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?

   We need to decide either to print the catalogs for 250 new customers or not print the catalog based on the prediction of the profit for the new customers being more than$10,000

2. What data is needed to inform those decisions?

   The catalog is printed only if the expected profit is greater than $10,000. In order to find the expected profit, we need to know probability of ordering the catalog, the cost of making the catalog and the revenues. The profit is calculated as the revenue multiplied by the probability of ordering minus the cost of making and sending out the catalog. We are given the cost of printing and the probability of ordering. To calculate revenues, we perform multiple linear regression of customer segment and number of products ordered.

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)

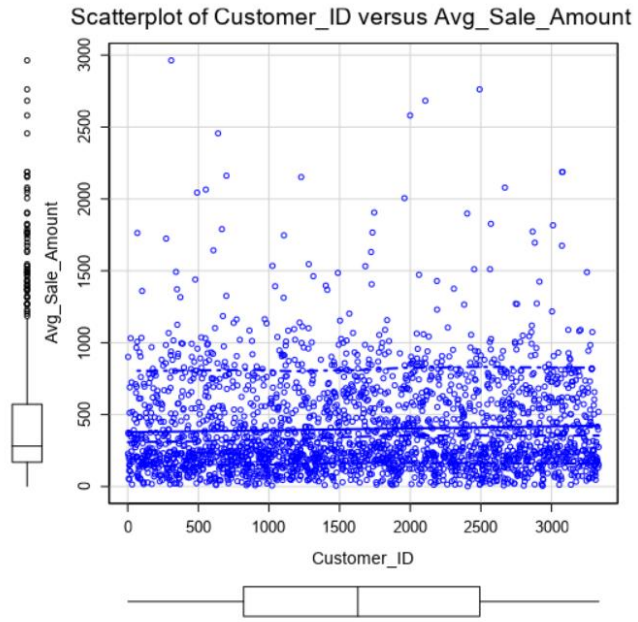
Important: Use the p1-customers.xlsx to train your linear model.

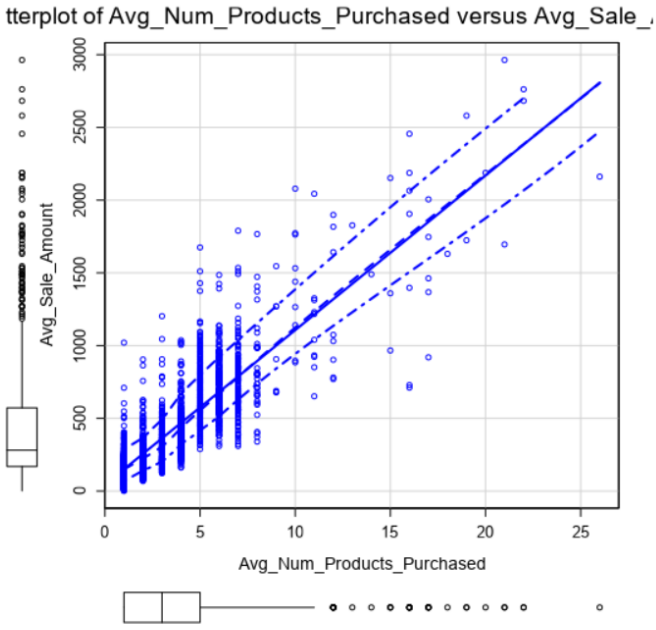
At the minimum, answer these questions:

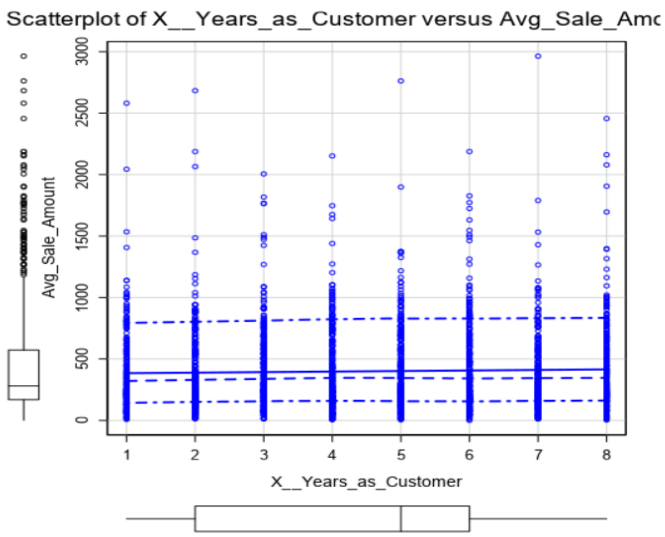
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

