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 company needs to send the print catalog to its 250 new customers from the mailing list. The decision that needs to be made is regarding the worth of sending those catalogs to new customers based on the profits this would bring.

2. What data is needed to inform those decisions?

To inform those decisions, the company needs a set of valuable information. First, the company needs the information about the expected profit from those 250 new customers who will receive the catalog. Second, the company will send the catalogs only if the expected profit exceeds $10,000.

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)

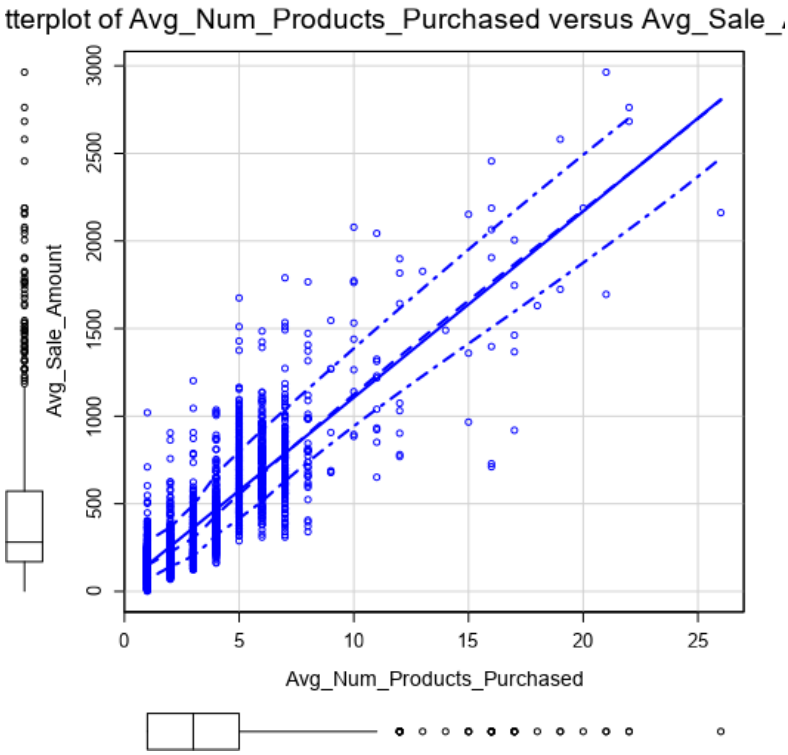
Important: Use the p1-customers.xlsx to train your linear model.

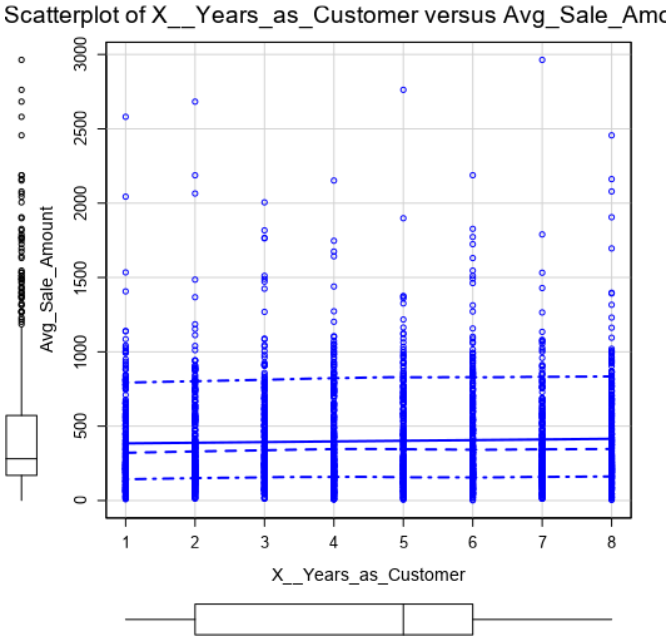
At the minimum, answer these questions:

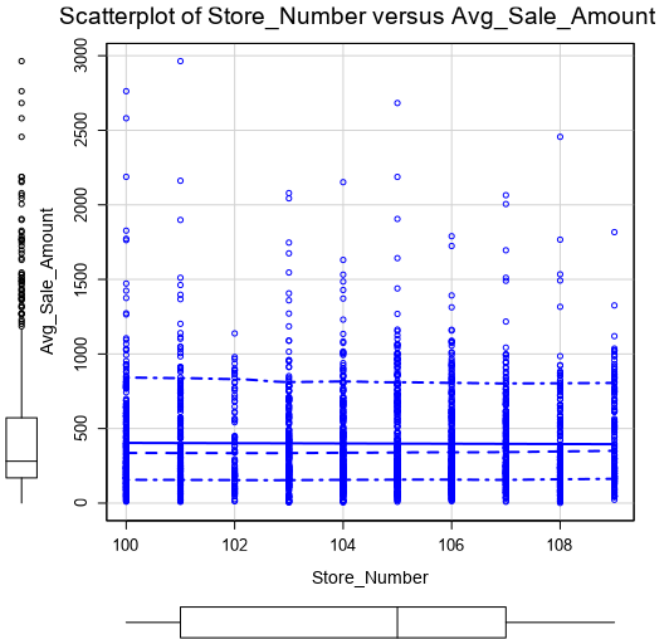
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

