Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Answer:

Using the existing customer data, Business Analyst needs to determine/predict the profit that company can expect from sending a catalog to 250 new customers. The catalog will be sent only if profit contribution exceeds $10,000.00.

2. What data is needed to inform those decisions?

Answer:

In order to predict potential profit, we will need –

- Current information about company’s customers (p1-customers.xlsx) to train linear regression model i.e. training data set
- Data about new customers, customer mailing list (p1-mailinglist.xlsx) to apply the model i.e. test dataset
- Base cost of printing and distributing catalog ($6.5/catalog)
- Average profit margin (50%)

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

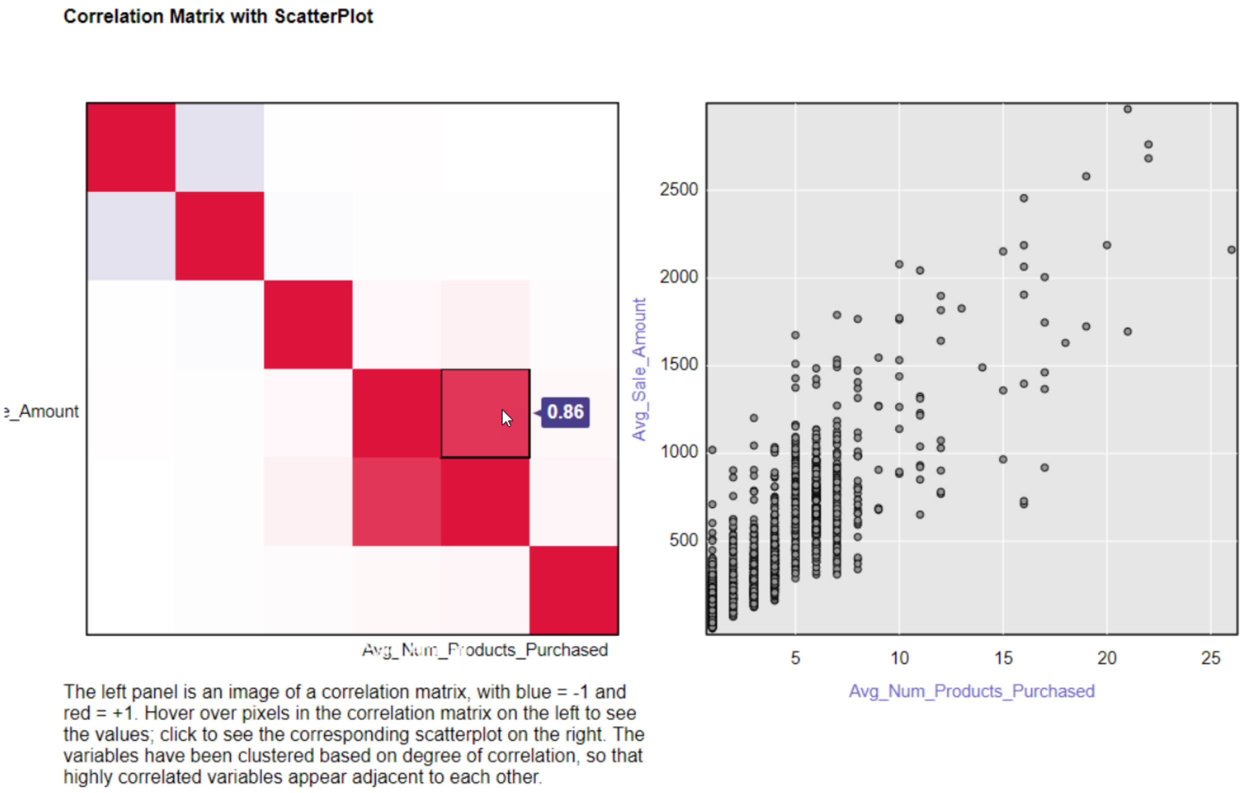
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer back to the “Multiple Linear Regression with Excel” lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

<u>Answer:</u>

Quantitative Variables:

Looking at the available fields in training dataset (specifically numeric fields), we can conclude that Store_Number, ZIP are attributes associated with store location and cannot be used as predictor variable for determining profit. Additionally, numeric field Customer_ID is a unique identifier for each customer and will not be used as a predictor variable. Therefore, my predictor variables of interest are Avg_Sale_Amount, Avg_Num_Products_Purchases and #_Years_as_Customer.