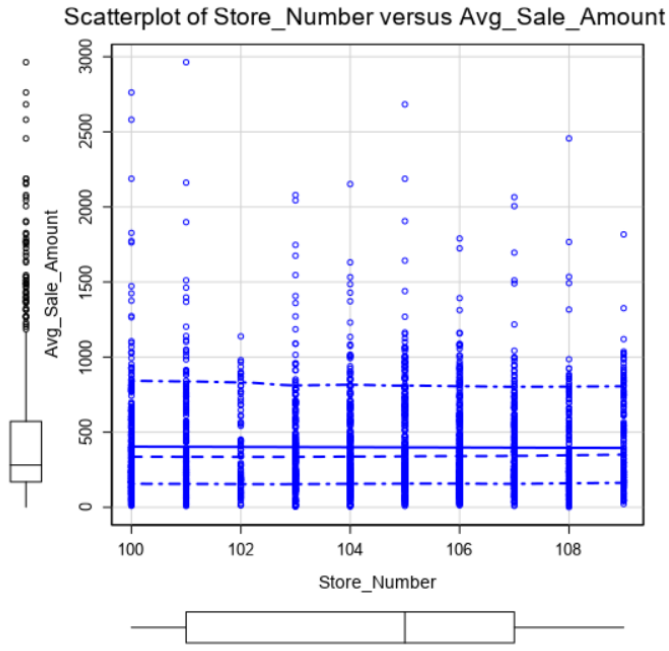
   ANSWER: To find out the linear relationship between the predictor variable and the target variable, I produced the scatter plots of numeric predictor variables vs Avg_dales_amount and observed it cafefully. Only **'avg_num_products'** tend to have

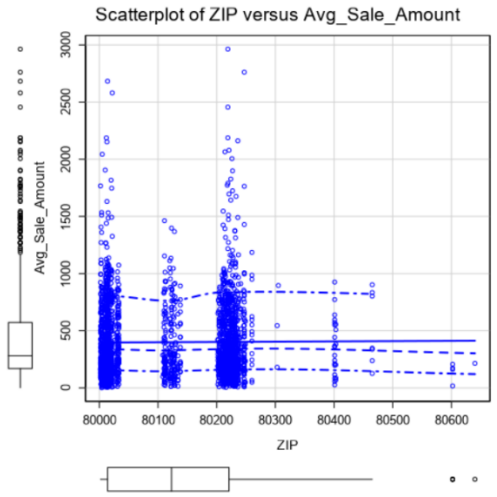
linear relationship. The scatter plots are given below. I used linear regression including non-numeric variables and the selected numeric variable. Then I selected only those where the p value is <=0.05. The selected predictor variables were ‘Customer-Segment’ with p value 0 and avg_num_products with p value 0, indicating to be the most significant.






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


Scatterplot of Store_Number versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
100
102
104
106
108
Store_Number



Only the scatter plot of avg_num_products vs Avg_sales_Amounts tends to have linear re

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.

   ANSWER: The linear model is a good fit as only the predictor variables with p <=0.05 were chosen and the high adjusted r-squared value is 0.837.

Report for Linear Model catalo_regressionLinear_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

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

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

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_Amount = 303.46 + 0.00 (If Type: Customer_SegmentCredit Card Only)
- 149.36 (If Type: Customer_SegmentLoyalty Club Only)
+ 281.84 (If Type: Customer_SegmentLoyalty Club and Credit Card)
- 245.42 (If Type: Customer_SegmentStore Mailing List)
*+ 66.98 * Avg_Num_Products_Purchased*

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

At the minimum, answer these questions:

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

ANSWER: Yes, the company should send catalog to the 250 customers as the predicted profit is much higher than $10,000.

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

ANSWER: In order to recommend whether to print catalogs or not, I followed the directive given to print catalog for 250 new customers if the predicted profit exceeds $10,000.
To get the predicted profit, the avg_sale_amount was calculated by running a linear regression with the predictive variables customer_segment and the avg_num_products which were chosen as mentioned above. Also, we had the probability of the customer ordering and not ordering. Using the probabitly of ordering with the avg_sale_amount(revenue) minus the cost of printing , I calculated the profit per customer.i.e. Profit = New_Customer_revenue * Probability of ordering -$6.50.Summing up the profts of 250 new cusomer gave the result $21,987.44.
Since, this value is greater than $10,000, the recommendation to make the catalog was made.

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

ANSWER: The expected profit is $21,987.43