Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The company did the last year mailling catalog compiegn by sending out catalogs to customers ,and they 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.

to determine this compeign will succeed by predict the expected profit from these 250 new customers and the management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   sending catalogs to new customer

2. What data is needed to inform those decisions?
   The data collected from last year first print catalog sended out.
   The data of the new customers from a mailling list.
   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%.

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)

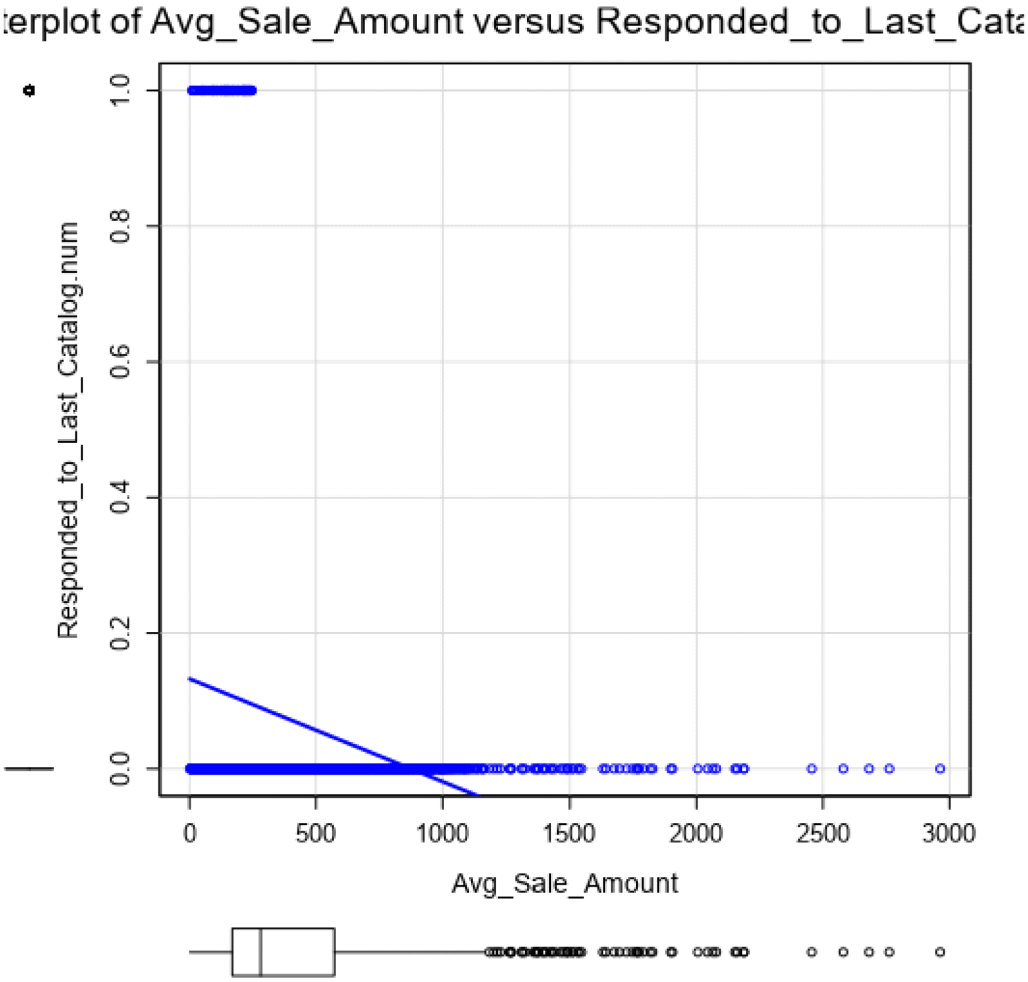
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

