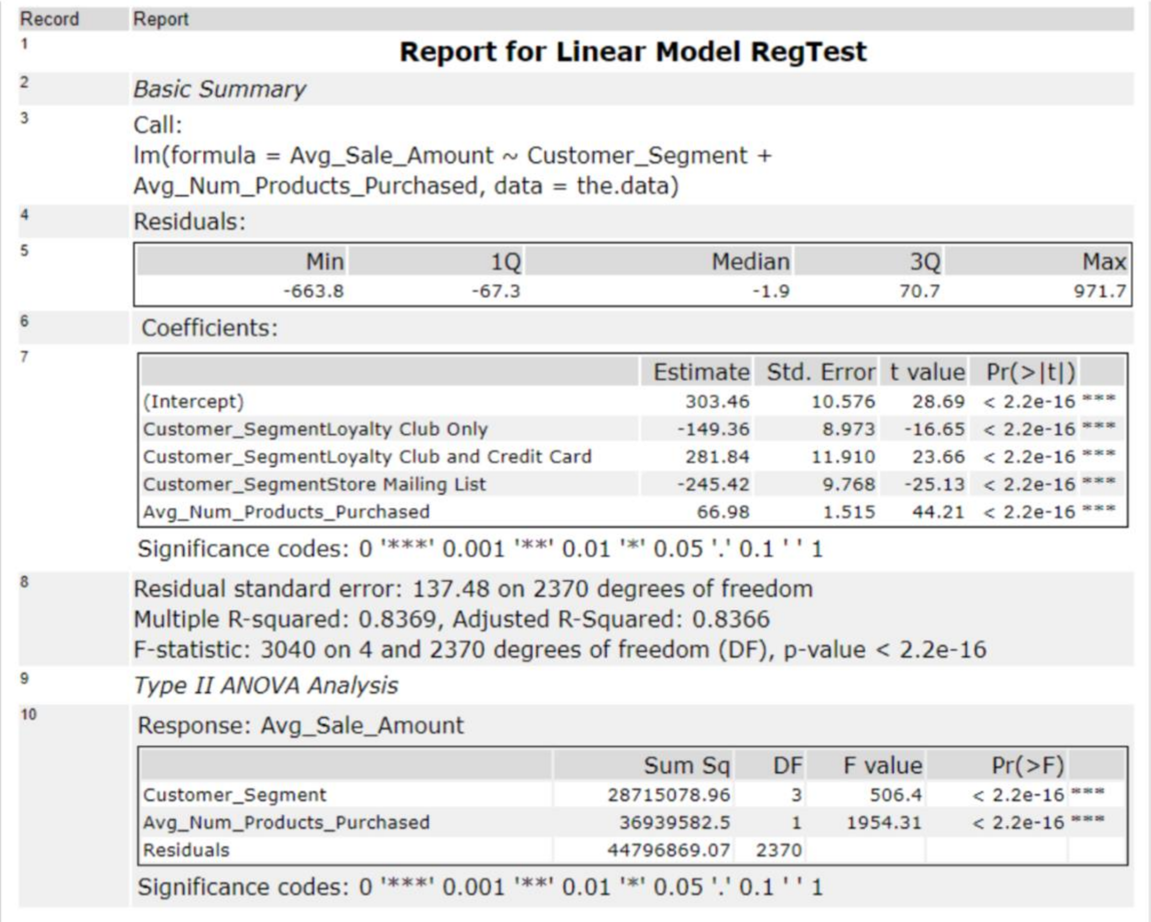


```
Coefficients:
                                             Estimate
(Intercept)                                   303.463
Customer_SegmentLoyalty Club and Credit Card  281.839
Customer_SegmentLoyalty Club Only            -149.356
Customer_SegmentStore Mailing List           -245.418
Avg_Num_Products_Purchased                     66.976
                                             Std. Error
(Intercept)                                      10.576
Customer_SegmentLoyalty Club and Credit Card     11.910
Customer_SegmentLoyalty Club Only                 8.973
Customer_SegmentStore Mailing List                9.768
Avg_Num_Products_Purchased                        1.515
                                             t value
(Intercept)                                    28.69
Customer_SegmentLoyalty Club and Credit Card   23.66
Customer_SegmentLoyalty Club Only             -16.64
Customer_SegmentStore Mailing List            -25.12
Avg_Num_Products_Purchased                     44.21
```



```
                                             Pr(>|t|)
(Intercept)                                    <2e-16 ***
Customer_SegmentLoyalty Club and Credit Card   <2e-16 ***
Customer_SegmentLoyalty Club Only              <2e-16 ***
Customer_SegmentStore Mailing List             <2e-16 ***
Avg_Num_Products_Purchased                     <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137.5 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
```



Record Report

1

Report for Linear Model RegTest

2 *Basic Summary*

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

4 Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 Coefficients:

7

	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
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