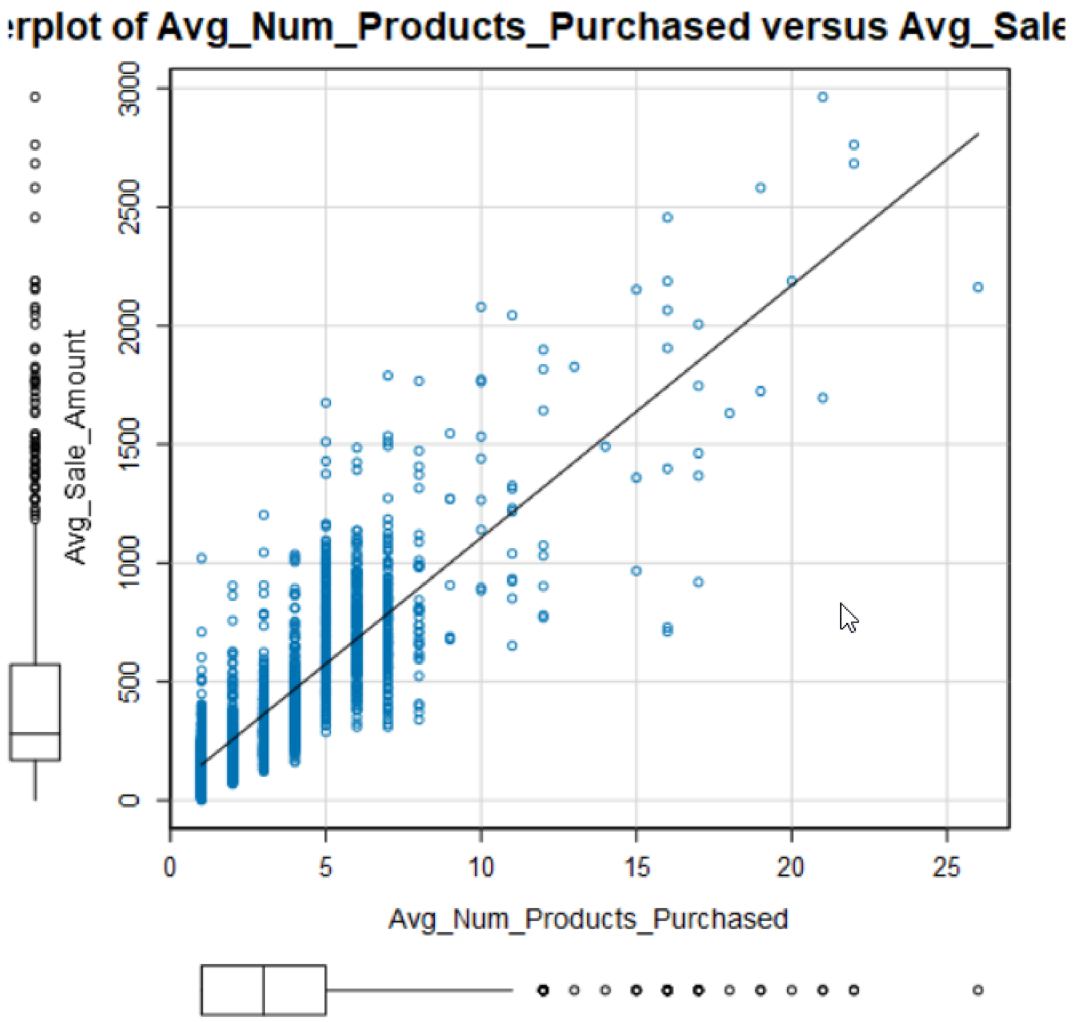
- Association Matrix – Following picture depicts association analysis shows correlation matrix for all the available variables within the training data set. As shown, we can conclude that Avg_Sale_Amount and Avg_Num_Products_Purchased has strong positive relationship.