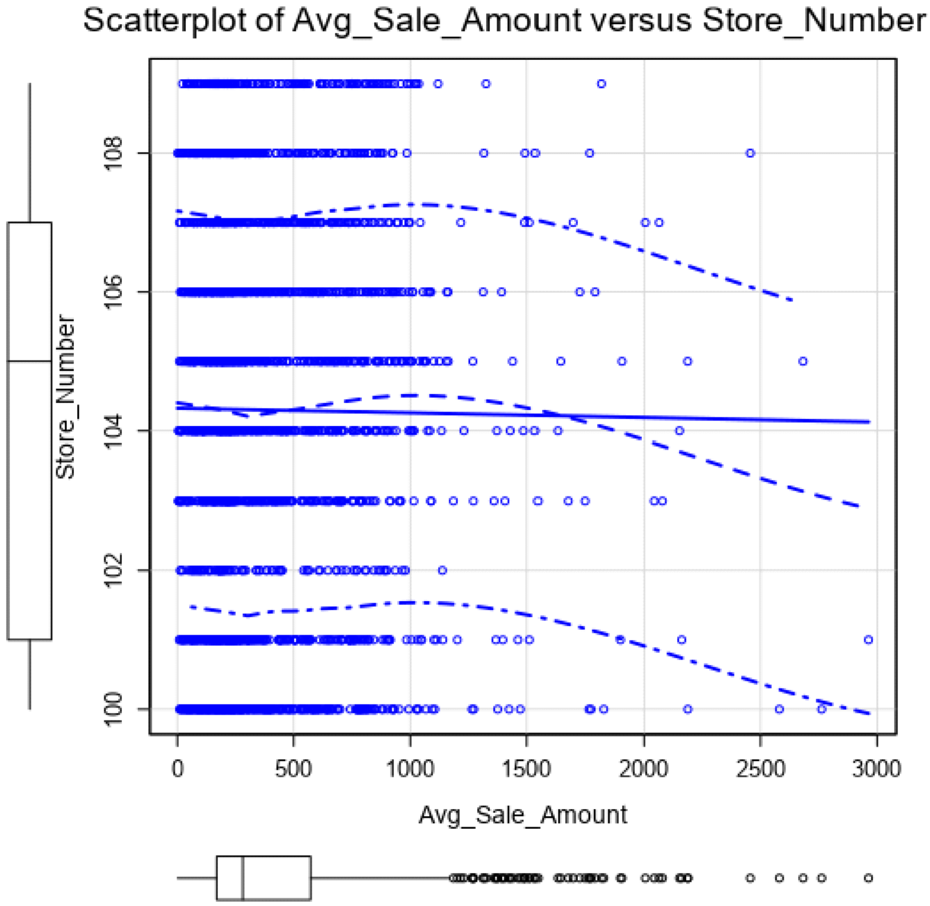
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.
   - The variable that we want to predict (target variable) is Avg-Sale-Amount for each customer

- predictor variables : initially exclude some variables ,Avg_Sale_Amount dose not depend on (name,customer_id,zip)
- check correlation between variables and target variable

