Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

- Predicting the estimated profit if a catalogue was sent to new customers and then on the basis of profit, decide whether the catalogue should be sent or not?

2. What data is needed to inform those decisions?

- Data about the sales occurred last year. (Given)
- Probability that a new customer will buy a catalogue and purchase items? (Given)
- Information about current customers, shopping behaviour, location etc. (Given)
- Cost structure (Cost for catalogue is given)
- Since, we have the past data about sales, we can predict the sales for current year.

Step 2: Analysis, Modeling, and Validation

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

- In the given variables, only few variables seem to be a good fit to predict the average sales for the current year.

- **Name : (Not Significant)** - Since sales doesn't depend on a user's name.

- **Customer_Id : (Not Significant**) - Customer id is a unique id assigned to a customer. It doesn't change average sales.

- **Address : (Not Significant) –** It is too detailed, instead of it we can use other variables like city or zip. **No. of years as customers vs Avg_Sales (Not linear)**

- 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.

Record	Report
1	**Report for Linear Model Linear_Regression**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + ZIP + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased, data = the.data)
4	Residuals:

5

Min	1Q	Median	3Q	Max
-667.9	-67.8	-3.4	70.0	968.8

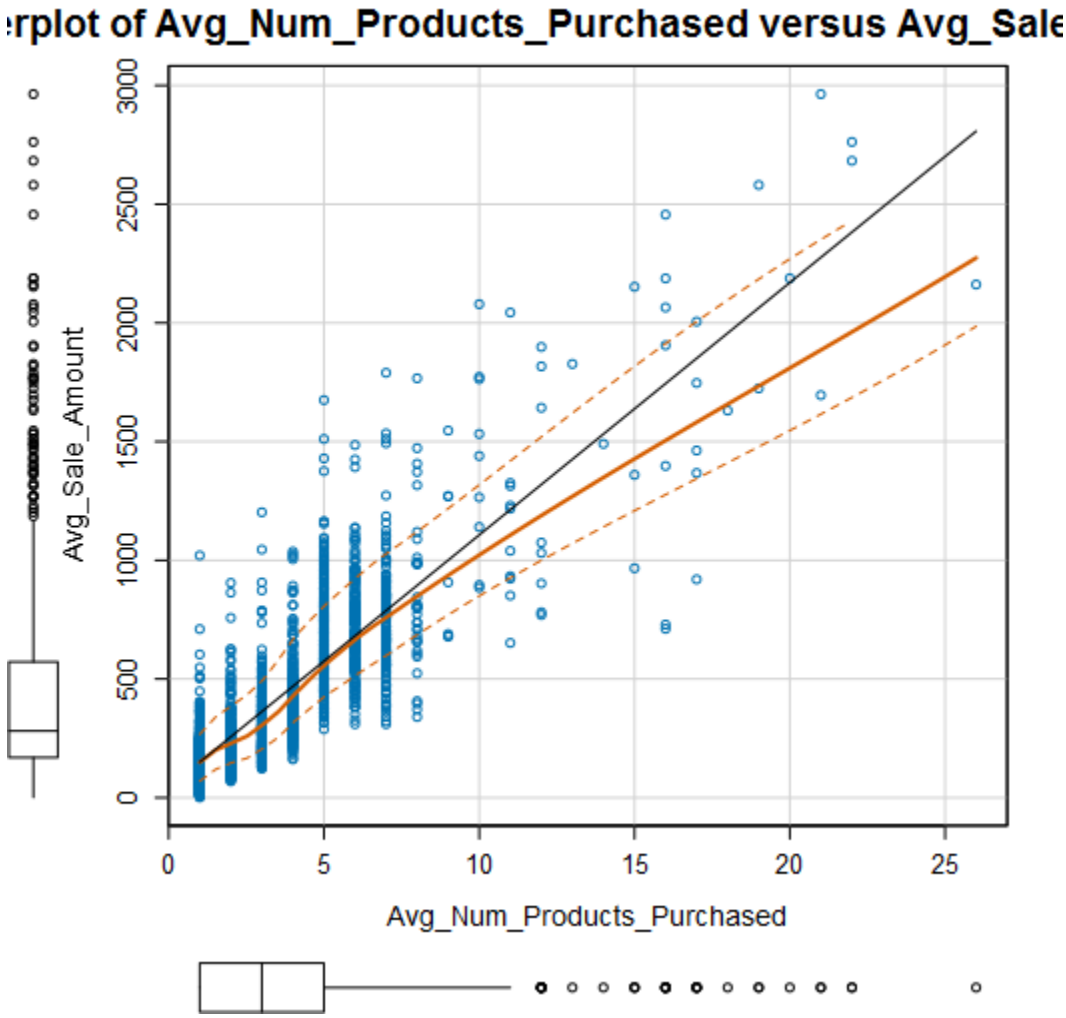
6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.698e+03	2.151e+03	-0.7895	0.42991	
Customer_SegmentLoyalty Club Only	-1.505e+02	8.975e+00	-16.7723	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.818e+02	1.190e+01	23.6838	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.430e+02	9.822e+00	-24.7438	< 2.2e-16	***
ZIP	2.629e-02	2.662e-02	0.9875	0.32352	
Store_Number	-9.819e-01	1.006e+00	-0.9763	0.32902	
Responded_to_Last_CatalogYes	-2.889e+01	1.128e+01	-2.5623	0.01046	*
Avg_Num_Products_Purchased	6.675e+01	1.516e+00	44.0363	< 2.2e-16	***

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

- **Store_Number vs Avg_Sales (Linear)**



- On the basis of p-values, **"Avg_Number_Of_Products_Sold", "Customer_Segment"**, both are significant.

2. Explain why you believe your linear model is a good model.

Report

Report for Linear Model Linear_Regression

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

- adjusted R-squared value of 0.8366 which is a high value.
  R-squared ranges from 0 to 1 and represents the amount of variation in the target variable explained by the variation in the predictor variables. The higher the rsquared, the higher the explanatory power of the model.
  here the adjusted r-squared value us 0.8366 and multiple r-Squared is 0.8369.
  It implies that I was able to improve the model.

- Customer Segment and Average Number of Products also have a p-value lower than 0.05, implying their statistical significance.
  For both the predictor variables, we used in our linear model creation, p-value (probability that the coefficient is going to be 0) is very less.
  The lower the p value the higher the probability that a relationship exists between the predictor and target variable.
  Thus, the model is considered a good one.

3. What is the best linear regression equation based on the available data?

- **Avg_Sales = 303 – 149.36Loyalty_Club_Only + 281.84Loyalty_Club_And_Credit_Card – 245.42Store_Mailing_List + 0Credit_Card_Only + 66.98*Avg_Number_Products_Purchased**

Step 3: Presentation/Visualization

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

- Yes, Company should send the catalog to these customers. Since the condition was that if the profit exceeds $10000, and it actually exceeds as calculated using linear regression model, hence catalog should be sent.

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 = $21,987.43**