Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

    - Decide if the company have to send this year's catalog to the 250 new customers or no.

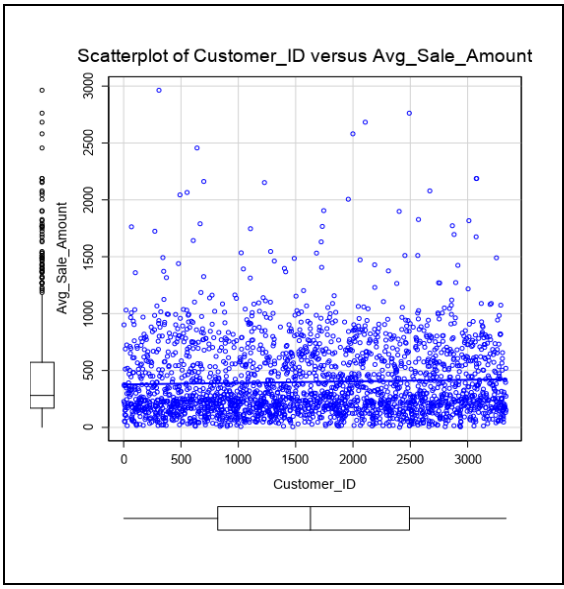
2. What data is needed to inform those decisions?

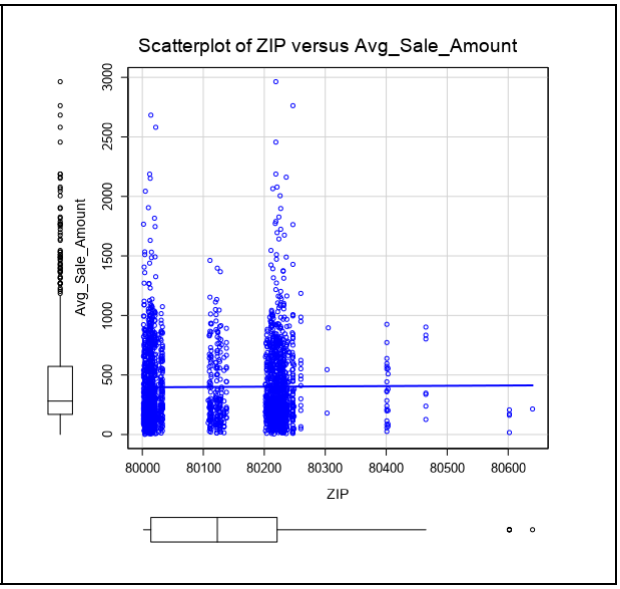
    - Data of the old customers.
    - Data of the 250 customers that the company would send a catalog to.
    - The costs of printing and distributing one catalog.
    - The average gross margin (price - cost) on all products sold through the catalog.

Step 2: Analysis, Modeling, and Validation

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

    - For numerical variables we use a scatterplot to see if the predictor variable has a linear relationship with the target variable (Avg_Sale_Amount).