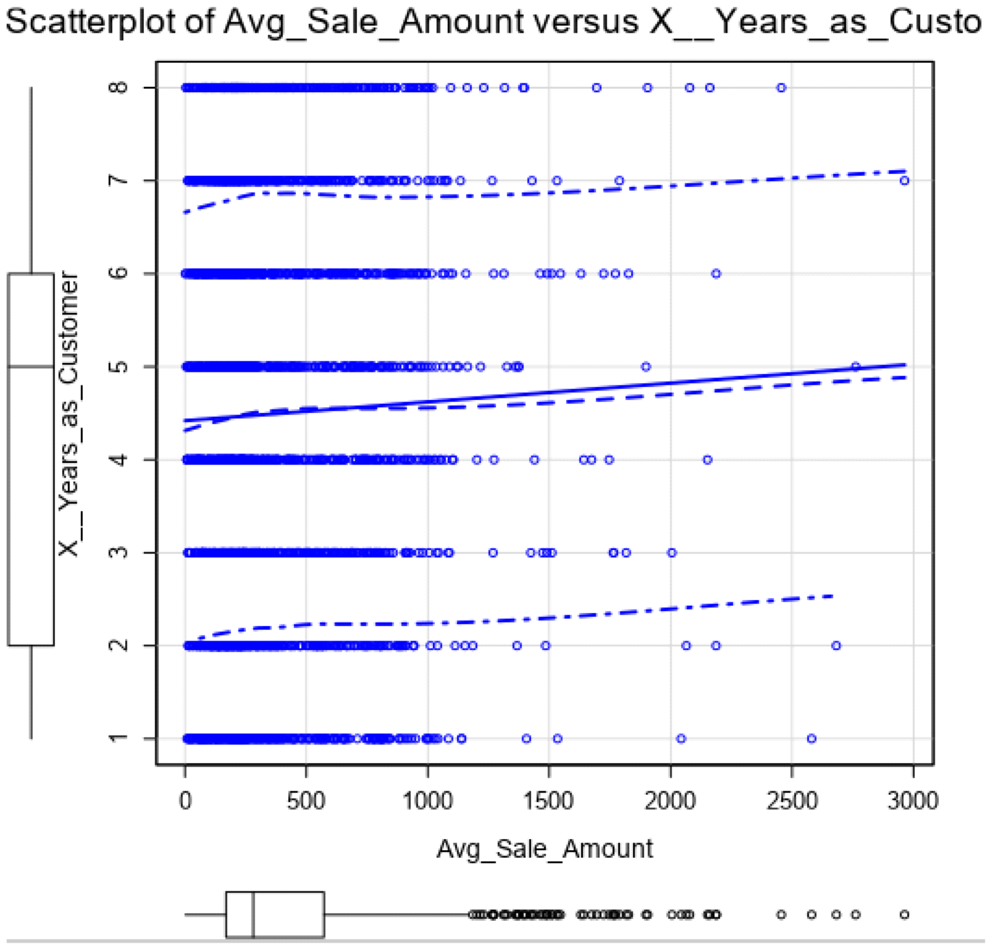
1- **with Responded_to_last_catalog:no linear correlation**



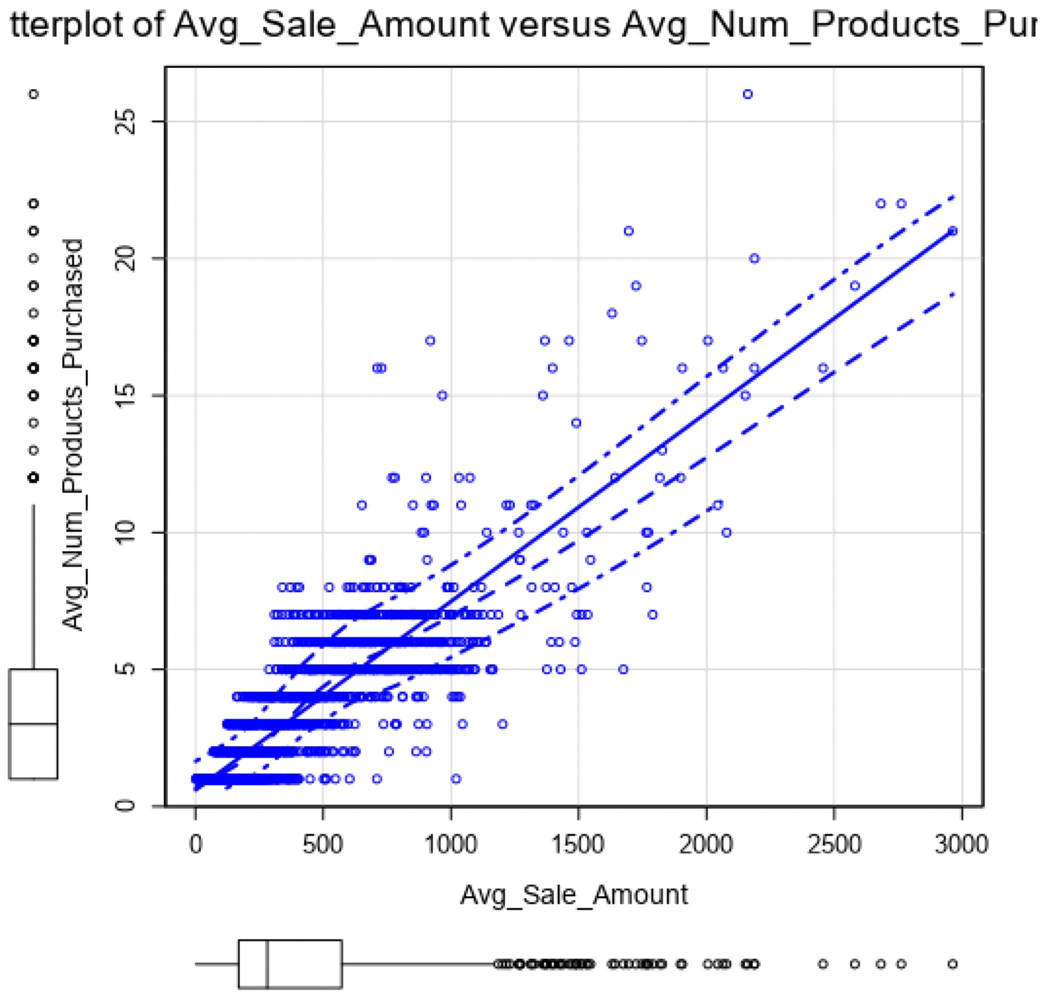
2-with store_number :no linear correlation