9 *Type II ANOVA Analysis*

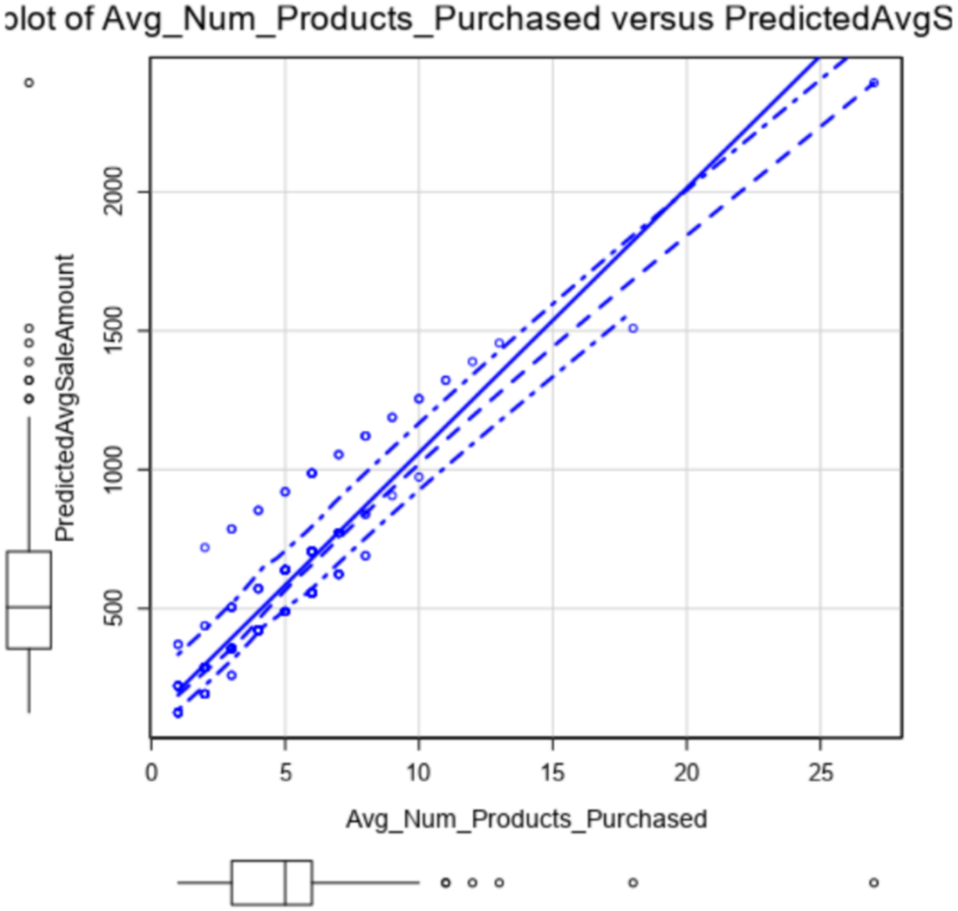
10 Response: Avg_Sale_Amount

	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

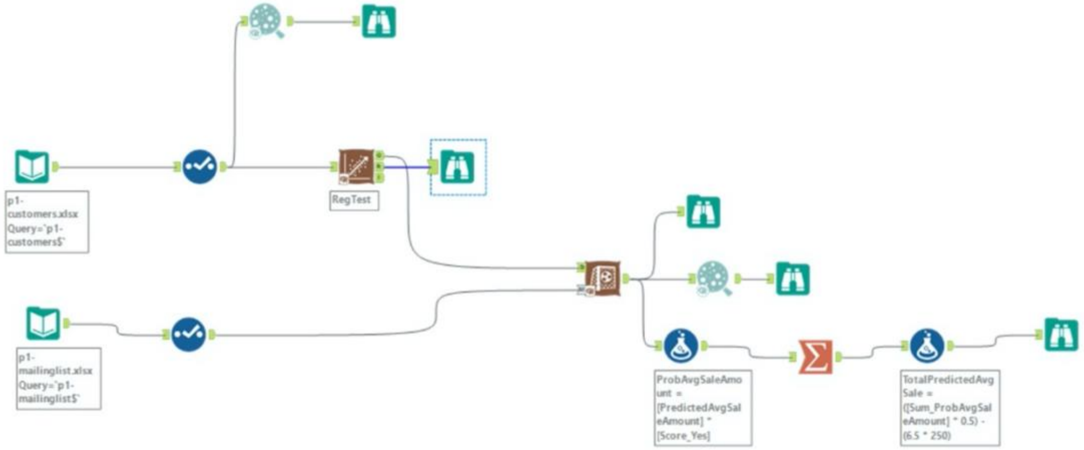
Step 3: Presentation/Visualization

We should send the catalog to the new customers, as we have estimated to earn a profit of $21,987.43. Which is substantially higher than $10,000 targeted by the management.



Steps Followed to come up with the recommendation.
We developed a linear regression model with an adjusted R_2 value of 0.83.
We predicted the average sale amount for each new customer by using the model.
We then multiplied the predicted average sale amount with the probability of a yes score for each customer.
Then calculated to total estimated average sale amount by summing the above.
*Then multiplied it 50% which is the gross margin and then subtracted the cost of manufacturing and distributing the catalog. (6.5 *250).*

Alteryx flow.


p1-
customers.xlsx
Query=`p1-
customers$`
RegTest
p1-
mailinglist.xlsx
Query=`p1-
mailinglist$`
ProbAvgSaleAmount =
[PredictedAvgSaleAmount] *
[Score_Yes]
TotalPredictedAvgSale =
([Sum_ProbAvgSaleAmount] * 0.5) -
(6.5 * 250)