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?

The decision the manager needs to make is, ”Do we design attractive catalog with least cost to new 250 customers for optimizing the chance of earning profit once send it” and “if yes, does the excepted profit from new customers pass the defined limit that was 10,000?”

2. What data is needed to inform those decisions?

According to calculate the predicted profit, we should consider the coming data to inform the decisions:

- The cost of printing and distributing.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Years as customer.
- Average number of purchased product.
- Average sale amount.

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.

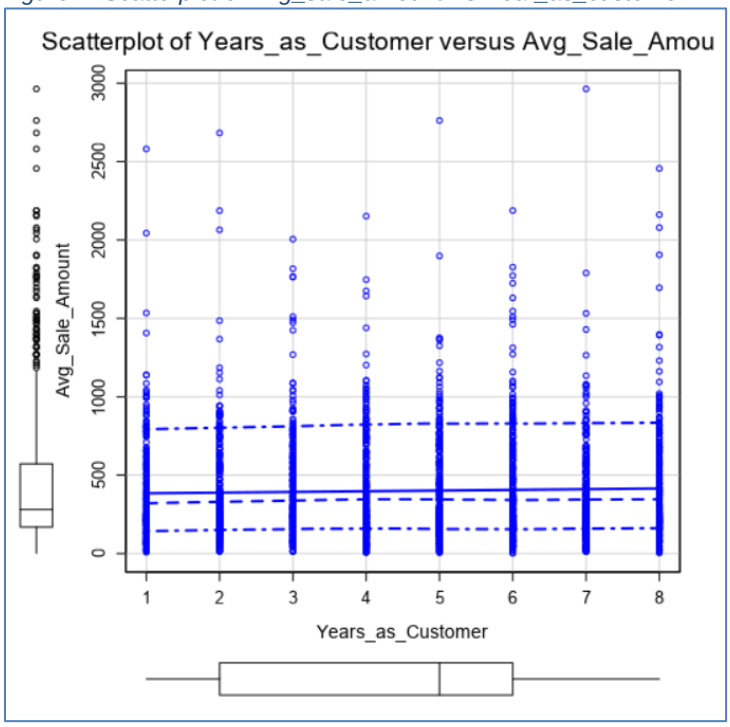
At the minimum, answer these questions:

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.

Firstly, I ignored some attributes after exploring the data set are Name, Customer_ID, Address, State, Store_Number and Responded_to_Last_Catalog due to some reasons which are some attribute has unique values and is information such as Name, Customer_ID, Store_Number and Address. Also, the attribute has only one value that could not provide any insight since it is the same on all records in the data set like State. As well as, the attribute exists in one data set only

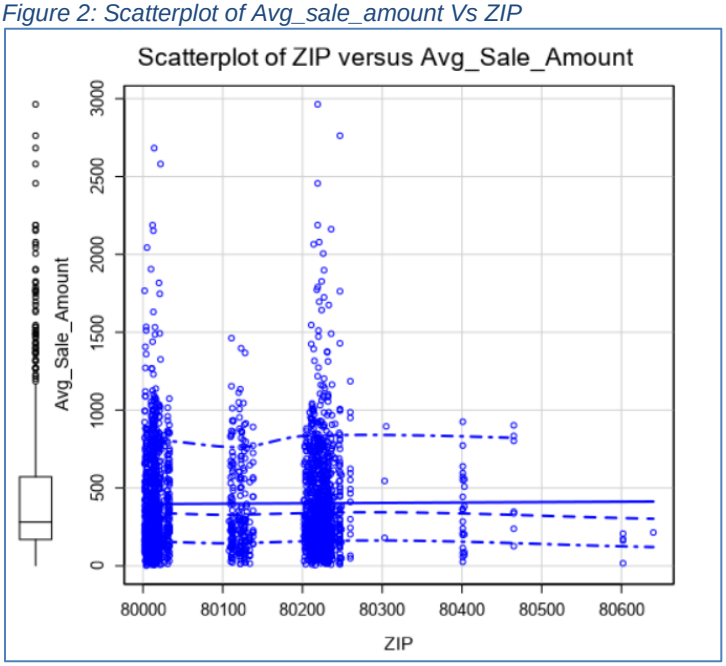
Responded_to_Last_Catalog. After that, I used the remaining attributes to show the linear relationship on the scatterplot in order to choose continuous predictor variables with avg_sale_amount as the target variable. And the all remaining attributes are continuous except the customer segment.

