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?
    - Predicting an estimated profit if the catalog was sent to new customers and then on the basis of profit, determine whether the catalog should be send to them or not.

2. What data is needed to inform those decisions?
    - Data about sales occurred last year when company sent out it's first catalog (Given)
    - Probability that a new customer will buy a catalog and purchase items (Given)
    - Information about current customers, shopping behavior, location, etc (Given)
    - Profit Margin(50%, Given)
    - Cost structure (Cost for catalog is given)
    - Since, we have the past data about sales, we can predict the sales for current year. And then multiplying the sales by probability that a new customer will respond to a catalog and make a purchase (Score_Yes), we get the sales for current year. On the basis of which profit can be calculated and then a decision could be taken if the catalog should be sent or not.

Step 2: Analysis, Modeling, and Validation

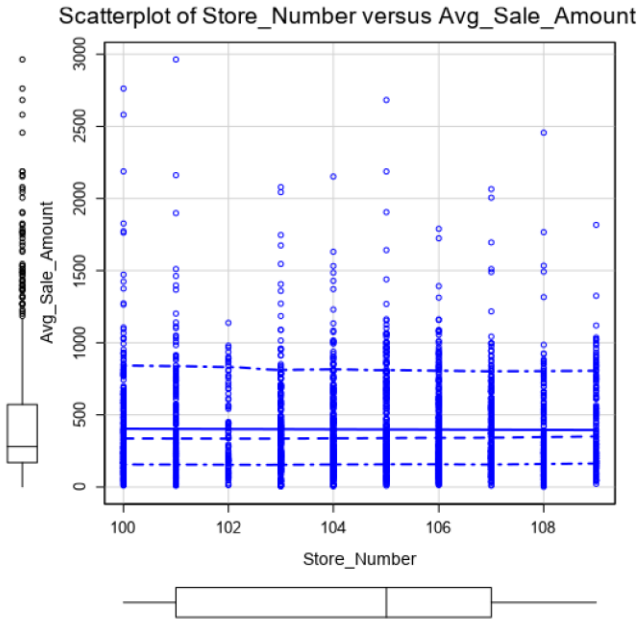
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

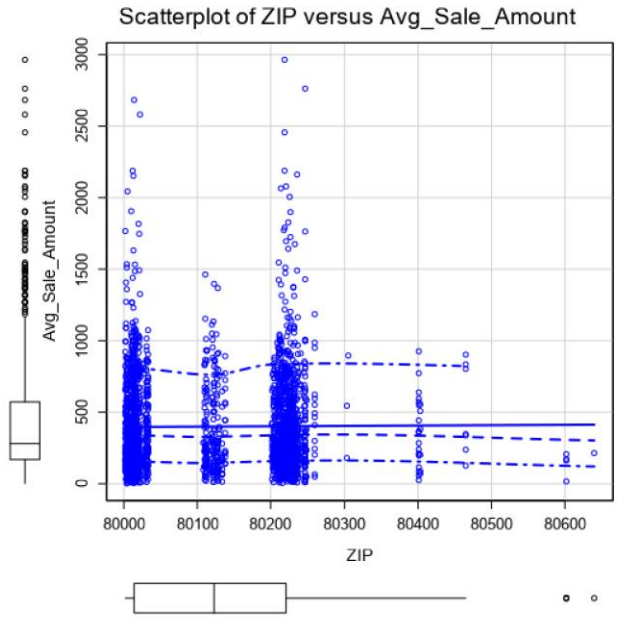
Important: Use the p1-customers.xlsx to train your linear model.

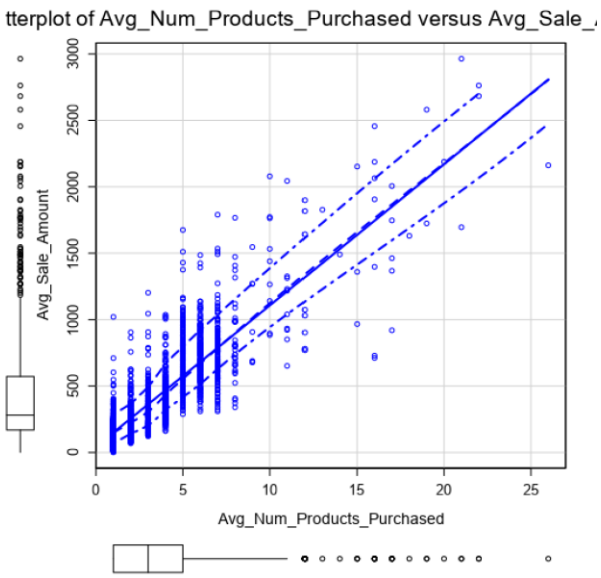
At the minimum, answer these questions:

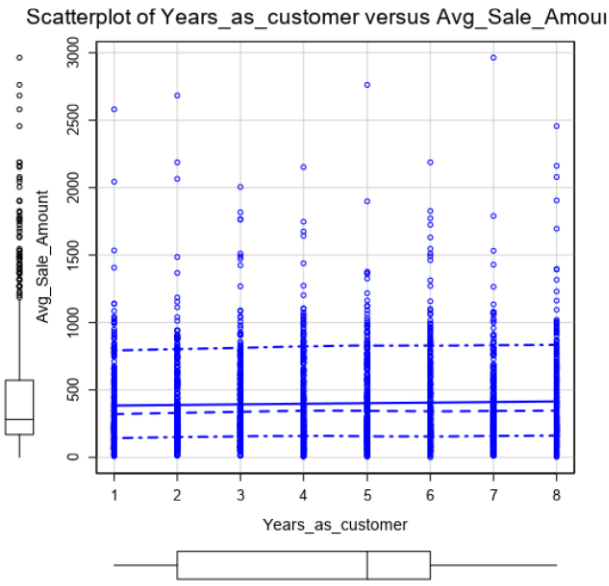
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.
    - Name: Sales doesn't depend on a user's name.