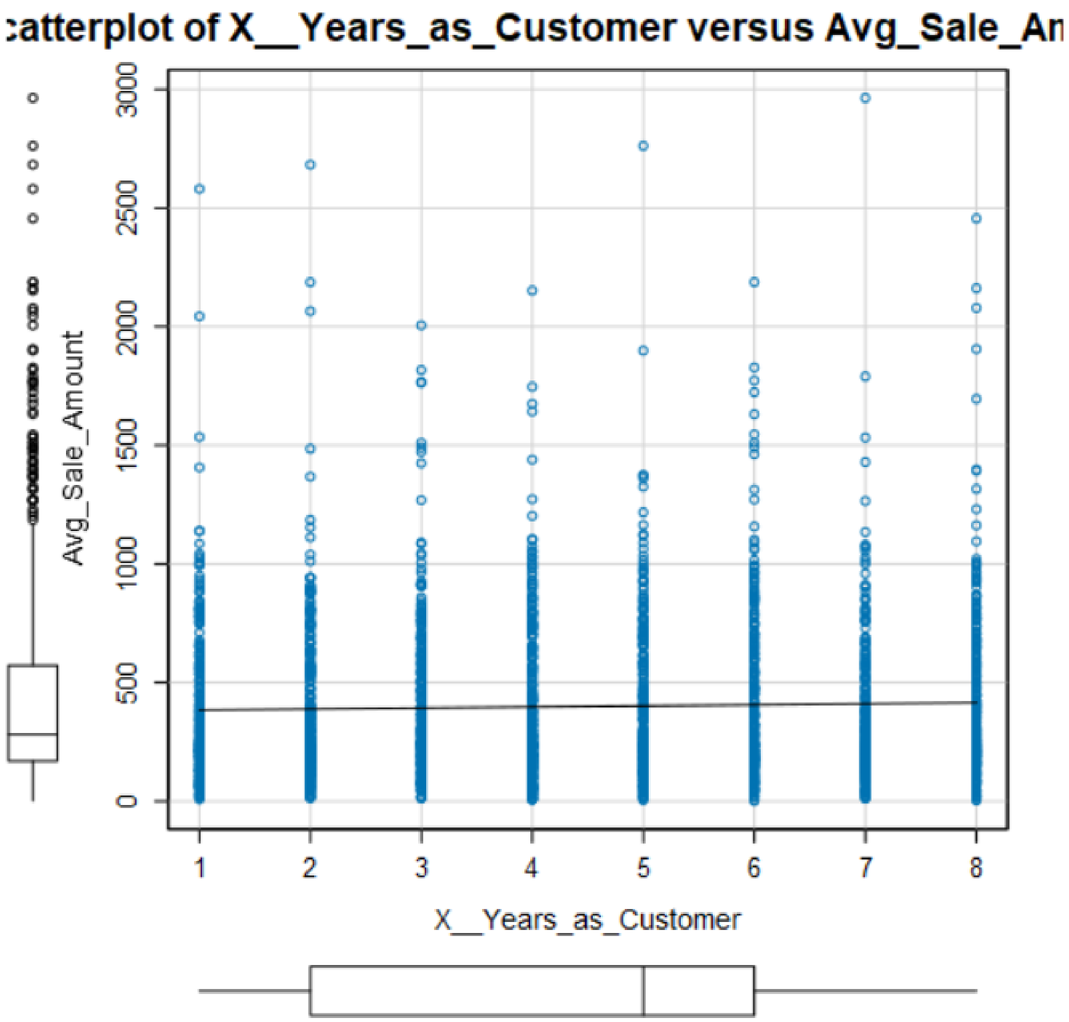
Correlation Matrix with ScatterPlot

The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

- Scatter plot –
  - Following visual depicts scatter plot between variables Avg_Num_Products_Purchased v/s Avg_Sale_Amount. The scatter plot indicates **strong positive correlation between these variables.**

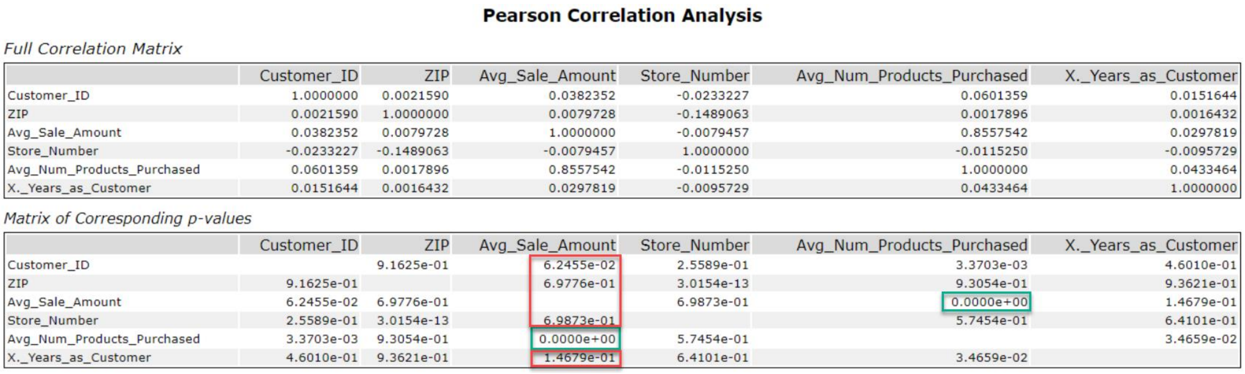


- Following visual depicts scatter plot between variables #_of_Years_as_Customer v/s Avg_Sale_Amount. The scatter plot does not indicate any clear relationship between the variables.