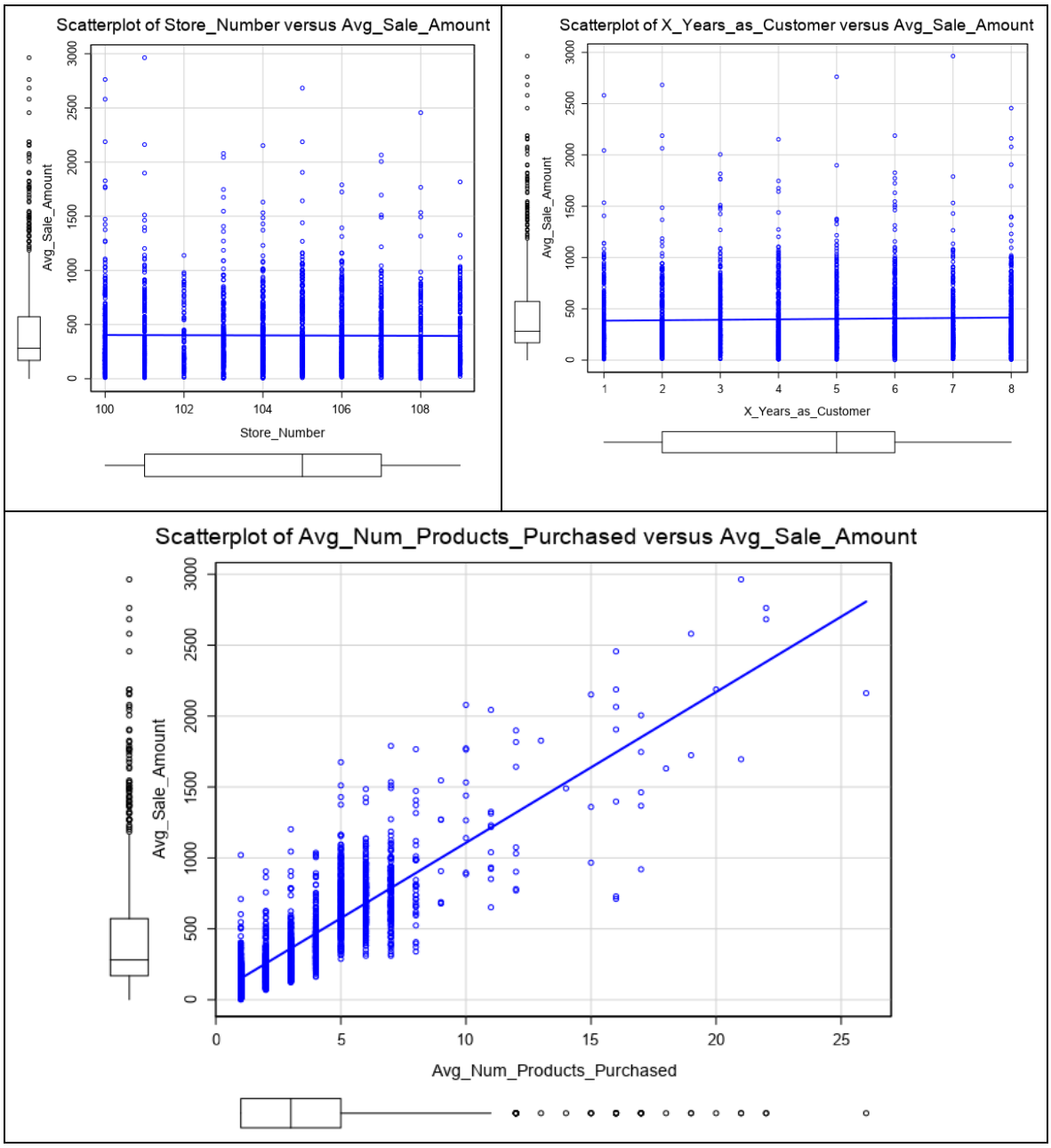


Table 1: Scatterplots of numeric variables with the target variable

- From the above Scatterplots we can see that ($Avg\_Num\_Products\_Purchased$) has a strong linear relationship with the target variable (Avg_Sale_Amount).

- For Categorical variables the best way to check for a linear relationship is to run the categorical variables through the regression model and see if the coefficients turn out to be significant with a high multiple-R-squared. After running the model with $Customer\_Segment$ the multiple-R-squared has gotten a higher value.

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

   - Multiple-R-squared=0.8369, which means that the explanatory power of the model is high.

   - p-values of all the predictor variables are below 0.05 which means that the relationship between Avg_Sale_Amount and ($Customer\_Segment$, $Avg\_Num\_Products\_Purchased$) is statistically significant.