- Customer_Id: It is a unique number assigned to each customer. Doesn't affect sales.
- Address: Its too detailed, instead of it we can use other variables like city or zip.
- Analyzed the scatter plot between numerical predictors and target variables









- Avg_Num_Products_Purchased versus Avg_Sale_Amount scatterplot shows not so strong linear relationship. Rest all of the other three scatterplots doesn't show linear relationship.
- Used 'Avg_Num_Products_Purchased' and 'Customer_Segment' to build the linear model.
- On the basis of p-value, both 'Avg_Num_Products_Purchased' and 'Customer_Segment' are significant

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.

Record	Report
1	**Report for Linear Model Linear_Regression_14**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	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

- For both the predictor variables, we used in our linear model creation, p-value (probability that the coefficient is going to be 0) is very less. Hence, both the predictors are significant in deciding the target variable.

- This model is strong since the Adjusted R-square value is very high (0.8366)

- There is no established association/relationship between p-value and R-square. This all depends on the data (i.e.; contextual).
- R-square value tells you how much variation is explained by your model. So 0.1 R-square means that your model explains 10% of variation within the data. The greater R-square the better the model. Whereas p-value tells you about the F statistic hypothesis testing of the "fit of the intercept-only model and your model are equal". So if the p-value is less than the significance level (usually 0.05) then your model fits the data well.
- Thus you have four scenarios:
- 1) low R-square and low p-value (p-value <= 0.05)
- 2) low R-square and high p-value (p-value > 0.05)
- 3) high R-square and low p-value
- 4) high R-square and high p-value
- Interpretation:
- 1) means that your model doesn't explain much of variation of the data but it is significant (better than not having a model)
- 2) means that your model doesn't explain much of variation of the data and it is not significant (worst scenario)
- 3) means your model explains a lot of variation within the data and is significant (best scenario)
- 4) means that your model explains a lot of variation within the data but is not significant (model is worthless)
- The **sum of squares total**, denoted **SST**, is the squared differences between the observed *dependent variable* and its **mean**.

- The second term is the **sum of squares due to regression**, or **SSR**. It is the sum of the differences between the *predicted* value and the **mean** of the *dependent variable*.
- The last term is the **sum of squares error**, or **SSE**. The error is the difference between the *observed* value and the *predicted* value.

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)

Average_sales = 303.46 – 149.36**Loyalty_Club_Only* + *281.84**Loyalty_Club_And_Credit_Card – 245.42**Store_Mailing_List* + *0**Credit_Card_Only + 66.98*Avg_Num_Products_Purchased

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:

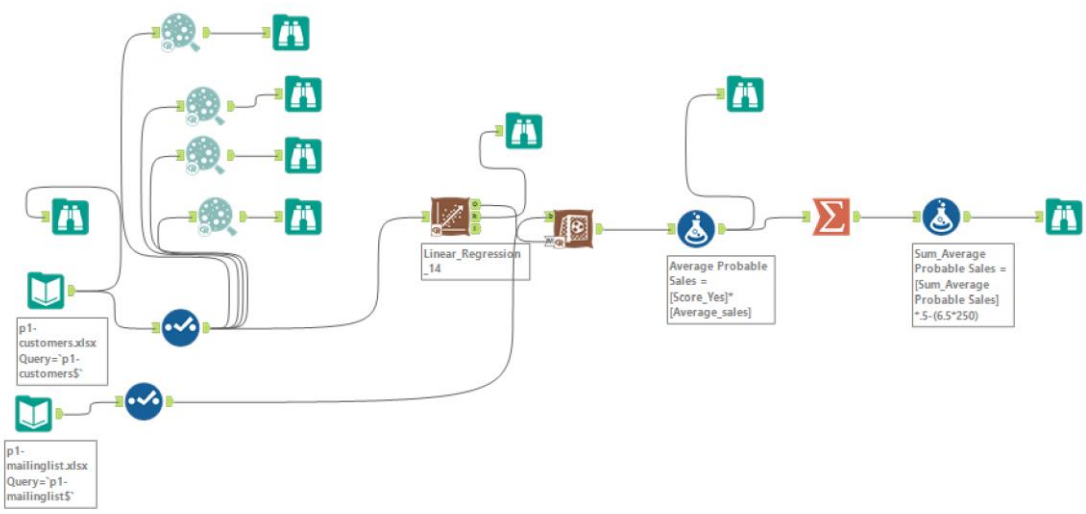
1. What is your recommendation? Should the company send the catalog to these 250 customers?
    - Yes, Company should send the catalog to these customers. Since the condition was that if the profit exceeds $10000, and it actually exceeds as calculated using linear regression model, hence catalog should be sent.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
    - Calculated Average_sales using linear regression model. Score_Yes : Probability that a customer will respond to catalog and make a purchase
    - Created a new column (Average Probable Sales = Average_sales * Score_Yes) Given profit margin is 50%, and cost for each catalog is $6.50, hence for all 250 customers

- Calculated the profit = Average 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 equals $21987.43

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.