Pearson Correlation Analysis:

As shown in the matrix below, Customer_ID, ZIP, Store_Number and #_Years_as_Customer has (absolute P-value) higher than 0.5. Therefore, we can exclude them as predictor in our model.



Pearson Correlation Analysis

Full Correlation Matrix

	Customer_ID	ZIP	Avg_Sale_Amount	Store_Number	Avg_Num_Products_Purchased	X._Years_as_Customer
Customer_ID	1.0000000	0.0021590	0.0382352	-0.0233227	0.0601359	0.0151644
ZIP	0.0021590	1.0000000	0.0079728	-0.1489063	0.0017896	0.0016432
Avg_Sale_Amount	0.0382352	0.0079728	1.0000000	-0.0079457	0.8557542	0.0297819
Store_Number	-0.0233227	-0.1489063	-0.0079457	1.0000000	-0.0115250	-0.0095729
Avg_Num_Products_Purchased	0.0601359	0.0017896	0.8557542	-0.0115250	1.0000000	0.0433464
X._Years_as_Customer	0.0151644	0.0016432	0.0297819	-0.0095729	0.0433464	1.0000000

Matrix of Corresponding p-values

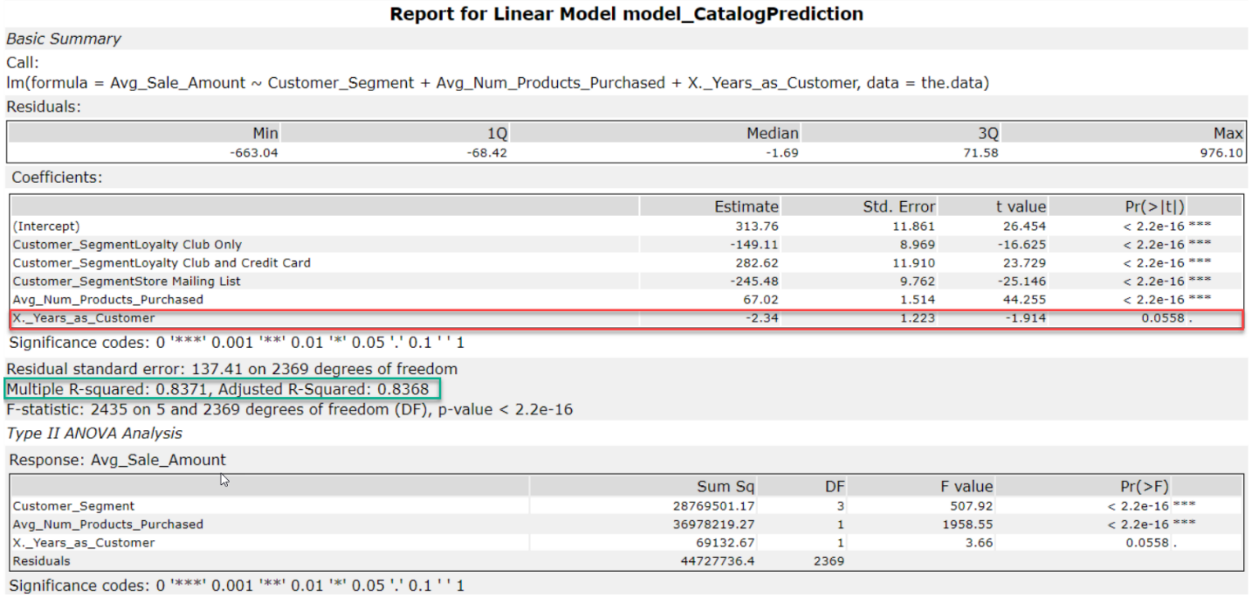
	Customer_ID	ZIP	Avg_Sale_Amount	Store_Number	Avg_Num_Products_Purchased	X._Years_as_Customer
Customer_ID		9.1625e-01	6.2455e-02	2.5589e-01	3.3703e-03	4.6010e-01
ZIP	9.1625e-01		6.9776e-01	3.0154e-13	9.3054e-01	9.3621e-01
Avg_Sale_Amount	6.2455e-02	6.9776e-01		6.9873e-01	0.0000e+00	1.4679e-01
Store_Number	2.5589e-01	3.0154e-13	6.9873e-01		5.7454e-01	6.4101e-01
Avg_Num_Products_Purchased	3.3703e-03	9.3054e-01	0.0000e+00	5.7454e-01		3.4659e-02
X._Years_as_Customer	4.6010e-01	9.3621e-01	1.4679e-01	6.4101e-01	3.4659e-02	

(*-red indicates predictor variables with higher p-value and green indicates lower p-value)

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.

