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?

Ans 1 - The decision that needs to be made is whether or not catalogs should be sent out to the new customers. Management does not want to send them unless the expected profit from the sale exceeds 10,000$

2. What data is needed to inform those decisions?

Ans 2 – Ultimately we need to calculate the expected profit from the sale. We know the average gross margin on all products sold through the catalog (50%) and the printing & distrubution cost for the each catalog but to get the expected profit we need to calculcate the expected sales from the list of customers.

To calculate expected sales from each new customer we need to build a predictive model. To construct that predictive model we will use the historical dataset in file p1-customers containing historical data of 2300 customers.

Using the model built we will calculate the expected sales from the 250 customers in p1-mailinglist file. Finally we will calculate the expected profit by multiplying the expected sale by gross margin and subtracting the cost of printing and distributing the catalog.

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:

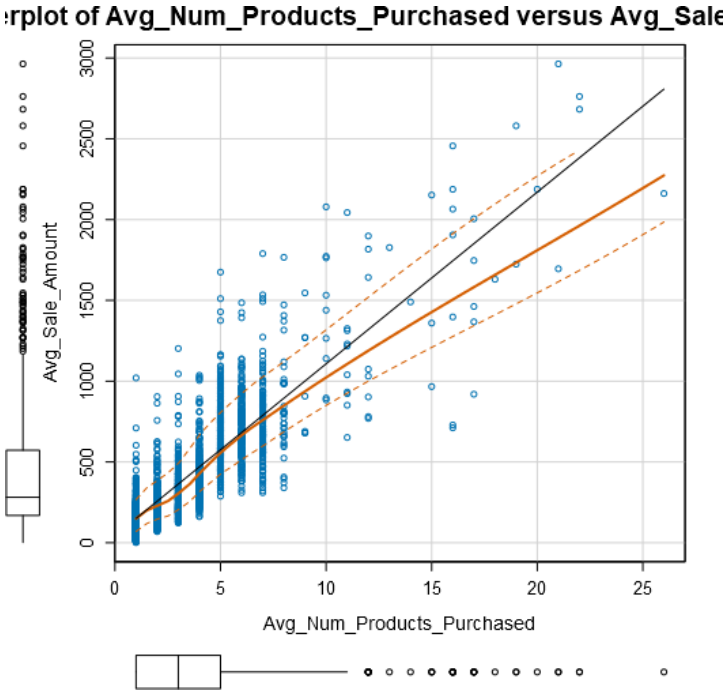
1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

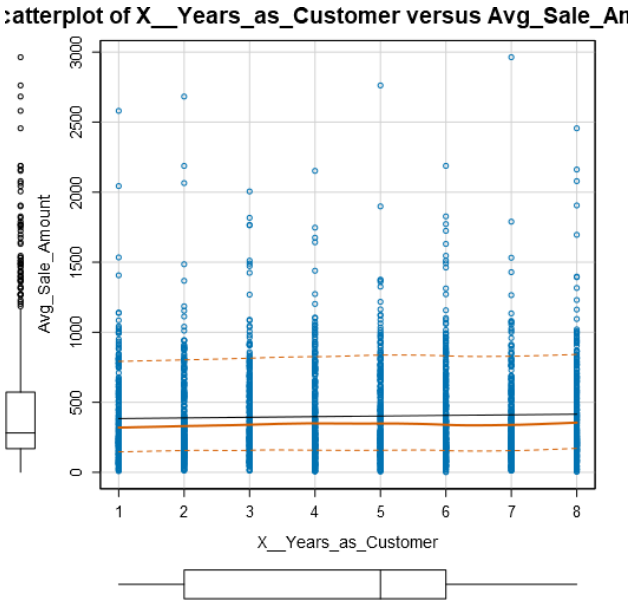
First step we input the data from the file p1 customers and create scatter plots between average selling price and other continous variables. We were trying to see which numeric variables had a linear relation ship with avg sales amt.

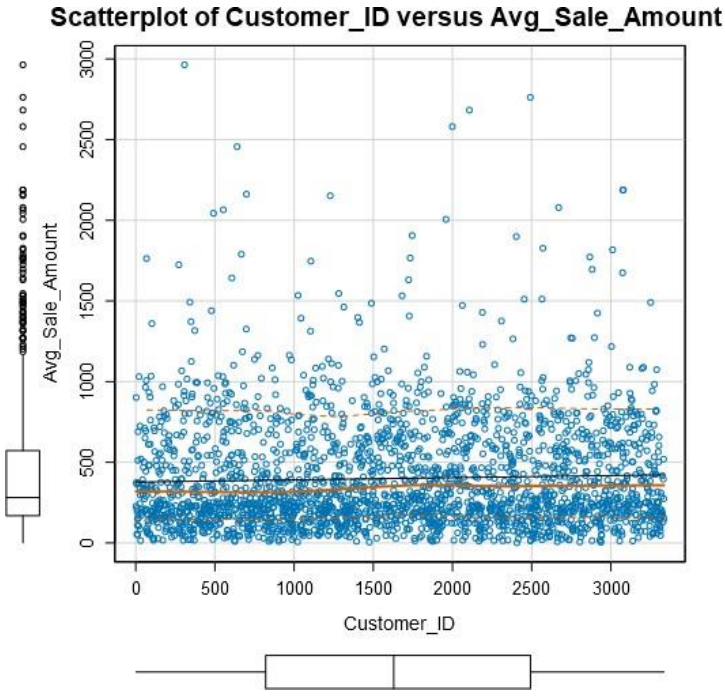
It turned out that only variable avg_num_products_purchased (average number of products purchased) had a reasonably strong correlation with avg sales amt. Hence we would include it in our predictive model