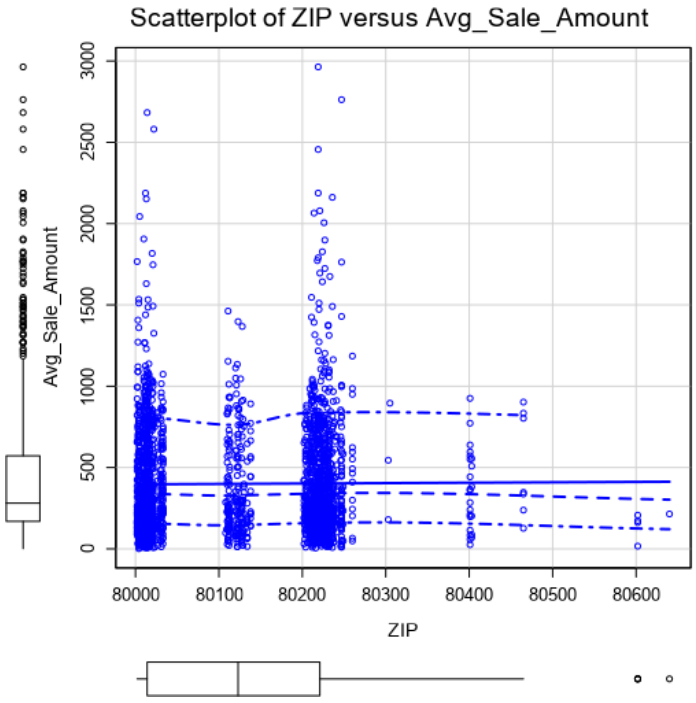
I have set up my regression model using the predictor variables that are statistically significant and have a relationship with target variable, which in our case is ***Average Sales Amount.*** *My predictor variables are:* ***Average Number of Products Purchased and Customer Segment****. After using scatterplots, I have come to the result that only* ***Average Number of Products Purchased (continuous variable),*** *has a linear relationship with the target variable. As the* ***Number of Products Purchased*** *increases, so does the* ***Sales Amount****. The Customer Segment predictor variable was chosen based on the report I got from my first Regression Model. Based on the report****, Average Number of Products Purchased and Customer***

Segment *have the most statistically significant with a p-value of* ***<u>< 2.2e-16.</u>*** *This means there is a relationship between the predictor values and target value.*









Record	Report
1	**Report for Linear Model Catalog_Linear_Regression**
2	*Basic Summary*
3	Call:

Record	Report
	lm(formula = Avg_Sale_Amount ~ Customer_Segment + ZIP + Store_Number + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)
4	Residuals:

5

Min	1Q	Median	3Q	Max
-668.09	-67.40	-2.23	72.15	971.30

6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958	
Customer_SegmentLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.8485	9.770e+00	-25.1625	< 2.2e-16	***
ZIP	0.0225	2.659e-02	0.8460	0.39761	
Store_Number	-1.0002	1.006e+00	-0.9939	0.32037	
Avg_Num_Products_Purchased	66.9646	1.515e+00	44.1928	< 2.2e-16	***
X._Years_as_Customer	-2.3528	1.223e+00	-1.9239	0.05449	.

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

8 Residual standard error: 137.41 on 2367 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8368
F-statistic: 1740 on 7 and 2367 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)	
Customer_Segment	28793567.64	3	508.35	< 2.2e-16	***
ZIP	13514.61	1	0.72	0.39761	
Store_Number	18651.26	1	0.99	0.32037	
Avg_Num_Products_Purchased	36873634.66	1	1953.01	< 2.2e-16	***
X._Years_as_Customer	69882.02	1	3.7	0.05449	.
Residuals	44690015.14	2367			

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

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.

*Based on the statistical result that my linear model created, I can say that this model is a good model. There are two predictor variables (Customer Segment and Average Number of Products Purchased) that have **p-value** less than **.05** which gives us the confidence that there is a strong relationship between predictor and target variables. R-squared values also have a big meaning in my model. As we know, the closer R-squared value is to 1, the better target variable is explained by the model. So, in my model, the R-squared value for my predictor variables is **0.8369,** which is a good sign of a good model.*

Record	Report
1	**Report for Linear Model Catalog_Linear_Regression**
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

8	Residual standard error: 137.48 on 2370 degrees of freedom Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366

Record	Report
	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)	
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)

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.