Regression Analysis:

- First model with all predictor variable – as shown below Customer_Segment (categorical variable) and Avg_Num_Products_Purchased has higher significance in the model, where has #_Years_as_Customer is insignificant in the model.



Report for Linear Model model_CatalogPrediction

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.04	-68.42	-1.69	71.58	976.10

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	313.76	11.861	26.454	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.11	8.969	-16.625	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.62	11.910	23.729	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.48	9.762	-25.146	< 2.2e-16 ***
Avg_Num_Products_Purchased	67.02	1.514	44.255	< 2.2e-16 ***
X._Years_as_Customer	-2.34	1.223	-1.914	0.0558 .

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

Residual standard error: 137.41 on 2369 degrees of freedom

Multiple R-squared: 0.8371, Adjusted R-Squared: 0.8368

F-statistic: 2435 on 5 and 2369 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

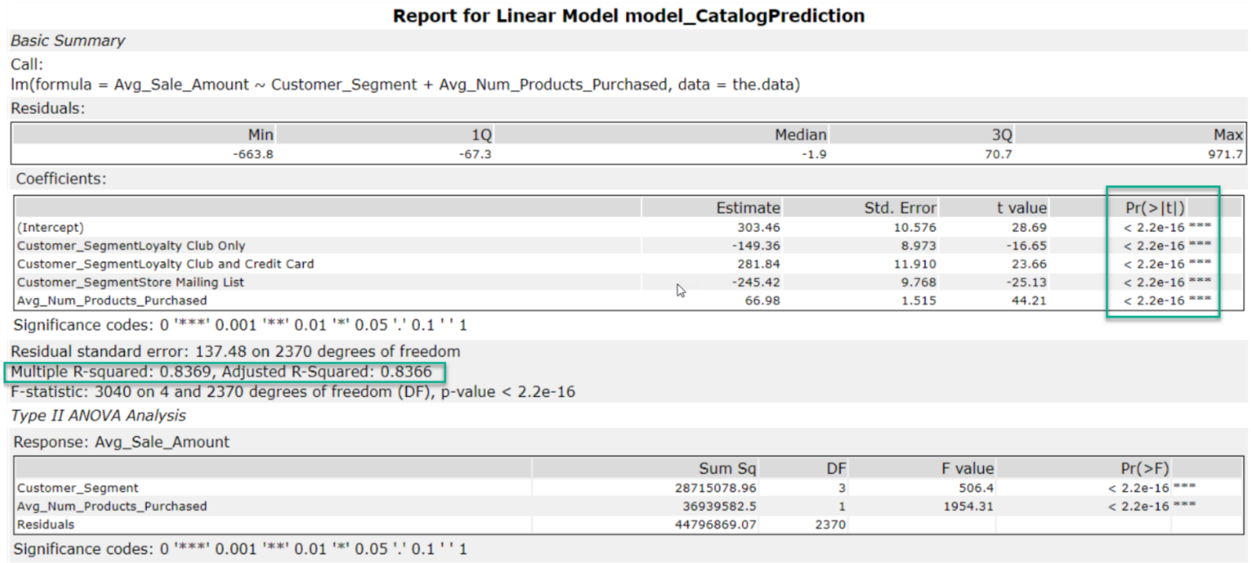
Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28769501.17	3	507.92	< 2.2e-16 ***
Avg_Num_Products_Purchased	36978219.27	1	1958.55	< 2.2e-16 ***
X._Years_as_Customer	69132.67	1	3.66	0.0558 .
Residuals	44727736.4	2369		

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

Given the lower insignificance, we will have to model again removing #_Years_as_Customer variable.

- Model with relevant and significant predictors – as shown below, this model has high adjusted R-squared value (0.83) as well as lower p-values for all the predictor variables. Therefore, this is model is more relevant and appropriate for prediction.



Report for Linear Model model_CatalogPrediction

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

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 Customer_Segment = Loyalty Club Only) + 281.84 x (if Customer_Segment = Loyalty Club and Credit Card) – 245.42 x (if Customer_Segment = Store Mailing List) + 0 x (if Customer_Segment = Credit Card Only + 66.98 x Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

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

Based on trained model and calculated potential profit, the company should send catalog to 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)

As per the requirement, you should only send this catalog to new customers if total profit margin (predicted) exceeds $10,000. (The cost of printing and distribution and average profit margin of 50% is considered in the calculation.) Calculated potential profit margin is ~ $21K hence the decision.

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

Expected profit is $21,987.44 – calculated as –

Per customer profit = 50% x (Score_Yes x Predicted_Avg_Sale) - 6.5
Sum total the per customer profit to get total profit

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.