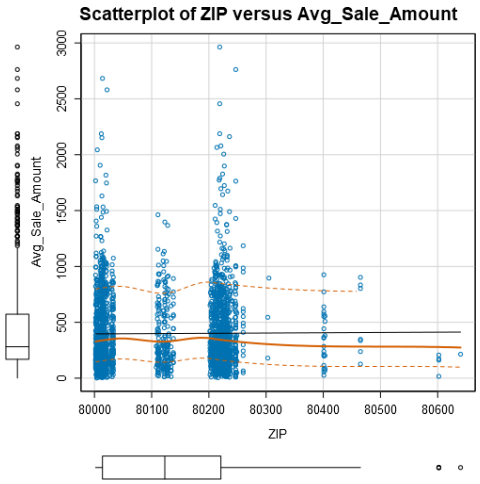
Other numeric variables (no. of years as a customer, store number, zip, customer id) had practically no correlation with the avg sales amt. So we wont include them in our model.

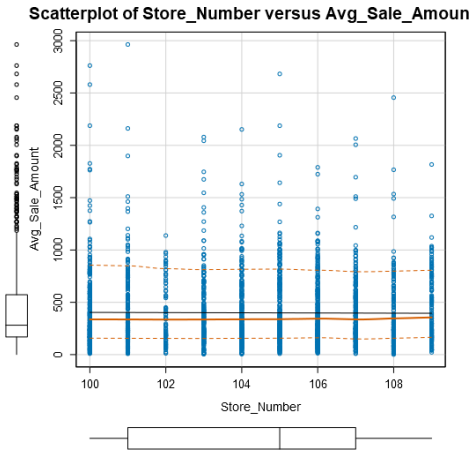
PFB the scatter plots which were used to come to this conclusion




catterplot of X__Years_as_Customer versus Avg_Sale_An
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
X__Years_as_Customer


Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
Customer_ID





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.

The linear model is a good model because of the following reasons

1. P values – for each of the variables the p values are close to zero. This indicates that the probability that the linear relationship exists by chance is extremely low. And since the values are close to zero, they are less than 0.05 and can be considered statistically significant.

2. Rsquared values – In this case the r squared value is very high 0.8366. This means that the model explains 83.66% of the variability which is very high. R-squared is a statistical measure of how close the data are to the fitted regression line

Coefficients:

7

	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	***

Record Report

1

Report for Linear Model Linear_Regression_3

2 *Basic Summary*

3 Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 Coefficients:

7

	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

8 Residual standard error: 137.48 on 2370 degrees of freedom

Record	Report
	Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366 F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

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)

Y (Avg_Sale_amount) = 303.46 + (66.98 * Avg_Num_Products_Purchased) -149.36 (IF type : customer_segmentloyalty club only) + 281.84 (IF type : Customer_segment Loyalty club and credit card) -245.42 (IF type : Customer segment Store mailing list) + 0 (IF : Customer_segment Credit card only)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

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?

Yes the company should send the catalog to these 250 customers

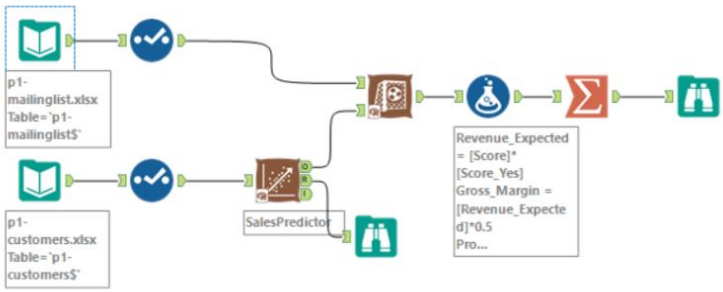
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

First the input data was used to prepare a regression model. Then after preparing the regression model, data from the mailing list was fed into the model using the score tool to calculate the expected sales . Finally the expected sales was used to calculate the expected profit , using the formula tool. Since the expected profit is 21987.44 $ which is greater than 10,000 $, the recommendation is to send the catalog to these 250 customers.

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

= 21987.44 **$**

Alteryx Workflow



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.