Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

There is a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

We need to determine how much profit the company can expect from sending a catalog to these customers.

We need to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Details

- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Make sure to multiply your revenue by the gross margin first before you subtract out the $6.50 cost when calculating your profit.

Key Decisions:

1. What decisions needs to be made?

   - Prediction of estimated profit if a catalog was sent to new customers
   - decide whether the catalog should be sent or not, based on profit

2. What data is needed to inform those decisions?
   Given
   - last year sales data when company sent out its first print catalog.
   - Probability that a new customer will buy a catalog and purchase items.
   - Information about current customers.
   - Profit Margin.
   - Cost for catalog.

Step 2: Analysis, Modeling, and Validation

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.

Non significant variables which can be ignored:

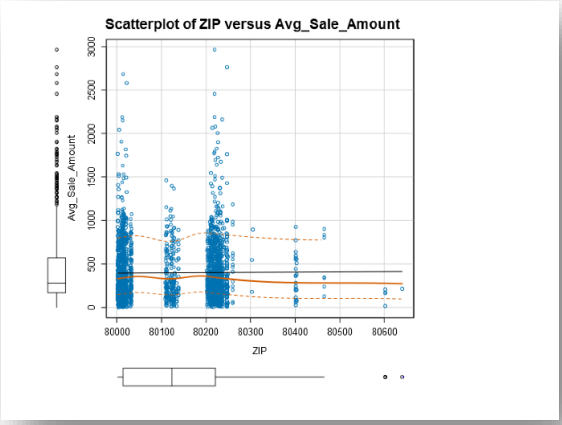
- Customer_Id : it wont change sales data
- Name : Sales not depend on names
- Address : complete information not required etc

Significant variables: based on p-value and linear relationship with target variable

- “Avg_Number_Of_Products_Sold”
- “Customer_Segment”

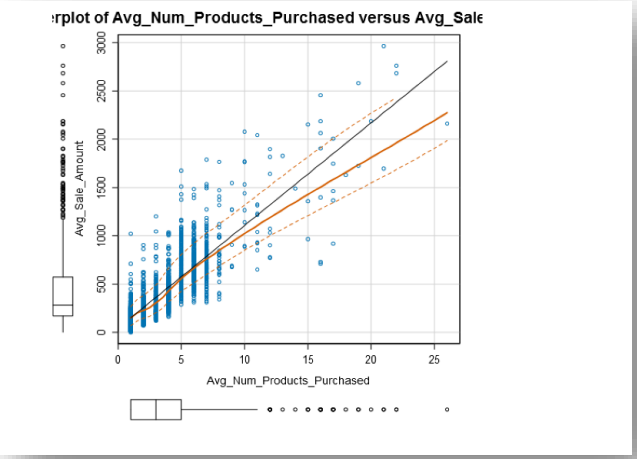
Plots: Scatterplot between numerical predictors and target variable.

Avg_Sales vs ZIP (Non linear)

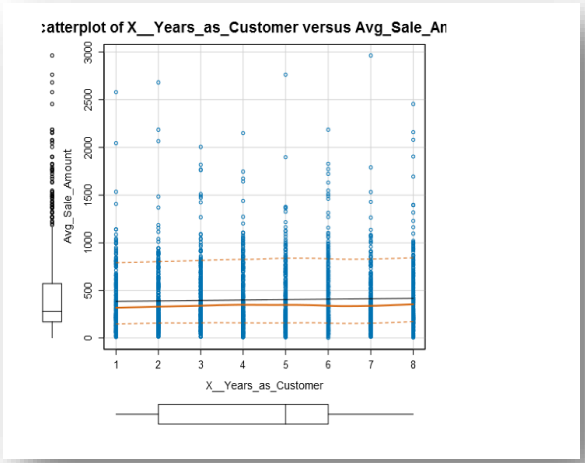


Avg_Sales vs Store_No (Non linear)

Avg_Sales vs Avg_Number_Of_Products_Sold (Linear)



Avg_Sales vs no.of_years_as_customer (non linear)



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.

Report for Linear Model Linear_Regression

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$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

- R value = 0.8366 , Higher value , Good model
- R Square = (1- SSE/SST)
- SSE – sum of squares of residuals using Predictive Model
- SST – sum of squares of residuals using Baseline Model

- SSE < SST, hence 0< R Square <1.
Higher the R Square means : Predicted values are very close to actual values which leads to smaller (SSE/SST).

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

Step 3: Presentation/Visualization

Questions:

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

As per the calculation using linear regression model profit exceeds $10000, hence I recommend that 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 , Calculated Avg_Sales.
Score_Yes : Probability that a customer will respond to catalog and make a purchase
Created a new column (Avg_Probable_Sales = Avg_Sales * Score_Yes)
Given profit margin is 50%, and cost for each catalog is $6.50, hence for all 250 customers
Calculated the profit = Avg_Probable_Sales*0.5-(6.50*250)

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
Profit = $21987.43

Alteryx Flow