3-with year_as_customer :no linear correlation

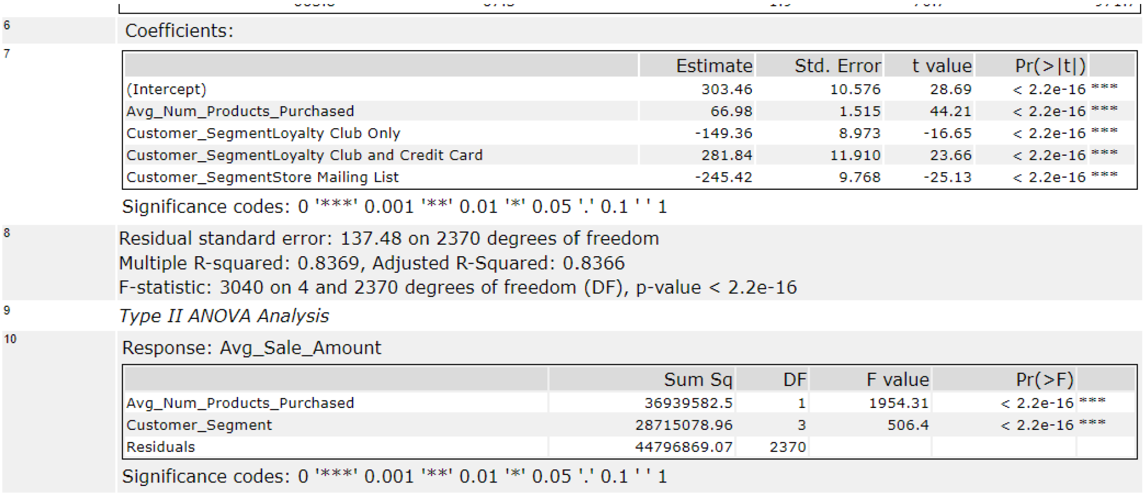


4-with Avg_number_product_purchased:positive linear correlation



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.

- check is the p-value on the predictor variables for the rest variables from linear regression tool



6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 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 ***

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

8 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

9 *Type II ANOVA Analysis*

10 Response: Avg_Sale_Amount

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

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

p-value of Avg_Num_Products_Purchased and Customer_Segment are smaller than 0.05 and R_Squared 0.8369

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)

Important: The regression equation should be in the form:

multiple linear regression model form:
*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

Avg_Sale_Amount = 303.46 + 66.98(Avg_Num_Products_Purchased) – 149.36(Customer_SegmentLoyalty Club Only) + 281.84(Customer_SegmentLoyalty Club and Credit Card) – 245.42(Customer_SegmentStore Mailing List)

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?

Since the profit =21987.4356 exceed $10,000 , yes sending catalogs recommended

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   1. first predict the Avg_Sale_Amount for the new mailling list dataset using score tool
   2. calculate the revenues =[Avg_Sale_Amount] *[Score_Yes]
   3. calculate profit or gross margin=[revenues]*0.5-6.5

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

find the sum of the profit using summarize tool profit =21987.4356