Project 1: Predicting Catalog Demand

Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**
   - Should the company send the catalogs to 250 new costumers from company mailing list?
   - If yes, what will be the predicted profit?

2. **What data is needed to inform those decisions?**

   - Historical customer profile and sales data are needed for the prediction of this year's mailing catalog.

Analysis, Modeling, and Validation

Variables:

Target variable is the “average sale amount”,
Predictor variables are “customer segment”, “average number of products purchased”, and “number of years as a customer”.
“customer segment” is a categorical variable with three categories:
a) customer segment loyalty club only
b) customer segment loyalty club and credit card
c) customer segment store mailing list
“average number of products purchased” and ‘number of years as a customer” are the continuous variables.

Correlation:

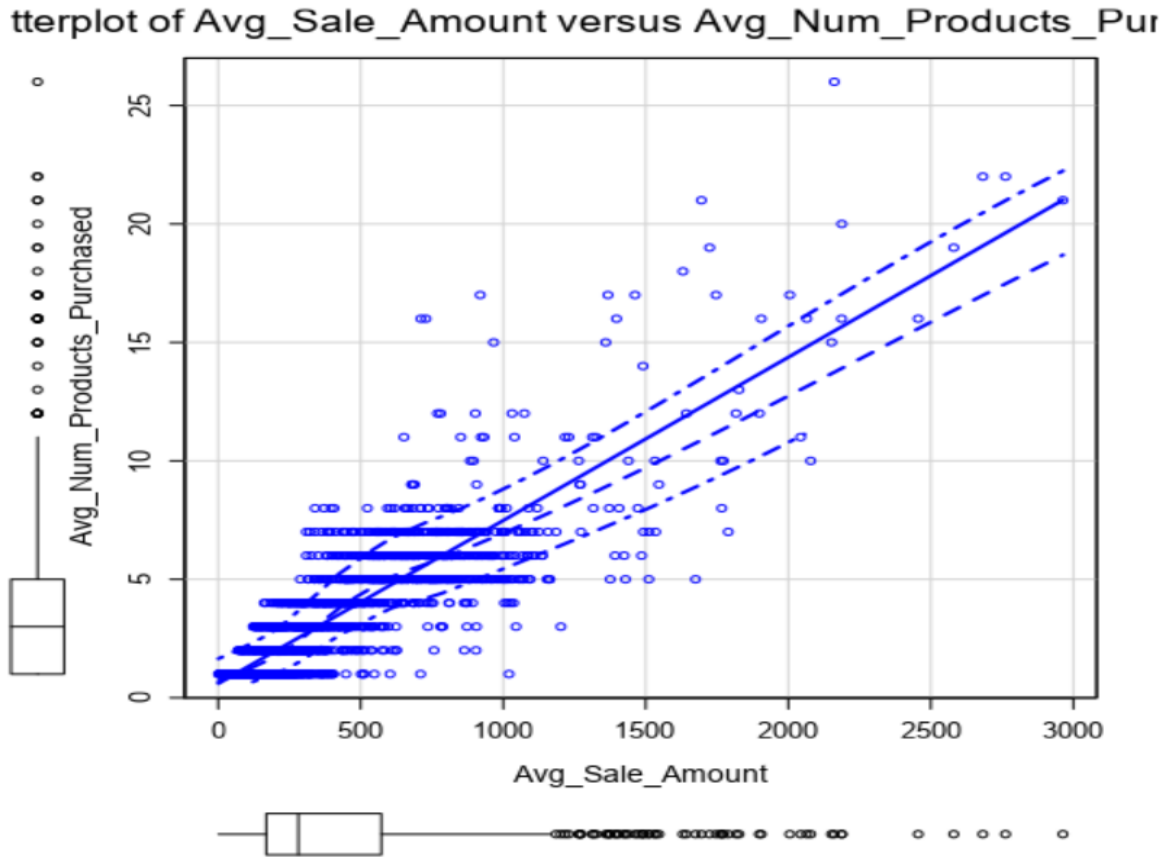
I observed a strong correlation (r=0.80) between target variable (average number of products purchased) and continuous predictor variables (average sale amount).

Table 1 shows Pearson correlation between target variable and continuous predictor variables that shows a strong correlation between “average number of products purchased” and “average sale amount” (r=0.80), but the correlation between ‘number of years as a customer” and “average sale amount” is weak (r=0.02). “average number of products purchased” used as a predictor variable, and eliminated the variable of the ‘number of years as a customer”.