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

Figure 1: Linear regression report

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

$$
\begin{aligned}
Avg\_Sale\_Amount \\
&= 303.46 - 149.36 * Customer\_Segment\_Loyalty\_Club\_Only \\
&+ 281.84 * Customer\_Segment\_Loyalty\_Club\_And\_Credi\_Card \\
&- 245.42 * Customer\_Segment\_Loyalty\_Mailing\_List \\
&+ 0 * Customer\_Segment\_Credi\_Only \\
&+ 66.98 * Avg\_Num\_Products\_Purchased
\end{aligned}
$$

Step 3: Presentation/Visualization

1. What is your recommendation?

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

2. How did you come up with your recommendation?

   - Build a model on the dataset that includes information on about 2,300 customers.
   - Predict sales on the dataset of the 250 new customers.
   - Calculate expected revenue from sending catalogue to each customer **(Avg_Sale_Amount * Score_Yes).**
   - Calculate expected profit **(Sum_revenue * 0.5 - 6.5 * 250).**
   - Compare if the expected profit exceeds $10,000.

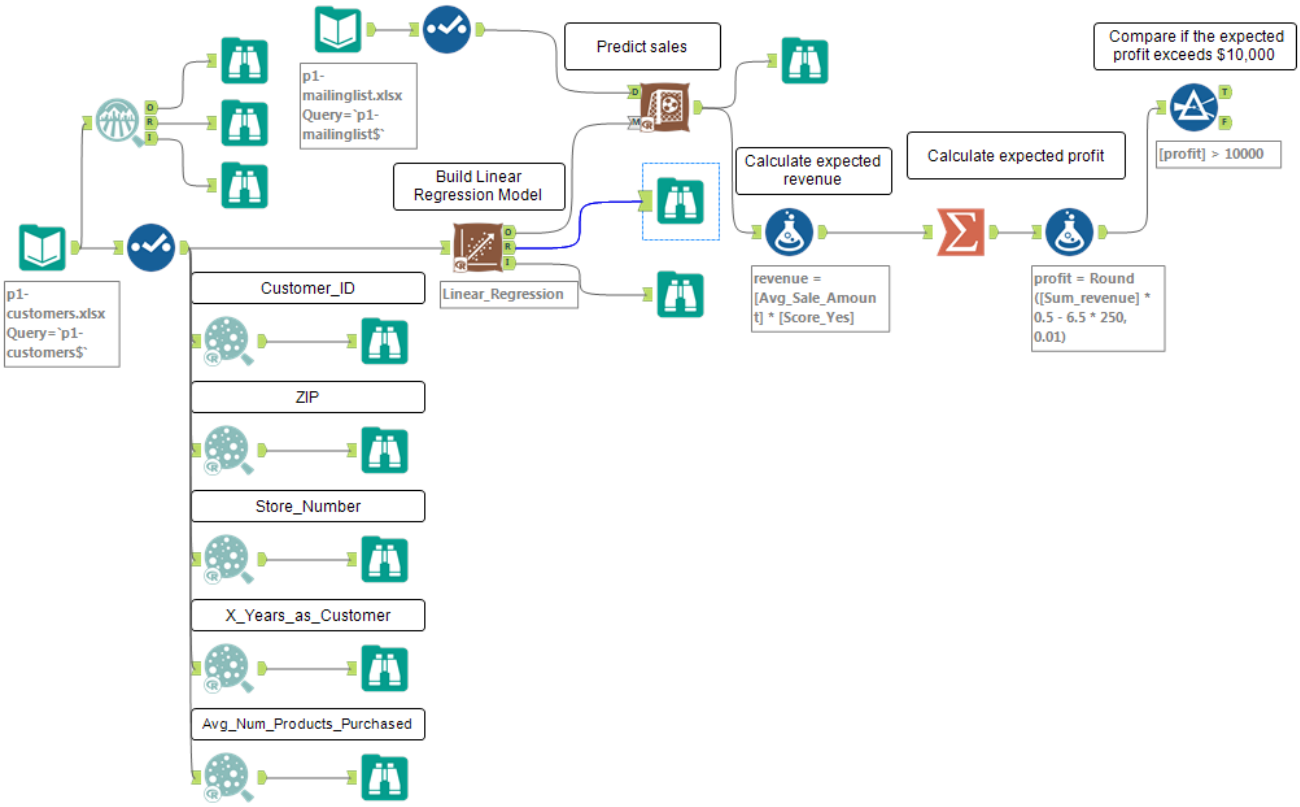


Figure 2: Determining profit workflow

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

- **21,987.44**