*Average Sale Amount (Y) = 303.46 – 149.36 (If Type: Loyalty club only) + 281.84 (If type: Loyalty Club and credit Card) – 245.42 (If Type: Store Mailing List) + 0 (If Type: Cash) + 66.98 * (Average Number of Products Purchased)*

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

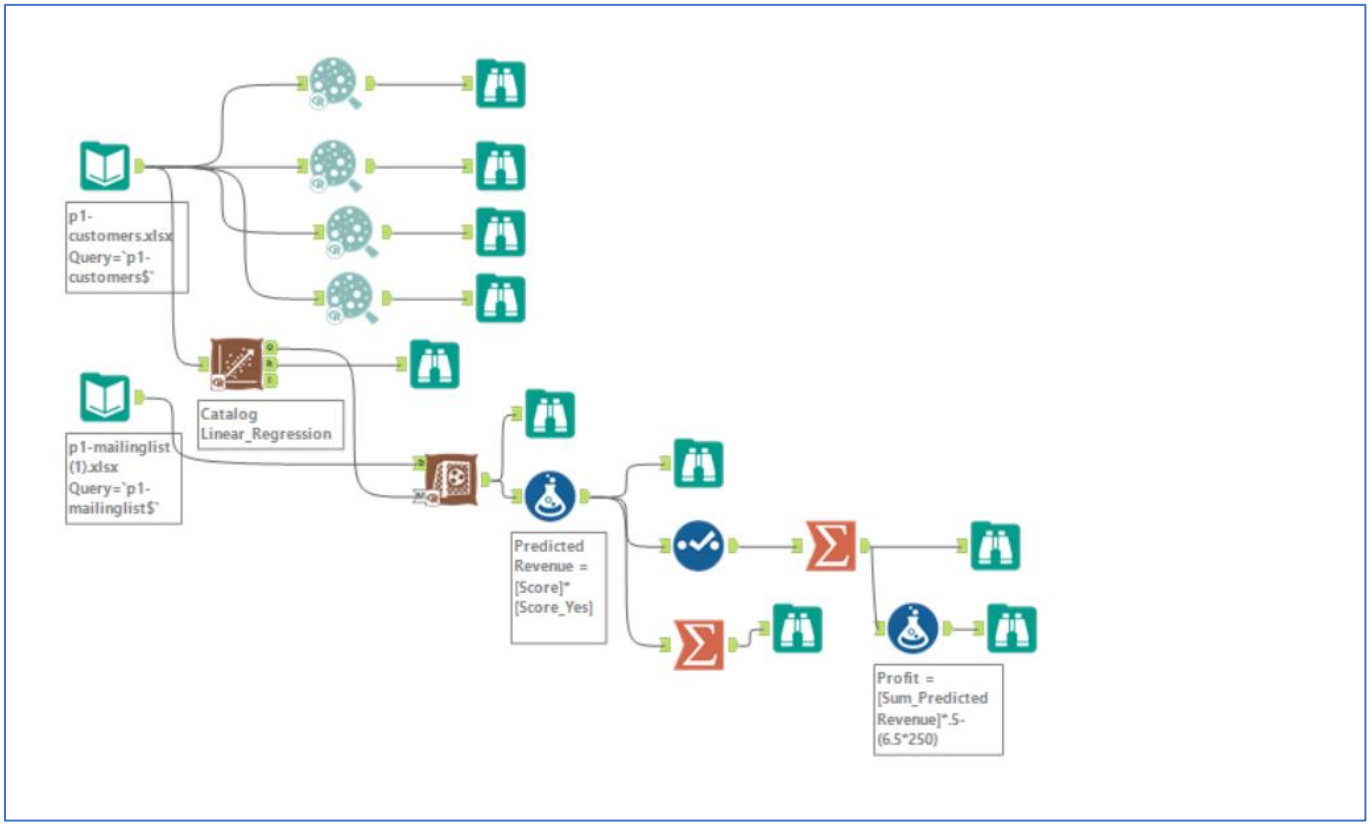
At the minimum, answer these questions:

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

My recommendation is that the company should send the catalog to its new 250 customers. The model has good enough results to be confident that the profit will go past the $10,000 set by the company. After applying the model, we can see that the expected profit goes higher than $10,000, and this is the reason why the company should consider sending those catalogs.

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

*I came up with this recommendation based on the results I have got from applying the model. The regression equation gave the chance to apply it to the mailing list data and get the **Average Sale Amount**. Then, based on this data, I found the **Predicted Revenue** by multiplying the **Average Sale Amount (Score)** with the probability to buy **(Score_Yes**). Then, I did the **SUM** of **Predicted Revenue,** multiplied it by the **gross margin** (which is 50%) and subtracted **the cost** of printing 250 catalogs from it, which allowed me to find out the profit for catalogs.*



Alteryx Workflow, Ludmila Turcan 1, Screenshot

***Predicted Revenue** = [Score] * [Score_Yes]*

Expected Profit** = [Sum_Predicted Revenue] * .5 – (6.50 * 250) = 47,224.8713 * .5 – (6.50 * 250) = **$21,987.44

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

Assuming that we send the new catalog to these 250 new customers, we expect to have the the profit of ***$21,987.44.***

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.