	FieldName	Avg_Sale_Amount	Avg_Num_Products_Purchased	NumberOfYearsAsACustumer
1	Avg_Sale_Amount	1	0.855754	0.029782
2	Avg_Num_Products_Purchased	0.855754	1	0.043346
3	NumberOfYearsAsACustumer	0.029782	0.043346	1

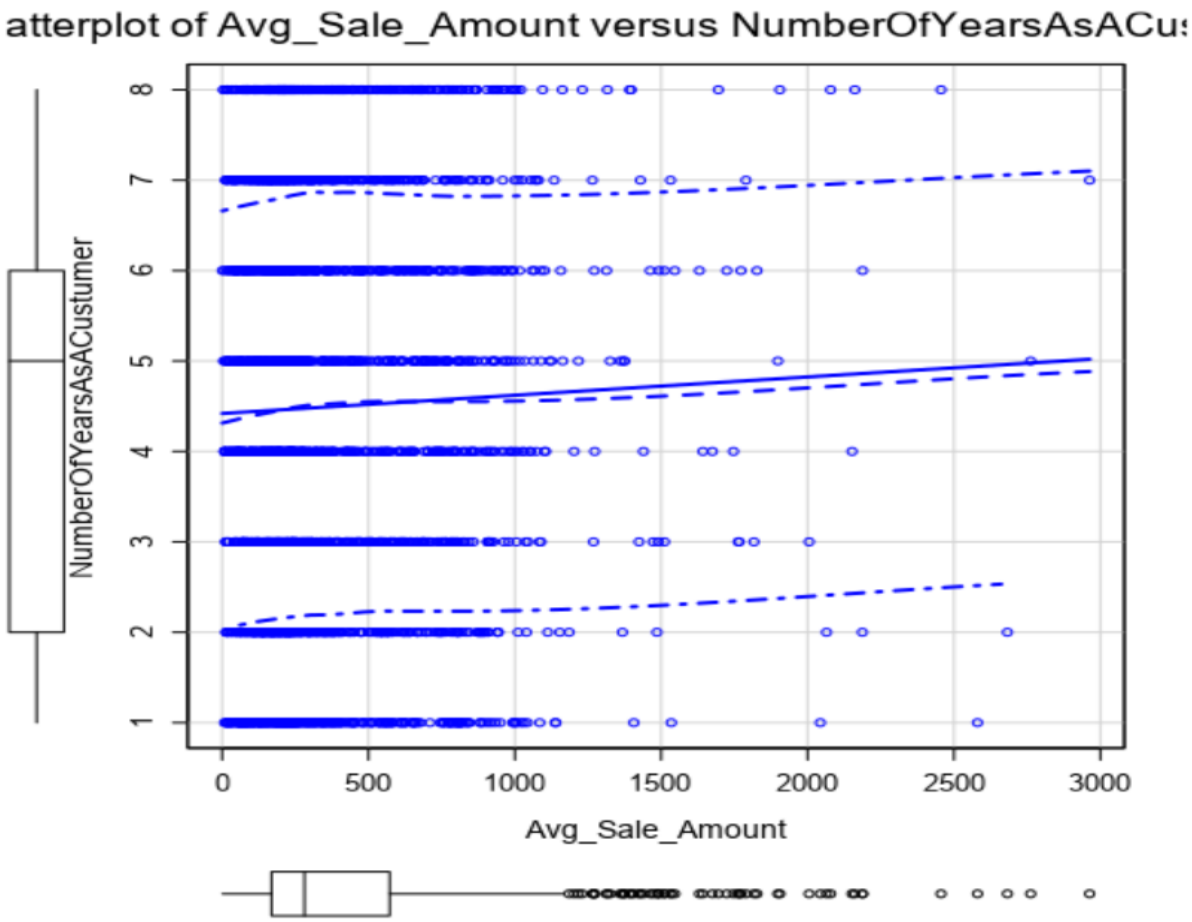
Table 1: Pearson Correlation

Scatterplot:



Graphic 1

Graphic 1 shows a strong positive relationship between Avg_Sale_Amount and Num_Products_Purchased.



Graphic 2

Graphic 2 shows no linear relationship between Avg_Sale_Amount and NumberOfYearsAsACustomer.

Regression Model:

Training-data-set: Last year's customer data
Test-data-set: List of this years' customers data which the company plans to send.

Target Variable: Average Sale Amount
Predictive Variables: "Customer Segment", and "Average Number Of Products Purchased"

The model shows which predictor variable(s) significantly explain the target variable.

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

Table 2

Table 2 shows that all categories of “Customer Segment” variables and “Average Number Of Products Purchased” significantly predicts the target variables (p-value<0.001), “Number of Years as a Customer” does not significantly predict the target variable.

R-squared is a goodness-of-fit measure for linear regression models. Multiple R-squared in this regression model is 0.8369, which indicates the percentage of the variance in the target variable that the predictor variables explain collectively.

The regression equation:

Y = 303.46 + (-149.36 * ‘customer segment loyalty club only') + (281.84 * ‘customer segment loyalty club and credit card’) + (-245.42 * ‘customer segment store mailing list’) + (66.98 * ‘Average Number Of Products Purchased’)

Name	Customer_Segment	Avg_Num_Products_Purchased	Predicted Sales
Alice Dewitt	**Loyalty Club Only**	**4**	**422.01**

“Average Sale Amount” = 303.46 + (-149.36*1) + (281.84*0) + (-245.42*0) + (66.98*4)

≈ 422.01

For Alice Dewitt, Score_Yes is 0.38, and 422.01*0.38 is 163.61.

Presentation/Visualization

Recommendation:

According to the results of the linear regression model, expected profit $21,987.43 is greater than $10,000.00

I recommend the executives of the company to send the catalog to new 250 customers.