Figure 1: Scatterplot of Avg_sale_amount Vs Year_as_customer



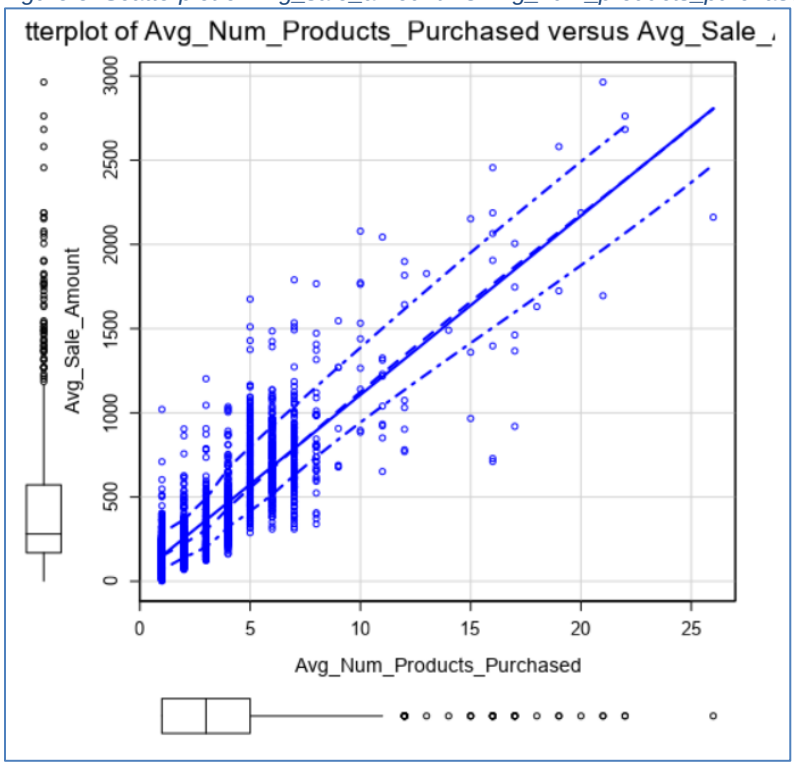
In the above figure, it represents no lineaar relationship with avg_sale_amount.

Figure 2: Scatterplot of Avg_sale_amount Vs ZIP



In the above figure, it represents no linear relationship with avg_sale_amount.

Figure 3: Scatterplot of Avg_sale_amount Vs Avg_num_products_purchased



In the above figure, it represents linear relationship with avg_sale_amount.

Figure 4: coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.804e+03	2107.1663	-0.8561	0.39203	
Customer_SegmentLoyalty Club Only	-1.492e+02	8.9692	-16.6357	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.827e+02	11.9104	23.7354	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.455e+02	9.7624	-25.1429	< 2.2e-16	***
ZIP	2.643e-02	0.0263	1.0050	0.31499	
Avg_Num_Products_Purchased	6.702e+01	1.5143	44.2550	< 2.2e-16	***
Years_as_Customer	-2.342e+00	1.2229	-1.9150	0.05561	.

Since the customer segment is categorial variable, I have used P Value to determine either use it or not in linear regression model. Therefore, it has low value which is less than 2.2e-16, so it is statistically significant. In the end, the predictor variables are Avg_num_products_purchased and customer segment, and the target variable is Avg_sale_amount.

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.

Figure 5: 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

The above figure, has shown the statistical result of linear regression model. The P-value for all variables is less than 2.2e-16 that represents their relationship with the target variable is statistically significant. As well as, the multiple R-squared value is 0.8369 and the adjusted R-squared is 0.8366. So, the model has higher explanatory power due to variation in the predictor variable. Therefore, this linear regression model is a good model based on P-value and R-squared.

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:

*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.

*Avg_sale_amount = 303.46 - 149.36 * (If customer_segment: Loyalty Club) + 281.84 * (If customer_segment: Loyalty Club and Credit Card) – 245.42 * (If customer_segment: Store Mailing List) + 0 * (If customer_segment: Credit Card Only) + 66.98 * (Avg_Num_Products_Purchased)*

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?

According to the result of analyzation, I recommend management unit send the catalog to 250 new customers since the excepted profit exceeds $10,000 and is **$21,987.44**. As result, the chance of pass profit limit is high due to the calculated excepted the profit is two times the limit.

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

The excepted profit is calculated by following steps:

1- Define the formula of avg_sale_amount by linear regression model that is Avg_sale_amount = 303.46 - 149.36 * (If customer_segment: Loyalty Club) + 281.84 * (If customer_segment: Loyalty Club and Credit Card) – 245.42 * (If customer_segment: Store Mailing List) + 0 * (If customer_segment: Credit Card Only) + 66.98 * (Avg_Num_Products_Purchased).
2- The formula applied on mailing list data to predict Avg_sale_amount.
3- The predicted Avg_sale_amount multiplied by score_yes to provide a profit of predicted Avg_sale_amount in all customer records.
4- Calculate the sum of profit of predicted Avg_sale_amount.
5- The summation multiply by gross margin that is 50%.
6- Subtract by multiplication cost of catalog and number of customers.

In simple form, excepted profit = sum(predicted Avg_sale_amount * score_yes) * 50% - ($6.50 * 250)

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

The expected profit is **$21,987.44**. It was calculated by the formula of excepted profit in pervious question.

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.